



**ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ**

**ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ  
ΥΠΟΛΟΓΙΣΤΩΝ**

**ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ & ΤΕΧΝΟΛΟΓΙΑΣ  
ΥΛΙΚΩΝ**

**Ανάπτυξη Μεθόδων Μηχανικής Μάθησης για την  
Εκτίμηση της Σοβαρότητας της Ρευματοειδούς  
Αρθρίτιδας**

**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

της

**ΛΥΔΙΑΣ ΣΦΟΝΤΟΥΡΗ**

**Επιβλέπουσα:** Κωνσταντίνα Νικήτα  
Καθηγήτρια Ε.Μ.Π.

Αθήνα, Ιούνιος 2022



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ

ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ &

ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

**Ανάπτυξη Μεθόδων Μηχανικής Μάθησης για την  
Εκτίμηση της Σοβαρότητας της Ρευματοειδούς  
Αρθρίτιδας**

**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

της

**ΛΥΔΙΑΣ ΣΦΟΝΤΟΥΡΗ**

**Επιβλέπουσα:** Κωνσταντίνα Νικήτα  
Καθηγήτρια Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την

*(Υπογραφή)*

.....  
Κωνσταντίνα Νικήτα  
Καθηγήτρια Ε.Μ.Π.

*(Υπογραφή)*

.....  
Γ.Στάμου  
Καθηγητής Ε.Μ.Π.

*(Υπογραφή)*

.....  
Α.-Γ.Σταφυλοπάτης  
Καθηγητής Ε.Μ.Π.

(Υπογραφή)

.....

**ΣΦΟΝΤΟΥΡΗ ΛΥΔΙΑ**

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

© 2022 – All rights reserved

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

## Σύνοψη

Η μηχανική μάθηση είναι ένα πεδίο της τεχνητής νοημοσύνης το οποίο εφαρμόζεται όλο και περισσότερο στην ιατρική για την υποστήριξη τόσο των ασθενών όσο και των γιατρών. Τα εκτενή ψηφιοποιημένα σύνολα δεδομένων που είναι πλέον διαθέσιμα παρέχουν την βάση για την χρήση αλγορίθμων μηχανικής μάθησης προκειμένου να δημιουργηθούν μοντέλα που εκπαιδεύονται με τα δεδομένα και παράγουν καινούργια γνώση.

Στην παρούσα διπλωματική εργασία χρησιμοποιήθηκαν δημογραφικά, κλινικά, εργαστηριακά και περιβαλλοντικά δεδομένα καθώς και πληροφορίες σχετικά με τη θεραπευτική αγωγή και τις συννοσηρότητες, από μια βάση δεδομένων τριακοσίων (300) ασθενών με ρευματοειδή αρθρίτιδα (ΡΑ) που προέρχεται από τα αρχεία ασθενών των ρευματολογικών κλινικών των νοσοκομείων Ευαγγελισμός και Λαϊκό Νοσοκομείο Αθηνών. Στόχος της εργασίας είναι η ανάπτυξη μοντέλων μηχανικής μάθησης για την εκτίμηση της σοβαρότητας της νόσου, με σκοπό τη διερεύνηση της δυνατότητας ανάπτυξης στο μέλλον ενός νέου δείκτη προσδιορισμού της σοβαρότητας της ΡΑ στους ασθενείς.

Στο πλαίσιο της παρούσας εργασίας, αναπτύχθηκαν πέντε μοντέλα μηχανικής μάθησης τα οποία εκπαιδεύτηκαν και ελέγχθηκαν με το διαθέσιμο σύνολο δεδομένων. Η επικύρωση των μοντέλων βασίστηκε στην «εκτίμηση σοβαρότητας» (Severity) που παράχθηκε από ειδικούς ρευματολόγους για κάθε ασθενή.

Οι αλγόριθμοι που χρησιμοποιήθηκαν στο πλαίσιο της παρούσας εργασίας περιλαμβάνουν αλγόριθμους συσταδοποίησης μη εποπτευόμενης μάθησης (kmeans, kmedoids, GaussianMixture) και τον αλγόριθμο εποπτευόμενης μάθησης K- nearest neighbors (KNN). Σύμφωνα με τη συγκριτική αξιολόγηση των αποτελεσμάτων που παρήχθησαν, το μοντέλο KNN πέτυχε την υψηλότερη ακρίβεια (90%). Οι σημαντικότερες μεταβλητές για την εκτίμηση της σοβαρότητας στο δείγμα ασθενών που εξετάστηκε περιλαμβάνουν τις εξής: Sharp-Score (αξιολόγηση ακτινολογικών ευρημάτων), Duration (διάρκεια νόσου), HLA-DRB1 (λευκοκυτταρικό αντιγόνο), CCP (αντισώματα ορού έναντι κιτρουλίνης) και RF (ρευματοειδή παράγοντα).

Τα αποτελέσματα της εργασίας υποδεικνύουν ότι τα μοντέλα μηχανικής μάθησης για την εκτίμηση της σοβαρότητας της ΡΑ που αναπτύχθηκαν είναι αξιόπιστα, παράγουν ακριβείς εκτιμήσεις και θα μπορούσαν να αξιοποιηθούν για την υλοποίηση μιας εφαρμογής λογισμικού που θα είναι χρήσιμη στον ειδικό γιατρό παρέχοντας έναν αξιόπιστο δείκτη της σοβαρότητας της ΡΑ.

**Λέξεις-κλειδιά:** Ρευματοειδής αρθρίτιδα, μηχανική μάθηση, συσταδοποίηση, κατηγοριοποίηση, μοντέλα πρόβλεψης της σοβαρότητας.

## Abstract

Machine learning is a field of artificial intelligence that is increasingly applied in medicine to help both patients and doctors. The large, digitized datasets that are now available provide the basis for the use of machine learning algorithms in order to create models that are trained with the data and produce new knowledge.

In the present thesis, we explore the potential of machine learning (ML) models to assess the severity of rheumatoid arthritis (RA) towards establishing a new biomarker for predicting the evolution of the disease. We use data from the medical records of three hundred (300) patients with RA collected from the rheumatology clinics at the Evangelismos Hospital and the Athens People's Hospital. The used dataset includes demographic, clinical, laboratory and environmental data as well as information on treatment and comorbidities.

In the context of the present work, five machine learning models were developed which were trained and tested with the available dataset. The validation of the models was performed in terms of produced "Severity" predictions against the "Severity" assessed for each patient by experienced rheumatologists.

Unsupervised learning aggregation algorithms (kmeans, kmedoids, GaussianMixture) and supervised learning algorithms (K-nearest neighbors (KNN)) were used and their performance was comparatively assessed. According to the obtained results, the KNN model provided superior performance achieving a 90% accuracy. The most important features for assessing the RA severity in the examined patient dataset included Sharp-Score, Disease duration, HLA-DRB1, CCP, and RF.

The results of the present work indicate that machine learning models can produce a reliable indicator for RA severity and can provide a powerful tool to support specialists in the assessment of disease severity.

**Keywords:** Rheumatoid arthritis, machine learning, clustering, RA severity prediction models.

## Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα της διπλωματικής μου εργασίας, καθηγήτρια κα. Κωνσταντίνα Νικήτα, χάρη στην οποία μου δόθηκε η ευκαιρία να έρθω σε επαφή με ένα θέμα εξαιρετικού ενδιαφέροντος για εμένα και να συνδέσω μέσω αυτής της εργασίας το αντικείμενο των σπουδών μου με την επιστήμη του πατέρα μου.

Ευχαριστώ ιδιαίτερω τον διδάκτορα Κώστα Μήτση για την άμεση και πολύτιμη βοήθειά του στην συγγραφή της εργασίας, τόσο σε επίπεδο καθοδήγησης και συμβουλών, καθ' όλη την εκπόνησή της, όσο και σε επίπεδο στήριξης και ενθάρρυνσης.

Επιπλέον θα ήθελα να ευχαριστήσω την υποψήφια διδάκτορα στην Ιατρική Σχολή του Εθνικού Καποδιστριακού Πανεπιστημίου Αθηνών Μολέ Ευαγγελία που μου παραχώρησε το σύνολο δεδομένων πάνω στο οποίο στηρίχτηκε η εκπόνηση της παρούσας διπλωματικής όπως και τους προϊσταμένους και συνεργάτες μου στην Artemis ITS που με διευκόλυναν με κάθε τρόπο τους τελευταίους μήνες ώστε να ολοκληρώσω τις σπουδές μου.

Κλείνοντας, θα ήθελα να εκφράσω την βαθιά μου ευγνωμοσύνη στην οικογένειά μου που είναι δίπλα μου και με στηρίζουν σε κάθε βήμα της ζωής και της εξέλιξής μου, με αγάπη και υπομονή, και τον Αντώνη με τον οποίο μοιράστηκα όλες τις όμορφες αλλά και δύσκολες στιγμές των φοιτητικών μου χρόνων και αποτέλεσε τον καλύτερο συνοδοιπόρο που θα μπορούσα να έχω.

## Περιεχόμενα

<b>1 Ρευματοειδής Αρθρίτιδα .....</b>	<b>13</b>
1.1 Γενική Προσέγγιση .....	13
1.2 Αιτιολογία- Παθογένεια.....	14
1.2.1 Γενετική βάση της ρευματοειδούς αρθρίτιδας .....	14
1.2.2 Μη γονιδιακοί παράγοντες κινδύνου .....	15
1.3 Παθολογική Ανατομική .....	15
1.3.1 Κλινική εικόνα .....	16
1.3.2 Αρθρικές -εξωαρθρικές εκδηλώσεις .....	16
1.3.3 Εκδηλώσεις από τις αρθρώσεις.....	17
1.3.4 Εξωαρθρικές εκδηλώσεις.....	20
1.4 Εργαστηριακά ευρήματα.....	24
1.4.1 Από το αίμα.....	24
1.4.2 Από το Αρθρικό υγρό.....	24
1.4.3 Από τα Ούρα .....	25
1.5 Ρευματοειδείς Παράγοντες (ΡΠ).....	25
1.6 Αντισώματα έναντι κιτρουλίνης anti-CCP (anticyclic citrullinated peptide antibody) .....	26
1.7 Απεικονιστικά ευρήματα.....	27
1.7.1 Ακτινολογικά ευρήματα με απλές ακτινογραφίες.....	27
1.7.2 Άλλες απεικονιστικές τεχνικές.....	28
1.8 Διάγνωση-Διαγνωστικά κριτήρια- Διαφορική διάγνωση .....	29
1.8.1 Διάγνωση.....	29
1.8.2 Διαγνωστικά κριτήρια .....	30
1.9 Πρόγνωση- θεραπεία .....	31
1.9.1 Πρόγνωση .....	31
1.9.2 Θεραπεία .....	31

<b>2 Εφαρμογές Μηχανικής Μάθησης στην Ρευματοειδή Αρθρίτιδα &amp; Συναφείς Μελέτες</b>	<b>37</b>
.....	
2.1 Η Μηχανική Μάθηση στην Ρευματολογία .....	37
2.2 Βασικές Αρχές Μηχανικής Μάθησης.....	43
2.2.1 Γενικά.....	43
2.2.2 Είδη Μηχανικής Μάθησης.....	44
2.3 Μεθοδολογίες και αλγόριθμοι που αξιοποιήθηκαν στην παρούσα εργασία.....	52
2.3.1 Συσταδοποίηση .....	53
2.3.2 Αλγόριθμοι μη επιβλεπόμενης μάθησης και μεθοδολογίες .....	58
2.3.3 Επιβλεπόμενη μάθηση με τον αλγόριθμο k πλησιέστερων γειτόνων (k-NN)	
63	
2.4 Αξιοποίηση της Μηχανικής Μάθησης στην Έρευνα της Ρευματοειδούς	
Αρθρίτιδας: Ανασκόπηση επιστημονικών άρθρων και δημοσιεύσεων .....	67
2.4.1 Επιτελική σύνοψη .....	67
2.4.2 Ενδεικτικά παραδείγματα μηχανικής μάθησης στην RA .....	74
<b>3 Περιγραφή Δεδομένων.....</b>	<b>83</b>
3.1 Εισαγωγή.....	83
3.2 Διερευνητική Ανάλυση Δεδομένων .....	87
3.2.1 Γενικές παρατηρήσεις .....	87
3.2.2 Δείκτες Σοβαρότητας και Ενεργότητας στην RA.....	93
3.3 Περιγραφή Κρίσιμων Μεταβλητών / Χαρακτηριστικών για τα Μοντέλα της	
Μελέτης .....	98
3.3.1 Δείκτης DAS28 .....	98
3.3.2 Λευκοκυτταρικά Αντιγόνα (Human Leukocyte Antigens, HLA).....	98
3.3.3 Αντισώματα Ορού έναντι κιτρουλίνης (CCP) .....	99
3.3.4 Ρευματοειδείς Παράγοντες (ΡΠ).....	99
3.3.5 C Αντιδρώσα Πρωτεΐνη (CRP).....	99
3.3.6 Ταχύτητα Καθίζησης Αίματος (ΤΚΕ).....	100



3.3.7	Sharp Score .....	100
<b>4</b>	<b>Μεθοδολογία Μελέτης.....</b>	<b>102</b>
4.1	Εισαγωγή.....	102
4.2	Προεπεξεργασία Δεδομένων.....	104
4.3	Μοντέλα Μηχανικής Μάθησης .....	109
4.3.1	Μοντέλο 1 : kMeans .....	109
4.3.2	Μοντέλο 2 : GaussianMixture.....	113
4.3.3	Μοντέλο 3 : K-Medoids.....	113
4.3.4	Μοντέλο 4 : k- Nearest Neighbors .....	114
<b>5</b>	<b>Αποτελέσματα.....</b>	<b>117</b>
5.1	Μοντέλο 1: Kmeans .....	117
5.1.1	Περίπτωση 1: SEVERITY .....	117
5.1.2	Περίπτωση 2: SEVERITY_binary .....	134
5.2	Μοντέλο 2: GaussianMixture.....	135
5.2.1	Περίπτωση 1: SEVERITY .....	136
5.2.2	Περίπτωση 2: SEVERITY_binary .....	151
5.3	Μοντέλο 3: K-Medoids.....	153
5.3.1	Περίπτωση 1: SEVERITY .....	153
5.3.2	Περίπτωση 2: SEVERITY_binary .....	170
5.4	Μοντέλο 4: K-Nearest Neighbors .....	172
5.4.1	Περίπτωση 1: SEVERITY .....	173
5.4.2	Περίπτωση 2: SEVERITY_binary .....	180
<b>6</b>	<b>Συμπεράσματα &amp; Επίλογος.....</b>	<b>189</b>
6.1	Συμπεράσματα .....	189
6.1.1	Μοντέλα Μη- Εποπτευόμενης Μηχανικής Μάθησης.....	190
6.1.2	Μοντέλα Μη- Εποπτευόμενης Μηχανικής Μάθησης.....	194
6.2	Περιορισμοί της έρευνας.....	195

6.3	Προτάσεις για μελλοντική έρευνα .....	196
6.4	Επίλογος.....	198
<b>7</b>	<b>Βιβλιογραφία .....</b>	<b>200</b>

# *Εισαγωγή*

Η ρευματοειδής αρθρίτιδα (ΡΑ) είναι χρόνια αυτόανοσο νόσημα και παρουσιάζει διάφορες διακυμάνσεις, οι οποίες περιλαμβάνουν πολύπλοκη υποβόσκουσα παθοφυσιολογία που περιπλέκει την θεραπεία. Οι ασθενείς με ΡΑ παρακολουθούνται τακτικά και η σωστή λήψη των αποφάσεων για την διαχείριση της νόσου από τον ειδικό ρευματολόγο απαιτεί την συνεκτίμηση και ερμηνεία πολύ διαφορετικών και μεγάλου όγκου δεδομένων, όπως εργαστηριακά δεδομένα, απεικονιστικά δεδομένα, συννοσηρότητες, βαθμολογίες δραστηριότητας και αυτοαξιολόγησης των ασθενών.

Οι τεχνολογίες υψηλής απόδοσης και τα μοντέλα μηχανικής μάθησης μπορούν να συνδράμουν το ειδικό ρευματολόγο σε πολλούς τομείς με σκοπό στον προσδιορισμό των κρίσιμων παραγόντων για την πορεία της νόσου σε έναν ασθενή. Μπορούν να αξιοποιηθούν για να βελτιώσουν την ταξινόμηση των ασθενών, να διευκολύνουν την έγκαιρη διάγνωση, να προβλέψουν την εξέλιξη της νόσου και τις εξάρσεις, να προτείνουν εξατομικευμένη φαρμακευτική προσέγγιση, οδηγώντας τελικά στην ιατρική ακριβείας.

Σκοπός της παρούσας διπλωματικής εργασίας είναι η αξιοποίηση βάσης δεδομένων με ψηφιοποιημένα στοιχεία των φακέλων υγείας 300 ασθενών με ΡΑ περιόδου 2017-2020, από το νοσοκομείο Ευαγγελισμός και Λαϊκό νοσοκομείο Αθηνών για την παρουσίαση μεθόδων και μοντέλων μηχανικής μάθησης που μπορούν να χρησιμοποιηθούν για την εκτίμηση της σοβαρότητας της ΡΑ στους ασθενείς.

Η δομή της παρούσας εργασίας είναι η ακόλουθη:

**Κεφάλαιο 1:** Συνοπτική περιγραφή της ΡΑ, που περιλαμβάνει τον ορισμό της, αιτιολογία και παθογένεια, την γενετική βάση της ασθένειας, την κλινική της εικόνα, τις εκδηλώσεις της νόσου, τα εργαστηριακά και τα απεικονιστικά ευρήματα. Στη συνέχεια περιγράφονται τα διάφορα κριτήρια για την ταξινόμηση της νόσου και την διαφορική διάγνωση. Τέλος γίνεται αναφορά των διαφορετικών θεραπειών και φαρμακευτικών προσεγγίσεων.

**Κεφάλαιο 2:** Στο κεφάλαιο 2 παρουσιάζονται οι βασικές αρχές και τα είδη της μηχανικής μάθησης, γίνεται ανασκόπηση των εφαρμογών της μηχανικής μάθησης στην ρευματοειδή αρθρίτιδα, παρουσιάζεται το θεωρητικό υπόβαθρο για τους βασικούς αλγορίθμους και μεθόδους που θα χρησιμοποιηθούν για την ανάπτυξη των μοντέλων της παρούσας εργασίας και περιγράφονται συνοπτικά σημαντικές συναφείς μελέτες που αφορούν στην ΡΑ και τη μηχανική μάθηση για διάγνωση, πρόβλεψη κινδύνων, ταξινόμηση αποτελεσμάτων απεικονιστικών μεθόδων, επιλογή θεραπείας κ.ο.κ.

**Κεφάλαιο 3:** Στο κεφάλαιο αυτό γίνεται περιγραφή της βάσης δεδομένων, ενώ παρουσιάζονται και επεξηγούνται συνοπτικά οι διάφορες μεταβλητές του συνόλου δεδομένων. Γίνεται διερευνητική ανάλυση των δεδομένων, παρουσιάζονται διάφορα στατιστικά στοιχεία που προκύπτουν και επηρεάζουν την λήψη της απόφασης για την επιλογή των μεταβλητών εισόδου των μοντέλων και αποτυπώνονται τα αποτελέσματα του ποιοτικού ελέγχου των δεδομένων. Τέλος, αναλύονται διεξοδικότερα οι μεταβλητές που θα χρησιμοποιηθούν ως εισόδοι στα μοντέλα μηχανικής μάθησης που θα αναπτυχθούν στη συνέχεια.

**Κεφάλαιο 4:** Στο κεφάλαιο 4 παρουσιάζονται τα βήματα για την δημιουργία των μοντέλων μηχανικής μάθησης, ο καθαρισμός, η προεπεξεργασία, η επιλογή των αλγορίθμων, τα χαρακτηριστικά εισόδου του κάθε μοντέλου.

**Κεφάλαιο 5:** Στο κεφάλαιο 5 παρουσιάζονται τα αποτελέσματα που προέκυψαν από τα μοντέλα που αναπτύχθηκαν και γίνεται συγκριτική αξιολόγηση αυτών με στόχο την επιλογή του πιο στιβαρού μοντέλου για περαιτέρω αξιοποίηση.

**Κεφάλαιο 6:** Στο κεφάλαιο αυτό παρουσιάζονται τα συμπεράσματα της εργασίας και παρατίθενται προτάσεις για μελλοντικές επεκτάσεις.

**Κεφάλαιο 7:** Βιβλιογραφία.

# 1

## *Ρευματοειδής Αρθρίτιδα*

Η Ρευματοειδής αρθρίτιδα είναι το πρότυπο αυτοάνοσης αρθρίτιδας και όπως συμβαίνει με τα περισσότερα αυτοάνοσα νοσήματα προκαλείται από ένα συνδυασμό γενετικών, περιβαλλοντικών και ανοσολογικών μηχανισμών. Η δυσλειτουργία της ανοσολογικής απάντησης περιλαμβάνει ειδική ανοσία των κυττάρων συμπεριλαμβανομένων των ουδετερόφιλων των μακροφάγων και αντιγονοπαρουσιαστικών κυττάρων, χαρακτηρίζεται από αλλαγές στον υμένα, στους χόνδρους και τα οστά [1]. Είναι επιμένουσα, συνεχής και καταστροφική μικρών και μεγάλων αρθρώσεων, ακολουθώντας την ένταση, τη διάρκεια της φλεγμονής και την περίοδο των εξάρσεων της νόσου.

### *1.1 Γενική Προσέγγιση*

Η ρευματοειδής αρθρίτιδα είναι αυτοάνοση φλεγμονώδης χρόνια συστηματική νόσος με προσβολή κυρίως περιφερικών αρθρώσεων. Οι βλάβες ξεκινούν από τον αρθρικό υμένα και επεκτείνονται στον χόνδρο, το οστό και τους συνδέσμους με αποτέλεσμα την καταστροφή της αρχιτεκτονικής της άρθρωσης. Είναι περισσότερο συχνή στις γυναίκες και στις ηλικίες μεταξύ 40-60 ετών. Οι εξωαρθρικές εκδηλώσεις από διάφορα όργανα εμφανίζονται στην πλειονότητα των ασθενών και είναι δείκτης νοσηρότητας και θνητότητας. Η σοβαρότητα της ΡΑ ποικίλει, περιέχει ένα ευρύ φάσμα εκδηλώσεων από αυτοπεριοριζόμενη, μέχρι χρόνια εξελισσόμενη καταστροφική αρθροπάθεια που προκαλεί αναπηρία. Η ετερογένεια αυτή καθορίζεται από γενετικούς περιβαλλοντικούς και άλλους παράγοντες που ελέγχουν την φλεγμονή [2].

## ***1.2 Αιτιολογία- Παθογένεια***

Η αιτιολογία προς το παρόν είναι άγνωστη. Φαίνεται ότι γενετικοί, λοιμώδεις και περιβαλλοντικοί παράγοντες διαδραματίζουν σημαντικό ρόλο στην παθογένεια της νόσου [3]. Όπως στα περισσότερα νοσήματα έτσι και στη ΡΑ η παθογένεια είναι μερικώς κατανοητή. Η διαδικασία της φλεγμονής συμβαίνει σε γενετικά καθορισμένα άτομα, αρχίζει στον αρθρικό υμένα, οδηγείται από τις προφλεγμονώδεις κυτταροκίνες όπως ο παράγοντας νέκρωσης των όγκων TNF $\alpha$  (tumor necrosis factor  $\alpha$ ). Καθώς ο αρθρικός υμένας φλεγμαίνει, εμπλουτίζεται με T και B κύτταρα και από ειδικά ανοσοκύτταρα όπως τα μακροφάγα, τα δενδριτικά και τα ιστιοκύτταρα. Επιπλέον, η κατά κανόνα λεπτή αρθρική επένδυση πολλαπλασιάζεται και υφίσταται έντονη αγγειογένεση, γεγονός που οδηγεί σε υπερπλασία του αρθρικού υμένα σε μία δομή που ονομάζεται “rannus”. Στο τέλος η ασυγκράτητη φλεγμονή οδηγεί σε καταστροφή του αρθρικού χόνδρου και σε διάβρωση των οστών περιαρθρικά. Ο “rannus” εξαπλώνεται πάνω στον αρθρικό χόνδρο και προωθεί την υποβάθμιση του με μηχανισμούς που περιλαμβάνουν άμεσες επιδράσεις των κυτταροκινών στα χονδροκύτταρα, καθώς και την παραγωγή ενζύμων αποδόμησης από τα κύτταρα του αρθρικού υμένα [4]. Οι οστεοκλάστες, πιστεύεται ότι μεσολαβούν στις διαβρώσεις στην ΡΑ, δεδομένου ότι αυτά τα κύτταρα οστικής απορρόφησης μπορούν να βρεθούν σε θέσεις όπου ο “rannus” ιστός έρχεται σε επαφή με το οστό [5]. Σε πειραματικά μοντέλα ποντικών με ανεπάρκεια οστεοκλαστών που αναπτύσσεται φλεγμονή στις αρθρώσεις, αυτά προστατεύονται από τις οστικές διαβρώσεις [6].

### ***1.2.1 Γενετική βάση της ρευματοειδούς αρθρίτιδας***

Μελέτες που έγιναν σε μονοζυγωτικούς και διζυγωτικούς διδύμους υποστηρίζουν την άποψη ότι οι γενετικοί παράγοντες είναι σημαντικοί στην επιδεκτικότητα για την ΡΑ. Στην προσπάθεια εξακρίβωσης γενετικών παραγόντων που συμβάλλουν στην εκδήλωση ΡΑ μελετήθηκαν τα αντιγόνα του μείζονος συμπλέγματος ιστοσυμβατότητας. Διαπιστώθηκε συσχέτιση της ΡΑ με το HLA DR4 [7], [8] αλλά τα αλληλία που σχετίζονται με τη νόσο είναι τα HLA-DRB1\*0401, \*0404 και \*0408 και για τον ελληνικό πληθυσμό HLA-DRB1\*10 [9] Εκτός από την αυξημένη επιδεκτικότητα στην εκδήλωση ΡΑ, τα παραπάνω αλληλία, πιθανό να συμβάλλουν στην εκδήλωση σοβαρότερης κλινικής εικόνας και μεγαλύτερης πιθανότητας υποτροπών της νόσου [10]

### **1.2.2 Μη γονιδιακοί παράγοντες κινδύνου**

Στους πιθανούς μη γονιδιακούς παράγοντες κινδύνου περιλαμβάνονται κλασσικοί περιβαλλοντικοί παράγοντες, τυχαία γεγονότα αθροιζόμενα κατά τη διάρκεια της ζωής, επίκτητη γονιδιακή μεταβλητότητα, λοιμώδεις παράγοντες. Ο πιθανός ρόλος κάποιου λοιμώδους παράγοντα ενισχύεται από την εποχική διακύμανση. Έχουν ενοχοποιηθεί ιογενείς και βακτηριδιακές λοιμώξεις [11]. Η άποψη ενισχύεται από παρατηρήσεις ότι αρκετοί λοιμώδεις παράγοντες μπορούν να προκαλέσουν σε ανθρώπους χρόνια αρθροπάθεια που ιστολογικώς είναι παρόμοια με τη RA. Διάφοροι παρβοϊοί, ο ιός της ερυθράς και ο HTLV-1 έχουν ενοχοποιηθεί, αλλά οι προσπάθειες εντοπισμού με τεχνικές απομόνωσης με ηλεκτρονικό μικροσκόπιο και μοριακές τεχνικές δεν έχουν αποδώσει. Τέλος το κάπνισμα έχει σχετιστεί με αυξημένο κίνδυνο ανάπτυξης RA και η ηλικία στο πλαίσιο της ανοσογήρανσης.

### **1.3 Παθολογική Ανατομική**

Τα ιστολογικά ευρήματα στην RA δεν είναι ειδικά, εξαρτώνται από το όργανο που έχει προσβληθεί. Στις αρθρώσεις η αρχική φλεγμονώδη βλάβη, αφορά τον αρθρικό υμένα. Οι πρωιμότερες αλλοιώσεις στη RA από το ηλεκτρονικό μικροσκόπιο είναι βλάβες του μικροαγγειακού δικτύου του αρθρικού υμένα, με απόφραξη του αυλού, οίδημα των ενδοθηλιακών κυττάρων και σχηματισμό οπών μεταξύ αυτών. Η φλεγμονώδης φάση σχετίζεται με υπεραιμία, οίδημα, ινώδες εξίδρωμα και ήπια υπερπλασία της επιφανειακής καλυπτήριας κυτταρικής στιβάδας. Και οι δύο τύποι των υμενοκυττάρων συντείνουν στην υπερπλασία του αρθρικού υμένα. Στην πορεία της νόσου εμφανίζονται κυτταρικές διηθήσεις από λεμφοκύτταρα και μακροφάγα, χαρακτηριστικό ασθενών με ενεργό RA. Σε αυτούς τους ιστούς τα T και B κύτταρα σχηματίζουν εστιακές συναθροίσεις με μια διάχυτη διήθηση. Αυτές οι συναθροίσεις είναι εμπλουτισμένες με CD4+ T κύτταρα, ενώ τα CD8+ T κύτταρα επικρατούν. Ο αρθρικός υμένας που επεκτείνεται και καταστρέφει τον χόνδρο και το οστό, ο γνωστός πάννος ‘‘rannus’’, αποτελείται από ινοβλάστες και μακροφάγα. Τα ρευματικά οζίδια έχουν ιστολογική εικόνα κοκκιωματώδους αντίδρασης. Η τενοντοελυτρίτιδα χαρακτηρίζεται από μη ειδική φλεγμονώδη διήθηση, σχηματισμό οζιδίων με ινιδοειδή νέκρωση. Η προσβολή των αγγείων περιορίζεται σε μικρά τμήματα των τελικών αρτηριών χωρίς ιδιαίτερα ιστολογικά χαρακτηριστικά. Σπανιότερα μπορεί να παρατηρηθεί εικόνα

νεκρωτικής αγγειίτιδας. Η διάμεση πνευμονική ίνωση δεν διακρίνεται από την ίνωση άλλων νοσημάτων του συνδετικού ιστού [12]

### ***1.3.1 Κλινική εικόνα***

Η RA είναι χρόνια αρθρίτιδα με συστηματικές εκδηλώσεις, πρότυπο φλεγμονωδών αρθροπαθειών. Συνήθως έχει βαθμιαία έναρξη, μερικές φορές εμφανίζεται ξαφνικά με πυρετό και γενικά φαινόμενα όπως αδιαθεσία, ανορεξία, απώλεια βάρους, εύκολη κόπωση, καταβολή και κατάθλιψη. Μπορεί να εκδηλωθεί σαν συμμετρική πολυαρθρίτιδα (η πλέον τυπική μορφή) και σπανιότερα σαν συμμετρική ολιγοαρθρίτιδα, ασύμμετρη ολιγοαρθρίτιδα ή μονοαρθρίτιδα. Προσβάλλονται όλες οι αρθρώσεις. Είναι χαρακτηριστική η ταυτόχρονη προσβολή μικρών και μεγάλων αρθρώσεων. Η πρωινή δυσκαμψία που κρατάει από λίγα λεπτά μέχρι ώρες και υποχωρεί κατά την διάρκεια της μέρας είναι το χαρακτηριστικότερο σύμπτωμα της νόσου. Το σύνδρομο καρπιαίου σωλήνα μπορεί να είναι το πρώτο ή συνοδό σύμπτωμα, λόγω φλεγμονής των μαλακών μορίων του καρπού. Τενοντοελυτρίτιδα συχνότερα στη ραχιαία και καμπτική επιφάνεια χεριών και καρπών. Η παρουσία ιγνυακής κύστης Baker είναι συχνή (περίπου 40%) σαν αρχική εκδήλωση ή σαν επιπλοκή της νόσου. Η μυϊκή ατροφία εμφανίζεται γρήγορα ιδιαίτερα στη ραχιαία επιφάνεια χεριών και καρπών καθώς και στους τετρακέφαλους γονάτων. Τα ρευματικά οζίδια εμφανίζονται σαν υποδόρια και ενδοδερμικά οζίδια σε περιοχές πίεσεως του δέρματος, όπως στην περιοχή του ωλέκρανου, στους αγκώνες, πάνω από αρθρώσεις, στις οστικές προεξοχές, στους τένοντες (εκτινασσόμενος δάκτυλος). Είναι δυσμενές προγνωστικό σημείο και σχετίζεται με ύπαρξη άλλων συστηματικών εκδηλώσεων και υποδηλώνει βαρεία οροθετική νόσο [13]

### ***1.3.2 Αρθρικές -εξωαρθρικές εκδηλώσεις***

Οι αρθρικές εκδηλώσεις της RA μπορούν να ταξινομηθούν σε δύο κατηγορίες. Αναστρέψιμες και μη αναστρέψιμες. Στην πρώτη περίπτωση τα σημεία και συμπτώματα σχετίζονται με φλεγμονώδη υμενίτιδα και στη δεύτερη, από μη αναστρέψιμες δομικές βλάβες προκαλούμενες από την επιμένουσα υμενίτιδα. Αν και η υμενίτιδα τείνει να ακολουθεί μια κυμαινόμενη πορεία η δομική βλάβη εξελίσσεται σαν γραμμική εξάρτηση από το σύνολο της προηγηθείσας υμενίτιδας. Η αρθρική φλεγμονή στο αρχικό στάδιο μπορεί να έχει ήπια συμπτωματολογία. Τα σημεία της υμενίτιδας μπορεί να είναι αμβληγρά και συχνά είναι υποκειμενικά. Θερμότητα, οίδημα και πόνος στην κίνηση, παρατηρούνται στην ενεργό φάση της φλεγμονώδους υμενίτιδας. Τα συμπτώματα αυτά περιορίζονται σε



επιφανειακές αρθρώσεις με κάψα, εύκολα διατατή, όπως γόνατα, καρποί, μετακαρποφαλαγγικές, μεσοφαλαγγικές αρθρώσεις. Οι καλυπτόμενες από μυϊκές μάζες αρθρώσεις όπως το ισχίο, σπανίως εμφανίζονται με ευαίσθητη διόγκωση, που να είναι εμφανής στη φυσική εξέταση. Το οίδημα της ποδοκνημικής έχει δυσκολίες να προσδιοριστεί, επειδή πολλοί τένοντες και σύνδεσμοι περιορίζουν την αρθρική κάψα. Βεβαίως αν η διόγκωση και ενεργός υμενίτιδα συνοδεύεται από πρωινή δυσκαμψία ωρών, υψηλή ΤΚΕ (ταχύτητα καθίζησης ερυθρών), αναιμία, υψηλή CRP (C-Reactive Protein) και εμφανίζονται διαβρώσεις και προοδευτική καταστροφή σε επαναλαμβανόμενες ακτινογραφίες, πρόκειται για προοδευτικά εξελισσόμενη λειτουργική και ανατομική επιδείνωση. Συχνά μπορεί να μην υπάρχει εμφανής διόγκωση ή υμενίτιδα και να εμφανίζονται από τις επαναλαμβανόμενες ακτινογραφίες διαβρώσεις, σημεία ενεργού νόσου και δομικής βλάβης. Η παρουσία κριγμών, ή υψίσυχνου ήχου, προσδιοριζόμενου με την ψηλάφηση ή την ακοή κατά την αντικειμενική εκτίμηση, και η στένωση του μεσάρθριου διαστήματος από την ακτινολογική εικόνα είναι αντικειμενική απόδειξη της καταστροφής του χόνδρου [2]

### **1.3.3 Εκδηλώσεις από τις αρθρώσεις**

#### **Άκρες χείρες**

Οι καρποί προσβάλλονται σε όλους τους ασθενείς με ΡΑ. Οι μετακαρποφαλαγγικές και οι μεσοφαλαγγικές προσβάλλονται συχνά. Αποτέλεσμα της προσβολής του καρπού μπορεί να είναι η πλήρη ακύλωση. Οι άπω φαλαγγικές δεν προσβάλλονται σχεδόν ποτέ. Μπορεί να παρατηρηθούν παραμορφώσεις στα δάκτυλα δίκην λαιμού κύκνου και παραμορφώσεις δίκην <<κουμπότρυπας>> (boutonniere) [εικ.1] που χαρακτηρίζονται από κάμψη των ΕΜΦ και υπερέκταση των ΑΜΦ. Η παραμόρφωση <<λαιμού κύκνου>> χαρακτηρίζεται με υπερέκταση των κεντρικών και κάμψη των περιφερικών φαλαγγοφαλαγγικών. Πολύ συχνή εκδήλωση είναι το σύνδρομο καρπιαίου σωλήνα, λόγω της υμενίτιδας και πίεσης του μέσου νεύρου. Παρόμοια νευροπάθεια αφορά το ωλένιο νεύρο το οποίο διέρχεται από το σωλήνα του Guyon, μπορεί να διακριθεί από νευροπάθεια παγίδευσης στον αγκώνα από την απουσία αδυναμίας του 5<sup>ου</sup> δακτύλου, όταν κάμπτεται με αντίσταση.

Η τενοντοελυτρίτιδα και ο σχηματισμός ρευματικών οζιδίων στα τενόντια έλυτρα μπορεί να εμποδίζουν την κάμψη των δακτύλων και συχνά να προκαλούν συμπτώματα εκτινασσόμενων δακτύλων.



**Εικόνα 1.1:** παραμόρφωση των ΜΦΦ δίκην λαιμού κύκνου, δίκην κομβιοδόχη, καθώς και ωλένια απόκλιση δεξιάς άκρας χείρας. (Rheumatology text book fifth edition 2011 chapter 82)

### Ισχία

Το ισχίο σπάνια προσβάλλεται στην αρχή. Συχνά η αρχική προσβολή είναι ασυμπτωματική αν και μπορεί να παρατηρηθεί περιορισμός του εύρους κίνησης της άρθρωσης. Ο πόνος εντοπίζεται στην βουβωνική χώρα ή τον μηρό, αλλά μπορεί να γίνει αισθητός και στον γλουτό. Εάν προκληθεί φθορά στον χόνδρο η εξέλιξη είναι περισσότερο γρήγορη από ότι σε άλλες αρθρώσεις.

### Άκροι πόδες-ποδοκνημικές

Οι βλάβες στους άκρους πόδες περιλαμβάνουν υπερξαρθρήματα μεταταρσοφαλαγγικών (ΜΤΦ) σχηματισμών, τήλων στα πέλματα, σύγκαμψη δακτύλων, βλαισός μέγας δάκτυλος (Hallux valgus).

Οι αρθρώσεις των κάτω άκρων, φέρνουν βάρος και η προσβολή προσδίδει μεγαλύτερη δυσλειτουργία και πόνο από την προσβολή των αρθρώσεων των άνω άκρων. Η αρθρίτιδα των ΜΤΦ προκαλεί παραμορφώσεις των δακτύλων των ποδών και υπερξαρθρήματα κεφαλών μεταταρσίων προς το πέλμα. Φλεγμονή της αστραγαλοσκαφοειδικής άρθρωσης προκαλεί πρηγισμό και αναστροφή του άκρου ποδός. Ο ταρσιαίος σωλήνας πίσω από τον έσω σφυρό, που διέρχεται το οπίσθιο κνημιαίο νεύρο συμπιέζεται συχνά από υμενίτιδα με αποτέλεσμα η παγίδευση του νεύρου με καυσalgία, και επιδείνωση στο βάδισμα και την ορθοστασία (σύνδρομο ταρσιαίου σωλήνα)[14].

### **Γόνατα**

Τα γόνατα προσβάλλονται συχνά με συλλογή υγρού, πάχυνση του αρθρικού υμένα και διόγκωση. Η κλινική εξέταση εύκολα αναδεικνύει τα παραπάνω. Παρατηρείται βλαισότητα και πολύ σπάνια ραιβότητα. Συχνά σχηματίζεται οπίσθια κήλη της αρθρικής κάψας με προβολή στην ιγνιακή χώρα, μια ιγνιακή κύστη που ονομάζεται κύστη Baker και μπορεί να διαπιστωθεί υπερηχογραφικά και με αρθρογράφημα. Μπορεί επίσης να παρατηρηθεί ατροφία τετρακέφαλου μυός.

### **Ωμοι**

Η άρθρωση του ώμου προσβάλλεται συχνά στη ΡΑ. Ιδιαίτερα στη Ρευματοειδή όψιμου έναρξης. Μια συλλογή υγρού είναι δύσκολο να εντοπιστεί με τη φυσική εξέταση γιατί η αρθρική κάψα βρίσκεται κάτω από το μυϊκό περίβλημα. Η συμπτωματολογία εκφράζεται με πόνο και περιορισμό στην κίνηση. Η επίταση των συμπτωμάτων οδηγεί σε παγωμένο ώμο. Τα συμπτώματα του παγωμένου ώμου είναι περισσότερο έντονα τη νύχτα, όταν οι κινήσεις πιέζουν την αρθρική κάψα. Το ιδιαίτερο της προσβολής της άρθρωσης αυτής είναι ότι τα συμπτώματα δεν ακολουθούν απαραίτητα πορεία παράλληλη με την καταστροφή του χόνδρου. Πιθανόν επειδή λειτουργεί σχεδόν χωρίς αντίσταση δεν φέρνει βάρος και η λειτουργικότητα του δεν είναι στενά συνυφασμένη με την ακεραιότητα του χόνδρου [15].

### **Αυχένιας**

Ενώ η προσβολή της θωρακικής και οσφυϊκής μοίρας της σπονδυλικής στήλης στην ΡΑ είναι εξαιρετικά σπάνια η προσβολή της αυχενικής μοίρας είναι πολύ συχνή. Πρόκειται όμως για όψιμη προσβολή και όχι στην αρχή της νόσου. Η τενοντοελυτρίτιδα του εγκάρσιου συνδέσμου του Α1 σπονδύλου που σταθεροποιεί την οδοντοειδή απόφυση του Α2 μπορεί να προκαλέσει αστάθεια και υπερξάρθρωμα. Η αυχενική μυελοπάθεια είναι αποτέλεσμα της διάβρωσης της οδοντοειδούς απόφυσης, της χαλάρωσης και ρήξης του συνδέσμου. Η προσβολή των αποφυσιακών αρθρώσεων συμβάλει στην αστάθεια του αυχένα. Γι αυτό είναι χρήσιμη και πολλές φορές σώζει ζωές η προσεκτική κλινική νευρολογική κλινική εξέταση του ασθενούς [16].

### **1.3.4 Εξωαρθρικές εκδηλώσεις**

Σημαντικές εξωαρθρικές εκδηλώσεις ενδέχεται να εμφανισθούν και από άλλα συστήματα με το δεδομένο της συστηματικής φλεγμονής. Πολλοί ασθενείς εμφανίζουν γενικά φαινόμενα όπως πυρετό, καταβολή, κακουχία, εύκολη κόπωση. Κυρίως οι ασθενείς με θετικό ρευματοειδή παράγοντα σχετίζονται με την ανάπτυξη εξωαρθρικών εκδηλώσεων.

Άλλοι παράγοντες κινδύνου που σχετίζονται με την ανάπτυξη εξωαρθρικών εκδηλώσεων είναι η σοβαρή νόσος και πιθανόν η ανεύρεση του αλληλίου τάξης II HLADRB1\* 0401 του μείζονος συστήματος ιστοσυμβατότητας. Οι σημαντικότερες εξωαρθρικές εκδηλώσεις αναλύονται διεξοδικά.

#### **Ρευματικά οζίδια**

Έχουν περιγραφεί σχεδόν σε όλες τις περιοχές του σώματος, μπορεί να εμφανιστούν και στα σπλάχνα. Κυρίως όμως αναπτύσσονται στις εκτατικές επιφάνειες των αντιβραχίων, στους αχίλλειους τένοντες, στην περιοχή των ισχίων, στις καμπτικές επιφάνειες των δακτύλων.

Στα σημεία που υφίστανται τριβή, σχηματίζονται υποδόρια και ενδοδερματικά, αναπτύσσονται κατά σωρούς στην διάρκεια των ενεργών φάσεων της νόσου. Μπορεί να αναπτυχθούν βαθμιαία ή αιφνίδια και σχετίζονται με φλεγμονή [εικ. 1.2].



**Εικόνα 1.2** : Ρευματοειδή οζίδια ωλεκράνου

### Αγγειίτιδα

Είναι συχνότερη στους άνδρες, σχετίζεται με οροθετική ΡΑ, εμφανίζεται σαν περιφερική νευροπάθεια ή με άλλες συστηματικές εκδηλώσεις. Η περισσότερο συχνή είναι η λευκοκυτταροκλαστική αγγειίτιδα με κλινική εμφάνιση ψηλαφητής πορφύρας. Τα κρύα άκρα, η ακροκυάνωση και φαινόμενο Raynaud, παρατηρούνται σπανιότερα. Εξελκώσεις (άτονα έλκη), συνήθως στην κνήμη περιγεγραμμένες με απουσία σκλήρυνσης του δέρματος μπορεί να παρατηρηθεί σε σοβαρή μορφή αγγειίτιδας με κακή πρόγνωση. Αγγειίτιδα εκτός από το δέρμα παρατηρείται σε περιφερικά νεύρα, οφθαλμούς, πνεύμονες και άλλα σπλάχνα. Είναι αξιοσημείωτο ότι δεν παρατηρείται προσβολή των νεφρών, εκτός από σπάνιες και αμφισβητούμενες περιπτώσεις [17].

### Πνευμονικές εκδηλώσεις

Η προσβολή του πνεύμονα στη ΡΑ είναι συχνή. Οι κυριότερες εκδηλώσεις είναι πλευρίτιδα (ξηρά ή υγρά), διάχυτη διάμεση ίνωση, σκληρυντική κυψελίτιδα, μονήρη ή πολλαπλά οζίδια πνευμόνων και υπεζωκότα. Η αποφρακτική πνευμονοπάθεια βρόγχων είναι συχνότερη σε καπνιστές, σε συνύπαρξη με σύνδρομο ξηρότητας (σύνδρομο Sjogren). Η προσβολή των βρογχιολίων με ή χωρίς ηωσινοφιλική πνευμονία ενδέχεται να είναι φαρμακευτικής αιτιολογίας. Τέλος έχει περιγραφεί αγγειίτιδα σε αρτηρίδια πνευμόνων και πνευμονική υπέρταση. Η πνευμονική προσβολή μπορεί να παραμένει για μεγάλο διάστημα ασυμπτωματική, αλλά η θνητότητα από πνευμονική νόσο είναι διπλάσια από εκείνη του γενικού πληθυσμού.

Τα ακτινολογικά ευρήματα αναδεικνύουν ευρήματα διάμεσης ίνωσης, που εντοπίζονται συνήθως στις βάσεις, μονήρη ή πολλαπλά οζίδια στο πνευμονικό παρέγχυμα. Από την αξονική τομογραφία υψηλής διακριτικότητας αναδεικνύεται εικόνα θολής υάλου ή σπηλαιοποίηση των οζιδίων. Έχει παρατηρηθεί να ραγεί υπεζωκοτικό οζίδιο και να δημιουργηθεί βρογχο-υπεζωκοτικό συρίγγιο, που μπορεί να εξελιχθεί σε πνευμοθώρακα ή εμπύημα. Πολύ συχνά η πνευμονική προσβολή δεν οφείλεται σε αυτή καθεαυτή τη νόσο, αλλά στη χρήση των φαρμάκων.

Παλαιότερα η χρήση D-πενικιλλαμίνης και χρυσού και σήμερα η χρήση μεθοτρεξάτης, ενοχοποιούνται για βρογχιολίτιδα. Τα συμπτώματα βελτιώνονται με την διακοπή των φαρμάκων [18].

### **Οφθαλμολογικές εκδηλώσεις**

Πολύ συχνά εμφανίζονται οφθαλμικές εκδηλώσεις είτε στο πλαίσιο της νόσου είτε στο πλαίσιο των επιπλοκών από την χρήση φαρμάκων. Η πιο συχνή προσβολή είναι η ξηρά κερατοεπιπεφυκίτιδα σαν εκδήλωση δευτεροπαθούς συνδρόμου Sjogren. Η επισκληρίτιδα έχει οξεία έκφραση είναι συνήθως καλοήθης δεν επιδρά στην όραση και αυτοιάται σε 2-3 βδομάδες. Η σκληρίτιδα (απλή ή ελκώδης) προκαλεί λέπτυνση του σκληρού χιτώνα και διακρίνεται σε πρόσθια και οπίσθια. Η πρόσθια μπορεί να είναι ελκώδης και νεκρωτική, μπορεί να προκαλέσει διάτρηση του σκληρού και τύφλωση [19]

Η ιρίτιδα και ιριδοκυκλίτιδα γνωστή σαν πρόσθια ραγοειδίτιδα μπορεί να είναι οξεία υποτροπιάζουσα και πολλές φορές χρόνια. Η φαρμακευτική οφθαλμοπάθεια τέλος είναι σημαντική και θα πρέπει ο γιατρός να την έχει πάντα στο νου του και να την υποψιάζεται. Συμβαίνει κυρίως από την χρήση κορτικοειδών και ανθελονοσιακών, μπορεί να είναι αμφιβληστροειδοπάθεια, κερατοπάθεια, καταρράκτης και γλαύκωμα.

### **Καρδιακές εκδηλώσεις**

Οι καρδιακές εκδηλώσεις είναι σχετικά σπάνιες. Προσβάλλονται το περικάρδιο, το μυοκάρδιο και το ενδοκάρδιο. Συμπτωματική περικαρδίτιδα είναι σπάνια, εκδηλώνεται με πόνο στον θώρακα και ταχυκαρδία. Στη φυσική εξέταση ακούγεται περικαρδιακός ήχος τριβής και συνυπάρχουν ηλεκτροκαρδιογραφικά ευρήματα. Χρόνια συμπίεστική περικαρδίτιδα που εκδηλώνεται με οιδήματα και σημεία δεξιάς καρδιακής ανεπάρκειας, αναπτύσσεται σπανιότερα. Διαταραχές αγωγιμότητας έχουν παρατηρηθεί από μυοκαρδίτιδα σαν παροδικές εκδηλώσεις που διαρκούν μερικές εβδομάδες.

Εκδηλώσεις βαλβιδοπάθειας λόγω προσβολής των γλωχίνων από βλάβες όμοιες με ρευματικά οζίδια, εκδηλώνονται σε οροθετικές ρευματοειδείς. Αορτίτιδα (τμηματική ή ολόκληρη) της αορτής έχει περιγραφεί στα πλαίσια της νόσου και εκδηλώνεται με ανεπάρκεια αορτής, με ανεύρυσμα ή ρήξη ανευρίσματος αορτής [18].

### **Νευρολογικές εκδηλώσεις**

Η προσβολή του κεντρικού νευρικού συστήματος είναι συχνή αλλά ήπια. Η περιφερική νευροπάθεια, μπορεί να οφείλεται σε προσβολή των αγγείων των νεύρων (vasa vasorum) και χαρακτηρίζεται σαν πραγματική ρευματοειδής περιφερική νευροπάθεια. Μπορεί να οφείλεται στη φλεγμονή και γεινίαση των νεύρων με τις προσβεβλημένες αρθρώσεις.

Νευροπάθεια που οφείλεται σε «εγκλωβισμό» όπως το σύνδρομο καρπιαίου σωλήνα και νευροπάθεια από μηχανική συμπίεση αυχενικών ριζών που εκδηλώνεται σαν αυχενικό σύνδρομο. Η αυχενική μυελοπάθεια εμφανίζεται βαθμιαία και δεν σχετίζεται με την εμφάνιση ή επίταση της αυχεναλγίας. Κλινικά μπορεί να εμφανίζεται αιφνιδίως παραισθησία τύπου αιμωδίας με την κάμψη του αυχένα.

Σε ασθενείς με βαριά καταστροφική αρθροπάθεια μπορεί να αναπτυχθεί αμφοτερόπλευρη αισθητική παραισθησία και διαταραχές κινητικότητας άκρων. Είναι το πιο σύνηθες εύρημα της αυχενικής μυελοπάθειας. Τέλος το σύνδρομο πολλαπλής μονονευρίτιδας που εμφανίζεται σε ασθενείς με RA, χαρακτηρίζεται από αιφνίδια επίμονη περιφερική νευροπάθεια που δεν μεταβάλλεται από αλλαγές της θέσης της άρθρωσης, ή της βελτίωσης της φλεγμονής. Οι βλάβες αυτές παραπέμπουν σε μηχανισμό αγγειίτιδας, η βιοψία δε του γαστροκνημιαίου (suPAI) νεύρου επιβεβαιώνει την διάγνωση και ενδεχομένως την τροποποίηση της φαρμακευτικής αγωγής [18].

### **Αμυλοείδωση**

Είναι δευτεροπαθής διάχυτη μορφή και το αμυλοειδές είναι τύπου Α. το κυρίαρχο σύμπτωμα είναι η λευκοματουρία. Στην αρχή μπορεί να είναι περιοδική μπορεί να καταλήξει σε πλήρες νεφρωσικό σύνδρομο και νεφρική ανεπάρκεια. Σπανιότερα παρατηρείται διόγκωση ήπατος σπληνός, προσβολή εντέρου με διάρροιες, ή σύνδρομο δυσαπορρόφησης, προσβολή καρδιάς, περιφερικών νεύρων, δέρματος, μυών, θυρεοειδούς. Συνήθως ακολουθεί την βαριά οροθετική καταστροφική RA με διάρκεια νόσου μεγαλύτερη από 5 (πέντε) χρόνια. Η διάγνωση πραγματοποιείται με βιοψία διαφόρων οργάνων κυρίως υποδόριου λίπους, ορθού, νεφρού, ούλων κλπ [20].

### **Σύνδρομο Felty**

Το σύνδρομο περιγράφηκε πρώτα σε ασθενή με RA, σπληνομεγαλία, λευκοπενία και έλκη στις κνήμες. Μεταγενέστερες παρατηρήσεις έχουν δείξει συσχετίσεις με λεμφαδενοπάθεια, θρομβοκυτταροπενία και τον απλότυπο HLA DR4 [21]. Η ακριβής παθολογία της λευκοπενίας ουσιαστικά της ουδετεροπενίας δεν έχει εξηγηθεί επαρκώς. Το σύνδρομο ακολουθεί σοβαρή οροθετική με οζίδια νόσο. Έχει περιγραφεί αυξημένη επίπτωση μη Hodgkin λεμφώματος σε ασθενείς με σύνδρομο Felty [22]. Θεωρείται ως συστηματική εξωαρθρική εκδήλωση της RA με διάρκεια 10-20 χρόνια τουλάχιστον, που εκτός από την διόγκωση σπληνός, υπάρχει γενικευμένη λεμφαδενοπάθεια, διόγκωση ήπατος και γκρίζα-

καφέ χρώση εκτεθειμένων σημείων δέρματος, συχνότερα κοντά στις αρθρώσεις. Τα άτομα αυτά, αναπτύσσουν συχνά σοβαρές λοιμώξεις, επιμολύνσεις δερματικών αλλοιώσεων και λοιμώδη αρθρίτιδα. Η διάγνωση του συνδρόμου Felty είναι εύκολη, όταν συνυπάρχουν μεγαλοσπληνία και ουδετεροπενία. Η διαφορική διάγνωση, εστιάζεται στον αποκλεισμό τυχαίας συνύπαρξης PA και άλλων αιτιών μεγαλοσπληνίας, όπως βακτηριδιακή ενδοκαρδίτιδα, λεμφώματα, κρυπτογενής κίρρωση ήπατος, φυματίωση, χρόνια λεμφογενής λευχαιμία, αμυλοείδωση σπληνός, συγγενή σφαιροκυττάρωση.

## ***1.4 Εργαστηριακά ευρήματα***

Τα εργαστηριακά ευρήματα αφορούν το αίμα, τα ούρα, το αρθρικό υγρό και τον αρθρικό υμένα η παθολογοανατομία του οποίου αναλύθηκε διεξοδικά.

### ***1.4.1 Από το αίμα***

Τα περισσότερα συχνά ευρήματα από το αίμα είναι η αναιμία. Μπορεί να είναι νορμόχρωμη, νορμοκυτταρική στο 50% και μπορεί να οφείλεται σε αμιγώς ερυθροκυτταρική απλασία. Υπόχρωμη, νορμοκυτταρική και μικροκυτταρική ενδέχεται να είναι λόγω απώλειας από το γαστρεντερικό ή λόγω σιδηροπενίας. Η συχνότερη αναιμία είναι η αναιμία της χρόνιας νόσου. Η ΤΚΕ είναι αυξημένη λόγω αύξησης του ινωδογόνου. Και αυξημένη ΤΚΕ μαζί με δείκτες φλεγμονής υποδηλώνει ενεργό νόσο. Υπάρχει αύξηση των πρωτεϊνών οξείας φάσης. Συγκεκριμένα το συστατικό C3 και C4 του συμπληρώματος, το ινωδογόνο, η σερούλοπλασμίνη, η  $\alpha_1$ -αντιθρυψίνη και κυρίως η CRP (C-Reactive Protein), είναι αυξημένες. Η CRP τα τελευταία χρόνια ελέγχεται και θεωρείται ο βασικός δείκτης ενεργότητας της νόσου και οδηγός για θεραπευτικούς χειρισμούς. Στη ηλεκτροφόρηση λευκωμάτων, παρατηρείται αύξηση της  $\alpha_2$  και  $\gamma$  σφαιρίνης, στην ανοσοηλεκτροφόρηση συχνή είναι η αύξηση της IgA, μερικές φορές της IgG. Τα αντιπυρηνικά αντισώματα ανιχνεύονται θετικά σε ποσοστό μεγαλύτερο του 25% με φθορισμό ομοιογενή κυρίως. Τα αντισώματα αντι-SSA (Ro) και αντι-SSB (La), ανιχνεύονται θετικά όταν συνυπάρχει σύνδρομο ξηρότητας. Η ανίχνευση ρευματοειδούς παράγοντα και αντι-CCP αντισωμάτων (έναντι κιτρουλίνης) θα αναπτυχθούν διεξοδικότερα.

### ***1.4.2 Από το Αρθρικό υγρό***

Το αρθρικό υγρό συνήθως είναι θολερό με χαμηλή γλοιότητα (στη δοκιμασία σταγόνας το υγρό στάζει από την σύριγγα σαν το νερό και όχι σχηματίζοντας λεπτή κλωστή). Ο αριθμός



των λευκοκυττάρων από την μικροσκόπηση είναι  $5.0-60 \times 10^9/l$  πολυμορφοπύρρηνα κατά 70-90%. Η παρουσία κυττάρων RA ή <<ραγοκυττάρων>> είναι συχνή. Τα κύτταρα αυτά είναι πολυμορφοπύρρηνα με κυτταροπλασματικά έγκλειστα μετά από χρώση Wright. Φαίνονται σε νωπά παρασκευάσματα και περιέχουν ανοσοσυμπλέγματα. Πολύ σπάνια ανιχνεύονται κρύσταλλοι χοληστερόλης. Οι τιμές του σακχάρου είναι αρκετά χαμηλές και παίρνεται υπόψη το πηλίκο σακχάρου αρθρικού υγρού/σακχάρου αίματος που βρίσκεται στο 0,5-0,8 (φυσ.1). Υπάρχει αύξηση του ολικού ποσού των λευκωμάτων. Η δοκιμασία ανίχνευσης του ρευματοειδή παράγοντα είναι θετική.

### **1.4.3 Από τα Ούρα**

Τα ευρήματα από τα ούρα είναι πτωχά. Μπορεί να ανευρεθεί λευκοματουρία και άλλα ευρήματα νεφρικής βλάβης στην πλειονότητα φαρμακευτικής αιτιολογίας είτε να οφείλεται σε δευτεροπαθή αμυλοείδωση.

## **1.5 Ρευματοειδείς Παράγοντες (ΡΠ)**

Ο ρευματοειδής παράγοντας [rheumatoid factor] (RF), είναι μια ανοσοσφαιρίνη (Ig) η οποία συνδέεται στο σταθερό (Fc) τμήμα της IgG. Διάφοροι ισότυποι των ανοσοσφαιρινών IgE, IgA και IgM μπορεί να δείχνουν δράση RF. Ο ισότυπος IgM ανιχνεύεται ευκολότερα. Στην RA παράγονται αυτοαντισώματα με εξιδίκευση για το Fc τμήμα της ανοσοσφαιρίνης G (immunoglobulin G- IgG). Οι ρευματοειδείς παράγοντες αντίθετα με ότι πιστεύεται δεν είναι ειδικοί για τη RA. μπορεί να ανιχνευθούν σε υγιή, σε ηλικιωμένα άτομα, σε **χρόνιες βακτηριακές λοιμώξεις** (πχ λέπρα, πνευμονική φυματίωση, σύφιλη, ν Lyme, Βρουκέλλωση), σε **ιογενείς λοιμώξεις** (ερυθρά, κυτταρομεγαλοϊός, λοιμώδης μονοπυρήνωση, ινφλουέντζα), **παρασιτικές λοιμώξεις** (τριχινίαση), **χρόνια φλεγμονώδη νοσήματα άγνωστης αιτιολογίας** (σαρκοείδωση, περιοδοντική νόσος, διάμεση πνευμονική νόσος, ηπατική νόσος, μεικτή κρουοσφαιριναιμία, υπεργαμμασφαιριναιμική πορφύρα), **σε ρευματικά νοσήματα εκτός RA** (σ. Sjogren, ρευματικός πυρετός, χρόνια νεανική αρθρίτιδα, αγκυλωτική σπονδυλαρθρίτιδα, σύνδρομο Reiter, ουρική αρθρίτιδα, εξωαρθρικός ρευματισμός, ΣΕΛ, μεικτή νόσος συνδετικού ιστού, σκληρόδερμα), **σε κακοήθειες** (πολλαπλό μυέλωμα, λευχαιμίες) σε **μεταμοσχεύσεις οργάνων** κλπ. Οι μηχανισμοί που εμπλέκονται στην έναρξη της παραγωγής ΡΠ δεν είναι γνωστοί. Εκτιμάται ότι η πολυκλωνική διέγερση και ενεργοποίηση των Β-κυττάρων εμπλέκεται στην πρόιμη παραγωγή ΡΠ. Από τους διεγέρτες περιλαμβάνονται οι λεκτίνες (lectins) και υπεραντιγόνα

(superAntigens). Οι ΡΠ αλληλεπιδρούν με το Fc μόριο της ανοσοσφαιρίνης και όχι με τη θέση τους για τη δέσμευση του αντιγόνου, η οποία μπορεί να καταλαμβάνεται από διαφορετικό αντιγόνο [23].

## ***1.6 Αντισώματα έναντι κιτρουλλίνης anti-CCP (anticyclic citrullinated peptide antibody)***

Τα αντισώματα έναντι μορίων της κιτρουλλίνης είναι αντισώματα που ανιχνεύονται στον ορό αρρώστων με ΡΑ, έχοντας χρησιμοποιηθεί σαν αντιγόνο κυκλικό κιτρουλλινιωμένο πεπτίδιο [24]. Η κιτρουλλίνη είναι ένα αμινοξύ, αποτελεί την απιμιδιωμένη μορφή της αργινίνης. Η μετατροπή της αργινίνης σε κιτρουλλίνη περιλαμβάνει αντικατάσταση μιας ομάδας αμινών από ένα άτομο οξυγόνου στην πλευρική αλυσίδα του αμινοξέους και σχετίζεται με την απώλεια θετικού φορτίου. Τα αντισώματα ανιχνεύονται με ανοσοενζυμική μέθοδο(ELISA). Είναι νέα αντισώματα ειδικά για την ΡΑ και βοηθούν σημαντικά στη διάγνωση της νόσου. Ανιχνεύονται στον ορό αρρώστων με ΡΑ σε ποσοστό μεγαλύτερο του 80%. Έχουν υψηλή ειδικότητα(98%) και βοηθούν να διακρίνουμε τη ΡΑ από άλλες αρθρίτιδες που την μιμούνται. Προβλέπουν καλύτερα την ΡΑ από τον ΡΠ. Είναι χρήσιμα για την πρόγνωση της νόσου.

Όπως έχει φανεί η κιτρουλλινοποίηση των πρωτεϊνών του υμένα είναι μια ενεργής διαδικασία κατά την διάρκεια της φλεγμονής και διάφορες κιτρουλλινικές πρωτεΐνες όπως η ινική μπορούν να βρεθούν στον υμένα ασθενών με ΡΑ. Μαζί με τις κιτρουλλινικές πρωτεΐνες ανιχνεύονται και Β-λεμφοκύτταρα που εκκρίνουν αντικιτρουλλινικά αντισώματα στο αρθρικό υγρο και στον υμένα των ασθενών με ΡΑ που δεν ανευρίσκονται στο περιφερικό αίμα ή στα υγεί κύτταρα [25]. Μπορούν να αναπτυχθούν και σε άλλα ρευματικά νοσήματα όπως παλίνδρομος ρευματισμός, ψωριασική αρθρίτιδα νεανική αρθρίτιδα σε χαμηλούς τίτλους. Γίνεται μεγάλη συζήτηση για το αν σχετίζεται η ύφεση της νόσου με την ελάττωση των τίτλων αντί-CCP αντισωμάτων στον ορό. Μένει να διευκρινιστεί στο μέλλον. Τα αντί-CCP1 και τα αντί-CCP2(δεύτερη γενιά αλλά και τα εντελώς νεότερα αντί-CCP3 που εκτός από IgG αντισώματα ανιχνεύονται και IgA, έχουν μεγάλη ειδικότητα στη ΡΑ και σχετίζονται με βαριές βλάβες της νόσου. Αυτό είναι σημαντικά χρήσιμη πληροφορία γιατί ξέροντας τις διαφορετικές υποομάδες των αντί-CCP στην πρώιμη ΡΑ και τη σχέση τους με τα αλλήλια

των HLA, ιδιαίτερα του DRB1\*0401 και DRB1\*0404 θα παρθούν σωστότερες θεραπευτικές αποφάσεις [26].

## 1.7 Απεικονιστικά ευρήματα

### 1.7.1 Ακτινολογικά ευρήματα με απλές ακτινογραφίες

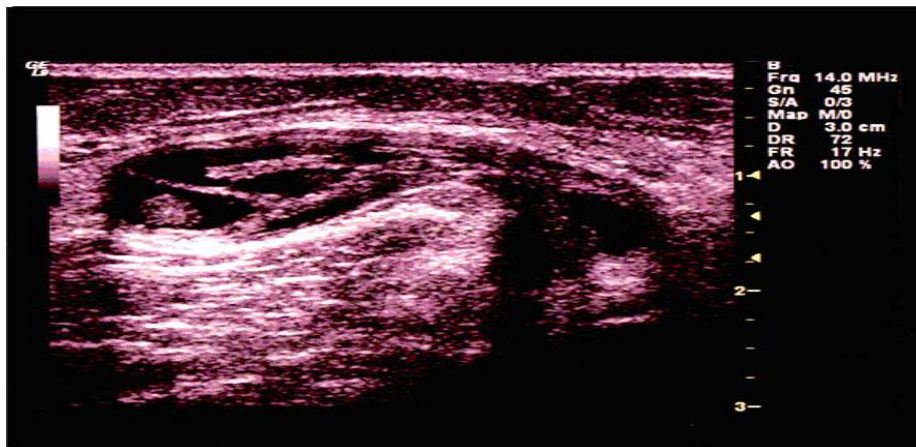
Στη PA υπερέχουν οι οστεοαπορροφητικές βλάβες από τις οστεοπαραγωγικές. Το αντίθετο ισχύει στην οστεοαρθρίτιδα. Δεν προσβάλλονται οι περιφερικές φ.φ. αρθρώσεις των άκρων χειρών. Ο απλός ακτινολογικός έλεγχος συνήθως είναι αρκετός για επιβεβαίωση της διάγνωσης της νόσου. Μια διάβρωση είναι αρκετή να δώσει τη διάγνωση. Υπάρχει μια διαβάθμιση των αλλοιώσεων από τις αρθρώσεις και ταυτόχρονα ακτινολογική ταξινόμηση ως εξής:

**Βαθμός 0:** φυσιολογική άρθρωση.

**Βαθμός I:** Μία από τις παρακάτω αλλοιώσεις:

- i) Περιαρθρική διόγκωση μαλακών μορίων. Ύπαρξη διόγκωσης του ορογόνου θυλάκου της ημιμεμβρανώδους κεφαλής του γαστροκνήμιου μυός (κύστη Baker, η ρήξη της οποίας στα μαλακά μόρια της κνήμης μπορεί να μιμείται θρομβοφλεβίτιδα. (διαγνώσκεται καλύτερα με υπερήχους και αρθρογράφημα).

[εικ. 3]



Εικόνα 1.3: Κύστη Baker με συνοδό υμενίτιδα και διαφραγμάτια.

- ii) Παρααρθρική οστεοπόρωση.
- iii) Ελαφρά στένωση μεσάρθριου διαστήματος.

**Βαθμός II:** Στένωση μεσάρθριου διαστήματος.

Κλασικές περιφερικές διαβρώσεις είναι περισσότερο ανιχνεύσιμες στις μικρές αρθρώσεις των χεριών και των ποδιών. (ΜΚΦ, ΜΦΦ).

**Βαθμός III:** Μετρίου βαθμού καταστροφικές αλλοιώσεις, μεγάλου μεγέθους διαβρώσεις, κυστικές διαβρώσεις, σημαντική στένωση μεσάρθριου διαστήματος.

**Βαθμός IV:** Σοβαρού βαθμού καταστροφικές αλλοιώσεις, οστικές απορροφήσεις, μεγάλου μεγέθους διαβρώσεις παραμορφώσεις οστών, υπερξαρθρήματα.

**Βαθμός V:** Πλήρης εξάλειψη αρθρικών επιφανειών, αγκύλωση, οστικές παραμορφώσεις, απορροφήσεις, υπερξαρθρήματα, εκφυλιστικές αλλοιώσεις. [εικ. 4].



Εικόνα 1.4: Ρευματοειδής αρθρίτιδα με καταστροφή της αρχιτεκτονικής των αρθρώσεων του καρπού, των μετακαρπίων, με υπερξαρθρήματα και ωλένιο απόκλιση.

### **1.7.2 Άλλες απεικονιστικές τεχνικές**

Η υπερηχογραφία δίνει σημαντικές πληροφορίες in vivo της κατάστασης των αρθρώσεων, των συνδέσμων και μάλιστα με τις προσεχείς τριςδιάστατες λήψεις πιστεύεται ότι οι εικόνες θα αποδίδουν εξαιρετικά αποτελέσματα. Χρησιμοποιούνται και στη διαγνωστική και στην παρακολούθηση των ασθενών. Η μαγνητική τομογραφία έχει τα τελευταία χρόνια εισβάλει στην διαγνωστική των νοσημάτων του ερειστικού συστήματος. Στην ΡΑ είναι χρήσιμη για την διάγνωση πρώιμων αλλοιώσεων, εκεί που πιθανόν η απλή ακτινογραφία είναι χωρίς ευρήματα. Χρησιμοποιούνται επίσης στην παρακολούθηση της επιτυχίας της φαρμακευτικής αγωγής με διάφορα φάρμακα.

Το σπινθηρογράφημα χρησιμοποιείται σπάνια και μόνο σε περιπτώσεις αβέβαιης διάγνωσης και φυσιολογικού ακτινολογικού ελέγχου. Η θερμογραφία χρησιμοποιείται παλαιότερα όταν γινόταν καταγραφή με ειδικές συσκευές, και σε ειδικό περιβάλλον, της θερμοκρασίας που εκπέμπεται από το σώμα.

## **1.8 Διάγνωση-Διαγνωστικά κριτήρια- Διαφορική διάγνωση**

### **1.8.1 Διάγνωση**

Η διάγνωση στις πρώτες εβδομάδες έναρξης της νόσου γίνεται κυρίως εξ αποκλεισμού. Εάν υπάρχει συμμετρική υμενίτιδα με ορολογικά ευρήματα RA και δεν είναι λοιμώδης, υπάρχουν ισχυρές ενδείξεις για RA. Εάν επιπλέον υπάρχουν ευρήματα από το αρθρικό υγρό όπως αύξηση κυττάρων, ελαττωμένο σάκχαρο, συνηγορούν υπέρ RA. Οι ακτινογραφικές ενδείξεις RA σε αυτή την φάση δεν είναι εμφανείς και δεν μπορούν να βοηθήσουν στη διάγνωση. Εμφανίζονται αργότερα μετά από μήνες ή χρόνο. Όταν ανευρεθούν ιδίως οι διαβρώσεις πιστοποιούν την νόσο. Το Αμερικανικό Κολέγιο Ρευματολογίας έχει θεσπίσει κριτήρια για τη διάγνωση της ταξινόμησης της σοβαρότητας της νόσου που αναθεωρήθηκαν [27] με σκοπό να καλυφθεί το κοινό της πρώιμης RA [**Πίνακας 1**].

Η διάγνωση της νόσου δεν μπορεί να γίνει με ασφάλεια εάν δεν περάσουν αρκετές εβδομάδες από την έναρξη των συμπτωμάτων. Η συμμετρική αρθρική φλεγμονή και τα ορολογικά ευρήματα είναι δυνατόν να μην είναι εμφανή τους πρώτους μήνες της νόσου. Έτσι χαρακτηρίζεται σαν πιθανή RA. Σε κάποιους ασθενείς ενδέχεται να πρωτοεμφανιστούν εξωαρθρικές εκδηλώσεις. Σε αυτή την περίπτωση η τεκμηρίωση της φλεγμονώδους υμενίτιδας είναι ουσιαστική για την διάγνωση της νόσου.

Η ανεύρεση λευκοκυττάρωσης στο αρθρικό υγρό με αριθμό μεγαλύτερο των 2000/mm<sup>3</sup>, η ιστολογική απόδειξη της υμενίτιδας και η ακτινολογική απόδειξη με χαρακτηριστικές διαβρώσεις τεκμηριώνουν την φλεγμονώδη υμενίτιδα [2].

### **Πίνακας 1.1: Τα Νέα Κριτήρια Ταξινόμησης της Ρευματοειδούς Αρθρίτιδας**

<b>A.Προσβολή άρθρωσης</b>	
1 μεγάλη άρθρωση	0
2-10 μεγάλες αρθρώσεις	1

1-3 μικρές αρθρώσεις	2
4-10 μικρές αρθρώσεις	3
>10 αρθρώσεις (τουλάχιστον 1 μικρή άρθρωση)	5
<b>B. Αυτοαντισώματα</b>	
Αρνητικός ΡΠ και αρνητικό ACPA	0
Θετικός με χαμηλό τίτλο ΡΠ ή ACPA	1
Θετικός με υψηλό τίτλο ΡΠ ή ACPA	2
<b>Γ. Πρωτείνες οξείας φάσης</b>	
Φυσιολογική CRP και φυσιολογική ΤΚΕ	0
Υψηλή CRP ή ΤΚΕ	1
<b>Δ. Διάρκεια συμπτωμάτων</b>	
<6 εβδομάδες	0
>6 εβδομάδες	1
-Οριστική ΡΑ όταν 6/10 κριτήρια στα πεδία Α-Δ.	
-Προσβολή άρθρωσης: διόγκωση (swelling) ή ευαισθησία στην πίεση (tenderness).	
-Μεγάλη άρθρωση : ώμος, αγκώνα, ισχίο, γόνατο, ποδοκνημική άρθρωση.	
-Μικρή άρθρωση: καρπός, μετα-καρποφαλαγγική (1-5), εγγύς φαλαγγική, μετα-ταρσοφαλαγγική (2-5) άρθρωση. Όχι ή άπω φαλαγγική (1-5), η 1 <sup>η</sup> καρπο-μετακάρπιος, η 1 <sup>η</sup> μετατάρσιος άρθρωση.	
-Ρευματοειδής παράγοντας ή αντι-CCP θετικός σε χαμηλό τίτλο:<3x ανώτερο φυσιολογικό όριο.	
-Ρευματοειδής παράγοντας ή αντι-CCP θετικός σε υψηλό τίτλο:>3x ανώτερο φυσιολογικό όριο.	
-Όταν υπάρχει ποιοτικός προσδιορισμός του ρευματοειδούς παράγοντα βαθμολογείται ως θετικός με χαμηλό τίτλο.	

### 1.8.2 Διαγνωστικά κριτήρια

Τα διαγνωστικά κριτήρια παλαιότερα έπαιζαν σημαντικό ρόλο στη διάγνωση, ενώ σήμερα παίζουν ρόλο στην ταξινόμηση και ομογενοποίηση του υλικού για επιδημιολογικές μελέτες. Από τα αναθεωρημένα κριτήρια για τη διάγνωση της νόσου απαιτούνται 5 κριτήρια για την οριστική διάγνωση με διάρκεια 6 εβδομάδων, ενώ είναι δυνατή ή αμφίβολή ΡΑ με 3 κριτήρια διάρκειας 4 εβδομάδων. Τα νέα κριτήρια [27] ουσιαστικά είναι κριτήρια κατάταξης.

Χρησιμοποιείται βαθμολογία που βασίζεται σε αλγόριθμο πρόσθεσης των κατηγοριών Α-Δ. Βαθμολογία > 6/10 είναι αναγκαία για την κατάταξη ασθενούς που έχει οριστική ΡΑ. Ασθενείς με διαβρωτική νόσο χαρακτηριστική της ΡΑ και ιστορικό συμβατό με την προηγούμενη εκπλήρωση κριτηρίων του 2010, πρέπει να χαρακτηριστούν ως ΡΑ. Ασθενείς με μακροχρόνια νόσο συμπεριλαμβανομένων και εκείνων των οποίων η νόσος είναι ανενεργός (με ή χωρίς θεραπεία) με βάση τα διαθέσιμα στοιχεία εκ των υστέρων, έχουν προηγουμένως εκπληρώσει τα κριτήρια του 2010, θα πρέπει να χαρακτηριστούν ΡΑ. Παρά το γεγονός ότι ασθενείς με βαθμολογία κάτω του 6/10 δεν μπορούν να καταταγούν ως ΡΑ, το καθεστώς τους μπορεί να αξιολογηθεί εκ νέου. Τα κριτήρια θα μπορούσαν να πληρούνται σωρευτικά με την πάροδο του χρόνου. Η χρήση αυτών των κριτηρίων, προβλέπει εκτός από την ταξινόμηση των ασθενών με ΡΑ, να αξιολογηθούν οι πρόωρες μορφές ΡΑ και να δημιουργηθούν θεραπευτικές στρατηγικές.

## **1.9 Πρόγνωση- θεραπεία**

### **1.9.1 Πρόγνωση**

Η ΡΑ είναι μια πολύμορφη νόσος με ευμετάβλητη έκφραση. Ένα μεγάλο ποσοστό ασθενών περισσότερο από 50% έχει βαθμιαία έναρξη των συμπτωμάτων. Το υπόλοιπο θα έχει μονοκυκλική πορεία που θα υφεθεί εντός διαιτίας και πολυκυκλική ή προοδευτική πορεία. Η μακροχρόνια πρόγνωση των ασθενών με αιφνίδια έναρξη και εκείνων με βαθμιαία είναι η ίδια. Εάν οι ασθενείς παραμείνουν χωρίς θεραπευτική παρέμβαση βαδίζουν σίγουρα σε προοδευτική μόνιμη αναπηρία. Όταν εμπλακούν εξωαρθρικές εκδηλώσεις και η νόσος αποκτήσει συστηματική προσβολή, μειώνεται το προσδόκιμο επιβίωσης. Σημαντικό ρόλο στην πιθανή πορεία της νόσου παίζουν οι γενετικοί παράγοντες το χαμηλό κοινωνικοοικονομικό και μορφωτικό επίπεδο, η σωστή εξατομικευμένη θεραπεία, η συμμόρφωση στη θεραπεία, η προθυμία των ασχολούμενων με την παροχή υπηρεσιών υγείας, η εκπαίδευση των ασθενών, η συνεργασία των ρευματολόγων με άλλους ειδικούς στη θεραπεία της ΡΑ.

### **1.9.2 Θεραπεία**

#### **Αρχές θεραπείας μη φαρμακευτικής προσέγγισης**

Οι βασικοί στόχοι της θεραπείας είναι: ανακούφιση από τον πόνο, το οίδημα και την κόπωση, βελτίωση της λειτουργικότητας της άρθρωσης, τερματισμός της αρθρικής

φλεγμονής και βλάβης των αρθρώσεων, πρόληψη της αναπηρίας. Οι στόχοι παραμένουν σε όλη την πορεία της νόσου. Η εκπαίδευση των ασθενών είναι αναγκαία και συνεχής για την πορεία της νόσου, τις επαγγελματικές και μη υποχρεώσεις, τον τρόπο ζωής, την οικογενειακή κατάσταση. Δίδονται συμβουλές για αύξηση αυτοεκτίμησης, για τις αλλαγές στην κατοικία και στον τρόπο διαβίωσης. Αυτό γίνεται με συμμετοχή ρευματολόγων και άλλων ειδικών περιλαμβανομένων νοσηλευτών, συμβούλων, εργασιοθεραπευτών και φυσικοθεραπευτών. Η εργασιοθεραπεία και η φυσικοθεραπεία παίζουν σημαντικό ρόλο στην βελτίωση της λειτουργικότητας των αρθρώσεων με χρήση ναρθήκων, βοηθητικών μηχανημάτων, κατάλληλων υποδημάτων και άλλων προσαρμοστικών μηχανισμών. Οι ασκήσεις πλήρους εύρους κινήσεων αλλά και ενίσχυσης μυών βοηθούν σημαντικά. Επίσης η ακινητοποίηση και η ανάπαυση της άρθρωσης. Τέλος η χειρουργική θεραπεία που περιλαμβάνει ολικές αρθροπλαστικές επανορθώσεις και μεταθέσεις τενόντων, απελευθέρωση μέσου νεύρου στον καρπιαίο σωλήνα και σταθεροποίηση αρθρώσεων έχουν βελτιώσει την λειτουργικότητα, τον πόνο και την ποιότητα ζωής των ασθενών. [28].

### **Φαρμακευτική θεραπεία**

Η φαρμακευτική θεραπεία της ΡΑ στην πορεία των χρόνων εμφανίζεται με διαφορετικές προσεγγίσεις. Άρχισε την δεκαετία του 30 σαν μονοθεραπεία με προσπάθεια ανακούφισης από τον πόνο. Από την δεκαετία του 60 και εντεύθεν προτάθηκε η θεραπευτική πυραμίδα όπου την βάση αποτελούσαν τα σαλικυλικά και αργότερα τα μη στεροειδή αντιφλεγμονώδη. Δειλά- δειλά άρχισαν να χορηγούνται τα κορτικοστεροειδή, που είχαν άμεση δράση στη φλεγμονή, αλλά δεν γνώριζαν την πρόποσα δοσολογία και τις παρενέργειες, ενώ λίγο αργότερα, ακολούθησαν τα τροποποιητικά φάρμακα. Στη δεκαετία του 80 προτάθηκε η αναστροφή της πυραμίδας, χορηγήθηκαν από την αρχή τροποποιητικά της νόσου φάρμακα DMARDS (disease-modifying-antirheumatic- drugs) μόνα ή σε συνδυασμό με ταυτόχρονη χορήγηση χαμηλής δόσης στεροειδών. Οι θεραπείες με τα συνδυαζόμενα τροποποιητικά δύο και τρία πολλές φορές χαρακτηρίστηκαν επιθετικές θεραπείες. Στα τέλη του 90 αρχές του 2000 άρχισε η θεραπεία με βιολογικούς τροποποιητές (μονοκλωνικά αντισώματα). Η πολυετής πείρα από την χρήση αυτών των φαρμάκων φαίνεται να αλλάζει το τοπίο. Όλο και περισσότερα φάρμακα προστίθενται στην θεραπευτική φαρέτρα. Καθώς η θεραπεία της ΡΑ γίνεται περισσότερο αποτελεσματική και περίπλοκη, απαιτείται αυξημένη προσοχή και εγρήγορση από τις ανεπιθύμητες ενέργειες των φαρμάκων, το κόστος, σε σχέση με την ύφεση της νόσου [29].



### **A. Μη στεροειδή αντιφλεγμονώδη φάρμακα (ΜΣΑΦ)**

Τα φάρμακα αυτά έχουν παρελθόν στη θεραπεία της ΡΑ συμπεριλαμβανομένων και των σαλικυλικών. Ανακουφίζουν από τον πόνο, μειώνουν τη φλεγμονή, αλλά δεν προφυλάσσουν από την οστική βλάβη και ούτε μπορούν να εξαλείψουν τα σημεία και τα συμπτώματα της ενεργού αρθρίτιδας. Αναστέλλουν τη δράση του ενός ή και των δύο ισομορφών του ενζύμου κυκλοοξυγενάσης (cyclooxygenase-COX). Το ισόενζυμο COX-1 εκφράζεται φυσιολογικά στο γαστρεντερικό βλεννογόνο, τα νεφρά, τα αιμοπετάλια και το ενδοθήλιο των αγγείων. Το ισόενζυμο COX-2 εκφράζεται επαγόμενο και προάγει την παραγωγή των προσταγλανδινών στους φλεγμαίνοντες ιστούς. Έχουν δημιουργηθεί εκλεκτικοί COX-2 αναστολείς φάρμακα τελευταίας γενιάς με λιγότερες παρενέργειες από το γαστρεντερικό, αλλά με αποτελεσματικότητα που δεν διαφέρει από αυτή των κλασικών ΜΣΑΦ. Η χρήση τους στη ΡΑ έχει περιοριστεί και υπάρχει η τάση να μην χρησιμοποιούνται.

### **B. Γλυκοκορτικοειδή**

Τα γλυκοκορτικοειδή χρησιμοποιήθηκαν και συνεχίζουν να χρησιμοποιούνται στη θεραπεία της ΡΑ. Είναι αποτελεσματικά στον έλεγχο του πόνου και της φλεγμονής, με ισχυρή κατασταλτική δράση αλλά δεν μπορούν να χορηγηθούν σαν μονοθεραπεία. Τα τελευταία χρόνια υπάρχει η τάση να χρησιμοποιούνται σε μικρές δόσεις μέχρι 10- 15mg για μικρό διάστημα, μέχρι να δράσουν τα DMARDS [30]. Σε μεγαλύτερες δόσεις χρησιμοποιούνται στην αγγειίτιδα της ΡΑ. Χορηγούνται επίσης υπό μορφή ενδοαρθρικών εγχύσεων στην έντονη φλεγμονή και τον εντοπισμένο πόνο σε επίμονη αρθρίτιδα, με καλά αποτελέσματα.

### **Γ. Τροποποιητικά της νόσου φάρμακα**

Αρκετά φάρμακα αυτής της κατηγορίας έχουν δοκιμασθεί κατά το παρελθόν με μηχανισμούς δράσης καλά γνωστούς, αλλά και λιγότερο κατανοητούς όπως της υδροξυχλωροκίνης και των αλάτων χρυσού. Το κυρίαρχο φάρμακο αυτής της κατηγορίας είναι η **μεθοτρεξάτη** (MTX), που συνεχίζει να χορηγείται και σήμερα και είναι απαραίτητη σε όλες τις θεραπείες με δόσεις 7,5-10mg αρχικά και maximum μέχρι 20-25mg εβδομαδιαίως. Θα πρέπει να χρησιμοποιείται όπως και όλα τα DMARDS όταν η διάγνωση είναι βέβαιη και πριν εμφανισθούν διαβρώσεις στις ακτινογραφίες. Η μεθοτρεξάτη σαν μονοθεραπεία παρέχει 60-70% αντικειμενική βελτίωση, χρησιμοποιείται σε συνδυασμό με κορτικοειδή ή άλλα

DMARDS συμπεριλαμβανομένων και των βιολογικών τροποποιητών [31]. Χορηγείται από το στόμα μια φορά την εβδομάδα ή παρεντερικά για μικρότερη συχνότητα εμφάνισης παρενεργειών, κυρίως στοματίτιδας και γαστρεντερικών διαταραχών. Άλλες διαταραχές όπως η ηπατοτοξικότητα, η πνευμονική ίνωση, η κίρρωση ήπατος παρατηρούνται από την χρήση της MTX, για το λόγο αυτό οι ασθενείς δεν θα πρέπει να κάνουν χρήση οινοπνευματωδών ποτών. Παραμένει η κύρια θεραπεία σε όλα τα ρευματολογικά κέντρα με προσθήκη κυρίως των βιολογικών τροποποιητών.

Άλλο φάρμακο με ισοδύναμη κλινική αποτελεσματικότητα είναι η **Λεφλουνομίδη (ARVA)**, που είναι αναστολέας της σύνθεσης της πυριμιδίνης. Από τη χρήση της λεφλουνομίδης μπορεί να παρατηρηθεί ηπατοτοξικότητα, αλλεργικό εξάνθημα, απώλεια βάρους, θρομβοκυτταροπενία και διάρροια. Συνχορηγείται σήμερα και με βιολογικούς τροποποιητές. Η **κυκλοσπορίνη** γνωστή από την μεταμόσχευση χρησιμοποιείται στη θεραπεία της PA σε μικρές δόσεις (2,5-5mg/kg/ημέρα) σαν μονοθεραπεία σε συνδυασμό με την MTX ή βιολογικούς τροποποιητές. Η **Αζαθειοπρίνη** χορηγείται σε ασθενείς με ήπια ή σοβαρή ανθεκτική νόσο, με σοβαρές εξωαρθρικές εκδηλώσεις.

Η **κυκλοφωσφαμίδη** έχει περιορισμένη χρήση στη PA. Χορηγείται στην αντιμετώπιση της συστηματικής αγγειίτιδας που ανθίσταται στη χρήση κορτικοειδών. Η **υδροξυχλωροκίνη** χορηγείται στη ήπια και οροαρνητική PA, είναι το φάρμακο πρώτης εκλογής σε πρόσφατης έναρξης ελαφρά νόσο ή όταν η διάγνωση ακόμη αμφιλέγεται. Η αμφιβληστροειδοπάθεια είναι σπάνια επιπλοκή και εμφανίζεται όταν δεν χρησιμοποιείται η κατάλληλη δόση. Η **σουλφασαλαζίνη** χρησιμοποιείται σαν θεραπεία σε πρώιμη έναρξη νόσου, με αρχική δόση 500-1000mg σταδιακά αυξανόμενη σε 1500mg δύο φορές ημερησίως. Χρησιμοποιείται σε συνδυασμό με άλλα φάρμακα και βιολογικούς τροποποιητές. Άλλα φάρμακα που χρησιμοποιήθηκαν είναι τα άλατα χρυσού, η D-πενικιλλαμίνη, η μινουκυκλίνη και η χλωραμβουκίλη που σήμερα δεν χρησιμοποιούνται.

#### **Λ. Βιολογικοί τροποποιητές**

Τα τελευταία χρόνια τα θεραπευτικά αδιέξοδα στην θεραπεία της PA έφερε στο προσκήνιο στοχευμένες θεραπείες εκτός των κλασικών τροποποιητικών φαρμάκων τις λεγόμενες βιολογικές θεραπείες (βιολογικά νοσοτροποποιητικά φάρμακα) που είναι μονοκλωνικά αντισώματα. Περιλαμβάνουν τους παλαιούς αντι-TNFα παράγοντες (infliximab, etanesept, adalimumab) τον ανταγωνιστή του υποδοχέα της IL-1 (anakinPA), τον αναστολέα της διέγερσης του T-λεμφοκυττάρου (CTL4-Ig, abadacept), τον παράγοντα αντι-B

λεμφοκυττάρων (αντι-CD20, rituximab), τον ανταγωνιστή του υποδοχέα IL-6R (IL-6PAtocilizumab) και άλλους δύο τελευταίας γενιάς αντι-TNFα τροποποιητές το golimumab και certolizumab Pegal. Σε μελέτες φάσης III βρίσκονται και άλλα μονοκλωνικά αντισώματα, με δεδομένες τις ανάγκες και τα θεραπευτικά προβλήματα στη ΡΑ. Η εικοσαεκαετής και πλέον χρήση του αντι-TNFα μας έδειξε ότι αυτή η ομάδα έπαιξε και συνεχίζει να παίζει σημαντικό ρόλο στην αλλαγή της θεραπευτικής προσέγγισης. Ο TNFα κατέχει κεντρικό ρόλο στη φλεγμονώδη διεργασία, μπορεί να είναι διαλυτός ή διαμεμβρανικός. Έχει τριπλή δράση: στους οστεοκλάστες απευθείας ή μέσω της αλληλεπίδρασης PANK-PANKL, προκαλώντας απορρόφηση και διάβρωση του οστού, στον αρθρικό υμένα προκαλώντας φλεγμονή πόνο και οίδημα και στα χονδροκύτταρα, προκαλώντας αλλοίωση του χόνδρου. Ο TNFα ενεργοποιεί την παραγωγή και άλλων προφλεγμονωδών κυτταροκινών όπως η ιντερλευκίνη-1 (IL-1), IL-6, IL-8, GM-CSF. Η πολυετής πείρα από την χρήση αυτών των φαρμάκων συνεχώς εμπλουτίζεται με καινούρια δεδομένα από την συνεχή χορήγηση και παρακολούθηση [32]. Άλλα μονοκλωνικά πρώτης γραμμής χορήγησης εκτός των αντι-TNFα είναι ο ανασυνδιασμένος ανταγωνιστής του υποδοχέα IL-1R, και IL-6. Ο ανασυνδυασμένος αναστολέας της ενεργοποίησης των CTLA-4 των T-λεμφοκυττάρων (Abatacept, CTLA-4IgG), ανθρώπινη πρωτεΐνη του CTLA-4 με IgG1 που ανταγωνίζεται το CD28 στη σύνδεση του με το CD80/86 έχει ένδειξη για τη θεραπεία της ΡΑ σε συνδυασμό με μεθοτρεξάτη. Το CTLA-4 αναστέλλει ευθέως το σχηματισμό οστεοκλαστών με αποτέλεσμα την ελάττωση των διαβρώσεων της νόσου [33].

Το μονοκλωνικό αντίσωμα έναντι του CD20 των B-λεμφοκυττάρων (Rituximab, anti-CD20) είναι χμιαρικό μονοκλωνικό αντίσωμα έναντι του CD20 μορίου επιφάνειας του B-λεμφοκυττάρου. Έχει ένδειξη στη θεραπεία της ΡΑ ανθεκτικής σε anti-TNFα θεραπεία, συνχορηγούμενο με μεθοτρεξάτη [34].

Για να χορηγηθούν τα μονοκλωνικά αντισώματα υπάρχουν κανόνες και ενδείξεις που πρέπει να τηρούνται. Έχουν θεσπιστεί από το ACR, την EULAR και την ελληνική ρευματολογική εταιρία. Χορηγούνται ως μονοθεραπεία ή σε συνδυασμό με μεθοτρεξάτη, λεφλουνομίδη και κυκλοσπορίνη, αλλά απαγορεύεται η συγχορήγηση μεταξύ τους [35]. Τα φάρμακα αυτά μπορεί να χορηγηθούν και στην πρώιμη ΡΑ με ορισμένες προϋποθέσεις. Να είναι πολυαρθρική μορφή με υψηλή ΤΚΕ, υψηλή CRP, αυξημένους τίτλους RF και anti-CCP και με παρουσία HLA-DRB1\*0401 ή HLA-DRB\*404. Δεν πρέπει να χορηγούνται σε κύηση, γαλουχία, ενεργό λοίμωξη, καρδιακή ανεπάρκεια, σε περίπτωση αρθροπλαστικής που έχει επιπλακεί με λοίμωξη εντός των τελευταίων μηνών, σε ασθενείς με HIV, HBV, HCV

λοιμώξη, σε ασθενείς με νεοπλασματική νόσο και φυματίωση. Θα πρέπει ο γιατρός να βρίσκεται σε συνεχή εγρήγορση και να παρακολουθεί τακτικά τον ασθενή για το ενδεχόμενο λοιμώξεων, TBC, κακοήθειας, απομυελινωτικής νόσου και άλλων επιπλοκών. Τα φάρμακα αυτά συνδυάζονται με DMARDS ή μόνα τους έχουν μοναδικό στόχο να ελέγξουν την φλεγμονή και να αποτρέψουν τις διαβρώσεις και την αρθρική καταστροφή [36].

# 2

## *Εφαρμογές Μηχανικής Μάθησης στην Ρευματοειδή Αρθρίτιδα & Συναφείς Μελέτες*

### *2.1 Η Μηχανική Μάθηση στην Ρευματολογία*

Τις τελευταίες τρεις δεκαετίες παρατηρείται μια έκρηξη της γνώσης σε όλους τους τομείς της ιατρικής. Η αυξανόμενη χρήση της επιστήμης των υπολογιστών, η μείωση των τιμών στις βιοτεχνολογικές μεθόδους, η βελτιωμένη κατανόηση των βιοδεικτών και των μεθόδων γενετικής αλληλουχίας, δημιουργούν απεριόριστες νέες δυνατότητες. Με την συνεχώς αυξανόμενη ροή επιστημονικών δεδομένων, η γνώση μας αυξάνεται ραγδαία αλλά επίσης αναδεικνύεται μια επαύξηση της πολυπλοκότητας [37]. Στο πλαίσιο αυτό γίνονται σημαντικές προσπάθειες να χρησιμοποιηθούν οι πόροι μάθησης που αντιπροσωπεύουν αυτά τα δεδομένα για την εκπαίδευση μοντέλων μηχανικής μάθησης για τη διάγνωση, πρόβλεψη και θεραπευτική προσέγγιση περισσότερο ή λιγότερο περίπλοκων παθολογιών [38]–[40].

Η ανάπτυξη των μεγάλων δεδομένων στην κλινική έρευνα και οι εξελίξεις στις υπολογιστικές προσεγγίσεις έχουν ανοίξει νέους δρόμους ιδιαίτερα για τη μελέτη πολύπλοκων ασθενειών όπως η ρευματοειδής αρθρίτιδα. Οι περισσότερες ρευματικές παθήσεις όπως και η ρευματοειδής αρθρίτιδα (ΡΑ) είναι χρόνιες και παρουσιάζουν διάφορες διακυμάνσεις, οι οποίες περιλαμβάνουν πολύπλοκη υποβόσκουσα παθοφυσιολογία που περιπλέκει την θεραπεία. Παρά την πρόοδο που έχει επιτευχθεί με στοχευμένες βιολογικές και συνθετικές θεραπείες, η παρατεταμένη ύφεση της ρευματοειδούς αρθρίτιδας επιτυγχάνεται σε μια μικρή μειοψηφία των ασθενών. Η μεγαλύτερη σήμερα πρόκληση για τους ασθενείς ΡΑ είναι η εφαρμογή μιας προσωποποιημένης προσέγγισης. Για το σκοπό

αυτό απαιτούνται μελέτες μεγάλου όγκου δεδομένων για την εύρεση των βέλτιστων θεραπευτικών στρατηγικών και προγνώσεων για μεμονωμένους ασθενείς [41], [42].

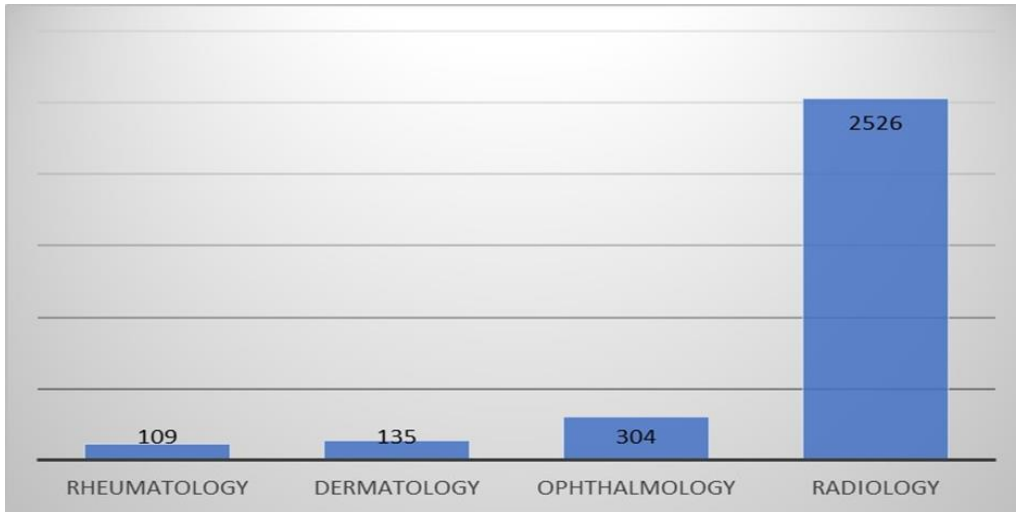
Η μηχανική μάθηση μπορεί να συμβάλει στην καλύτερη διαχείριση της ΡΑ, να μειώσει το χρόνο που απαιτείται για την λήψη αποφάσεων ως προς την θεραπευτική προσέγγιση, να μειώσει το φυσική και οικονομική επιβάρυνση του ασθενούς αλλά και του συστήματος υγείας γενικότερα. [43].

Πρόσφατες μελέτες υποστηρίζουν ότι στο άμεσο μέλλον η μηχανική μάθηση θα συνδράμει τους ρευματολόγους στην πρόβλεψη της εξέλιξης της ΡΑ και στον προσδιορισμό των κρίσιμων παραγόντων για την πορεία της νόσου. Είναι πολύ πιθανόν η μηχανική μάθηση να μπορεί να προτείνει την φαρμακευτική αγωγή και να εκτιμήσει το προσδοκώμενο αποτέλεσμα. Έτσι η λήψη αποφάσεων δεν θα στηρίζεται αποκλειστικά στην εμπειρία του γιατρού για την ερμηνεία των διάφορων ευρημάτων και την αυτοαξιολόγηση του ασθενούς αλλά θα μπορεί επίσης να επηρεάζεται από τα επιπλέον στοιχεία που θα προκύπτουν από μοντέλα μηχανικής μάθησης [37].

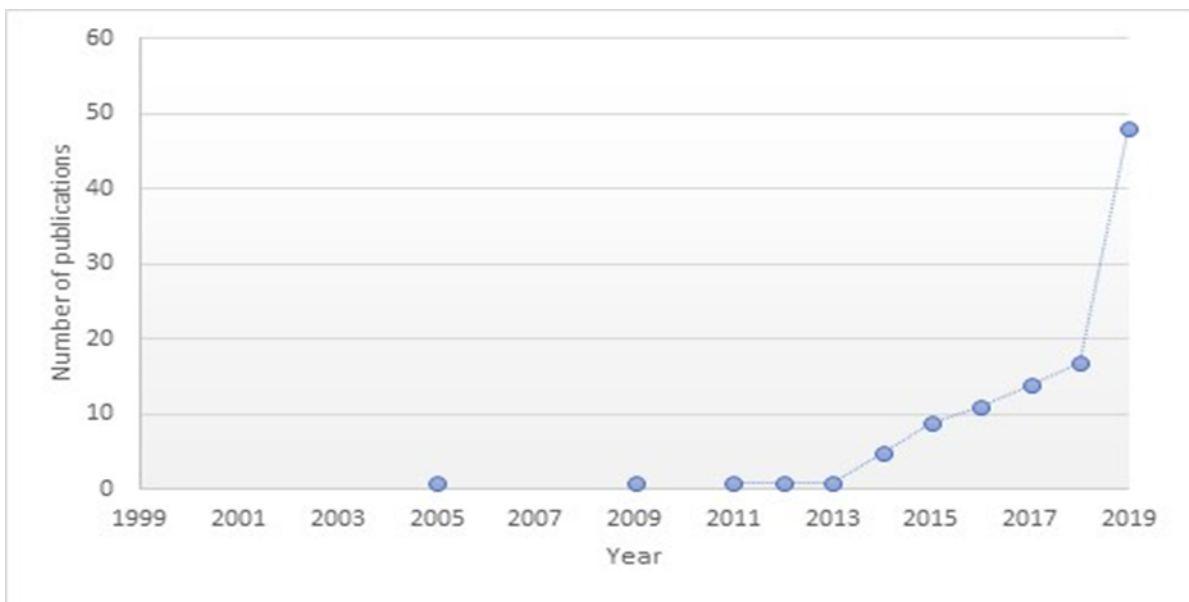
Μεγάλος αριθμός πρόσφατων μελετών επιβεβαιώνουν ότι η μηχανική μάθηση, ως κλάδος της τεχνητής νοημοσύνης, μπορεί να διαδραματίσει σημαντικό ρόλο στην ανάπτυξη εξατομικευμένης ιατρικής για την ΡΑ και να συμβάλλει στην εξατομικευμένη θεραπεία, παρέχοντας στα υπολογιστικά συστήματα την δυνατότητα να μάθουν εμπειρικά χωρίς κανόνες που να έχουν προσδιοριστεί με απόλυτη σαφήνεια από τον άνθρωπο. Επιπλέον, η βαθιά μάθηση ως κλάδος της μηχανικής μάθησης εκτός από την δυνατότητα να βελτιώσει την ποιότητα της διάγνωσης, της θεραπείας και της φροντίδας των ασθενών, έχει επίσης την δυνατότητα να μειώσει τα σχετικά κόστη [41], [42].

Η μηχανική μάθηση έχει ήδη πολύ σημαντικές εφαρμογές στην ιατρική όπως στις ιατρικές απεικονιστικές μεθόδους, στις συσκευές προσομοίωσης εγκεφάλου και στην πρόγνωση του καρκίνου, ενώ ο τομέας της ρευματολογίας ακολουθεί [44]. Στα Σχήματα 2.1 έως 2.3, παρουσιάζονται συγκριτικά αποτελέσματα για σχετικές μελέτες σε διάφορους τομείς της ιατρικής για την περίοδο 1999 και 2019 (Pubmed.org), στο οποίο διαπιστώνουμε ότι παρότι στην ρευματολογία έχουμε συγκριτικά λιγότερες σχετικές μελέτες με άλλους τομείς της ιατρικής, υπάρχει συνεχής αυξητική τάση [42].

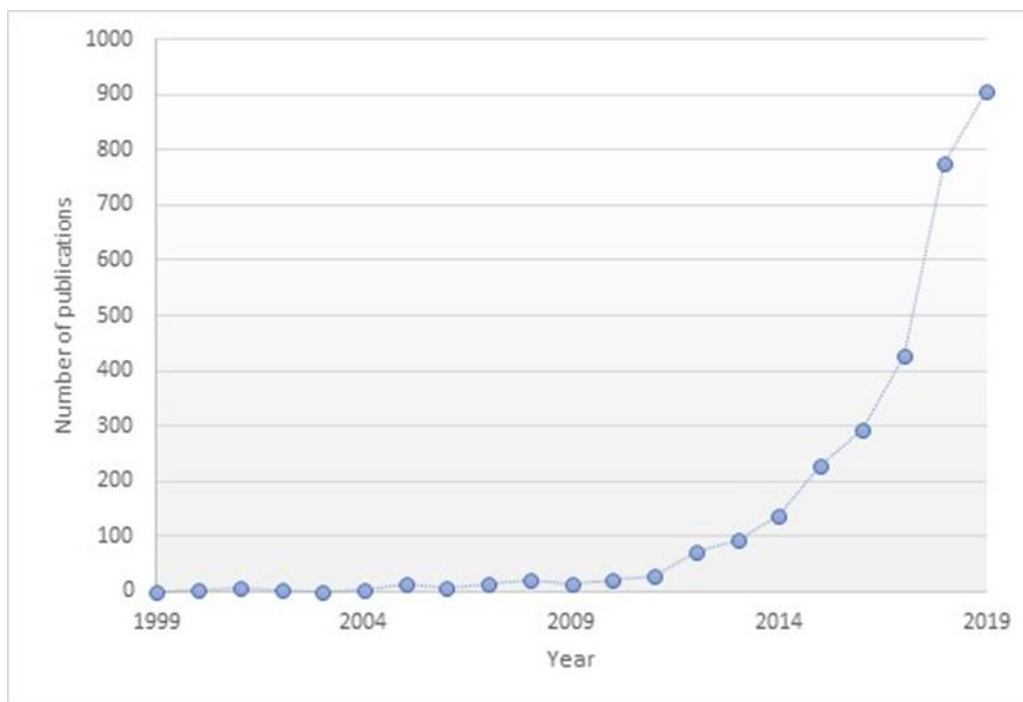
Ειδικότερα, σε ότι αφορά τον τομέα της διάγνωσης έχει υπάρξει μια έκρηξη του αριθμού των άρθρων για ακριβέστερες διαγνώσεις στις ρευματολογικές ασθένειες, δεδομένου ότι τουλάχιστον δύο άρθρα δημοσιεύονται ετησίως για αυτό το θέμα τα τελευταία χρόνια [45].



**Σχήμα 2.1:** Αριθμός μελετών μηχανικής μάθησης σε διάφορους τομείς της ιατρικής (1999-2019) [42]



**Σχήμα 2.2:** Συνολικός αριθμός δημοσιεύσεων σχετικών με την μηχανική μάθηση στην ρευματολογία [42]



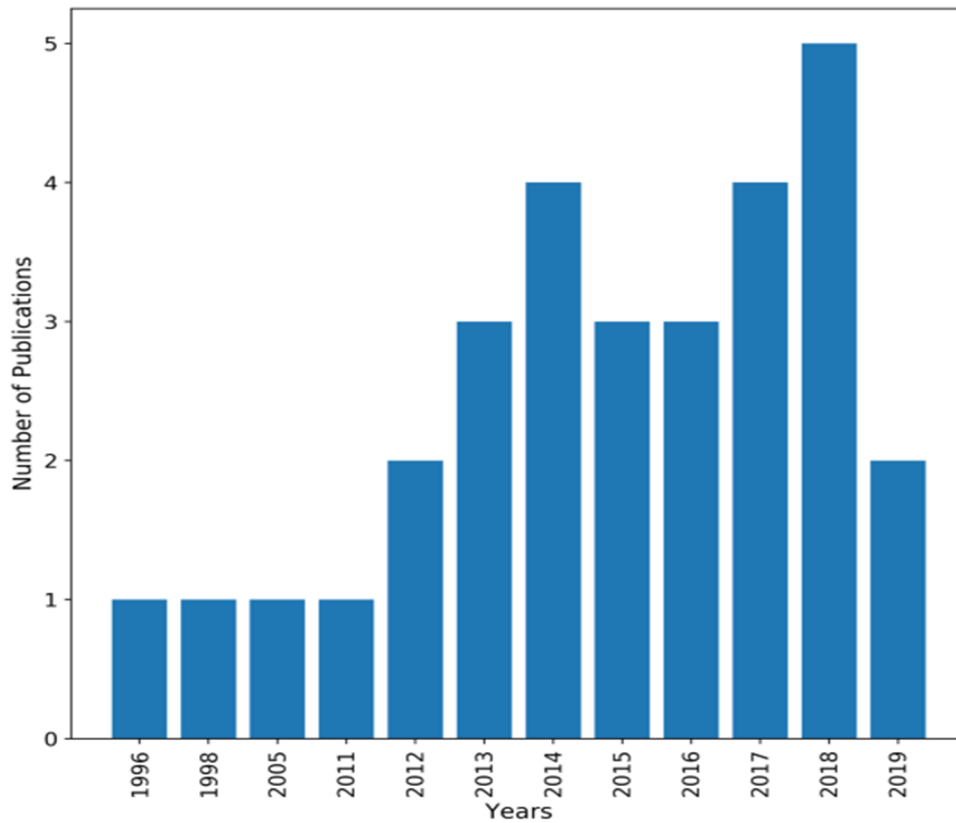
**Σχήμα 2. 3:** Συνολικός αριθμός μελετών μηχανικής μάθησης στην ιατρική [39]

Σε μια πρόσφατα δημοσιευμένη μελέτη των [45] αξιοποιώντας το λογισμικό BIBOT πραγματοποίησαν αναζήτηση σε μεγάλο αριθμό περιλήψεων ερευνητικών δημοσιεύσεων στον τομέα της ρευματολογίας με αυτόματη εξαγωγή λέξεων-κλειδιών που υπάρχουν σε αυτές τις περιλήψεις και πραγματοποίησαν μια ανάλυση της κατανομής τους. Η έρευνα αυτή αποκαλύπτει ότι η πιο μελετημένη παθολογία μέσω προσεγγίσεων μηχανικής μάθησης είναι η οστεοαρθρίτιδα. (Σχήμα 2.4, 2.6).

Δεν προκαλεί έκπληξη το γεγονός ότι η παθολογία αυτή έχει μελετηθεί περισσότερο μεταξύ των ρευματικών παθήσεων, καθώς χρησιμοποιούνται για τη διάγνωση κυρίως απεικονιστικές μέθοδοι και η μηχανική μάθηση έχει τεράστια επιτυχία στην επεξεργασία και ταξινόμηση εικόνας για πάνω από δέκα χρόνια (Σχήμα 2.1.).

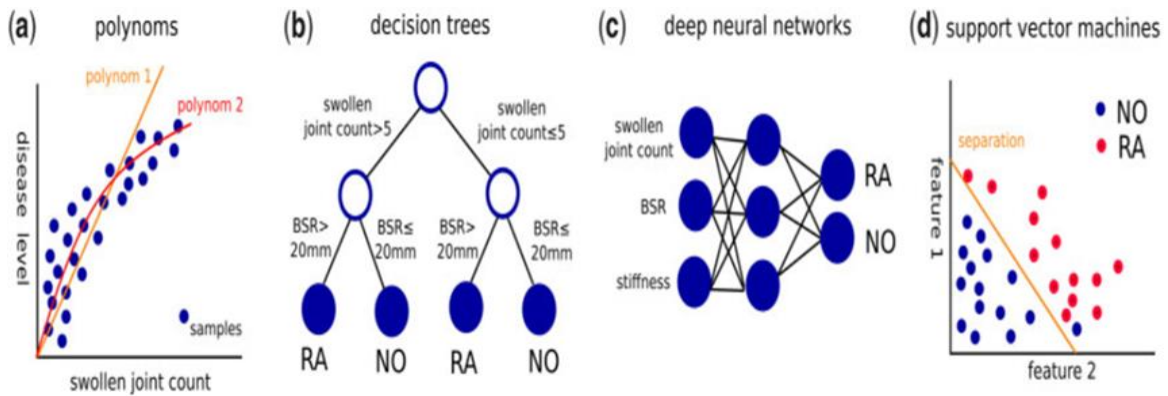
Παρόλα αυτά, ο αριθμός μελετών είναι περιορισμένος και υπάρχει μεγάλο έδαφος για περαιτέρω έρευνα στον τομέα. Οι διαφορές που παρατηρούνται σε σχέση με τα στοιχεία της μελέτης των Kruche και συνεργατών ενδεχομένως προκύπτουν από τον διαφορετικό τρόπο προσέγγισης και πραγματοποίησης της κάθε έρευνας.



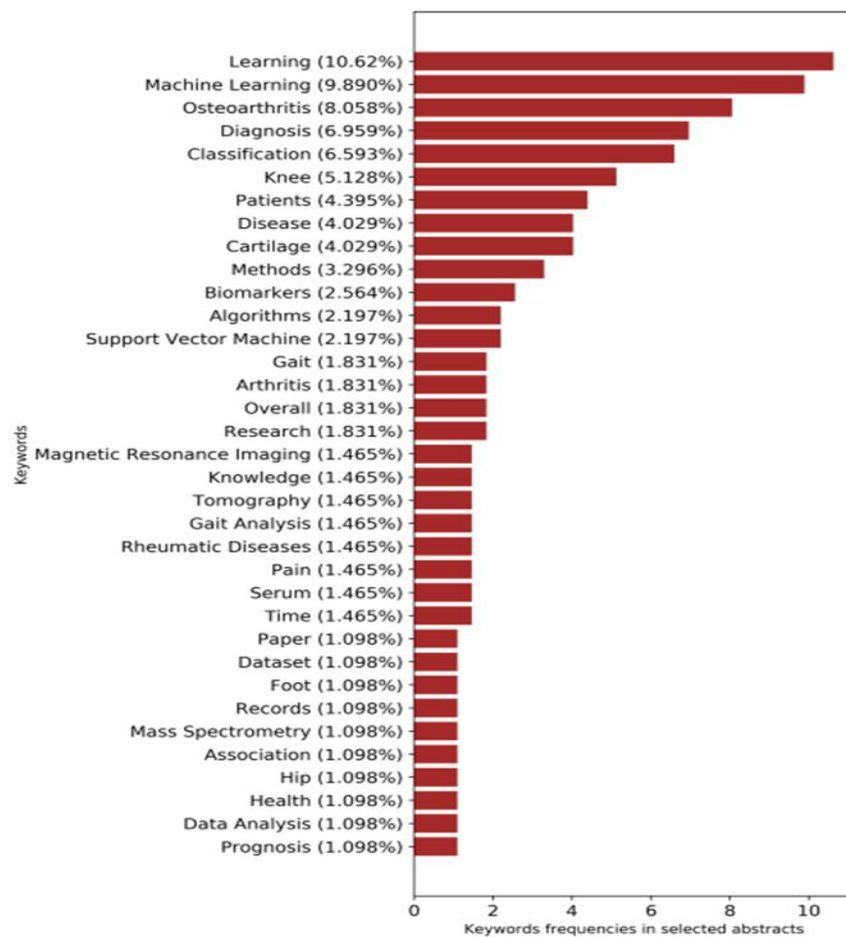


**Σχήμα 2.4:** Εξέλιξη του αριθμού των δημοσιευμένων άρθρων που αναφέρονται σε προσεγγίσεις μηχανικής μάθησης για την πρόιμη διάγνωση ρευματολογικών παθήσεων (τα αποτελέσματα προέκυψαν με το λογισμικό BIBOT)[45]

Οι M. Hugel και συνεργάτες το 2020 συμπεριέλαβαν σε επιστημονική τους εργασία [37] μια σχηματική παρουσίαση παραδειγμάτων αξιοποίησης μοντέλων μηχανικής μάθησης σε προβλήματα που σχετίζονται με την διάγνωση στην ΡΑ. Στο Σχήμα 2.5 παρουσιάζονται διάφορες απεικονίσεις σχετικών αλγορίθμων που χρησιμοποιούνται σε βασικά μοντέλα μηχανικής μάθησης για την αντιστοίχιση των χαρακτηριστικών εισόδου σε συγκεκριμένες εξόδους / αποτελέσματα στην ρευματοειδή αρθρίτιδα.



**Σχήμα 2.5.:** Παραδείγματα Εφαρμογής Μοντέλων Μηχανικής Μάθησης στην ΡΑ [37].



**Σχήμα 2.6:** Λέξεις κλειδιά που συναντώνται σε δημοσιευμένα άρθρα σχετικά με εφαρμογές της μηχανικής μάθησης στην πρόωση διάγνωση των ρευματολογικών παθήσεων (BIBOT software) [45]

## 2.2 Βασικές Αρχές Μηχανικής Μάθησης

### 2.2.1 Γενικά

Η μηχανική μάθηση (ML) είναι μια επέκταση της επιστήμης των υπολογιστών και της στατιστικής που περιλαμβάνει τη χρήση αλγορίθμων για την αναγνώριση σχέσεων στα δεδομένα ή και προβλέψεων για τα αποτελέσματα. Τα μοντέλα ML δεν είναι ρητά προγραμματισμένα, δηλαδή κάθε απόφαση που παίρνει το μοντέλο δεν καθορίζεται από μια δήλωση «αν-τότε» που έχει εισαχθεί από τον χρήστη. Επειδή τα μοντέλα ML δεν έχουν ρητές οδηγίες σχετικά με το πώς να προσεγγίσουν ένα πρόβλημα, τα μοντέλα «μαθαίνουν» ή μάλλον εξάγουν γνώση από τα δεδομένα εισόδου.

Ενώ η στατιστική ανάλυση στοχεύει σε ένα στατιστικό μοντέλο βασισμένο σε κανόνες εξηγώντας συγκεκριμένες σχέσεις ή σχέσεις που βασίζονται σε υποθέσεις επί των διαθέσιμων δεδομένων, η ML προσπαθεί να ανακαλύψει τις υποκείμενες συνδέσεις στα δεδομένα και λαμβάνει αποφάσεις βάσει αυτών. Μπορεί να χρησιμοποιηθεί σε αναλύσεις μεγάλων δεδομένων για τον εντοπισμό συσχετισμών όταν δεν υπάρχει εμφανής υπόθεση που πρέπει να ελεγχθεί. Ειδικότερα, η ML χρησιμοποιεί συχνά άγνωστες σχέσεις που δεν μπορούν να ταυτιστούν με άλλες στατιστικές τεχνικές για να προβλέψει τα αποτελέσματα. Η χρήση των εφαρμογών ML παρουσιάζει ραγδαία αύξηση λόγω της ικανότητας των μοντέλων να αναλύουν δεδομένα υψηλών διαστάσεων (δηλαδή δεδομένα στα οποία υπάρχουν περισσότερα σημεία δεδομένων από δείγματα). Αυτά τα «μεγάλα δεδομένα» συχνά καθιστούν ανέφικτες τις αμιγώς στατιστικές αναλύσεις.

Η μηχανική μάθηση είναι μια τεχνολογία πολύ παλαιότερη από τις πρόσφατες και ευρέως γνωστές εφαρμογές που σχετίζονται με αυτήν, όπως οι μηχανές αναζήτησης ιστού, email και anti-spam filters, κρατήσεις αεροπορικών πτήσεων και ξενοδοχείων, αυτόματοι μεταφραστές και πολλές άλλες που χρησιμοποιούμε στην καθημερινή μας ζωή. Στην πραγματικότητα, οι θεμελιώδεις αρχές που αποτελούν την καρδιά της αρχιτεκτονικής των πιο ισχυρών τεχνητών νευρωνικών δικτύων που είναι γνωστά σήμερα, χρησιμοποιούνταν ήδη στα τέλη της δεκαετίας του 1960. Η απόδοση αυτής της τεχνολογίας βασίζεται σε μεγάλο βαθμό στον όγκο και την ποιότητα των δεδομένων που είναι διαθέσιμα για την εκτέλεση της μάθησης [45].

Όπως αναφέρεται και στην προηγούμενη ενότητα, η πλέον προφανής χρήση της τεχνητής νοημοσύνης στην ιατρική είναι η διάγνωση. Στο πλαίσιο της μηχανικής μάθησης, αυτό

σημαίνει ότι ένα σύνολο δεδομένων, που αποτελείται από μια ομάδα ασθενών για τους οποίους είναι διαθέσιμη μια συλλογή κλινικών ή βιολογικών παραμέτρων και οι διαγνώσεις τους, χρησιμοποιείται για την εκπαίδευση ενός προγράμματος. Το πρόγραμμα εκπαιδεύεται να αναγνωρίζει, εντός των παραμέτρων που χαρακτηρίζουν τον ασθενή στο σύνολο δεδομένων, έναν περισσότερο ή λιγότερο περίπλοκο συνδυασμό μεταβλητών στους οποίους μπορεί να βασιστεί για να προβλέψει τη σωστή διάγνωση. Αυτή η προσέγγιση βασίζεται στην επίλυση προβλημάτων ταξινόμησης. Τα εκπαιδευμένα προγράμματα ονομάζονται «ταξινομητές» και εκπαιδεύονται να ταξινομούν τους ασθενείς σε σχέση με τη διάγνωση τους.

Συμπερασματικά, παρότι υπάρχουν επικαλύψεις μεταξύ των μεθόδων στατιστικής και μηχανικής μάθησης, οι στόχοι εφαρμογής και η επεκτασιμότητα των λύσεων είναι γενικά διαφορετικοί. Η παραδοσιακή στατιστική έχει μια αυστηρή επικέντρωση σε ακριβή δείγματα δεδομένων και συνόλων, ερμηνεύοντας τις στατιστικές σχέσεις μεταξύ των μεταβλητών και εκτιμώντας με ακρίβεια τις παραμέτρους του πληθυσμού [37], [45]. Αντίθετα, ο βασικός στόχος των μεθόδων ML, είναι η ικανότητα πρόβλεψης που παρέχουν δεδομένα που ενδεχομένως δεν έχουν εντοπιστεί (unseen data).

Επιπλέον, οι αλγόριθμοι ML μπορούν αυτομάτως να μαθαίνουν χρήσιμες αναπαραστάσεις δεδομένων και να επεξεργάζονται διαφορετικού τύπου δεδομένα εισόδου, πχ. cohorts ασθενών, ιατρικές απεικονίσεις και γενετικές πληροφορίες. Έτσι η μηχανική μάθηση καλύπτει ένα σημαντικό κενό στην μάθηση από την κλινική εμπειρία. Ιδανικά, μεταφράζει την γνώση που αποκτάται από την κλινική εικόνα του ασθενούς μέσω υπολογιστικών συστημάτων, επιτυγχάνοντας προβλέψεις για τα κλινικά αποτελέσματα, αναγνωρίζοντας τα μοτίβα της ασθένειας (disease patterns), ανιχνεύοντας χαρακτηριστικά της νόσου και βελτιστοποιώντας τις στρατηγικές θεραπείας [37], [44].

### **2.2.2 Είδη Μηχανικής Μάθησης**

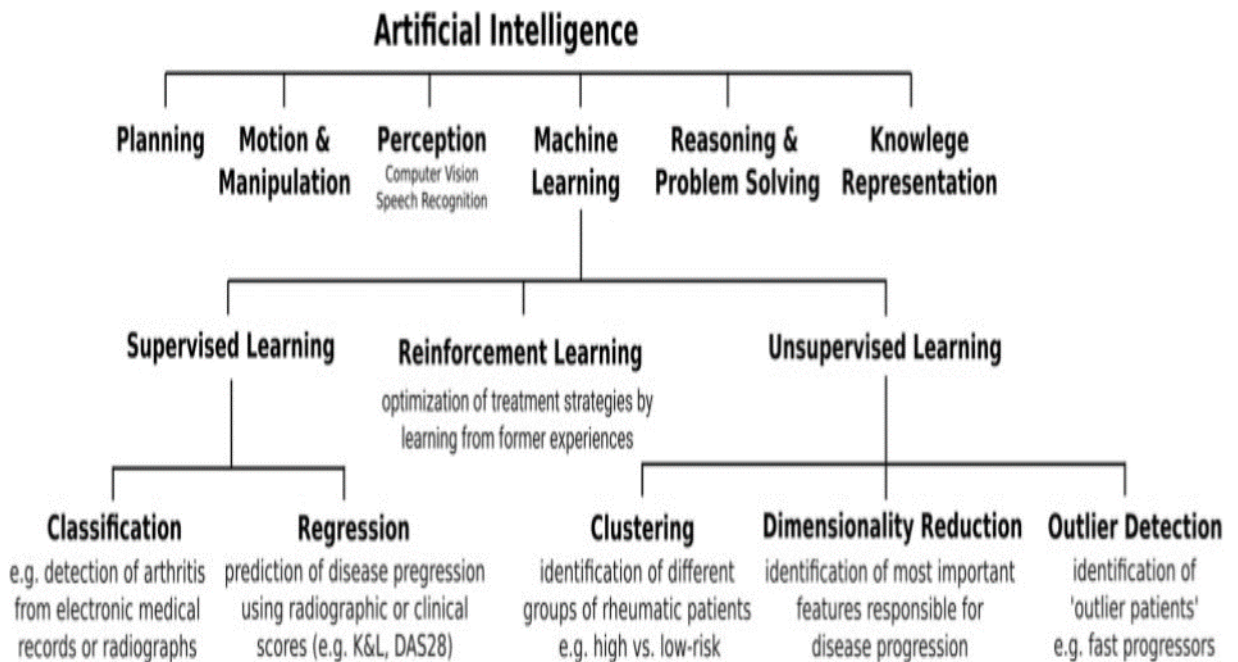
Η μηχανική μάθηση αποτελεί κλάδο της τεχνητής νοημοσύνης που παρέχει αλγόριθμους, ήτοι ακολουθίες καλά προσδιορισμένων εντολών που μπορούν να επιλύσουν ένα συγκεκριμένο πρόβλημα αναπτύσσοντας μαθηματικά μοντέλα βασισμένα σε δεδομένα δείγματος. Τα μαθηματικά μοντέλα (συναρτήσεις) αντιστοιχίζουν τα δεδομένα εισόδου στα επιθυμητά αποτελέσματα.

Τα δεδομένα εισόδου μπορεί να είναι εικόνες είτε μια τυχαία σειρά αριθμητικών ή κατηγορηματικών δεδομένων. Τα δεδομένα που έχουν συλλεχθεί αναφέρονται στην

συνέχεια ως χαρακτηριστικά εισόδου (features). Για την απεικόνιση των αντιστοιχίσεων μπορούν να χρησιμοποιηθούν διάφορα μοντέλα όπως οι πολυωνμικές συναρτήσεις, δέντρα αποφάσεων, μηχανές διανυσμάτων υποστήριξης (SVMs) και βαθιά τεχνητά νευρωνικά δίκτυα [37].

Η μηχανική μάθηση συνίσταται από διάφορους υποκλάδους ή είδη, όπως η επιβλεπόμενη μάθηση, η μη επιβλεπόμενη μάθηση και η ενισχυτική μάθηση, στις οποίες περιγράφονται αναλυτικά στη συνέχεια (Σχήμα 2.7.).

Η βαθιά μάθηση αποτελεί έναν επίσης σημαντικά υποκλάδο της ML, ο οποίος επεκτείνει τις δυνατότητες όλων αυτών των τεχνικών με την αξιοποίηση τεχνητών βαθιών νευρωνικών δικτύων για την αντιστοίχιση των δεδομένων εισόδου με τα επιθυμητά αποτελέσματα.



**Σχήμα 2.7:** Είδη και μοντέλα μηχανικής μάθησης και παραδείγματα αξιοποίησής τους στην RA [37]

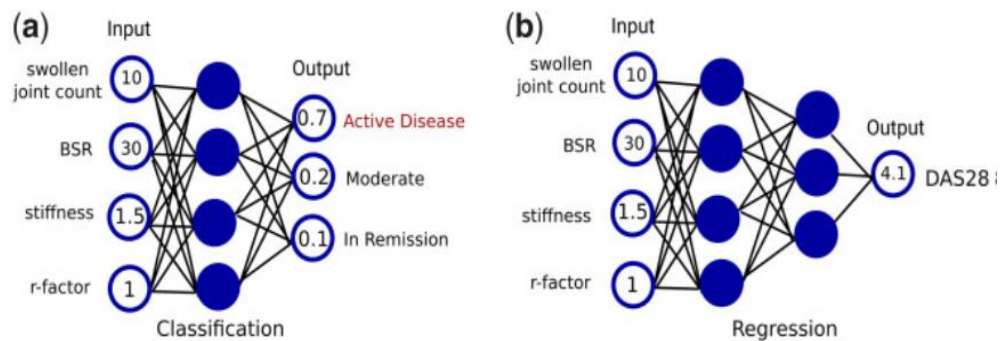
**A. Επιβλεπόμενη μάθηση (Supervised Learning)**

Είναι η διαδικασία όπου ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει δεδομένα εισόδου (σύνολο δεδομένων εκπαίδευσης) σε γνωστές επιθυμητές εξόδους, με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο. Χρησιμοποιείται σε προβλήματα ταξινόμησης (classification), πρόγνωσης (prediction) και διερμηνείας (interpretation).

Στην επιβλεπόμενη μάθηση το σύστημα πρέπει να "μάθει" επαγωγικά μια συνάρτηση που ονομάζεται συνάρτηση στόχος (target function) και αποτελεί έκφραση του μοντέλου που περιγράφει τα δεδομένα. Η συνάρτηση στόχος χρησιμοποιείται για την πρόβλεψη της τιμής μιας μεταβλητής, που ονομάζεται εξαρτημένη μεταβλητή ή μεταβλητή εξόδου, βάσει των τιμών ενός συνόλου μεταβλητών, που ονομάζονται ανεξάρτητες μεταβλητές ή μεταβλητές εισόδου ή χαρακτηριστικά. Η επαγωγική μάθηση στηρίζεται στην "υπόθεση επαγωγικής μάθησης" (inductive learning hypothesis), σύμφωνα με την οποία κάθε υπόθεση που προσεγγίζει καλά τη συνάρτηση στόχο για ένα αρκετά μεγάλο σύνολο παραδειγμάτων, θα προσεγγίζει το ίδιο καλά τη συνάρτηση στόχο και για περιπτώσεις που δεν έχει εξετάσει.

Στην επιβλεπόμενη μάθηση διακρίνουμε δυο βασικά είδη προβλημάτων (learning tasks), τα προβλήματα ταξινόμησης και τα προβλήματα παλινδρόμησης. Τα προβλήματα ταξινόμησης (classification) που αφορούν στη δημιουργία μοντέλων πρόβλεψης διακριτών τάξεων (κλάσεων/κατηγοριών) και τα προβλήματα παλινδρόμησης (regression) που αφορούν στη δημιουργία μοντέλων πρόβλεψης αριθμητικών τιμών. Τα μοντέλα εκπαιδεύονται βασίζόμενα σε γνωστά παραδείγματα που αποτελούνται από εισόδους και επιθυμητές εξόδους που παρέχονται από έναν ειδικό, στην περίπτωση μας τον ρευματολόγο.

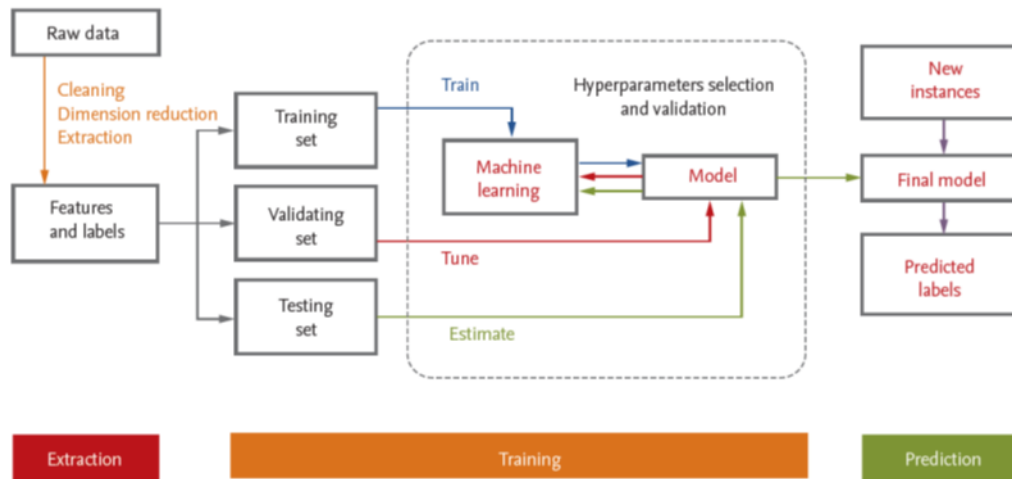
Οι έξοδοι μπορούν να είναι είτε μια ομάδα από κατηγορίες (π.χ. ενεργή νόσος, μέτρια νόσος, νόσος σε ύφεση) ή αριθμοί (π.χ. ο δείκτης DAS28). Τα μοντέλα που εκπαιδεύονται να εξάγουν πραγματικούς αριθμούς είναι τα μοντέλα παλινδρόμησης. Οι διαφορές μεταξύ των μοντέλων ταξινόμησης και παλινδρόμησης για την πρόβλεψη νόσου στην ρευματοειδή αρθρίτιδα, παρουσιάζονται στο Σχήμα 2.8.



**Σχήμα 2.8:** Σχηματική απεικόνιση διαφορών ταξινόμησης – παλινδρόμησης στην πρόβλεψη νόσου RA [37]

Στο Σχήμα 7 αποτυπώνεται ένα τυπικό διάγραμμα ροής για την ανάπτυξη ενός μοντέλου πρόβλεψης με επιβλεπόμενη μηχανική μάθηση [44]. Την εκπαίδευση ακολουθεί η φάση της πιστοποίησης της παραγόμενης νέας γνώσης. Συνήθως, η πιστοποίηση πραγματοποιείται

καταρχάς από τον ίδιο τον αλγόριθμο μέσω διαδικασιών ανάκλησης (recall) με τη βοήθεια δεδομένων ελέγχου (test data) και στη συνέχεια μέσω κριτικής που κάνει ο ειδικός βάσει των γνώσεων που διαθέτει για το πρόβλημα που επιχειρεί να λύσει ο αλγόριθμος. Τέλος, η νέα γνώση δίνεται προς χρήση σε εφαρμογές στις οποίες είναι απαραίτητη, για να λυθούν πραγματικά προβλήματα.



**Σχήμα 2.9** Τυπικό Διάγραμμα ροής επιβλεπόμενης μηχανικής μάθησης [44]

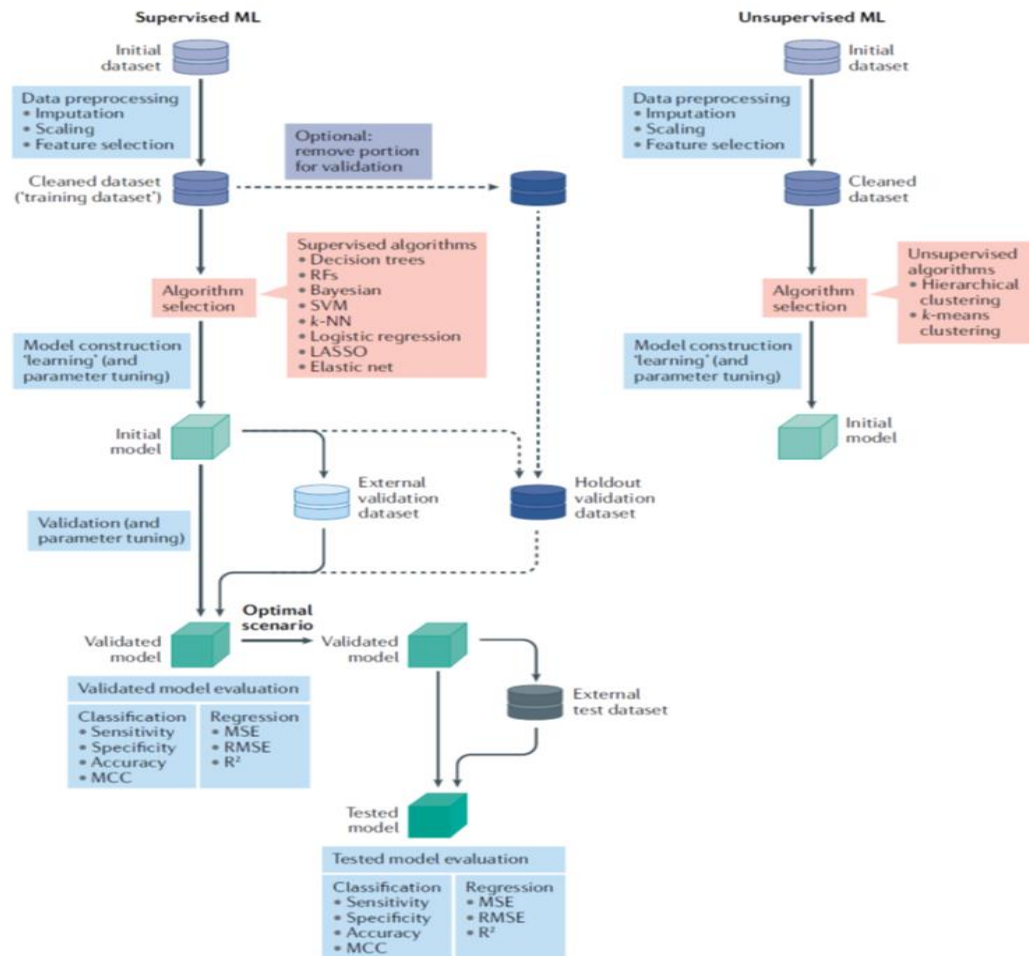
### ***B. Μη επιβλεπόμενη μάθηση (Unsupervised Learning)***

Στην περίπτωση της μη επιβλεπόμενης μάθησης ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων υπό μορφή παρατηρήσεων χωρίς να γνωρίζει τις επιθυμητές εξόδους. Χρησιμοποιείται κυρίως σε προβλήματα ανάλυσης συσχετισμών (association analysis), συσταδοποίησης (clustering) και σε μοντέλα παλινδρόμησης. Στην μη επιβλεπόμενη μάθηση το σύστημα έχει στόχο να ανακαλύψει συσχετίσεις και ομάδες από τα δεδομένα, βασιζόμενο μόνο στις ιδιότητές τους. Σαν αποτέλεσμα προκύπτουν πρότυπα (περιγραφές), κάθε ένα από τα οποία περιγράφει ένα μέρος από τα δεδομένα. Παραδείγματα προτύπων πληροφόρησης είναι οι κανόνες συσχέτισης (association rules) και οι ομάδες (clusters), οι οποίες προκύπτουν από τη διαδικασία της συσταδοποίησης (clustering). Η επιβλεπόμενη μάθηση λόγω της εκτενούς αξιοποίησής της στο πλαίσιο της παρούσας εργασία παρουσιάζεται αναλυτικότερα στην συνέχεια.

### ***Γ. Διαδικασία επιλογής του κατάλληλου τύπου Μηχανικής Μάθησης***

Για την επιλογή του κατάλληλου τύπου μηχανικής μάθησης που πρέπει να αξιοποιηθεί για την επίλυση ενός προβλήματος πρέπει να ληφθεί υπόψη το σύνολο δεδομένων που είναι

διαθέσιμα και τον σκοπό της ανάλυσης. Στο Σχήμα 2.10 περιγράφεται μια συσχέτιση των παραπάνω τύπων μηχανικής μάθησης για την λήψη απόφασης αναφορικά με το είδος καθώς και τα μοντέλα που είναι κατάλληλα για την κάθε περίπτωση [41].

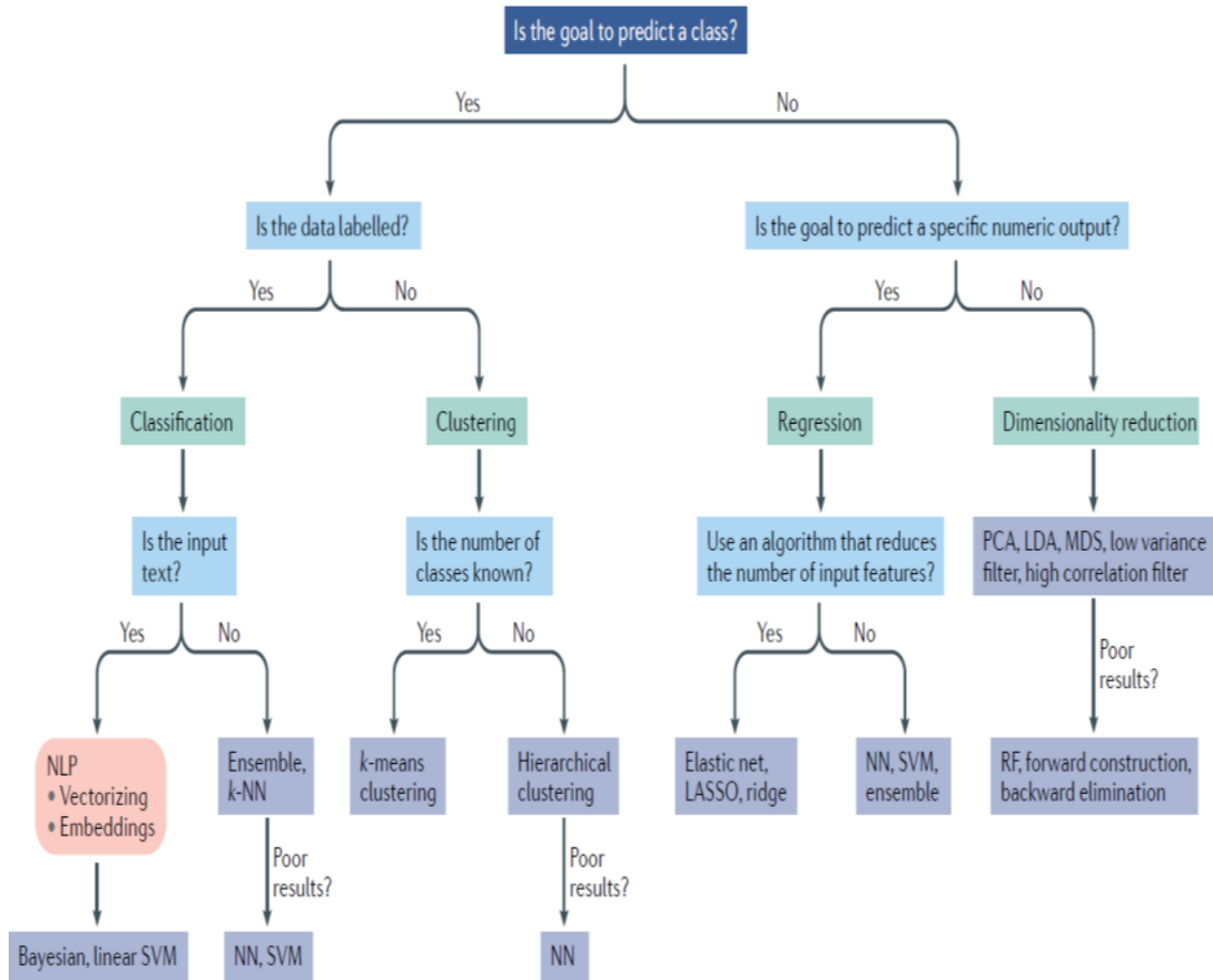


**Σχήμα 2.10:** Διάγραμμα ροής για λήψη αποφάσεων σχετικά με το κατάλληλο είδος μηχανικής μάθησης για ένα πρόβλημα [41]

Τα μοντέλα στην επιβλεπόμενη μηχανική μάθηση περιλαμβάνουν πρόσθετα βήματα επικύρωσης και αξιολόγησης σε σχέση με τα μοντέλα στην μηχανική μάθηση χωρίς επίβλεψη. Ο στόχος της μηχανικής μάθησης (ML) είναι η ταξινόμηση, η παλινδρόμηση, η συσταδοποίηση ή η μείωση διαστάσεων. Με βάση τα χαρακτηριστικά των δεδομένων εισόδου (για παράδειγμα, με ετικέτα έναντι χωρίς ετικέτα) και το επιθυμητό αποτέλεσμα (δηλαδή, πρόβλεψη μιας ομάδας ή πρόβλεψη τιμής), υπάρχουν συχνά πιο κατάλληλοι αλγόριθμοι για εφαρμογή. Ωστόσο εξακολουθεί να συνιστάται η χρήση πολλαπλών αλγορίθμων οι οποίοι στη συνέχεια συγκρίνονται μεταξύ τους για να προκύψει το μοντέλο με τη μεγαλύτερη προγνωστική δύναμη. Στο Σχήμα 9 παρουσιάζεται ένα διάγραμμα με



οδηγίες για την λήψη απόφασης σχετικά με το μοντέλο που θα επιλεγεί αναλόγως των συνόλων δεδομένων εισόδου, της επεξεργασίας τους και της γνώσης ή όχι των επιθυμητών εξόδων.



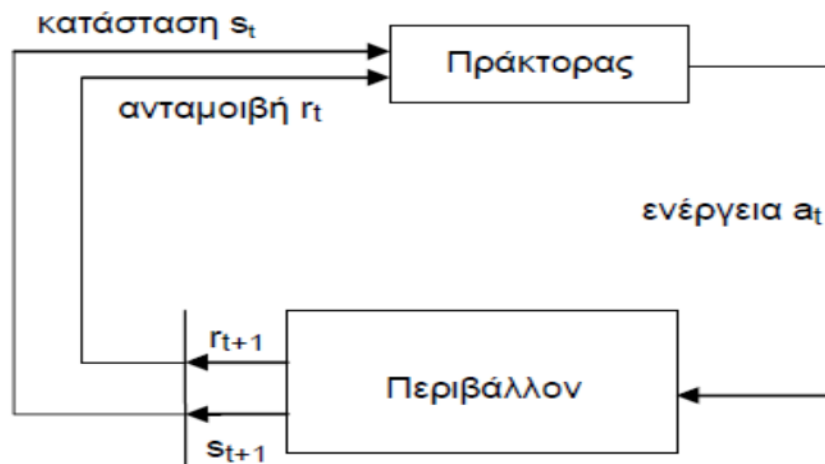
**Σχήμα 2.11:** Οδηγίες για την επιλογή του κατάλληλου αλγορίθμου μηχανικής μάθησης [41]

#### 4. Ενισχυτική Μάθηση (Reinforcement Learning)

Η ενισχυτική μάθηση είναι ένα είδος μηχανικής μάθησης που βασίζεται στην θεωρία της συμπεριφορικής ψυχολογίας και περιλαμβάνει ανταμοιβές. Στην ενισχυτική μάθηση ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών μέσα από άμεση αλληλεπίδραση με το περιβάλλον. Χρησιμοποιείται κυρίως σε προβλήματα σχεδιασμού (planning), όπως για παράδειγμα ο έλεγχος κίνησης ρομπότ και η βελτιστοποίηση εργασιών σε εργοστασιακούς

χώρους. Στην ενισχυτική μάθηση το σύστημα μαθαίνει στρατηγικές για το ποιες είναι οι βέλτιστες ενέργειες σε ορισμένες καταστάσεις σε σχέση με ένα δεδομένο κριτήριο (π.χ. αναμενόμενη βελτίωση της υγείας). Ένας αλγόριθμος RL λαμβάνει την ανατροφοδότηση για το πόσο καλά αποδίδει με αυτό το κριτήριο κατά τη διάρκεια της εκπαίδευσης. Στον τομέα της ρευματολογίας, το RL θα μπορούσε να χρησιμοποιηθεί, για παράδειγμα, για βελτιστοποίηση των στρατηγικών θεραπείας. Με βάση τα δεδομένα που παρατηρήθηκαν, μπορεί να εκπαιδεύσει ένα μοντέλο που προσδιορίζει πόσο πολύτιμη μια ορισμένη θεραπεία αναμένεται να γίνει στο μέλλον [37], [44].

Η ενισχυτική μάθηση είναι εμπνευσμένη από την μάθηση με επιβράβευση και τιμωρία που συναντάται στα έμβια όντα. Σκοπός του συστήματος μάθησης είναι να μεγιστοποιήσει μια συνάρτηση του σήματος ενίσχυσης (ανταμοιβή), για παράδειγμα την αναμενόμενη τιμή του σήματος ενίσχυσης στο επόμενο βήμα. Το σύστημα δεν καθοδηγείται από κάποιον εξωτερικό επιβλέποντα για το ποια ενέργεια θα πρέπει να ακολουθήσει αλλά πρέπει να ανακαλύψει μόνο του ποιες ενέργειες είναι αυτές που θα του αποφέρουν το μεγαλύτερο κέρδος. Στο βασικό πλαίσιο της ενισχυτικής μάθησης, η οντότητα που μαθαίνει και παίρνει αποφάσεις ονομάζεται πράκτορας (agent), ενώ οτιδήποτε άλλο εκτός του πράκτορα ονομάζεται περιβάλλον. Ο πράκτορας και το περιβάλλον αλληλεπιδρούν συνεχώς, με τον πρώτο να επιλέγει ενέργειες και το δεύτερο να αποκρίνεται σε αυτές και να του παρουσιάζει καινούριες καταστάσεις. Το περιβάλλον δίνει στον πράκτορα ανταμοιβές (rewards), ειδικές αριθμητικές τιμές οποίες αυτός προσπαθεί μακροπρόθεσμα να μεγιστοποιήσει. Ο πράκτορας και το περιβάλλον αλληλεπιδρούν σε μια ακολουθία διακριτών χρονικών στιγμών  $t=0, 1, 2, \dots$  [46]



**Σχήμα 2.12:** Αλληλεπίδραση του πράκτορα με το περιβάλλον [46].

Αυτή η συνάρτηση μπορεί να αναπαρασταθεί από μοντέλα παλινδρόμησης, όπως δέντρα παλινδρόμησης ή βαθιά νευρωνικά δίκτυα. Στη συνέχεια, από αυτή τη συνάρτηση μπορεί να εξαχθεί η στρατηγική, για παράδειγμα στην περίπτωση της ρευματοειδούς αρθρίτιδας να προτείνει την καλύτερη θεραπεία με το υψηλότερο αναμενόμενο αποτέλεσμα.

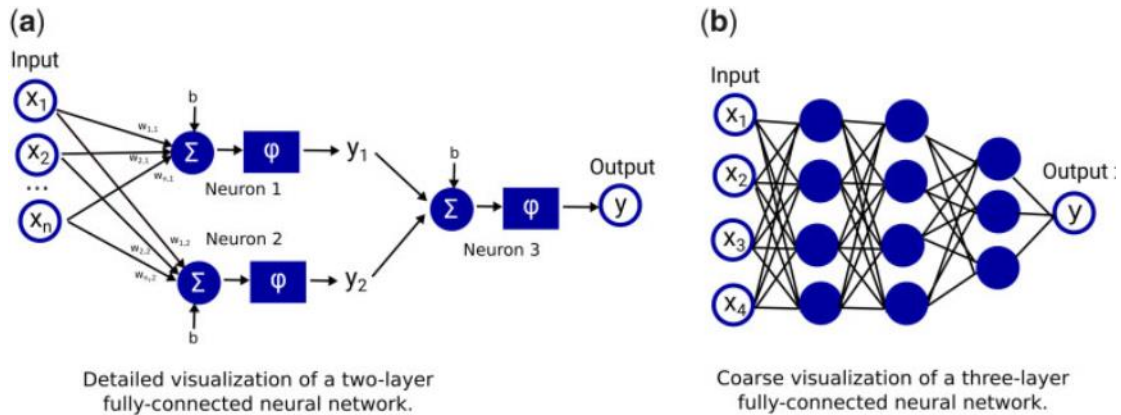
Παρότι η ενισχυτική μάθηση έχει πολλές δυνατότητες, δεν εφαρμόζεται συχνά στον τομέα της ιατρικής γιατί απαιτεί αυστηρά καθορισμένες κλινικές καταστάσεις, παρατηρήσεις (π.χ. εργαστηριακά αποτελέσματα, ενέργειες (θεραπείες) και ανταμοιβές, τα οποία στις περισσότερες περιπτώσεις είναι δύσκολο να προσδιοριστούν [37], [44].

### ***E. Βαθιά Μάθηση***

Σε αυτό το είδος μηχανικής μάθησης οι αναπαραστάσεις δεδομένων μαθαίνονται αυτόματα από βαθιά νευρωνικά δίκτυα. Τα βαθιά νευρωνικά δίκτυα μπορεί να μάθουν πολύ περίπλοκες, μη γραμμικές μαθηματικές συναρτήσεις. Αποτελούνται από πολλά διαδοχικά στρώματα που συντίθενται από πολλές απλές μη γραμμικές λειτουργίες, τους λεγόμενους νευρώνες. Μπορούν να χρησιμοποιηθούν τόσο στην επιβλεπόμενη όσο και στη μη επιβλεπόμενη μάθηση [37]. Σε ένα τεχνητό νευρωνικό δίκτυο, τα χαρακτηριστικά εισόδου τροφοδοτούνται στο πρώτο στρώμα των νευρώνων και στη συνέχεια διαδίδονται μέσω του δικτύου στο επίπεδο εξόδου. Οι αρχιτεκτονικές δικτύου μπορούν να αντιμετωπίσουν διαφορετικούς τύπους δεδομένων εισαγωγής (π.χ. ιατρικές εικόνες, κείμενο ή οποιοδήποτε άλλο είδος δεδομένων ασθενούς). Υπάρχουν διαφορετικές αρχιτεκτονικές για διαφορετικά είδη δεδομένων εισόδου, π.χ. πλήρως συνδεδεμένα νευρωνικά δίκτυα, συνελκτικά νευρωνικά δίκτυα και επαναλαμβανόμενα νευρωνικά δίκτυα (Σχήμα 2.13).

Τα συνελκτικά νευρωνικά δίκτυα χρησιμοποιούνται συνήθως για εικόνες και άλλα δεδομένα με δομή σαν πλέγμα. Εκτελούνται μαθηματικές πράξεις συνέλιξης σε κάθε μέρος των εισροών τους για να μάθουν όλο και πιο αφηρημένα χαρακτηριστικά σε πολλαπλά στρώματα. Ένα χαρακτηριστικό παράδειγμα είναι τα συνελκτικά νευρωνικά δίκτυα που εκπαιδεύονται από φυσικές εικόνες. Τα χαρακτηριστικά που προκύπτουν μέσα από την εκπαιδευτική διαδικασία στα πρώτα στρώματα συνήθως ειδικεύονται στην ανίχνευση των περιγραμμάτων και των άκρων, ενώ τα χαρακτηριστικά σε μεταγενέστερα στρώματα είναι σε θέση να συνδυάσουν αυτά τα προηγούμενα χαρακτηριστικά για να εντοπίσουν πιο περίπλοκα αντικείμενα, όπως αρθρώσεις ή χέρια. Για να αυξηθεί η ερμηνευτικότητα, μπορούν να χρησιμοποιηθούν συνελκτικά νευρωνικά δίκτυα που δημιουργούν θερμικούς χάρτες, τονίζοντας περιοχές ανά χρώμα και συμβάλλουν περισσότερο στην απόφαση του

δικτύου. Έτσι, μη φυσιολογικές περιοχές, όπως δομικές βλάβες, μπορούν να απεικονιστούν για διαγνωστική χρήση.



**Σχήμα 2.13:** Οπτικοποιήσεις πλήρως συνδεδεμένων νευρωνικών δικτύων [37]

Επαναλαμβανόμενα νευρωνικά δίκτυα, όπως τα μακροπρόθεσμα και πρόσφατες αναμνήσεις (LSTMs) είναι μια κατηγορία τεχνητών νευρικών δικτύων με εσωτερική μνήμη για την επεξεργασία ακολουθιών δεδομένων, όπως γραπτά κείμενα, ομιλία κ.ο.κ.

Στον τομέα της ιατρικής χρησιμοποιούνται LSTM π.χ. για πρόβλεψη έκβασης σε μονάδες εντατικής θεραπείας, πρόβλεψη καρδιακής ανεπάρκειας ή πρόβλεψη εξέλιξης υγειονομικής περίθαλψης από ιατρικά αρχεία [37].

Στην ενότητα 2.4. παρουσιάζονται παραδείγματα εφαρμογών διάφορων μοντέλων και τύπων μηχανικής μάθησης και των αποτελεσμάτων τους στην μελέτη της ρευματοειδούς αρθρίτιδας.

## 2.3 Μεθοδολογίες και αλγόριθμοι που αξιοποιήθηκαν στην

### παρούσα εργασία

Με βάση την ανάλυση της προηγούμενης ενότητας προκύπτει ότι για κάθε πρόβλημα προς επίλυση στο χώρο της Μηχανικής Μάθησης υπάρχει ένας κατάλληλος τρόπος μάθησης, και για κάθε τρόπο μάθησης υπάρχει τουλάχιστον ένας κατάλληλος αλγόριθμος που μπορεί να χρησιμοποιηθεί. Οι αλγόριθμοι ML που χρησιμοποιούνται για την κατασκευή των μοντέλων ποικίλουν ως προς την πολυπλοκότητά τους.

Συνήθεις αλγόριθμοι που χρησιμοποιούνται για την ταξινόμηση στην επιβλεπόμενη μάθηση περιλαμβάνουν δέντρα αποφάσεων, τυχαία δάση (RF), μηχανή ενίσχυσης κλίσης (Gradient Boosting Machine), απλή Bayes, μηχανές διανυσμάτων υποστήριξης (Support Vector

Machines), k-πλησιέστεροι γείτονες (k-NN) και λογιστική παλινδρόμηση (logistic regression). Ειδικότερα για την παλινδρόμηση χρησιμοποιείται μεταξύ άλλων ο LASSO (Least Absolute Shrinkage and Selection Operator).

Οι πιο συνηθισμένοι αλγόριθμοι στην μη επιβλεπόμενη μάθηση που χρησιμοποιούνται για συσταδοποίηση είναι το ιεραρχικό clustering, το k-means clustering και το Gaussian Mixture clustering. Η επιλογή αλγορίθμων εξαρτάται από τις ερωτήσεις που θέτουμε σχετικά με τα δεδομένα και στα ποιοτικά χαρακτηριστικά των δεδομένων. Ένα γενικό πλαίσιο για την λήψη αποφάσεων σχετικά με τους κατάλληλους αλγορίθμους έχει δοθεί στο Σχήμα 9.

Οι αλγόριθμοι που επιλέγονται είναι αυτοί που με βάση την εμπειρία κρίνονται πιο κατάλληλοι για το συγκεκριμένο σύνολο δεδομένων. Για παράδειγμα οι αλγόριθμοι Bayes και SVM χρησιμοποιούνται περισσότερο σε δεδομένα κειμένου. Επίσης τα δέντρα αποφάσεων είναι ταξινομητές υψηλής ερμηνείας, αλλά μπορεί να είναι επιρρεπείς σε υπερβολική προσαρμογή. Ως εκ τούτου, μέθοδοι συνόλου που περιλαμβάνουν συγκεντρωτικά δέντρα απόφασης όπως τα τυχαία δάση (RF) μπορεί να είναι καλύτερες ως σημείο εκκίνησης για την ταξινόμηση [47], [48].

Στην παρούσα διπλωματική εργασία για την κατασκευή των μοντέλων μηχανικής μάθησης αρχικά εφαρμόστηκε η διαδικασία της μη επιβλεπόμενης μάθησης με συσταδοποίηση και η κατασκευή των μοντέλων βασίστηκε στους αλγορίθμους k-means, k-medoids και Gaussian Mixture, ενώ εφαρμόστηκε η τεχνική Elbow για τον προσδιορισμό του βέλτιστου αριθμού των συστάδων. Στην συνέχεια, εφαρμόστηκε η διαδικασία της επιβλεπόμενης μάθησης χρησιμοποιώντας τον αλγόριθμο των k-πλησιέστερων γειτόνων. Στις επόμενες ενότητες του παρόντος κεφαλαίου γίνεται μια θεωρητική περιγραφή για τα μοντέλα και τις μεθοδολογίες που αξιοποιήθηκαν.

### **2.3.1 Συσταδοποίηση**

Μια μέθοδος μάθησης χωρίς επίβλεψη είναι μια μέθοδος στην οποία αντλούνται αναφορές από σύνολα δεδομένων που αποτελούνται από δεδομένα εισόδου χωρίς αποκρίσεις με ετικέτα. Σκοπός της διαδικασίας είναι η εύρεση ουσιαστικής δομής, επεξηγηματικών υποκείμενων διαδικασιών, γενετικών χαρακτηριστικών και ομαδοποιήσεων που είναι εγγενείς σε ένα σύνολο παραδειγμάτων.

Συσταδοποίηση είναι η διαδικασία εκείνη κατά την οποία ένα σύνολο από «αντικείμενα», διαχωρίζονται σε ένα σύνολο από λογικές ομάδες. Στη βιβλιογραφία, η μη επιβλεπόμενη μάθηση συναντάται και ως ομαδοποίηση. Η καταχώρηση αντικειμένων στην ίδια ομάδα

μεταφράζεται ως ομοιότητα των αντικειμένων αυτών και αντίστροφα, δηλαδή αντικείμενα που ανήκουν σε διαφορετικές ομάδες θεωρούνται ανόμοια. Η ομοιότητα ή μη, μεταξύ των αντικειμένων, ουσιαστικά εξαρτάται από το συγκεκριμένο πρόβλημα και τη μορφή των «αντικειμένων». Τα αντικείμενα μπορούν να αναφερθούν και αυτά με διαφορετικούς όρους όπως πρότυπα ή διανύσματα.

Διακρίνεται η αυστηρή συσταδοποίηση όπου κάθε διάνυσμα ανήκει σε μία και μόνο ομάδα, ενώ εναλλακτικά μπορεί να οριστεί η ασαφής συσταδοποίηση. Κάνοντας χρήση των ασαφών συνόλων μπορούν να οριστούν  $m$  συναρτήσεις συμμετοχής με τις οποίες ποσοτικοποιείται η βεβαιότητα για το αν κάποιο διάνυσμα  $i$  ανήκει στην ομάδα  $j$  [49].

Η συσταδοποίηση ως έννοια είναι σημαντική όχι μόνο σε πολλές επιστήμες όπως η κοινωνιολογία, η βιολογία και η στατιστική, αλλά και σε πολλούς τομείς της πληροφορικής όπως η αναγνώριση προτύπων, η εξόρυξη γνώσης, η ανάκτηση δεδομένων, η τεχνητή νοημοσύνη και η μηχανική μάθηση. Επίσης, όσον αφορά την επεξεργασία δεδομένων η συσταδοποίηση μπορεί να συμβάλει στην μείωση των δεδομένων, στην παραγωγή και έλεγχο μίας υπόθεσης και στην πραγματοποίηση μιας πρόβλεψης βασισμένης στις συστάδες. Όλα αυτά ουσιαστικά αναφέρονται και ορίζουν το ίδιο πρόβλημα με διαφορετική γλώσσα ορισμού. Τα τελευταία χρόνια οι τομείς αυτοί αποκτούν όλο και περισσότερο ενδιαφέρον, καθιστώντας έτσι την συσταδοποίηση αντικείμενο εντατικής έρευνας.

Οι εφαρμογές της συσταδοποίησης στην ιατρική είναι ευρύτατες και σημαντικές. Χαρακτηριστικό παράδειγμα αποτελεί η εφαρμογή της συσταδοποίησης σε ένα σύνολο δεδομένων που αφορά ασθενείς μιας συγκεκριμένης νόσου. Έστω ότι το επιθυμητό αποτέλεσμα είναι συστάδες ασθενών με βάση την αντίδρασή τους στα ίδια φάρμακα. Με την συγκεκριμένη διαδικασία, ένας νέος ασθενής μπορεί να κατηγοριοποιηθεί στην κατάλληλη συστάδα και να πάρει την σωστή φαρμακευτική αγωγή.

#### ***A. Μέθοδοι και Διαδικασία Συσταδοποίησης***

Το πρώτο στάδιο κάθε μεθοδολογίας μηχανικής μάθησης είναι η προεπεξεργασία των δεδομένων μας που αποτελεί μια θεμελιώδη πρώτη φάση για την ανάπτυξη ενός συστήματος. Επειδή η φύση των δεδομένων εισόδου θα επηρεάσει την ακρίβεια του μοντέλου μας, η προσεκτική προεπεξεργασία των δεδομένων πριν την εφαρμογή ενός αλγορίθμου μηχανικής μάθησης είναι απαραίτητη για την βελτίωση της χρησιμότητας και της αξιοπιστίας του μοντέλου [50].

Η προεπεξεργασία αφορά στη διαχείριση των ατελών δεδομένων, στον μετασχηματισμό και την κανονικοποίησή τους και στην επιλογή των πιο κατάλληλων χαρακτηριστικών ή

μεταβλητών που θα χρησιμοποιήσουμε σαν είσοδο στο μοντέλο. Τα μοντέλα μηχανικής μάθησης αποδίδουν καλύτερα όταν ο αριθμός των μεταβλητών εισόδου βελτιστοποιείται. Η επιλογή χαρακτηριστικών είναι μια τεχνική μείωσης των διαστάσεων που χρησιμοποιείται για την επιλογή την πιο κατάλληλων μεταβλητών που θα χρησιμοποιηθούν σαν είσοδοι στο μοντέλο.

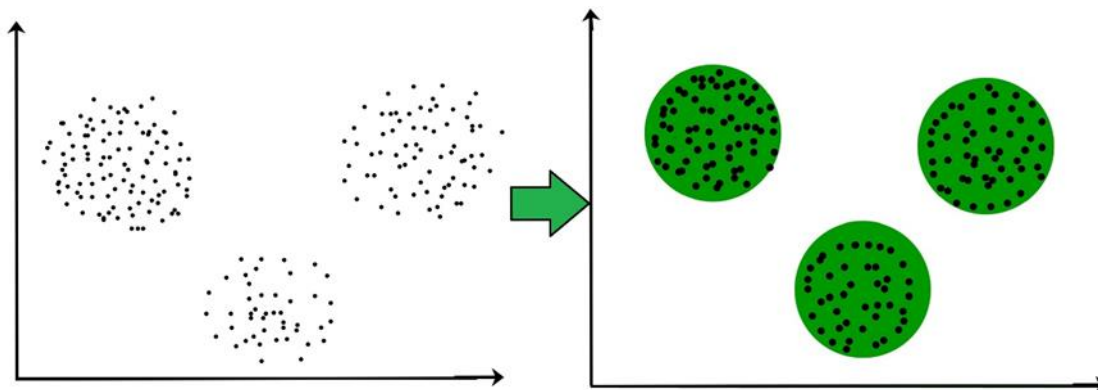
Ουσιαστικά, όλες οι μέθοδοι συσταδοποίησης χρησιμοποιούν την ίδια προσέγγιση. Πρώτα υπολογίζουμε τις ομοιότητες και μετά τις χρησιμοποιούμε για να συγκεντρώσουμε τα σημεία δεδομένων σε ομάδες. Οι βασικές μέθοδοι που χρησιμοποιούνται είναι οι παρακάτω:

**Μέθοδοι που βασίζονται στην πυκνότητα:** Αυτές οι μέθοδοι θεωρούν τις συστάδες ως την πυκνή περιοχή που έχει κάποιες ομοιότητες και διαφορές από την χαμηλότερη πυκνή περιοχή του χώρου. Αυτές οι μέθοδοι έχουν καλή ακρίβεια και τη δυνατότητα συγχώνευσης δύο συστάδων. Τέτοιες μέθοδοι μεταξύ άλλων είναι οι DBSCAN (Density-based spatial clustering of applications with noise) και OPTICS (Ordering points to identify the clustering structure).

**Μέθοδοι με βάση την ιεραρχία:** Τα συμπλέγματα που σχηματίζονται σε αυτήν τη μέθοδο σχηματίζουν μια δομή δέντρου που βασίζεται στην ιεραρχία. Νέες συστάδες σχηματίζονται χρησιμοποιώντας τις προηγούμενες. Χωρίζεται σε δύο κατηγορίες (1) συγκεντρωτική (προσέγγιση από κάτω προς τα πάνω) (2) διαιρετική (προσέγγιση από πάνω προς τα κάτω). Τέτοιες μέθοδοι μεταξύ άλλων είναι οι μέθοδοι CURE (Clustering Using Representatives) και BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies) [51].

**Μέθοδοι με βάση το διαχωρισμό:** Αυτές οι μέθοδοι χωρίζουν τα αντικείμενα σε  $k$  συμπλέγματα και κάθε διαμέρισμα σχηματίζει ένα σύμπλεγμα. Αυτή η μέθοδος χρησιμοποιείται για τη βελτιστοποίηση μιας συνάρτησης ομοιότητας αντικειμενικού κριτηρίου, όπως όταν η απόσταση είναι μια κύρια παράμετρος. Παράδειγμα αποτελούν οι αλγόριθμοι K-means και CLARANS (Clustering Large Applications based on Radomized Search) που χρησιμοποιείται κυρίως για χωρικά δεδομένα.

**Μέθοδοι που βασίζονται σε πλέγμα:** Σε αυτή τη μέθοδο, ο χώρος δεδομένων διαμορφώνεται σε έναν πεπερασμένο αριθμό κελιών που σχηματίζουν μια δομή που μοιάζει με πλέγμα. Όλες οι λειτουργίες ομαδοποίησης που γίνονται σε αυτά τα πλέγματα είναι γρήγορες και ανεξάρτητες από τον αριθμό των αντικειμένων δεδομένων, όπως στις μεθόδους STING (Statistical INformation Grid) και CLIQUE (Clustering In QUEst).



**Σχήμα 2.14:** Τα σημεία δεδομένων συγκεντρωμένα μαζί μπορούν να ταξινομηθούν σε μία ενιαία ομάδα. (Διακρίνονται τα συμπλέγματα και μπορούν να αναγνωριστούν 3 συστάδες στην εικόνα) [50].

Μια διαδικασία συσταδοποίησης ακολουθεί τα παρακάτω βασικά βήματα:

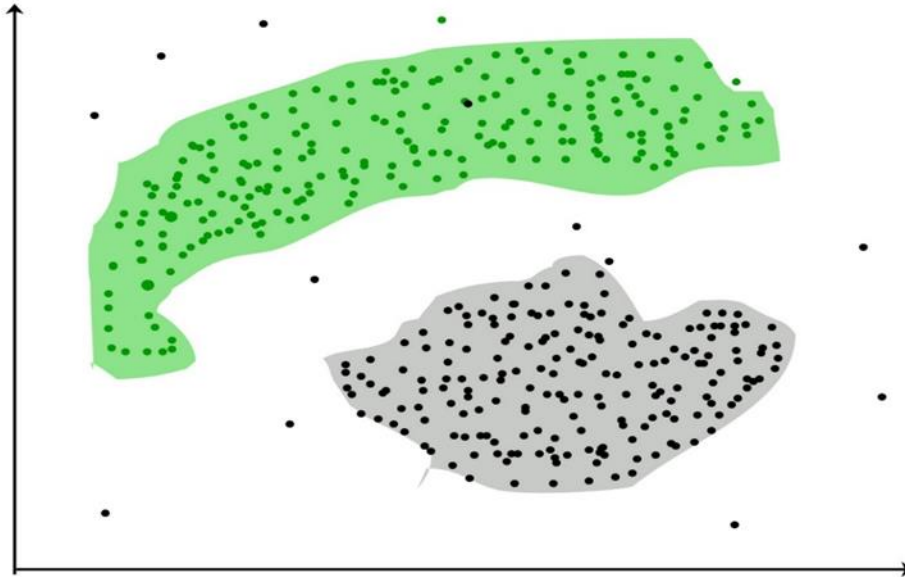
**Επιλογή χαρακτηριστικών γνωρισμάτων.** Ο στόχος είναι να επιλεγούν τα καταλληλότερα γνωρίσματα στα οποία πρόκειται να εφαρμοστεί η συσταδοποίηση ώστε να επιτυγχάνεται η βέλτιστη ομοιογένεια σε κάθε συστάδα. Έτσι η προεπεξεργασία των δεδομένων πριν την εφαρμογή της διαδικασίας συσταδοποίησης κρίνεται απαραίτητη.

**Επιλογή Αλγορίθμων συσταδοποίησης.** Σε αυτό το στάδιο γίνεται η επιλογή ενός αλγορίθμου που θα οδηγήσει σε ένα καλό σχήμα συσταδοποίησης για ένα σύνολο δεδομένων. Για τη επιλογή του αλγορίθμου χρησιμοποιείται το μέτρο γειννιάσης και το κριτήριο συσταδοποίησης τα οποία ορίζουν απόλυτα τον αλγόριθμο, καθώς επίσης και η δυνατότητά του να καθορίσει ένα σχήμα συσταδοποίησης που να προσαρμόζεται στο συγκεκριμένο σύνολο δεδομένων.

**Το μέτρο γειννιάσης (proximity measure)** αναφέρεται στην ομοιότητα δύο αντικειμένων (δηλαδή διανύσματα γνωρισμάτων). Η επιλογή των γνωρισμάτων πρέπει να γίνεται προσεχτικά ώστε η συμβολή τους να είναι ίση κατά τον υπολογισμό του μέτρου γειννιάσης και να μην υπερισχύει το ένα έναντι του άλλου.

**Το κριτήριο συσταδοποίησης (clustering criterion)** εκφράζεται βάσει μιας συνάρτησης κόστους ή κάποιου άλλου τύπου κανόνων. Είναι σημαντικό να γνωρίζουμε τον τύπο των συστάδων που θα προκύψουν στο σύνολο δεδομένων, για να διαλέξουμε το κατάλληλο κριτήριο που θα ταιριάζει στο σύνολο δεδομένων και θα έχει ως αποτέλεσμα μία επιτυχημένη τμηματοποίηση.





**Σχήμα 2.15:** Οι συστάδες δεν είναι απαραίτητο να είναι σφαιρικές [49].

**Επικύρωση αποτελεσμάτων.** Σε αυτή τη φάση αξιολογούνται τα αποτελέσματα του αλγορίθμου συσταδοποίησης σύμφωνα με κατάλληλα κριτήρια ορθότητας συσταδοποίησης και τεχνικές. Παράδειγμα ενός τέτοιου κριτηρίου είναι η σύγκριση των αποτελεσμάτων της ανάλυσης με κάποια ήδη γνωστά αποτελέσματα ή η σύγκριση των αποτελεσμάτων δύο διαφορετικών συσταδοποιήσεων. Η ποιότητα της συσταδοποίησης εξαρτάται από την ομοιότητα (δηλαδή μεγάλη ομοιότητα εντός της συστάδας - μικρή ομοιότητα μεταξύ των συστάδων) και την μέθοδο υλοποίησης της συσταδοποίησης.

**Ερμηνεία των αποτελεσμάτων.** Αποτελεί το τελευταίο στάδιο της διαδικασίας συσταδοποίησης, όπου οι αναλυτές καλούνται να εξάγουν γνώση από τις παραχθείσες συστάδες, συνδυάζοντας κι άλλα στοιχεία, αναλύσεις, με σκοπό το καλύτερο και εγκυρότερο αποτέλεσμα.

### ***B. Επικύρωση αποτελεσμάτων και έλεγχος αλγορίθμων***

Η επικύρωση της εγκυρότητας της συσταδοποίησης είναι η διαδικασία εκείνη κατά την οποία εκτιμάται ότι η έξοδος του αλγορίθμου είναι ουσιαστικά ένα σύνολο από ομάδες διανυσμάτων, οι οποίες πληρούν τον ορισμό της συσταδοποίησης καθώς και ορισμένα επιπλέον κριτήρια. Γενικά ο έλεγχος μπορεί να επιτευχθεί με τρεις διαφορετικούς τρόπους ανάλογα με τα κριτήρια που χρησιμοποιούνται:

**Εσωτερικά κριτήρια.** Ο έλεγχος ενός αλγορίθμου A επιτυγχάνεται λαμβάνοντας υπ' όψη μόνο τα δεδομένα εισόδου του, αυτά που χρησιμοποιήθηκαν για να επιτευχθεί η συσταδοποίηση. Για παράδειγμα, αν ο A αποτελεί έναν ιεραρχικό αλγόριθμο, τότε η

εκτίμηση του συγκεκριμένου αλγορίθμου μπορεί να γίνει αποκλειστικά από τον πίνακα εγγύτητας.

**Εξωτερικά κριτήρια.** Σε αυτή την περίπτωση υπολογίζεται μία δομή των διανυσμάτων  $M$  εξωτερικά από τον χρήστη. Στη συνέχεια απλά γίνεται σύγκριση της συσταδοποίησης  $M = \{m_1, m_2 \dots m_s\}$  με την έξοδο του αλγορίθμου  $C = \{c_1, c_2 \dots c_k\}$ .

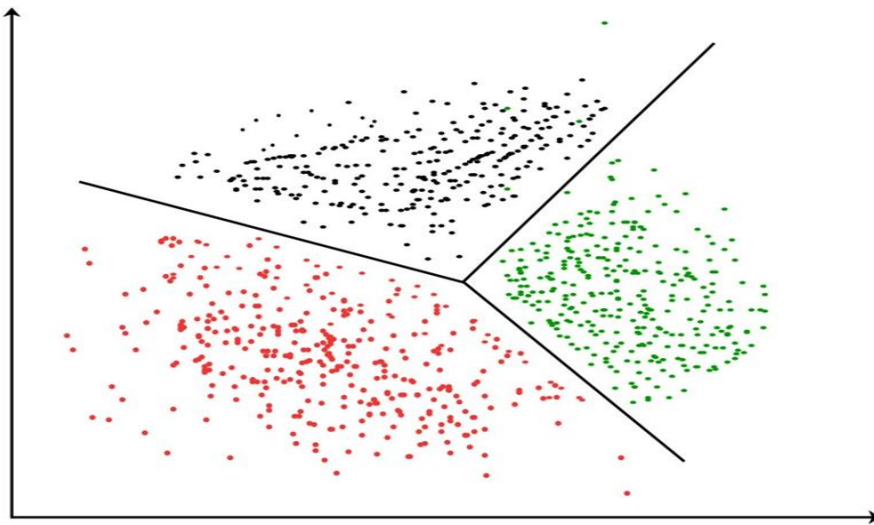
**Σχετικά κριτήρια.** Τέλος υπάρχει και η διαδικασία ελέγχου που βασίζεται σε σχετικά κριτήρια. Στη συγκεκριμένη περίπτωση γίνονται αλλαγές των παραμέτρων εισόδου του αλγορίθμου και εκτιμάται η νέα συσταδοποίηση  $C'$ . Αυτό μπορεί να γίνει επαναληπτικά έτσι ώστε να καταλήξουμε στην καλύτερη δυνατή συσταδοποίηση. Οι αλλαγές που μπορούν να γίνουν είναι (υποθέτουμε ότι τα παρακάτω είναι παράμετροι εισόδου): διαφορετική διάταξη των διανυσμάτων, διαφορετικό πλήθος ομάδων συσταδοποίησης, διαφορετικό διάστημα πλήθους συσταδοποιήσεων  $[m_{min}, m_{max}]$ , διαφορετικοί αρχικοί αντιπρόσωποι (για παράδειγμα τα αρχικά μέσα διανύσματα που χρησιμοποιεί ο k-means), κλπ.

### **2.3.2 Αλγόριθμοι μη επιβλεπόμενης μάθησης και μεθοδολογίες**

Υπάρχει ένα μεγάλο πλήθος από αλγόριθμους συσταδοποίησης που έχουν προταθεί και ο καθένας τους βασίζεται σε διαφορετική φιλοσοφία. Σχεδόν όλοι τους δέχονται ένα σύνολο παραμέτρων που μπορεί να είναι το πλήθος των ομάδων, διανύσματα αρχικοποίησης που απαιτούνται από τον αλγόριθμο κάποιες υποθέσεις για την πυκνότητα των διανυσμάτων στο χώρο και άλλες διάφορες παραμέτρους. Διαφοροποιώντας αυτές τις παραμέτρους προκύπτει ένα σύνολο από αλγόριθμους σε κάθε βασική κατηγορία. Οι αλγόριθμοι που χρησιμοποιήθηκαν στο πλαίσιο της παρούσας εργασίας περιλαμβάνουν τους παρακάτω:

**Διαμεριστικός αλγόριθμος K-μέσων (k-means)** – Είναι ο απλούστερος αλγόριθμος μάθησης χωρίς επίβλεψη που λύνει το πρόβλημα ομαδοποίησης. Ο αλγόριθμος K-means χωρίζει  $n$  παρατηρήσεις σε  $k$  συστάδες όπου κάθε παρατήρηση ανήκει στο σύμπλεγμα με τον πλησιέστερο μέσο όρο να χρησιμεύει ως πρωτότυπο του συμπλέγματος [50]. Ο συγκεκριμένος αλγόριθμος έχει πάρα πολλές εφαρμογές και είναι η βάση για πολλούς άλλους όπως ο k-medoids. Ανήκει στην κατηγορία της επίπεδης ομαδοποίησης διότι παράγει ένα σύνολο ομαδοποιήσεων οι οποίες δεν έχουν κάποια ιδιαίτερη δομή-σχέση μεταξύ τους. Ο αλγόριθμος k-means έχει ως αποτέλεσμα μια διαμέριση του χώρου των δεδομένων, σε αντίθεση με τους ιεραρχικούς αλγορίθμους που δημιουργούν πιο πολύπλοκες δομές που περιγράφονται από δενδρογράμματα. Σε αντίθεση με τις ιεραρχικές μεθόδους κατηγοριοποίησης που βασίζονται στην πληροφορία που αποθηκεύεται στον πίνακα

ομοιότητας, οι διαμεριστικοί μέθοδοι εφαρμόζονται απευθείας στα δεδομένα με αποτέλεσμα να έχουν μικρότερες απαιτήσεις σε χώρο και να διαχειρίζονται ευκολότερα πολυπληθή σύνολα δεδομένων. Το κυρίως πρόβλημα των αλγορίθμων αυτών είναι η απόφαση για τον αριθμό των τελικών συστάδων. Ο αριθμός αυτός καθορίζεται κυρίως από την προσπάθεια βελτιστοποίησης. Στην πραγματικότητα αυτό που γίνεται είναι να τρέχει ο αλγόριθμος για διάφορους αριθμούς από συστάδες και να επιλέγεται εκείνη η τελική κατάσταση η οποία βελτιστοποιεί μια συνάρτηση. Το κριτήριο που κυρίως χρησιμοποιείται σε διαμεριστικούς αλγορίθμους για την τελική απόφαση του αριθμού των συστάδων είναι το κριτήριο του τετραγωνικού λάθους ή η συνάρτηση τετραγωνικού λάθους (squared error function).



**Σχήμα 2.16:** Γραφική απεικόνιση διαμέρισης  $n$  παρατηρήσεων σε  $k$  συστάδες με  $k$ -means [50]

Ο αλγόριθμος K-μέσων (K-means) είναι ένας πολύ απλός και πολύ διαδεδομένος διαμεριστικός αλγόριθμος που χρησιμοποιεί το κριτήριο του τετραγωνικού λάθους. Ο αλγόριθμος ξεκινά με μια τυχαία διαμέριση και συνεχώς τοποθετεί στοιχεία στις συστάδες με βάση την απόσταση των στοιχείων από το κεντροειδές (γεωμετρικός μέσος) της συστάδας.

Αυτό σταματάει όταν ικανοποιηθεί κάποιο κριτήριο όπως η ελαχιστοποίηση της συνάρτησης τετραγωνικού λάθους ή η μη διαφοροποίηση των συστάδων από κάποια επανάληψη και μετά [47], [50], [52], [53].

Αναλυτικότερα ο αλγόριθμος K-μέσων (K-means) περιγράφεται ως εξής:

1. Επιλογή  $k$  συστάδων. Τα αντικείμενα της βάσης εντάσσονται αρχικά σε κάποια συστάδα είτε με βάση κάποια συνάρτηση είτε με τυχαίο τρόπο.

2. Τοποθέτηση κάθε αντικειμένου στη πιο κοντινή συστάδα μετά από υπολογισμό της απόστασης του σημείου του αντικειμένου από το κεντροειδές της συστάδας.
3. Υπολογισμός του νέου κεντροειδούς μετά την αλλαγή των αντικειμένων της κάθε συστάδας που έγινε στο βήμα 2.
4. Αν το κριτήριο τερματισμού (αν δηλαδή υπάρχουν αλλαγές στη δομή όλων των συστάδων) δεν ικανοποιείται επανάληψη των βημάτων 2 και 3.

Ο αλγόριθμος K-means κλιμακώνεται καλά σε μεγάλο αριθμό δειγμάτων και έχει χρησιμοποιηθεί σε πολλούς διαφορετικούς τομείς.

Το κύριο μειονεκτήματα της μεθόδου είναι ότι σε χώρους πολύ υψηλών διαστάσεων, οι Ευκλείδειες αποστάσεις τείνουν να διογκώνονται (αυτό είναι ένα παράδειγμα της λεγόμενης «κατάρας της διάστασης»). Η εκτέλεση ενός αλγορίθμου μείωσης διαστάσεων πριν από την ομαδοποίηση k-means, μπορεί να μετριάσει αυτό το πρόβλημα και να επιταχύνει τους υπολογισμούς [50], [52], [53].

**K-Medoids:** Ο αλγόριθμος k-Means είναι ευαίσθητος στην ύπαρξη εξαιρέσεων. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι η χρήση ως κέντρου, όχι ενός υπολογιζόμενου μέσου σημείου, αλλά ενός υπαρκτού σημείου δεδομένων. Ο αλγόριθμος k-Medoids ακολουθεί αυτήν την προσέγγιση.

Αναλυτικότερα, στον αλγόριθμο k-Medoids επιλέγονται αρχικά k σημεία ως κέντρα (medoids). Τα υπόλοιπα σημεία κατατάσσονται στη συστάδα του πλησιέστερου κέντρου. Μια συνάρτηση κόστους μετρά το άθροισμα των αποστάσεων όλων των σημείων από το κέντρο της συστάδας τους. Σε μια επαναληπτική διαδικασία, σημεία τα οποία δεν είναι κέντρα δοκιμάζονται ως πιθανά κέντρα. Εάν για ένα σημείο το κόστος γίνεται μικρότερο, τότε το σημείο αυτό γίνεται το νέο κέντρο στη θέση του προηγούμενου. Ο αλγόριθμος k-Medoids λειτουργεί πιο αποτελεσματικά από τον k-Means, όταν στα δεδομένα υπάρχουν αντικείμενα με ακραίες τιμές. Ωστόσο, το κόστος υπολογισμού των medoids είναι σημαντικά μεγαλύτερο από το κόστος υπολογισμού των μέσων τιμών. Για τον λόγο αυτό, ο k-Medoids υπολείπεται του k-Means ως προς τον χρόνο επεξεργασίας μεγάλων συνόλων δεδομένων [48].

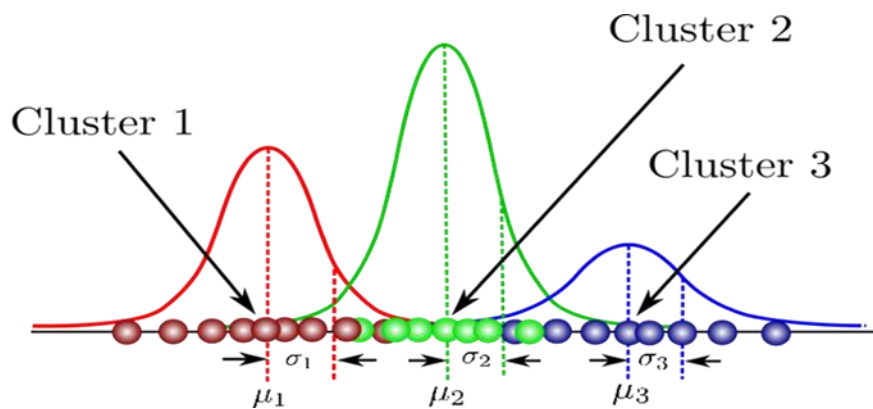
**Μοντέλο μείξης Γκαουσιανών GMM (Gaussian mixture model)** – Το μοντέλο GMM είναι ένα πιθανολογικό μοντέλο που υποθέτει ότι όλα τα σημεία δεδομένων δημιουργούνται από ένα σύμπλεγμα πεπερασμένου αριθμού Gaussian κατανομών με άγνωστες παραμέτρους. Κάποιος μπορεί να σκεφτεί τα μοντέλα GMM ως γενικευμένη ομαδοποίηση k-means για να

ενσωματώσει πληροφορίες σχετικά με τη δομή συνδιακύμανσης των δεδομένων καθώς και τα κέντρα των λανθανουσών Gaussians.

Ένα GMM μοντέλο είναι μια συνάρτηση που αποτελείται από πολλές Γκουσιανές, κάθε μία από τις οποίες προσδιορίζεται με  $k \in \{1, \dots, K\}$ , όπου  $K$  είναι ο αριθμός των συστάδων του συνόλου δεδομένων μας. Κάθε Γκουσιανή  $k$  αποτελείται από τις ακόλουθες παραμέτρους [49], [54]:

- Την μέση τιμή  $\mu$  που ορίζει το κέντρο της.
- Την συνδιακύμανση  $\sigma$  που ορίζει το πλάτος του.
- Την πιθανότητα ανάμειξης  $\pi$  που ορίζει πόσο μεγάλη ή μικρή θα είναι η Γκουσιανή συνάρτηση.

Οι παράμετροι αυτοί απεικονίζονται γραφικά στο σχήμα που ακολουθεί.



**Σχήμα 2.17:** Γραφική απεικόνιση των παραμέτρων του μοντέλου GMM[59]

Εδώ φαίνεται ότι υπάρχουν τρεις Gaussian συναρτήσεις, άρα  $K = 3$ . Κάθε Γκουσιανή ερμηνεύει τα δεδομένα που περιλαμβάνονται σε κάθε ένα από τα τρία διαθέσιμα clusters. Οι συντελεστές ανάμειξης είναι πιθανότητες που πρέπει να ικανοποιούν την παρακάτω συνθήκη:

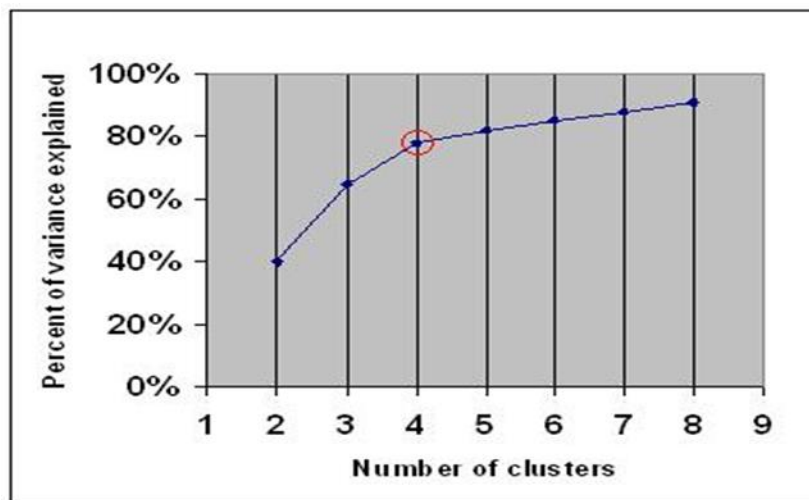
$$\sum_{k=1}^K \pi_k = 1 \quad (1)$$

Για να προσδιοριστούν οι βέλτιστες τιμές αυτών των παραμέτρων πρέπει να επιβεβαιωθεί ότι η κάθε Γκουσιανή ταιριάζει με τα σημεία που ανήκουν σε κάθε cluster με την μέγιστη πιθανότητα. Γενικά, η γκουσιανή συνάρτηση πυκνότητας δίνεται από την παρακάτω σχέση:

$$\mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right)$$

Στην οποία το  $\mathbf{x}$  αντιπροσωπεύει τα σημεία δεδομένων μας, το  $D$  είναι ο αριθμός διαστάσεων του κάθε σημείου,  $\mu$  και  $\Sigma$  είναι η μέση τιμή και συνδιακύμανση, αντίστοιχα. Αν ένα σύνολο δεδομένων συνίσταται από  $N = 1000$  σημεία τριών διαστάσεων ( $D = 3$ ), τότε το  $\mathbf{x}$  θα είναι ένας πίνακας  $1000 \times 3$ , το  $\mu$  θα είναι ένα διάνυσμα  $1 \times 3$ , και το  $\Sigma$  θα είναι ένας πίνακας  $3 \times 3$  [49].

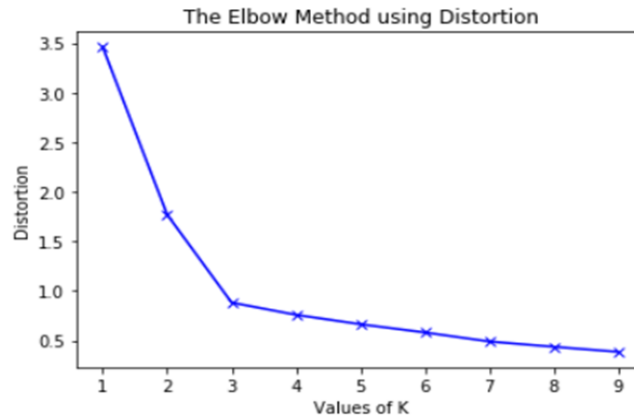
**Μεθοδολογία Elbow** – Στην ανάλυση συσταδοποίησης η μέθοδος elbow χρησιμοποιείται για τον προσδιορισμό των clusters σε ένα σύνολο δεδομένων. Η μέθοδος συνίσταται στον σχεδιασμό της διακύμανσης ως μιας συνάρτησης του αριθμού των clusters και την επιλογή του αγκώνα της καμπύλης ως τον αριθμό των clusters που πρέπει να χρησιμοποιήσουμε. Η ίδια μέθοδος μπορεί να χρησιμοποιηθεί για να επιλέξουμε τον αριθμό των παραμέτρων. Η μέθοδος αυτή εντοπίζεται πρώτη φορά σε μελέτη του Robert L. Thorndike in 1953 [55]. Η χρήση του αγκώνα “elbow” ή του γόνατου μιας καμπύλης “knee of a curve” ως ένα σημείο αποκοπής είναι συνηθισμένη ευρετική στην μαθηματική βελτιστοποίηση. Στην συσταδοποίηση έχει την έννοια ότι κάποιος πρέπει να επιλέξει έναν αριθμό συστάδων τέτοιοι ώστε η προσθήκη μιας επιπλέον συστάδας να μην προσφέρει πλέον κάποια σημαντική βελτίωση στην μοντελοποίηση [56]



**Σχήμα2.18:** Μέθοδος Elbow [56]

Η γενική προσέγγιση είναι ότι η αύξηση του αριθμού των συστάδων θα βελτιώσει την προσαρμογή (θα εξηγήσει μεγαλύτερο μέρος της διακύμανσης), δεδομένου ότι υπάρχουν

περισσότερες συστάδες προς χρήση, αλλά στην πραγματικότητα μετά από κάποιο σημείο αυτό αποτελεί υπερβολική προσαρμογή και ο αγκώνας το αντανακλά αυτό.



**Σχήμα 2.19:** Η μέθοδος Elbow με χρήση της παραμόρφωσης [57]

Η παραμόρφωση (distortion) υπολογίζεται ως ο μέσος όρος των τετραγωνικών αποστάσεων από τα κέντρα των αντίστοιχων συστάδων. Συνήθως χρησιμοποιείται η μετρική της ευκλείδειας απόστασης.

Η ιδέα είναι ότι οι πρώτες συστάδες θα προσθέσουν πολλή πληροφορία (θα εξηγήσουν μεγάλη διακύμανση), αφού τα δεδομένα αποτελούνται στην πραγματικότητα από τόσες ομάδες (άρα αυτές οι συστάδες είναι απαραίτητες), αλλά μόλις ο αριθμός των συστάδων υπερβεί τον πραγματικό αριθμό των ομάδων στα δεδομένα, η προστιθέμενη πληροφορία θα μειωθεί απότομα, επειδή απλώς υποδιαιρούνται οι πραγματικές ομάδες. Υποθέτοντας ότι αυτό συμβαίνει, θα υπάρχει ένας απότομος αγκώνας στο γράφημα της εξηγούμενης διακύμανσης σε σχέση με τις συστάδες: αυξάνεται γρήγορα μέχρι το  $k$  (περιοχή υποπροσαρμογής) και στη συνέχεια αυξάνεται αργά μετά το  $k$  (περιοχή υπερπροσαρμογής). Στην πράξη μπορεί να μην υπάρχει απότομος αγκώνας, και ως ευρετική μέθοδος, ένας τέτοιος "αγκώνας" δεν μπορεί πάντα να εντοπιστεί με σαφήνεια [57].

### 2.3.3 Επιβλεπόμενη μάθηση με τον αλγόριθμο $k$ πλησιέστερων γειτόνων ( $k$ -NN)

Η μέθοδος  $k$ -πλησιέστερου γείτονα έχει χρησιμοποιηθεί εκτενώς ως ταξινομητής αναφοράς στην περιοχή της μηχανικής μάθησης. Αυτή η μέθοδος ταξινομεί τα δεδομένα συγκρίνοντας δεδομένα δοκιμής με δεδομένα εκπαίδευσης παρόμοια με αυτήν. Κάθε στιγμιότυπο δεδομένων είναι ένα σημείο σε έναν  $n$ -διάστατο χώρο. Έτσι, όλα τα στιγμιότυπα εκπαίδευσης διατηρούνται σε αυτόν τον χώρο. Μόλις ένα στιγμιότυπο πρόκειται να ταξινομηθεί, το  $k$ -NN αναζητά τον  $n$ -διάστατο χώρο για τα  $k$  στιγμιότυπα εκπαίδευσης που

είναι γειτονικά με τη δεδομένη είσοδο. Αυτές οι  $k$  περιπτώσεις εκπαίδευσης ονομάζονται «πλησιέστεροι γείτονες» της παρουσίας εισόδου. Η «εγγύτητα» των καθορισμένων περιπτώσεων εισόδου και εκπαίδευσης ορίζεται με βάση μια μέτρηση απόστασης, όπως η Ευκλείδεια απόσταση. Η δεδομένη παρουσία εισόδου μπορεί να ταξινομηθεί ως η πιο κοινή κατηγορία μεταξύ των «γειτόνων» της [58]. Υπάρχουν τρία κύρια στοιχεία αυτής της μεθόδου: ένα σύνολο περιπτώσεων εκπαίδευσης που επισημαίνονται, μια απόσταση μεταξύ των παρουσιών και η τιμή των πλησιέστερων γειτόνων  $k$ . Προκειμένου να ταξινομηθεί το στιγμιότυπο εισόδου, υπολογίζεται η απόστασή του από το επισημασμένο αντικείμενο, στη συνέχεια προσδιορίζεται ο αριθμός των πλησιέστερων γειτόνων και, τέλος, η κλάση των πλησιέστερων γειτόνων εκχωρείται ως ετικέτα στο δεδομένο στιγμιότυπο [59].

Για λόγους απλότητας παρατίθεται αρχικά ένα πρόβλημα κατηγοριοποίησης, όπου οι παρατηρήσεις αποτελούνται από δύο αριθμητικά πεδία και το γνώρισμα της κλάσης. Κάθε παρατήρηση μπορεί να θεωρηθεί ως ένα σημείο στον χώρο των δύο διαστάσεων. Μια παρατήρηση  $X$  απέχει από μια άλλη παρατήρηση  $Y$ , απόσταση  $d(X,Y)$  μέσα στον δισδιάστατο χώρο.

Η απόσταση  $d(X,Y)$  μπορεί να υπολογιστεί ως η Ευκλείδεια απόσταση σύμφωνα με την εξίσωση:

$$d(X, Y) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

όπου  $x_1, y_1$  οι τιμές των  $X$  και  $Y$  για την πρώτη διάσταση και  $x_2, y_2$  οι τιμές των  $X$  και  $Y$  για τη δεύτερη διάσταση. Σύμφωνα με τον αλγόριθμο  $k$ -NN, ο χρήστης προκαθορίζει την τιμή της σταθερής παραμέτρου  $k$ .

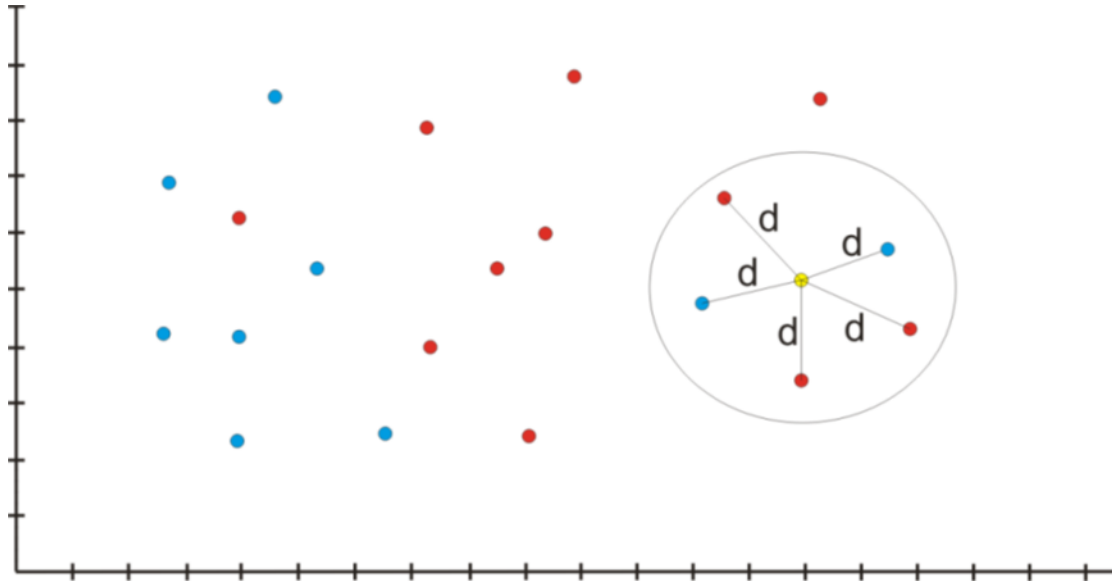
Ο αλγόριθμος αναζητά μέσα στον δισδιάστατο χώρο τα  $k$  σημεία-παρατηρήσεις που βρίσκονται πλησιέστερα στη νέα παρατήρηση. Ο κατηγοριοποιητής εκχωρεί τη νέα παρατήρηση στην κλάση που πλειοψηφεί μεταξύ των  $k$  πλησιέστερων γειτόνων. Εάν οριστεί ότι  $k=1$ , τότε η νέα παρατήρηση εκχωρείται στην κλάση της πιο όμοιας παρατήρησης εκπαίδευσης.

Όταν εφαρμόζεται στο άγνωστο σύνολο, το  $k$ -NN αναζητά τα  $k$  σύνολα εκπαίδευσης που είναι τα πιο κοντά στο άγνωστο σύνολο στον χώρο προτύπων του. Αυτά τα  $k$  σετ εκπαίδευσης αντιπροσωπεύουν το  $k$  «NN» των άγνωστων συνόλων. Η «εγγύτητα» των συνόλων ορίζεται με μια μετρική απόστασης, όπως η Ευκλείδεια απόσταση. Για την ταξινόμηση  $k$ -NN, στο άγνωστο σύνολο εκχωρείται η πιο κοινή κατηγορία μεταξύ του  $k$ -



NN. Για  $k=1$ , δίνεται στο άγνωστο σύνολο η κλάση του συνόλου εκπαίδευσης που είναι πιο κοντά σε αυτό στο χώρο προτύπων.

Γενικά, όταν ο αριθμός του συνόλου εκπαίδευσης είναι μεγαλύτερος, η τιμή του  $k$  θα είναι επίσης μεγαλύτερη [58]. Τα παραπάνω παρουσιάζονται διαγραμματικά στο Σχήμα 2.20.



**Σχήμα 2.20:** Εφαρμογή του  $k$ -NN αλγορίθμου Σχήμα: Κατηγοριοποιητής με  $K=5$

Στο παράδειγμα που παρουσιάζεται υπάρχουν δύο δυνατές τιμές κλάσης, οι οποίες συμβολίζονται με το χρώμα των σημείων. Το κίτρινο σημείο συμβολίζει τη νέα παρατήρηση που θα κατηγοριοποιηθεί. Στο παράδειγμα η τιμή του  $k$  έχει οριστεί να είναι 5. Εντοπίζονται τα 5 πλησιέστερα σημεία. Παρατηρούμε ότι τρία από αυτά είναι κόκκινα και δύο είναι μπλε. Η νέα παρατήρηση εκχωρείται στην «κόκκινη» κλάση [48].

Παρά την απλότητά της, η μέθοδος έχει μια ισχυρή θεωρητική βάση στη μη παραμετρική εκτίμηση της πυκνότητας και μπορεί συχνά να ξεπεράσει πολύ πιο εξελιγμένες μεθόδους. Η μέθοδος απαιτεί μόνο την επιλογή του  $k$ , ο αριθμός των γειτόνων που πρέπει να λαμβάνεται υπόψη κατά την ταξινόμηση. Οι μικρές τιμές του  $k$  θα επιλέξουν τα πλησιέστερα σημεία εκπαίδευσης που είναι καλύτερα σε θέση να εκτιμήσουν τη σωστή ταξινόμηση στο σημείο δοκιμής. Ωστόσο, λόγω των μικρών αριθμών, αυτή η εκτίμηση θα είναι επιρρεπής σε μεγάλες στατιστικές διακυμάνσεις. Αντίθετα, μεγάλες τιμές του  $k$  μειώνουν τα στατιστικά σφάλματα, αλλά επιτρέπουν σε μακρινά σημεία να συμβάλλουν στην ταξινόμηση, η οποία μπορεί να εξομαλύνει ορισμένες από τις λεπτομέρειες των κατανομών κλάσεων. Συνήθως το  $k$  επιλέγεται ως η τιμή που ελαχιστοποιεί το σφάλμα ταξινόμησης σε ορισμένα

ανεξάρτητα δεδομένα επικύρωσης ή με διαδικασίες διασταυρούμενης επικύρωσης. Όπως και με το k-means, το kNN παράγει απλώς μια ταξινόμηση για κάθε αντικείμενο χωρίς καμία σχετική οπτικοποίηση ή ερμηνεία, και ως εκ τούτου δεν χρησιμοποιείται συχνά μόνο του. Ωστόσο, η ανταγωνιστική του απόδοση στην ταξινόμηση έχει ενθαρρύνει τη χρήση του ως βάσης με την οποία μπορούν να συγκριθούν άλλες τεχνικές. [59], [60]

Η μέθοδος k-NN εκτός από κατηγοριοποίηση μπορεί να χρησιμοποιηθεί και για παλινδρόμηση, δηλαδή για πρόβλεψη αριθμητικών τιμών. Για την πρόβλεψη της τιμής της εξαρτημένης μεταβλητής μιας νέας παρατήρησης, ο αλγόριθμος εντοπίζει τις k πλησιέστερες παρατηρήσεις και επιστρέφει ως πρόβλεψη τη μέση τιμή των εξαρτημένων μεταβλητών των επιλεγμένων παρατηρήσεων.

Οι κατηγοριοποιητές k-NN διαθέτουν αξιολογικά πλεονεκτήματα [48]:

- είναι αποτελεσματικοί όταν υπάρχουν σύνθετες εξαρτήσεις μεταξύ των μεταβλητών
- διαθέτουν απλό αλγόριθμο
- σε πολλές περιπτώσεις επιτυγχάνουν υψηλές επιδόσεις κατηγοριοποίησης

Ωστόσο, παρά την αποτελεσματικότητά της, οι χρόνοι υπολογισμού μπορεί να είναι μεγάλοι με μεγάλες βάσεις δεδομένων, ο προσδιορισμός του αριθμού των γειτόνων που θα χρησιμοποιηθούν (k) απαιτεί δοκιμές για προσδιορισμό σφάλματος και ο αλγόριθμος είναι αδύναμος με ακραίες τιμές που μπορούν να επηρεάσουν έντονα την απόδοσή του. Στην λίστα που ακολουθεί συνοψίζονται ορισμένα από τα βασικά μειονεκτήματα των αλγορίθμων αυτών [48]:

- γίνονται πολλές συγκρίσεις μεταξύ παρατηρήσεων το οποίο απαιτεί πολύ αποτελεσματικές τεχνικές καταλογοποίησης (indexing)
- η κατηγοριοποίηση νέων παρατηρήσεων διαρκεί πολύ περισσότερο χρόνο, ειδικά στις περιπτώσεις όπου ο αριθμός των εν δυνάμει «γειτόνων» είναι μεγάλος
- τα αποτελέσματα τους μπορούν να επηρεαστούν σε σημαντικό βαθμό από το πλήθος των γειτόνων k
- είναι ευαίσθητοι σε τοπικά χαρακτηριστικά των δεδομένων
- είναι ευαίσθητοι στην ύπαρξη μη σημαντικών μεταβλητών εισόδου

## **2.4 Αξιοποίηση της Μηχανικής Μάθησης στην Έρευνα της Ρευματοειδούς Αρθρίτιδας: Ανασκόπηση επιστημονικών άρθρων και δημοσιεύσεων**

### **2.4.1 Επιτελική σύνοψη**

Στην ενότητα αυτή θα παρουσιάσουμε συνοπτικά το θέμα, τις μεθοδολογίες, τα μοντέλα και τα συμπεράσματα σημαντικών άρθρων και επιστημονικών δημοσιεύσεων που αναφέρονται στην αξιοποίηση της μηχανικής μάθησης στην ρευματοειδή αρθρίτιδα και παρουσιάζουν συνάφεια με το θέμα της παρούσας διπλωματικής εργασίας. Στην συνέχεια θα σχολιάσουμε ενδεικτικά ορισμένες αντιπροσωπευτικές εργασίες, τους τομείς που αφορούν και τις μελλοντικές προκλήσεις για περαιτέρω έρευνα στον τομέα. Για την παρουσίαση που ακολουθεί μελετήθηκαν τριάντα δύο (32) σχετικές εργασίες [43], [61]–[91] που έχουν δημοσιευτεί σε διεθνή επιστημονικά περιοδικά από το 2010 έως σήμερα.

Οι εργασίες καλύπτουν ένα ευρύ φάσμα ερευνητικών αντικειμένων, όπως την πρόγνωση νόσου (ύφεση, έξαρση, θνησιμότητα), εξατομικευμένη φαρμακευτική αγωγή σε ασθενείς με RA, μελέτη γονδιακού υπόβαθρου, διαφορική διαγνωστική της ρευματοειδούς αρθρίτιδας κ.ο.κ. Τα δεδομένα εισόδου των μελετών είναι κυρίως δημογραφικά, εργαστηριακά, δεδομένα εικόνας και κλινικά δεδομένα. Στον **Πίνακα 2.1** καταγράφονται τα βασικά στοιχεία για έντεκα (11) σημαντικές εργασίες, οι οποίες αναφέρονται συχνότερα σε πρόσφατες μελέτες ανασκόπησης εφαρμογών της μηχανικής μάθησης στην ρευματοειδή αρθρίτιδα [37], [41], [42], [44], [45], [92], [93].

Στην ενότητα 3.3.2. γίνεται παρουσίαση αντιπροσωπευτικών μελετών ανά κατηγορία ερευνητικού αντικειμένου, με έμφαση σε εργασίες που παρουσιάζουν συνάφεια σε ότι αφορά τα σύνολα δεδομένων που αξιοποιούν με τα στοιχεία της βάσης δεδομένων της παρούσας μελέτης, ώστε να ληφθούν υπόψη στο τεχνικό μέρος αυτής. [37], [41], [42]. Τέλος, υπάρχει ένας σχετικά μεγάλος αριθμός μελετών μηχανικής μάθησης στις οποίες ως δεδομένα εισόδου αξιοποιούνται εικόνες από απεικονιστικές μεθόδους (υπέρηχοι, αξονικές τομογραφίες, μαγνητικές τομογραφίες και ακτινογραφίες). Ενδεικτικές μελέτες με αντικείμενο την διάγνωση της ρευματοειδούς αρθρίτιδας με αναγνώριση εικόνας σχολιάζονται στην επόμενη ενότητα 3.3.2. [67], [68], [70]–[76], [91].

A/A	ΔΗΜΟΣΙΕΥΣΗ	ΤΙΤΛΟΣ	ΒΑΣΙΚΕΣ ΠΛΗΡΟΦΟΡΙΕΣ ΓΙΑ ΤΗΝ ΜΕΛΕΤΗ
1	Gossec et al. (2019) [79]	Detection of flares by decrease in physical activity, collected using wearable activity trackers in rheumatoid arthritis or axial spondyloarthritis: an application of Machine-Learning analyses in rheumatology	<p><b><u>Μέγεθος Δείγματος:</u></b> 82 ασθενείς με ΡΑ 73 ασθενείς με ΑΣ</p> <p><b><u>Τύπος Ανάλυσης:</u></b> Ταξινόμηση</p> <p><b><u>Τύπος Δεδομένων Εισόδου:</u></b> Δεδομένα δραστηριότητας (βήματα το λεπτό) Δημογραφικά δεδομένα Κλινικά δεδομένα</p> <p><b><u>Αλγόριθμοι:</u></b> Naïve Bayes Random Forests</p> <p><b><u>Έλεγχος Μοντέλου:</u></b> Διαχωρισμός του dataset 70% για εκπαίδευση και 30% για επικύρωση</p> <p><b><u>Αξιολόγηση:</u></b> Sensitivity: 96% , Specificity: 97%</p>
2	Norgeot et al (2019) [62]	Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis.	<p><b><u>Μέγεθος Δείγματος:</u></b> 578 ασθενείς Πανεπιστημιακό Νοσοκομείο, Τμήμα Ρευματολογίας, San Francisco 243 ασθενείς Γενικό Νοσοκομείο San Francisco.</p> <p><b><u>Τύπος Ανάλυσης:</u></b> Ταξινόμηση Βαθιά Μάθηση</p> <p><b><u>Τύπος Δεδομένων Εισόδου:</u></b> Demographics, prior CDAI score, ESR, CRP, DMARTs, glucocorticoids, autoantibodies</p> <p><b><u>Αλγόριθμοι:</u></b> Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD), Outcome posterior classifier, Change posterior classifier</p> <p><b><u>Έλεγχος Μοντέλου:</u></b> 117 ασθενείς</p>

A/A	ΔΗΜΟΣΙΕΥΣΗ	ΤΙΤΛΟΣ	ΒΑΣΙΚΕΣ ΠΛΗΡΟΦΟΡΙΕΣ ΓΙΑ ΤΗΝ ΜΕΛΕΤΗ
			<p><b><u>Αξιολόγηση:</u></b> 0.91 AUROC* SCORE</p>
3	Guan Y. et al., (2019) [63]	Machine Learning to Predict Anti-Tumor Necrosis Factor Drug Responses of Rheumatoid Arthritis Patients by Integrating Clinical and Genetic Markers	<p><b><u>Μέγεθος Δείγματος:</u></b> 1.892 PA ασθενείς (provided by DREAM Challenge Organizers)</p> <p><b><u>Τύπος Ανάλυσης:</u></b> Ταξινόμηση</p> <p><b><u>Τύπος Δεδομένων Εισόδου:</u></b> Δημογραφικά στοιχεία Θεραπευτική αγωγή DAS28</p> <p><b><u>Αλγόριθμοι:</u></b> Guassian process regression model (GPR) Το μοντέλο αυτό συγκρίθηκε με γραμμικά μοντέλα, δέντρα αποφάσεων και SVM.</p> <p><b><u>Έλεγχος Μοντέλου:</u></b> 680 ασθενείς (Corona Registry)</p> <p><b><u>Αξιολόγηση:</u></b> Accuracy: 77,8% 0,653 AUROC SCORE</p>
4	Odgers, D. et al, (2016) [43]	Using LASSO regression to predict rheumatoid arthritis treatment efficiency	<p><b><u>Μέγεθος Δείγματος:</u></b> 383 ασθενείς κατανεμημένοι σε τρεις κλάσεις βάσει θεραπευτικής αγωγής. (Stanford Hospital and Clinics)</p> <p><b><u>Τύπος Ανάλυσης:</u></b> Ταξινόμηση</p> <p><b><u>Τύπος Δεδομένων Εισόδου:</u></b> Δημογραφικά χαρακτηριστικά Εργαστηριακά δεδομένα Θεραπευτική αγωγή</p> <p><b><u>Αλγόριθμοι:</u></b> LASSO regression model</p> <p><b><u>Έλεγχος Μοντέλου:</u></b> Διασταυρωτικοί έλεγχοι στο 10% των δεδομένων σε κάθε κύκλο (10 επαναλήψεις)</p>

A/A	ΔΗΜΟΣΙΕΥΣΗ	ΤΙΤΛΟΣ	ΒΑΣΙΚΕΣ ΠΛΗΡΟΦΟΡΙΕΣ ΓΙΑ ΤΗΝ ΜΕΛΕΤΗ
			<p><b><u>Αξιολόγηση:</u></b></p> <p>AUC values μεταξύ 0.73-0.79</p> <p>Sensitivity: 0,43-0,67</p> <p>Specificity: 0.61-0,95</p>
5	Zhou S. et al. 2016 [64]	Defining disease phenotypes in primary care electronic health records by a machine learning approach: a case study in identifying rheumatoid arthritis	<p><b><u>Μέγεθος Δείγματος:</u></b></p> <p>15.459 ασθενείς από το πρωτοβάθμιο σύστημα υγείας* που διασταυρώθηκαν στο αρχείο του ρευματολογικού τμήματος του νοσοκομείου ABMU</p> <p>Αντίστοιχα 5.208 ασθενείς που διασταυρώθηκαν στο αρχείο του ρευματολογικού τμήματος του νοσοκομείου του CARDIFF.</p> <p>*(Farr Institute of Health Informatics Research database)</p> <p><b><u>Τύπος Ανάλυσης:</u></b></p> <p>Ταξινόμηση</p> <p><b><u>Τύπος Δεδομένων Εισόδου:</u></b></p> <p>43.100 εν δυνάμει κωδικοί μειώθηκαν σε 37 ομάδες κωδικών</p> <p><b><u>Αλγόριθμοι:</u></b></p> <p>Data mining algorithms</p> <p>Random Forest</p> <p>C5 Decision Tree</p> <p><b><u>Έλεγχος Μοντέλου:</u></b></p> <p>Το dataset AMBU χρησιμοποιήθηκε για την εκπαίδευση του αλγορίθμου</p> <p>Το dataset Cardiff χρησιμοποιήθηκε για την επικύρωση του αλγορίθμου</p> <p><b><u>Αξιολόγηση:</u></b></p> <p>94,6% specificity</p> <p>86,2% sensitivity</p> <p>92% overall accuracy</p>
6	Kalweit, M. et al. 2021 [78]	Personalized prediction of disease activity in patients with rheumatoid arthritis using an adaptive deep neural network.	<p><b><u>Μέγεθος Δείγματος:</u></b></p> <p>9.500 ασθενείς και 65.000 επισκέψεις από την βάση δεδομένων SCQM (Swiss Quality Management DB)</p> <p><b><u>Τύπος Ανάλυσης:</u></b></p> <p>Ταξινόμηση</p> <p><b><u>Τύπος Δεδομένων Εισόδου:</u></b></p>

A/A	ΔΗΜΟΣΙΕΥΣΗ	ΤΙΤΛΟΣ	ΒΑΣΙΚΕΣ ΠΛΗΡΟΦΟΡΙΕΣ ΓΙΑ ΤΗΝ ΜΕΛΕΤΗ
			<p>Δημογραφικά χαρακτηριστικά</p> <p>Εργαστηριακά δεδομένα</p> <p>Κλινική εικόνα</p> <p>Φαρμακευτική αγωγή</p> <p><b><u>Αλγόριθμοι:</u></b></p> <p>Deep neural network AdaptiveNet</p> <p>Συγκριτική αξιολόγηση με:</p> <p>Linear and logistic regression</p> <p>Random forest</p> <p>Support vector machine</p> <p><b><u>Έλεγχος Μοντέλου:</u></b></p> <p>DAS28-BSR&gt;2.6</p> <p><b><u>Αξιολόγηση:</u></b></p> <p>Ακρίβεια: 75,6%</p> <p>Ευαισθησία: 84,2%</p> <p>Εξειδίκευση:61,5%</p>
7	Alfredo Madrid-García et al. 2019 [80]	Outpatient Readmission in Rheumatology: A Machine Learning Predictive Model of Patient's Return to the Clinic	<p><b><u>Μέγεθος Δείγματος:</u></b></p> <p>&gt;18,000 γεγονότα εξιτηρίων</p> <p><b><u>Τύπος Ανάλυσης:</u></b></p> <p>Ταξινόμηση</p> <p><b><u>Τύπος Δεδομένων Εισόδου:</u></b></p> <p>Clinical Treatment</p> <p>Demographic variables</p> <p>Previous health care use</p> <p><b><u>Αλγόριθμοι:</u></b></p> <p>Radom forests</p> <p><b><u>Έλεγχος Μοντέλου:</u></b></p> <p>&gt; 200 γεγονότα</p> <p><b><u>Αξιολόγηση:</u></b></p> <p>AUC-ROC 0,653</p> <p>Ευαισθησία: 0,385</p> <p>Εξειδίκευση: 0,794</p>
8	Orange et al. 2018 [81]	Identification of Three Rheumatoid Arthritis Disease Subtypes by Machine Learning	<p><b><u>Μέγεθος Δείγματος:</u></b></p> <p>123 ασθενείς με RA</p>

A/A	ΔΗΜΟΣΙΕΥΣΗ	ΤΙΤΛΟΣ	ΒΑΣΙΚΕΣ ΠΛΗΡΟΦΟΡΙΕΣ ΓΙΑ ΤΗΝ ΜΕΛΕΤΗ
		Integration of Synovial Histologic Features and RNA Sequencing Data	<p>6 ασθενείς με OA (HSS NY)</p> <p><b>Τύπος Ανάλυσης:</b> Ταξινόμηση Συσταδοποίηση Στατιστική Ανάλυση</p> <p><b>Τύπος Δεδομένων Εισόδου:</b> Ιστολογικά χαρακτηριστικά (αρθρικού ιστού) Κλινικά Χαρακτηριστικά</p> <p><b>Αλγόριθμοι:</b> Kmeans, Support Vector Machines (SVMs), CIBERSORT ML system (Cell-type Identification by Estimating Relative Subsets of known RNA Transcripts)</p> <p><b>Έλεγχος Μοντέλου:</b> Σύγκριση των κλινικών χαρακτηριστικών των 123 ασθενών που είχαν ταξινομηθεί σε τρεις υποτύπους από το ιστολογικό μοντέλο μηχανικής μάθησης.</p>
9	Andersen et al. 2019 [67]	Neural networks for automatic scoring of arthritis disease activity on ultrasound images.	<p><b>Μέγεθος Δείγματος:</b> 40 ασθενείς με RA (παρακολούθηση για 6 μήνες) The hand with biopsy was US scanned at the baseline, 3 and 6<sup>th</sup> months (1342 Doppler Ultrasound Images) ΠΗΓΗ:RA Study: Clinical Trials.gov:Anct02652299)</p> <p><b>Τύπος Ανάλυσης:</b> Ταξινόμηση</p> <p><b>Τύπος Δεδομένων Εισόδου:</b> 1694 DUS Images with OESS Scores divided into a training, validation and test set. [An experienced rheumatologist scored the images according to the 4 point DUS OESS system scale]</p> <p><b>Αλγόριθμοι:</b> Convolutional Neural Networks</p> <p><b>Έλεγχος Μοντέλου:</b> 176 Doppler Ultrasound Images as a validation set 176 Doppler Ultrasound Images as a testing set</p> <p>The highest accuracies compared with an expert rheumatologist were 86.4% and 86.9% with a</p>



A/A	ΔΗΜΟΣΙΕΥΣΗ	ΤΙΤΛΟΣ	ΒΑΣΙΚΕΣ ΠΛΗΡΟΦΟΡΙΕΣ ΓΙΑ ΤΗΝ ΜΕΛΕΤΗ
			sensitivity of 0.864 and 0.875 and specificity of 0.864 and 0.864, respectively.
10	Vodencarevic A. et 2021 [77]	Advanced machine learning for predicting individual risk of flares in rheumatoid arthritis patients tapering biologic drugs.	<p><b><u>Μέγεθος Δείγματος:</u></b>  Δεδομένα από 135 επισκέψεις 41 ασθενών  ΠΗΓΗ:RETRO study.</p> <p><b><u>Τύπος Ανάλυσης:</u></b>  Ταξινόμηση</p> <p><b><u>Τύπος Δεδομένων Εισόδου:</u></b>  Δοσολογίες bDMARTs  Δείκτης ενεργότητας νόσου DAS-28 ESR  Διάρκεια νόσου  Δείκτες φλεγμονής</p> <p><b><u>Αλγόριθμοι:</u></b>  Logistic Regression  k-Nearest Neighbors  Naïve Bayes  Random Forest  Meta-Stacking Classifier</p> <p><b><u>Έλεγχος Μοντέλου:</u></b>  5X3 Nested cross – validation  Outer loop: The data is split into the training set (80%) and test set (20%)</p> <p><b><u>Αξιολόγηση:</u></b>  AUROC 0,81  (Meta-Stacking Classifier)</p>
11	Tao et al. 2021 [61]	Multimomics and Machine Learning Accurately Predict Clinical Response to Adalimumab and Etanercept Therapy in Patients With Rheumatoid Arthritis.	<p><b><u>Μέγεθος Δείγματος:</u></b>  80 ασθενείς με ΡΑ.</p> <p><b><u>Τύπος Ανάλυσης:</u></b>  Multidimensional scaling analysis (MDS)</p> <p><b><u>Τύπος Δεδομένων Εισόδου:</u></b>  .RNA-sequencing, DNA methylation  (from BLOOD SAMPLES)</p> <p><b><u>Αλγόριθμοι:</u></b>  Random Forest Algorithms</p> <p><b><u>Έλεγχος Μοντέλου:</u></b></p>

A/A	ΔΗΜΟΣΙΕΥΣΗ	ΤΙΤΛΟΣ	ΒΑΣΙΚΕΣ ΠΛΗΡΟΦΟΡΙΕΣ ΓΙΑ ΤΗΝ ΜΕΛΕΤΗ
			<p>9 εξωτερικοί ασθενείς επιλέχθηκαν για να επιβεβαιώσουν την πρόβλεψη από τα μοντέλα μηχανικής μάθησης.</p> <p><b>Αξιολόγηση:</b> Ακρίβεια πρόβλεψης ανταπόκρισης σε ADA και ETN 85,9%, (84,7%) και 79% (88%) αντίστοιχα.</p>

**ΠΙΝΑΚΑΣ 2.1:** Συναφείς μελέτες Μηχανικής Μάθησης στην Ρευματοειδή Αρθρίτιδα

\* Οι καμπύλες ROC (Receiver Operating Characteristics) είναι μέτρο για την εκτίμηση της ανά κλάση ακρίβειας του κατηγοριοποιητή και σχεδιάζονται σε έναν δυοδιάστατο επίπεδο χώρο. Ο οριζόντιος άξονας εκφράζει το μέγεθος  $1 - \text{specificity}$  το οποίο ονομάζεται και FalsePositiveRate, ενώ ο κατακόρυφος άξονας εκφράζει το μέγεθος  $2 - \text{sensitivity}$  που ονομάζεται και TruePositiveRate [48].

#### 2.4.2 Ενδεικτικά παραδείγματα μηχανικής μάθησης στην PA

##### A. Ανάλυση απεικονιστικών δεδομένων

Ο ρυθμός γήρανσης του πληθυσμού αυξάνεται ταχύτατα παγκοσμίως, ενώ η ποιότητα ζωής του ηλικιωμένου πληθυσμού επηρεάζεται από διάφορες ασθένειες και κυρίως από τα ρευματικά νοσήματα. Με τον αυξανόμενο ρυθμό εκδήλωσης της ρευματοειδούς αρθρίτιδας στον ηλικιωμένο πληθυσμό, οι τεχνολογικές εξελίξεις στον τομέα της επεξεργασίας και ανάλυσης εικόνας έχουν ανοίξει τον δρόμο για την αυτόματη ανίχνευση και διάγνωση της αρθρίτιδας. Τα νευρωνικά δίκτυα και η βαθιά μάθηση έχουν αξιοποιηθεί σε αρκετές μελέτες που στοχεύουν στην διάγνωση της PA με βάση απεικονιστικά δεδομένα [67], [68], [70]–[76], [91]. Μια πρόσφατα δημοσιευμένη μετα-ανάλυση του Liu [76] έδειξε ότι η διαγνωστική επίδοση της βαθιάς μάθησης στην ιατρική απεικόνιση είναι ισοδύναμη με αυτή ενός επαγγελματία γιατρού. Γενικώς θεωρείται ότι είναι ζήτημα χρόνου η βαθιά μάθηση να ξεπεράσει τον άνθρωπο στην αναγνώριση εικόνας.

Στην ανωτέρω εργασία η αρθρική περιοχή προσδιορίζεται χρησιμοποιώντας την τεχνική του ενεργού περιγράμματος. Η αρθρίτιδα κατηγοριοποιείται σε τέσσερις διαφορετικούς βαθμούς με βάση την μελέτη του υγρού στην αρθρική περιοχή. Οι διαφορετικοί βαθμοί ορίζονται και αναλύονται μέσω της βαθιάς μάθησης. Για τη διάγνωση του συγκεκριμένου βαθμού αρθρίτιδας χρησιμοποιήθηκε ένα συνελκτικό νευρωνικό δίκτυο ως αλγόριθμος βαθιάς μάθησης. Με την διαδικασία αυτή ορίστηκε μια μέθοδος για την αυτόματη ανίχνευση της φύσης της αρθρίτιδας.

Οι Liu και συνεργάτες με χρήση δισδιάστατων συνελκτικών νευρωνικών δικτύων πέτυχαν ευαισθησία 81% και εξειδίκευση 88% στην αυτόματη ανίχνευση των βλαβών στον αρθρικό χόνδρο. Αντίστοιχα, τρισδιάστατα συνελκτικά νευρωνικά δίκτυα ανίχνευσαν βλάβες σε μηνίσκους και επιγονατιδομηριαίους χόνδρους από σύνολα δεδομένων MRI με ευαισθησία/εξειδίκευση 90%/82% και 80%/80%. Τα ποσοστά αυτά είναι πολύ κοντά στα ποσοστά επιτυχίας των διαγνώσεων που γίνονται από κλινικούς γιατρούς [76].

### ***B. Ανάλυση δεδομένων ηλεκτρονικών φακέλων ασθενών***

Η μηχανική μάθηση εφαρμόζεται ολοένα και περισσότερο στους ηλεκτρονικούς φακέλους ασθενών (EMR) σε διάφορους ιατρικούς τομείς επειδή περιέχουν μεγάλα, ετερογενή σύνολα δεδομένων που μπορούν να χρησιμοποιηθούν για την εκπαίδευση αλγορίθμων για την ανίχνευση ασθενειών και ταξινόμηση ασθενών με χρήση μεθόδων εποπτευόμενης μάθησης. Μέχρι σήμερα, πολλές από τις προσεγγίσεις για την ανίχνευση της νόσου έχουν χρησιμοποιήσει SVM ή RF και εκτελούν ταξινόμηση με ανάθεση κατηγοριών όπως RA/non-RA [37].

Οι Carroll και συνεργάτες [66] χρησιμοποίησαν SVM για την ανίχνευση ασθενών με PA σε ένα EMR με διερεύνηση των κωδικών χρέωσης στα φάρμακα. Με ένα πολύ μικρό σύνολο δεδομένων (100 δείγματα εκπαίδευσης), πέτυχαν εξαιρετική βαθμολογία AUC του 0,97. Ο Lin και συνεργάτες [87] πρότειναν μια αυτόματη προσέγγιση για την ανίχνευση της δραστηριότητας της νόσου PA από ένα EMR. Διάφορα μοντέλα μηχανικής μάθησης εκπαιδεύτηκαν σε ένα σετ εκπαίδευσης >2500 κλινικών σημειώσεων και εργαστηριακών τιμών και στη συνέχεια αξιολογήθηκαν σε δύο σετ δοκιμών συνολικού αριθμού >2000 σημειώσεων. Τα μοντέλα εξήγαγαν όρους όπως αρθρίτιδα, πόνος ή δυσκαμψία ως δυνατότητες εισαγωγής χρησιμοποιώντας ένα εργαλείο ανάλυσης κειμένου στις κλινικές σημειώσεις, και χρησιμοποίησαν επίσης εργαστηριακές τιμές του CRP ή TKE. Με αυτές τις κλινικές και εργαστηριακές πληροφορίες, προβλέφθηκε η δραστηριότητα της νόσου κάθε ασθενή με PA χρησιμοποιώντας τις ακόλουθες κατηγορίες: υψηλή (DAS28>5.1), μέτρια (DAS28 3.2–5.1), χαμηλή (DAS28 2,6–3,2) και ύφεση (DAS28<2,6). Το καλύτερο SVM πέτυχε βαθμολογία AUC 0,831. Οι συγγραφείς κατέληξαν ότι η ανίχνευση της ενεργότητας της νόσου PA από τα δεδομένα EMR είναι μια πολλά υποσχόμενη μεθοδολογία με αποτελέσματα που προσεγγίζουν την ανθρώπινη απόδοση.

Ο Norgeot και συνεργάτες [62] μπόρεσαν να προβλέψουν την ενεργότητα της νόσου κατά τη στιγμή της επόμενης ρευματολογικής επίσκεψης σε μια δοκιμαστική ομάδα 116 ασθενών με ρευματοειδή αρθρίτιδα. Εφαρμόστηκε ένα μοντέλο βαθιάς γνώσης, το οποίο βασίστηκε

σε ηλεκτρονικούς φακέλους 820 ασθενών από δύο διαφορετικές ρευματολογικές κλινικές. Η μελέτη έδειξε ότι οι ηλεκτρονικοί φάκελοι υγείας μπορούν να χρησιμοποιηθούν για να αναπτυχθούν ακριβή μοντέλα που θα προβλέπουν σύνθετα αποτελέσματα της νόσου με την χρήση μοντέλων βαθιάς γνώσης. Για την πρόβλεψη της νόσου, η δραστηριότητα της κλινικής νόσου (CDAI) μεταφράζεται σε δύο καταστάσεις δραστηριότητας ύφεσης /χαμηλή (CDAI <10) ή μέτρια/υψηλή (CDAI>10). Τα χαρακτηριστικά που εξετάστηκαν ήταν δημογραφικά (ηλικία, φύλο και εθνικότητα), προηγούμενη βαθμολογία CDAI, επίπεδο ΤΚΕ και CRP, θεραπείες με ανοσοτροποποιητικά φάρμακα (DMARDs), από του στόματος και ενέσιμα, γλυκοκορτικοειδή, βιολογικοί παράγοντες και αυτοαντισώματα (PPI και/ή ACPA). Οι συγγραφείς ανέφεραν μια AUC 0,91 σε μια δοκιμαστική ομάδα 114 ασθενών. Το CDAI ήταν το πιο σημαντικό χαρακτηριστικό για την πρόβλεψη της νόσου, ακολουθούμενο από την χορήγηση / δοσολογία κορτιζόνης και την CRP. Τα DMARDs ήταν λιγότερο σημαντικά για την πρόβλεψη της δραστηριότητας της νόσου, δυνητικά λόγω του μικρού αριθμού ασθενών που έκαναν χρήση αυτών και οκτώ διαφορετικών DMARDs.

Ο Shiezadeh και συνεργάτες [88] πραγματοποίησαν ανίχνευση της PA σε ένα σύνολο δεδομένων 2500 ασθενών που παραπέμφθηκαν σε ρευματολόγο σε κλινική στο Ιράν. Επέλεξαν έντεκα (11) χαρακτηριστικά χρησιμοποιώντας τεχνικές επιλογής χαρακτηριστικών. Επώδυνες αρθρώσεις, φύλο, αριθμός των προσβεβλημένων αρθρώσεων και τα αποτελέσματα των δοκιμών της ESR είχαν τον μεγαλύτερο αντίκτυπο σχετικά με την ανίχνευση. Ως μέθοδο ML χρησιμοποίησαν ένα σύνολο δέντρων απόφασης και τα συνέκριναν με k-πλησιέστερους γείτονες και τεχνικές SVM. Το καλύτερό τους μοντέλο απέδωσε 85% ακρίβεια και ευαισθησία/εξειδίκευση 44% και 74% αντίστοιχα.

Ο Zhou [64] δημοσίευσε επιστημονική εργασία στην οποία ταξινομήσε ασθενείς με PA βασιζόμενος σε γονιδιακά δεδομένα που εξήχθησαν από τους ηλεκτρονικούς φακέλους ασθενών και επιλεγμένους κλινικούς κωδικούς (χαρακτηριστικά) που παρουσίαζαν μεγαλύτερη συχνότητα σε ασθενείς με PA σε σχέση με τον υγιή πληθυσμό. Αξιοποιήθηκε η μέθοδος RF για την μείωση του αριθμού δεικτών πρόβλεψης με κατάταξη της αξίας του κάθε χαρακτηριστικού ώστε να επιλεγούν τα σημαντικότερα χαρακτηριστικά. Με τον τρόπο αυτό μειώθηκαν τα 43100 χαρακτηριστικά σε 37 ομάδες κλινικών κωδικών. Τα πιο σημαντικά χαρακτηριστικά που επιλέχθηκαν με βάση το μοντέλο συγκεντρώθηκαν με άλλους κωδικούς, στις περιπτώσεις που προέκυψε ότι παρουσιάζουν μεγάλη ομοιότητα. Στη συνέχεια εκπαιδεύτηκε ένα δέντρο απόφασης πάνω σε αυτά τα χαρακτηριστικά για την ταξινόμηση ασθενών με PA. Ένα εξωτερικό σύνολο δεδομένων χρησιμοποιήθηκε για την

επικύρωση του μοντέλου. Το μοντέλο παρουσίασε 94,6% εξειδίκευση, 86,2% ευαισθησία και 92,3% ακρίβεια στην ταξινόμηση των ασθενών. Συμπερασματικά, η μελέτη αυτή αξιοποιεί αποτελεσματικά την μείωση των διαστάσεων για την επιλογή των χαρακτηριστικών και χρησιμοποιεί ένα εξωτερικό σύνολο δεδομένων για επαλήθευση.

Ο Feldman και συνεργάτες [90] το 2019 εφάρμοσαν το μοντέλο LASSO για την ταξινόμηση ασθενών ως προς την ενεργότητα νόσου PA, βασιζόμενοι σε δεδομένα αιτημάτων / διοικητικά δεδομένα, που είναι επίσης ένα είδος EMRs, αλλά σε πολύ μεγαλύτερη κλίμακα. Οι σχετικές βάσεις δεδομένων (claims data) συλλέγουν πληροφορίες για εκατομμύρια ραντεβού γιατρών, λογαριασμούς, ασφαλιστικές πληροφορίες και άλλες επικοινωνίες ασθενών- παρόχων υπηρεσιών. Επίσης διέθεταν δεδομένα φαρμακευτικής αγωγής και εργαστηριακά δεδομένα. Ο αριθμός των μεταβλητών διέφερε μεταξύ των μοντέλων. Το μοντέλο που βασίστηκε σε διοικητικά δεδομένα, φαρμακευτική αγωγή και εργαστηριακά δεδομένα πέτυχε την υψηλότερη ευαισθησία (83,1%), αλλά το μοντέλο με τα διοικητικά δεδομένα πέτυχε την μεγαλύτερη εξειδίκευση (74,7%) αποδεικνύοντας τα πλεονεκτήματα το να συγκρίνουμε διαφορετικά μοντέλα.

Στη μελέτη της Kalweit [78] χρησιμοποιήθηκαν δημογραφικά χαρακτηριστικά των ασθενών, κλινικά αποτελέσματα, αναφορές των ασθενών, εργαστηριακές τιμές και φαρμακευτική αγωγή ως είσοδος σε ένα βαθύ νευρωνικό δίκτυο για την πρόβλεψη της ενεργότητας νόσου με καλά αποτελέσματα.

### ***Γ. Ανάλυση βιομετρικών & ιστολογικών δεδομένων***

Ο Gossec και συνεργάτες [79] δημοσίευσαν μια μελέτη που βασίστηκε σε μοντέλα μηχανικής μάθησης για να αναπτύξει ένα φορητό διαγνωστικό αισθητήρα που μπορεί να διακρίνει την κίνηση από την ακινησία και την δραστηριότητα σε ασθενείς με PA. Εφάρμοσαν προεπεξεργασία των δεδομένων για την επιλογή των χαρακτηριστικών για την κατασκευή του μοντέλου. Επίσης συνέκριναν διαφορετικά μοντέλα, αρχικά με Αφελείς Μπαϊεσιανές (Naïve Bayes) και RF και στη συνέχεια RF, SMV και Βαθιά Μάθηση για να συγκρίνουν την ικανότητά τους να ταξινομούν αυτά τα δεδομένα δραστηριότητας. Αν η προσέγγιση επαληθεύονταν σε ένα μεγάλο ανεξάρτητο πληθυσμό ασθενών η παρακολούθηση της δραστηριότητας θα μπορούσε να παρέχει ένα μέσο πρόβλεψης εξάρσεων της νόσου ή να επιτρέψει την ταξινόμηση της ενεργότητας της νόσου.

Η Orange και συνεργάτες [81] εφάρμοσαν μηχανική μάθηση σε ακόμα πιο πολύπλοκα datasets. Συνδυάζοντας κλινικά και ιστολογικά δεδομένα με γενετικούς τύπους/ δείκτες προσδιόρισαν τρεις διακριτούς αρθρικούς υποτύπους στους ασθενείς με ρευματοειδή

αρθρίτιδα. Για την μελέτη αυτή αξιολογήθηκαν 20 ιστολογικά χαρακτηριστικά και 129 δείγματα αρθρικού ιστού. Η μηχανική μάθηση εφαρμόστηκε στα ιστολογικά χαρακτηριστικά με τους γενετικούς δείκτες να αποτελούν τις ετικέτες (labels), και αναπτύχθηκε ένας αλγόριθμος για την βαθμολόγηση των ιστολογικών χαρακτηριστικών.

Οι ασθενείς με τον υψηλά φλεγμονώδη αρθρικό υποτύπο εμφάνισαν υψηλότερα επίπεδα C-αντιδρώσας πρωτεΐνης (CRP) και ΤΚΕ, θετικό τίτλο ρευματοειδούς παράγοντα και θετικά αντισώματα αντι-CCP (αντι-κυκλικού κιτρουλινωμένου πεπτιδίου). Επιπλέον, τα επίπεδα CRP συσχετίστηκαν σημαντικά με την ένταση πόνου στην υποομάδα υψηλής φλεγμονής.

Αρθρικά ιστολογικά χαρακτηριστικά των ίδιων 45 δειγμάτων βιοψίας χρησιμοποιήθηκαν ως δεδομένα εισόδου για την εκπαίδευση αρκετών δυαδικών ταξινομητών SVM για την πρόβλεψη ενός επισημασμένου φλεγμονώδους υποτύπου από το άλλα δύο. Ο έλεγχος του μοντέλου είχε καλή επίδοση (AUC = 0,88) όταν γινόταν ταξινόμηση του υποτύπου υψηλής φλεγμονής από τους άλλους δύο. Το μοντέλο ήταν λιγότερο επιτυχημένο στην ταξινόμηση του υποτύπου μικτής φλεγμονής (AUC = 0,59).

Η εργασία αυτή δείχνει ότι η ML μπορεί να χρησιμοποιηθεί για τον προσδιορισμό των ιστολογικών χαρακτηριστικών που είναι σε θέση να προβλέψουν καλύτερα τον μοριακό φαινότυπο αρθρικών δειγμάτων και πώς τα ιστολογικά χαρακτηριστικά μπορούν στη συνέχεια να χρησιμοποιηθούν για την πρόβλεψη του γονιδιακού προφίλ έκφρασης σε ασθενείς με PA. Παρότι η συγκεκριμένη μελέτη έχει ορισμένους περιορισμούς (π.χ. το dataset ήταν σχετικά μικρό) αποδεικνύει την δυναμική των εφαρμογών μηχανικής μάθησης στην ρευματοειδή αρθρίτιδα. Είναι ένας πρώτος δείκτης για το πως η μηχανική μάθηση μπορεί να μας βοηθήσει να προσδιορίσουμε νέες υποομάδες ασθενών στην ρευματοειδή αρθρίτιδα στο μέλλον.

#### ***Δ. Ανάλυση κινδύνου για πρόβλεψη εξέλιξης της νόσου***

Ο Briggs και συνεργάτες, [83] σε μελέτη τους χρησιμοποίησαν RF και λογιστική παλινδρόμηση για την διερεύνηση γενετικών αλληλεπιδράσεων σε γονίδια της PA με στόχο τον εντοπισμό του κινδύνου εμφάνισης της PA.

Η διερεύνηση των γενετικών αλληλεπιδράσεων έχει αποδειχθεί δύσκολη παρά τις πρόσφατες προόδους τόσο στις εργαστηριακές μεθόδους όσο και στα στατιστικά μοντέλα. Χωρίς να υπάρχει διαθέσιμη «καλύτερη» στατιστική προσέγγιση, ο συνδυασμός πολλών αναλυτικών μεθόδων μπορεί να είναι η βέλτιστη μέθοδος στην ανίχνευση γενετικών αλληλεπιδράσεων. Χρησιμοποιώντας μια ανάλυση πολλαπλών σταδίων που ενσωμάτωσε εποπτευόμενη μηχανική εκμάθηση και μεθόδους ελέγχων συσχέτισης, διερευνήθηκαν οι

γενετικές αλληλεπιδράσεις με έναν καλά εδραιωμένο γενετικό παράγοντα (PTPN22 1858T) σε ένα πολύπλοκο αυτόανοσο νόσημα όπως η ρευματοειδής αρθρίτιδα. Η ανάλυση αποτελείται από τέσσερα κύρια στάδια: Στάδιο I (μείωση δεδομένων)- ταυτοποίηση υποψήφιων χρωμοσωμικών περιοχών με τον εποπτευόμενο αλγόριθμο μηχανικής μάθησης Τυχαία Δάση (RF). Στάδιο II (επέκταση ανάλυσης) — έλεγχος λεπτομερών γενετικών δεδομένων σε υποψήφιες χρωμοσωμικές περιοχές με PTPN22 1858T σε 677 περιπτώσεις και με 750 ελέγχους με χρήση λογιστικής παλινδρόμησης. Στάδιο III (επιβεβαίωση)— επιβεβαίωση γενετικών αλληλεπιδράσεων σε 947 περιπτώσεις και με 1756 ελέγχους. Στάδιο IV (συνδυασμένη ανάλυση)— μια συγκεντρωτική ανάλυση που περιλαμβάνει και τις 1624 περιπτώσεις PA και 2506 αντικείμενα ελέγχου για τις τελικές εκτιμήσεις του αποτελέσματος. Ταυτοποιήθηκαν συνολικά επτά επαναλαμβανόμενες αλληλεπιδράσεις.

Στη παρουσίαση των αποτελεσμάτων αναφέρεται ότι αν και τα μοντέλα RF θα μπορούσαν να χρησιμοποιηθούν για την πρόβλεψη κινδύνου, είναι απαραίτητη και η στατιστική ανάλυση με λογιστική παλινδρόμηση. Επιπλέον, είναι δύσκολο να αξιολογηθεί αν η ML είτε οι περισσότερες τυπικές προσεγγίσεις είναι πιο ακριβείς για την πρόβλεψη των γονιδίων κινδύνου, αλλά η μέθοδος RF μπορεί να είναι σε θέση να προσδιορίσει τη σημασία του κάθε γονιδίου μέσω της χρήσης πολλών διαφορετικών δέντρων αποφάσεων [83]. Παρότι η ML χρησιμοποιήθηκε για αυτό το σκοπό, παραμένει ασαφές εάν η RF ή η λογιστική παλινδρόμηση είναι η καταλληλότερες μέθοδοι για αναγνώριση των γονιδίων κινδύνου. Αντ' αυτού, θα μπορούσε κανείς να εξετάσει τη χρήση της ML για την ταξινόμηση των ασθενών με βάση τα γονίδια κινδύνου που έχουν προσδιοριστεί με τυπικές αναλυτικές μεθόδους [41]. Ο Chin και συνεργάτες [94] πραγματοποίησαν μια μεγάλης κλίμακας αξιολόγηση κινδύνου εκδήλωσης PA σε πρώιμο στάδιο με αξιοποίηση των δεδομένων EMR στην Ταϊβάν. Ο στόχος της έγκαιρης αξιολόγησης κινδύνου ήταν η ανακάλυψη κρυμμένων παραγόντων και η δημιουργία μοντέλων αξιολόγησης με βάση κωδικούς διάγνωσης που προηγούνταν της επίσημης διάγνωσης της PA.

Το σύνολο δεδομένων αποτελείτο από 1000 ασθενείς με PA και 500.000 ασθενείς χωρίς PA, και χρήση των κωδικών διάγνωσης ως χαρακτηριστικά εισόδου. Στόχος ήταν ο υπολογισμός των λανθανόντων παραγόντων κινδύνου. Χρησιμοποιώντας ένα SVM, εντόπισαν τη PA πρώιμου σταδίου με ευαισθησία και ειδικότητα 74 και 70%, αντίστοιχα [37].

Ένα μοντέλο πρόβλεψης θνησιμότητας που βασίζεται σε μια Τυχαία Δάσος Επιβίωσης χρησιμοποιήθηκε από τους Lezcano-Valverde και συνεργάτες [86] για την πρόβλεψη της θνησιμότητας ασθενών PA, χρησιμοποιώντας ένα εκπαιδευτικό σετ κοόρτης με >1400

ασθενείς από μια κλινική στο Σαν Κάρλος και ένα σετ επικύρωσης από 280 ασθενείς με ρευματοειδή αρθρίτιδα από κλινική στη Μαδρίτη. Το δέντρο επιβίωσης εντόπισε πέντε ομάδες κινδύνου θνησιμότητας. Επιτεύχθηκε ευαισθησία 0,79–0,80 και ειδικότητα 0,43–0,48 για ένα και επτά χρόνια παρακολούθησης αντίστοιχα.

Ο Huggle και συνεργάτες [37] χρησιμοποίησαν ένα νέο δυναμικό βαθύ νευρωνικό δίκτυο, σχεδιασμένο για πολυτροπικά κλινικά δεδομένα, για πρόβλεψη της εξέλιξη της νόσου σε επισκέψεις παρακολούθησης ασθενών με PA. Η δυναμική αρχιτεκτονική βαθιών νευρωνικών δικτύων ξεπέρασε τα τυχαία δάση και τα πλήρως συνδεδεμένα νευρωνικά δίκτυα, επιτυγχάνοντας μέσο τετραγωνικό σφάλμα 0,9, το οποίο αντιστοιχεί σε σφάλμα 8% στο εύρος της τιμής στόχου.

### ***E. Πρόβλεψη της κλινικής ανταπόκρισης στη φαρμακευτική αγωγή***

Η αξιολόγηση της θεραπευτικής ανταπόκρισης μέσω της μηχανικής μάθησης είναι αντικείμενο πολλά υποσχόμενων μελετών. Στην εποχή των ακριβών βιολογικών θεραπειών, η ικανότητα της πρόβλεψης με ακρίβεια της ανταπόκρισης σε στοχευμένες θεραπείες είναι επωφελής για τους ασθενείς και τους κλινικούς γιατρούς, έτσι ώστε η εστίαση να μπορεί να στραφεί σε θεραπείες που είναι πιθανότερο να είναι αποτελεσματικές.

Η δυνατότητα να προβλέψεις πως μπορεί να ανταποκριθεί ο ασθενής σε ένα φάρμακο θα απομάκρυνε την απόφαση για την θεραπευτική προσέγγιση από την λογική δοκιμής και λάθους, θα μείωνε τις ασθένειες από παρενέργειες φαρμάκων καθώς επίσης και τα οικονομικά βάρη του συστήματος. Οι προσεγγίσεις της μηχανικής μάθησης είναι πολλά υποσχόμενες για την παροχή προσωποποιημένες θεραπευτικής αγωγής αλλά η εφαρμογή της πρέπει να βελτιστοποιηθεί [84].

Το 2014 δημοσιεύτηκε η ανοικτή πρόσκληση Dream Challenge [85], σε ενδιαφερόμενους ερευνητές για να σχεδιάσουν μοντέλα που μπορεί να προβλέψουν την ανταπόκριση στη θεραπεία εξουδετέρωσης του TNF σε ασθενείς με PA [84], [85]. Για την εκπαίδευση των μοντέλων χρησιμοποιήθηκαν ολοκληρωμένα δημογραφικά δεδομένα, ιστορικά θεραπείας, δεδομένα SNP (single nucleotide polymorphism) και δραστηριότητα της νόσου για 1.892 ασθενείς με ρευματοειδή αρθρίτιδα.

Στο πλαίσιο αυτής της πρόσκλησης, ο Guan και συνεργάτες [63] παρουσίασαν το 2019 μια μελέτη μηχανικής μάθησης για την πρόβλεψη της anti-TNF απόκρισης. Για την υλοποίηση της μελέτης 1892 ασθενείς με PA από 13 κοόρτες ασθενών επιλέχθηκαν τυχαία. Αυτά τα δεδομένα χρησιμοποιήθηκαν ως σύνολα δεδομένων εκπαίδευσης (training data set) και στη συνέχεια αξιολογήθηκαν από ένα ανεξάρτητο σύνολο δεδομένων 680 ατόμων από το



μητρώο CORRONA. Το μοντέλο αποδείχθηκε χρήσιμο στον εντοπισμό υποκλάσεων μη ανταποκρινόμενων στην αντι- TNF θεραπεία. Η μελέτη είναι πολλά υποσχόμενη ως μια πρώτη προσέγγιση στην καθοδήγηση για χορήγηση φαρμάκων στην κλινική πρακτική με βάση την συγκεκριμένη προσέγγιση.

Η ερευνητική ομάδα χώρισε τους ασθενείς με βάση τη θεραπεία εξουδετέρωσης του TNF (adalimumab, etanercept ή infliximab) και κατασκεύασαν πολλαπλά εποπτευόμενα μοντέλα ML για κάθε θεραπεία με στόχο την πρόβλεψη της ποσοτικής αλλαγής στη βαθμολόγηση ενεργότητας της νόσου PA. Η τιμή της βαθμολογίας χρησιμοποιήθηκε στη συνέχεια για την ταξινόμηση των ασθενών ως ανταποκρινόμενους ή μη ανταποκρινόμενους στη θεραπεία. Τα μοντέλα που δημιουργήθηκαν από παλινδρόμηση Gauss, gradient-boosting, συγκρίθηκαν με άλλους αλγόριθμους παλινδρόμησης ή RF. Το μοντέλο μπόρεσε να προβλέψει τη μη ανταπόκριση στη θεραπεία μεταξύ ασθενών με PA ευρωπαϊκής καταγωγής με ακρίβεια 78%. Σε ανεξάρτητο έλεγχο δοκιμάζοντας ένα σύνολο δεδομένων 680 ασθενών, το μοντέλο πέτυχε ένα AUC 0,615.

Από την εργασία αυτή προκύπτει το συμπέρασμα ότι η μηχανική μάθηση έχει την δυνατότητα να διευκολύνει την καλύτερη διαφοροποίηση μεταξύ των εξατομικευμένων περιπτώσεων και οι ασθενείς μπορούν να επωφεληθούν από μια περισσότερο εντατική θεραπεία και φροντίδα. Επίσης ασθενείς με καλύτερη έκβαση της νόσου μπορούν να επωφεληθούν από την έγκαιρη τροποποίηση της φαρμακευτικής τους αγωγής ώστε να αποφύγουν τις παρενέργειες μιας υπερβατικής θεραπευτικής προσέγγισης.

Ο Ταο και συνεργάτες κατασκεύασαν μοντέλα ML για να προβλέπουν την ανταπόκριση της σε αντι-TNF βιολογικές θεραπείες ασθενών με PA [58]. Το μοντέλο RF που κατασκευάστηκε, ήταν αρκετά ακριβές στην πρόβλεψη της ανταπόκρισης στο adalimumab. Οι θέσεις CpG ήταν πιο ακριβείς στην πρόβλεψη ανταπόκρισης στην ετανερσέπτη. Ωστόσο, απαιτείται εξωτερική επικύρωση με μεγαλύτερο μέγεθος δείγματος, δεδομένου ότι οι συγγραφείς πραγματοποίησαν επικύρωση με μόνο εννέα εξωτερικά δείγματα και η μελέτη περιορίζεται από ένα μικρό μέγεθος δείγματος, δηλαδή με μόνο 40 ασθενείς σε κάθε ομάδα θεραπείας που χρησιμοποιήθηκε στην κατασκευή μοντέλων.

Άλλοι ερευνητές έχουν προτείνει επίσης την εφαρμογή της ML στη θεραπεία της PA με βιολογικά φάρμακα. Για παράδειγμα, σε μια παρουσίαση στο πλαίσιο συνεδρίου το 2005 παρουσιάστηκε η χρήση των μοντέλων SVM και βαθιάς μάθησης για την πρόβλεψη εάν οι ασθενείς με PA λαμβάνουν αυξημένες δόσεις infliximab [86]

Στην εργασία του Vodencarevic και συνεργατών η ML χρησιμοποιήθηκε για την πρόβλεψη των εξάρσεων που ακολουθούν τη μείωση του φαρμάκου σε 41 ασθενείς με PA. Ως δεδομένα εισόδου χρησιμοποιήθηκαν δεδομένα ενεργότητας της νόσου, δεδομένα φαρμακευτικής αγωγής και εργαστηριακά δεδομένα που προέκυψαν από μια κλινική δοκιμή. Αν και περιορίζεται από ένα μικρό μέγεθος δείγματος η μελέτη αξιολόγησε πολλαπλά μοντέλα ML (συμπεριλαμβανομένων λογιστικής παλινδρόμησης, απλής Bayes, k-NN και RF) για την ικανότητά τους να προβλέπουν με ακρίβεια τις εξάρσεις της νόσου. Επιπλέον, η μελέτη χρησιμοποίησε ένα μοντέλο πολλαπλών βημάτων (στοίβαξη), στο οποίο τα αποτελέσματα των τεσσάρων ταξινομητών χρησιμοποιήθηκαν ως μεταβλητές σε ένα μοντέλο λογιστικής παλινδρόμησης. Μετά από διασταυρούμενη επικύρωση, το στοίβαγμα λογιστικής παλινδρόμησης πέτυχε το υψηλότερο AUC 0,81. Είναι σημαντικό ότι αυτή η μελέτη δείχνει ότι η ML έχει τη δυνατότητα να βοηθήσει στην ανάλυση κλινικών δεδομένων, ωστόσο, όπως προτείνουν οι συγγραφείς, τα μοντέλα χρειάζονται να επικυρωθούν σε μια μεγαλύτερη κούρτη και το μοντέλο να δοκιμαστεί σε περιβάλλον κλινικής περίθαλψης, για να προσδιοριστεί εάν είναι πράγματι εφαρμόσιμο σε πραγματικές συνθήκες εκτός ενός αυστηρά ελεγχόμενου περιβάλλοντος.

Δεδομένου ότι η μακροπρόθεσμη ανάπτυξη της νόσου μπορεί να επηρεαστεί από πολλούς απρόβλεπτους παράγοντες, η πρόβλεψη της εξέλιξης της ασθένειας είναι πολύ πιο περίπλοκη από την ανίχνευση της ασθένειας. Ωστόσο, η ML μπορεί να χρησιμοποιηθεί για μάθηση μοντέλων πρόβλεψης εξέλιξης ασθενειών, που παρουσιάζουν μεγάλο ενδιαφέρον για την επιλογή θεραπείας και την λήψη αποφάσεων.

# 3

## *Περιγραφή Δεδομένων*

### *3.1 Εισαγωγή*

Το αρχικό σύνολο δεδομένων που χρησιμοποιήθηκε στην παρούσα διπλωματική εργασία, συγκεντρώθηκε και καταχωρίστηκε σε φύλλο excel από ειδικευμένη ρευματολόγο στο πλαίσιο διδακτορικής διατριβής που βρίσκεται σε εξέλιξη στην Ιατρική Σχολή του Εθνικού Καποδιστριακού Πανεπιστημίου Αθηνών.

Το σύνολο δεδομένων αφορά σε τριακόσιους (300) ασθενείς με ρευματοειδή αρθρίτιδα από το Νοσοκομείο Ευαγγελισμός και Λαϊκό Νοσοκομείο Αθηνών, που συλλέχθηκαν την περίοδο 2017-2020. Το σύνολο δεδομένων αξιοποιήθηκε για την εφαρμογή μεθοδολογιών και μοντέλων μηχανικής μάθησης και την εξαγωγή σχετικών συμπερασμάτων.

Για τον κάθε ασθενή έχουν καταχωριστεί σαράντα τέσσερα (44) συνολικά χαρακτηριστικά, με βάση τα στοιχεία που τηρούνται στους ιατρικούς φακέλους των ασθενών σε αναλογική μορφή (φυσικοί φάκελοι) στις ρευματολογικές κλινικές των παραπάνω νοσοκομείων. Ειδικότερα, έγινε ανασκόπηση των αρχείων των ασθενών και κατεγράφησαν δημογραφικά, κλινικά, εργαστηριακά, ανοσολογικά και ακτινολογικά χαρακτηριστικά για τον κάθε ασθενή του δείγματος. Επιπλέον, καταχωρήθηκαν χαρακτηριστικά που αφορούν στο ιατρικό ιστορικό του ασθενούς και τις υποκείμενες συννοσηρότητες καθώς και τις θεραπείες που έχει λάβει ο ασθενής μέχρι την στιγμή της συλλογής των στοιχείων. Η πρώτη στήλη του πίνακα αφορά στον μοναδικό χαρακτηριστικό αριθμό (RecordID) για κάθε ασθενή και πρόκειται για τον αύξοντα αριθμό της κάθε γραμμής του πίνακα. Δεδομένου ότι ο αριθμός αυτός δεν περιέχει κάποια χρήσιμη πληροφορία και υπάρχει για λόγους διάκρισης, δεν χρησιμοποιείται σε κανένα από στα μοντέλα που αναπτύχθηκαν.

Κατά την επεξεργασία του συνόλου δεδομένων, για τις ανάγκες της διπλωματικής εργασίας προστέθηκαν δύο επιπλέον στήλες προκειμένου στη συνέχεια να αξιοποιηθούν για τον έλεγχο και την επαλήθευση των μοντέλων. Η πρώτη στήλη που προστέθηκε αφορά την δραστηριότητα /ενεργότητα της νόσου (activity) και αξιολογείται από το 1 έως το 3, όπου το 1 αντιπροσωπεύει νόσο σε ύφεση, το 2 μερικώς ενεργή (σταθερή) νόσο και το 3 νόσο σε έξαρση. Ομοίως, η στήλη της σοβαρότητας (severity) αξιολογήθηκε στην κλίμακα από 1 έως 3, όπου όταν έχει τιμή 1 πρόκειται για μια ελαφριά ρευματοειδή, όταν έχει τιμή 2 πρόκειται για μια ρευματοειδή μεσαίας σοβαρότητας και όταν έχει τιμή 3 πρόκειται για μια σοβαρή ρευματοειδή. Η αξιολόγηση των ασθενών και η συμπλήρωση της σωστής τιμής έγινε από δύο ανεξάρτητους κλινικούς ιατρούς του Νοσοκομείου Ευαγγελισμός, με αποτέλεσμα οι στήλες αυτές να απεικονίζουν τις απόψεις του ειδικού ρευματολόγου για τον κάθε ασθενή.

Θα πρέπει να αναφέρουμε ότι το σύνολο των δεδομένων έχει συλλεχθεί από δύο από τις μεγαλύτερες ρευματολογικές κλινικές στην Ελλάδα και για το λόγο αυτό περιλαμβάνει κατά κύριο λόγο ασθενείς που έχουν μέτρια ή/και σοβαρή ρευματοειδή αρθρίτιδα, ώστε να οδηγηθούν στην τριτοβάθμια ιατρική περίθαλψη. Σίγουρα θα είχε προκύψει ένα αντιπροσωπευτικότερο δείγμα ασθενών αν επιπλέον περιείχε δεδομένα και από μικρά διαγνωστικά κέντρα ή ιδιωτικά ιατρεία και κλινικές.

Τα χαρακτηριστικά της βάσης δεδομένων περιγράφονται στους παρακάτω πίνακες, με μια σύντομη αναφορά σε κάθε ένα από τα αυτά τα χαρακτηριστικά που αφορά στις τιμές του δείγματος των ασθενών.

A/A	Δημογραφικά και Λοιπά Δεδομένα Ιστορικού Ασθενούς	Εύρος τιμών Δείγματος
1	Φύλο	1- Άνδρας, 2- Γυναίκα
2	Ηλικία	Min: 18 (χρονών) Max: 90 (χρονών)
3	Διάρκεια Νόσου	Min: 10 (μήνες) ή 0.8 (χρόνια) Max: 660 (μήνες) ή 55 (χρόνια)
4	BMI (Δείκτης Μάζας Σώματος)	Min: 17.3 Max: 48.7
5	Κάπνισμα	Ναι ή Όχι
6	DM2 (Διαβήτης Τύπου 2)	Ναι ή Όχι
7	ILD (Διάμεσες Πνευμονοπάθειες)	Ναι ή Όχι

8	THYROID (Προσβολή θυροειδούς)	Ναι ή Όχι
9	TBC (Φυματίωση)	Ναι ή Όχι
10	CANCER (Καρκίνος)	Ναι ή Όχι
11	OSTEOPOROSIS (Οστεοπόρωση)	Ναι ή Όχι
12	DEPRESSION (Κατάθλιψη)	Ναι ή Όχι
13	HYPERURICEMIA (Υπερουριχαιμία)	Ναι ή Όχι
14	HBV (Ηπατίτιδα Β)	Ναι ή Όχι
15	ΑΥ(Αρτηριακή Υπέρταση)	Ναι ή Όχι
16	DYSLIPIDAEMIA (Δυσλιπιδαιμία)	Ναι ή Όχι
17	ARTHROPLASTY (Αρθροπλαστική)	Ναι ή Όχι
18	H.ZOSTER (Ερπης Ζωστήρ)	Ναι ή Όχι

**ΠΙΝΑΚΑΣ 3.1:** Δημογραφικά Δεδομένα και Συνυπάρχουσες Νόσοι

A/A	Εργαστηριακά Δεδομένα	Φάσμα Τιμών Δείγματος	Φυσιολογικές Τιμές (Εργαστήρια Ευαγγελισμού)
1	HCT (Αιματοκρίτης)	Min: 25.9 Max: 48.8	37-47
2	HGB (Αιμοσφαιρίνη)	Min: 9 Max: 16.7	12-15
3	ΤΚΕ (Ταχύτητα Καθιζήσεως Αίματος)	Min: 2 Max: 120	<30 mm/h
4	CRP (C-Αντιδρώσα Πρωτεΐνη)	Min: 0.01 Max: 19	<0,5 mg/dl
5	GLU (Γλυκόζη)	Min: 61 Max: 201	70-110 mg/dl
6	CHOL (Ολική Χοληστερίνη)	Min: 95 Max: 321	< 200 mg/dl
7	LDL (λιποπρωτεΐνη χαμηλής πυκνότητας)	Min: 42 Max: 208	< 115 mg/dl

8	HDL (λιποπρωτεΐνη υψηλής πυκνότητας)	Min: 24 Max: 116	>45 mg/dl
9	RFi	(+) ή (-)	Θετικό, Αρνητικό
10	RF (Ρευματοειδής Παράγοντας)	Min: 1.3 Max: 4370	< 20 IU/ml
11	CCPi	(+) ή (-)	Θετικό - Αρνητικό
12	CCP (αντισώματα έναντι του κυκλικού κιτρογλυκωμένου πεπτιδίου)	Min: 0.5 Max: 7000	0,1-30
13	HLA_DRB1 (αντιγόνα ιστοσυμβατότητας)	(+) ή (-)	Θετικό - Αρνητικό

**ΠΙΝΑΚΑΣ 3.2:** Δεδομένα Εργαστηριακών Εξετάσεων

A/A	Δείκτες σοβαρότητας και δραστηριότητας ΡΑ	Τιμές Δείγματος
1	TENDER (αριθμός ευαίσθητων στις πίεση αρθρώσεων)	Min: 0 Max: 28
2	SWOLLEN (αριθμός διογκωμένων αρθρώσεων)	Min: 0 Max: 28
3	VAS: visual analogue scale of handicap from rheumatoid arthritis of the hand (δείκτης αναπηρίας)	Min: 0 Max: 90
4	HAQ: health assessment questionnaire (δείκτης αναπηρίας)	Min: 0 Max: 2.63
5	DAS28_TKE: Disease Activity Score Calculator for Rheumatoid Arthritis με TKE	Min: 0.25 Max: 8.74
6	DAS28_CRP: Disease Activity Score Calculator for Rheumatoid Arthritis με CRP	Min: 1.08 Max: 8.35
7	Sharp Score: A scoring system used to quantify the Radiological changes in patients with rheumatoid arthritis.	Min: 3 Max: 156

**ΠΙΝΑΚΑΣ 3.3:** Δείκτες Σοβαρότητας και Ενεργότητας Νόσου

A/A	Δεδομένα θεραπευτικής αγωγής	Τιμές Δείγματος
2	STEROID (Χορήγηση στεροειδών)	(+) ή (-)
1	STEROID DOSE (Δοσολογία στεροειδών)	Min: 2.5 (mg) Max: 15 (mg)
5	cDMARD: τροποποιητικά της νόσου αντιρευματικά φάρμακα. Αυτά που χρησιμοποιούνται συχνότερα είναι η μεθοτρεξάτη, η σουλφασαλαζίνη, η υδροξυχλωροκίνη και η λεφλουνομίδη. DMARDs που χρησιμοποιούνται λιγότερο συχνά είναι τα άλατα χρυσού, η αζαθειοπρίνη και η κυκλοσπορίνη.	(+) ή (-)
6	bDMARD: Μία άλλη τάξη φαρμάκων που χρησιμοποιούνται στους ασθενείς με ρευματοειδή αρθρίτιδα και τις σχετιζόμενες με αυτή φλεγμονώδεις ρευματικές παθήσεις είναι οι βιολογικοί παράγοντες, οι οποίοι λέγονται και βιολογικά DMARDs και στους οποίους περιλαμβάνονται η ετανεрсέπη, η ανταλιμουμάμπη, η ινφλιξιμάμπη, η κερτολιζουμάμπη πεγόλη και η γκολιμουμάμπη. Τα παραπάνω φάρμακα ανήκουν στους αναστολείς του TNF. Άλλοι βιολογικοί παράγοντες με διαφορετικούς στόχους είναι το ανακίρα, η αμπατασέπη, η ριτουξθμάμπη και η τοσιλιζουμάμπη. Μία άλλη ομάδα DMARDs που λέγονται αναστολείς κινάσης, περιλαμβάνουν την τοφασινιτινίμη και τη μπαρισιτινίμη. Τα βιολογικά DMARDs ή οι αναστολείς κινάσης συνδυάζονται συχνά με μεθοτρεξάτη ή άλλα DMARDs για να βελτιωθεί η αποτελεσματικότητά τους.	(+) ή (-)
3	MTX (Μεθοτρεξάτη)	Min: 7.5 (mg) Max: 25 (mg)
4	LEFLUNOMIDE (Λεφλουνομίδη)	Min: 10 (mg) Max: 20 (mg)

**ΠΙΝΑΚΑΣ 3.4:** Χορηγούμενη Φαρμακευτική αγωγή

## 3.2 Διερευνητική Ανάλυση Δεδομένων

### 3.2.1 Γενικές παρατηρήσεις

Η ενότητα αυτή αναφέρεται σε μια πρώτη διερεύνηση των δεδομένων, προκειμένου να κατανοηθούν σε μεγαλύτερο βάθος τα διαθέσιμα στοιχεία και να αποφασιστεί το είδος ανάλυσης που θα εφαρμοστεί στην συνέχεια. Σκοπός είναι η ανακάλυψη και απεικόνιση εξαρτή-

σεων μεταξύ των δεδομένων, η εξαγωγή βασικών συμπερασμάτων και η επιλογή των χαρακτηριστικών τα οποία θα αξιοποιήσουμε για την ανάπτυξη των μοντέλων μηχανικής μάθησης. Επίσης, η διερευνητική ανάλυση θα συνεισφέρει στην διατύπωση υποθέσεων, ενώ θα επιτρέψει τον εντοπισμό τυχόν θορυβώδη δεδομένα και χαμένες τιμές στο dataset που αξιοποιήθηκε.

Για κάθε ασθενή του συνόλου δεδομένων της παρούσας διπλωματικής εργασίας, έχουν συλλεγεί και καταχωρηθεί οι τιμές για περισσότερες από σαράντα ποιοτικές και ποσοτικές μεταβλητές. Όλες οι παραπάνω μεταβλητές είναι σημαντικές για τον ειδικό ρευματολόγο ιατρό και αποτελούν κριτήρια για τον προσδιορισμό της δραστηριότητας και της σοβαρότητας της ρευματοειδούς αρθρίτιδας σε έναν ασθενή, καθώς και για την λήψη αποφάσεων σχετικά με την φαρμακευτική αγωγή. Για το λόγο αυτό οι σχετικές πληροφορίες συλλέγονται και καταχωρούνται αμελλητί στους ιατρικούς φακέλους των ασθενών των ρευματολογικών κλινικών. Είναι προφανές ότι τα διαθέσιμα στοιχεία είναι πολλά και μπορούν να αναλυθούν με ποικίλους τρόπους και να αναπτυχθούν διαφορετικά μοντέλα. Για παράδειγμα, ένα ενδιαφέρον θέμα που μπορεί να μελετηθεί αξιοποιώντας το συγκεκριμένο σύνολο δεδομένων, είναι η συσχέτιση της RA με άλλες ασθένειες (καρκίνος, πνευμονοπάθειες, μεταβολικό σύνδρομο κ.λ.π.) σε έναν ασθενή, τυχόν εμπλοκή της θεραπευτικής αγωγής της RA στην εκδήλωση άλλων σοβαρών ασθενειών, μελέτη της επίπτωσης της κάθε θεραπείας στο δείκτη DAS28 κοκ.

Ορισμένες από αυτές τις συσχετίσεις μεταβλητών θα παρουσιαστούν ενδεικτικά, ενώ ειδικότερη έμφαση θα δοθεί στην συγκριτική αξιολόγηση της σοβαρότητας και της δραστηριότητας σε άνδρες / γυναίκες και στην λήψη απόφασης σχετικά με τα χαρακτηριστικά του δείγματος και τα ερωτήματα που αποτελούν αντικείμενο της παρούσας διπλωματικής εργασίας.

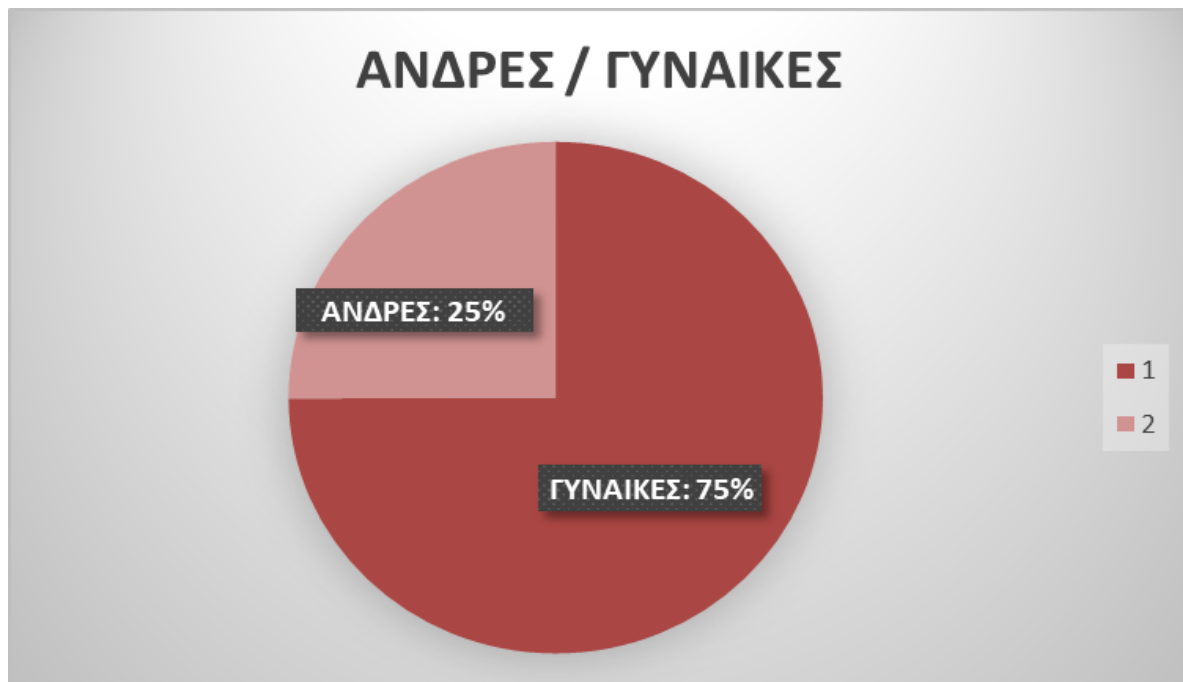
Για την στατιστική επεξεργασία και την οπτικοποίηση των αποτελεσμάτων χρησιμοποιήθηκαν τα εργαλεία για την ανάλυση δεδομένων του Excel 2016 και του στατιστικού πακέτου Gretl. Το Gretl είναι ένα στατιστικό πακέτο ανοιχτού κώδικα, κυρίως για την οικονομετρία. Το όνομα είναι αρκτικόλεξο για το Gnu Regression, Econometrics και Time-series Library. Διαθέτει γραφικό περιβάλλον εργασίας χρήστη και διεπαφή γραμμής εντολών.

Αρχικά διερευνήθηκαν τα δημογραφικά στοιχεία του δείγματος, όπως το φύλο των ασθενών, η ηλικία των ασθενών και το αν είναι καπνιστές ή όχι. Το δείγμα επιβεβαιώνει το γνωστό από πληθώρα επιδημιολογικών μελετών εύρημα ότι η ρευματοειδής αρθρίτιδα εμφανίζεται

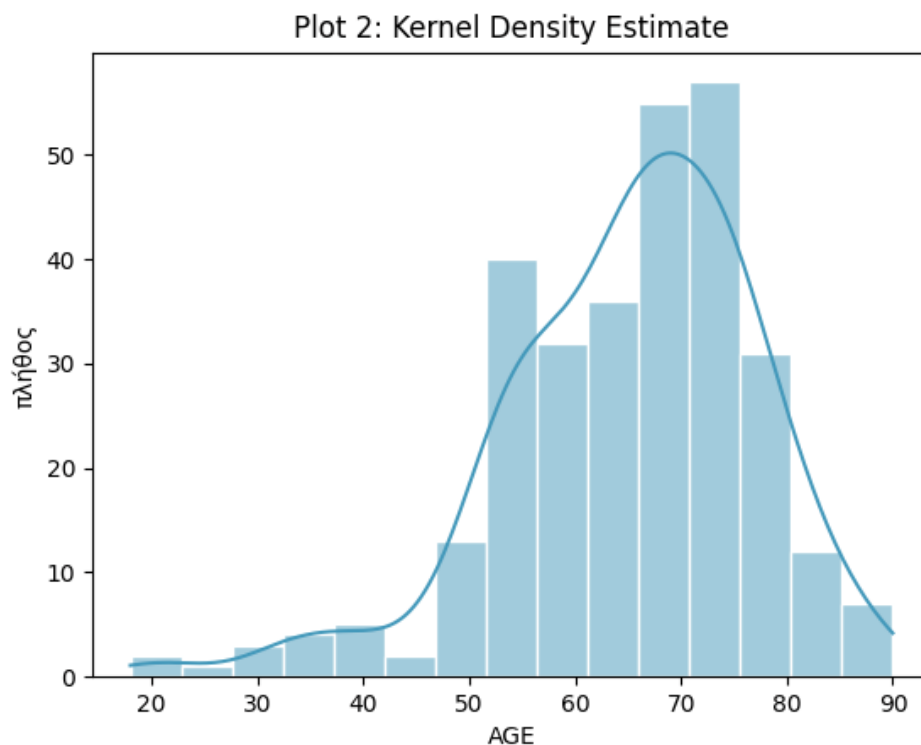


πιο συχνά στις γυναίκες σε αναλογία περίπου 3:1 ως προς τους άνδρες [95]. Η πάθηση προσβάλλει όλες τις ηλικιακές ομάδες, αλλά εκδηλώνεται συνήθως στις ηλικίες μεταξύ 35 και 55 ετών. Πληθυσμιακή μελέτη που παρουσιάστηκε στο Ετήσιο Ευρωπαϊκό Συνέδριο Ρευματολογίας (EULAR 2017) έδειξε ότι οι γυναίκες που είναι παχύσαρκες ή υπέρβαρες, κινδυνεύουν περισσότερο από την πάθηση, συσχετισμός που δεν φάνηκε να ισχύει για τους άνδρες.

Στο δείγμα των 300 ασθενών της παρούσας εργασίας, τα 222 άτομα είναι γυναίκες, 77 άνδρες και σε μία περίπτωση δεν υπάρχει τιμή (Σχήμα 1).



**Σχήμα 3.1:** Δημογραφικά στοιχεία δείγματος ασθενών ΡΑ ως προς το φύλο



**Σχήμα 3.2:** Ηλικιακή κατανομή του δείγματος

Στην ρευματοειδή αρθρίτιδα το κάπνισμα έχει χαρακτηριστεί ένας από τους πιο σημαντικούς παράγοντες κινδύνου για την εκδήλωση και την σοβαρότητα της νόσου, γεγονός που επιβεβαιώνεται και από τα στοιχεία του δείγματος.

Πρόσφατες μελέτες έχουν περιγράψει την παθοφυσιολογία της PA στους καπνιστές, συμπεριλαμβανομένου του οξειδωτικού στρες, της φλεγμονής, των επιγενετικών αλλαγών και της ανάπτυξης αυτοαντισωμάτων. Η συσχέτιση του καπνίσματος με την ανάπτυξη της PA έχει αποδειχτεί με σειρά επιδημιολογικών μελετών [96].

Στην παρούσα φάση αντικείμενο έρευνας αποτελεί ο τρόπος με τον οποίο επηρεάζει το κάπνισμα την αντίδραση των ασθενών στην φαρμακευτική αγωγή της PA και ειδικότερα στους βιολογικούς παράγοντες και τα DMARDs [96].

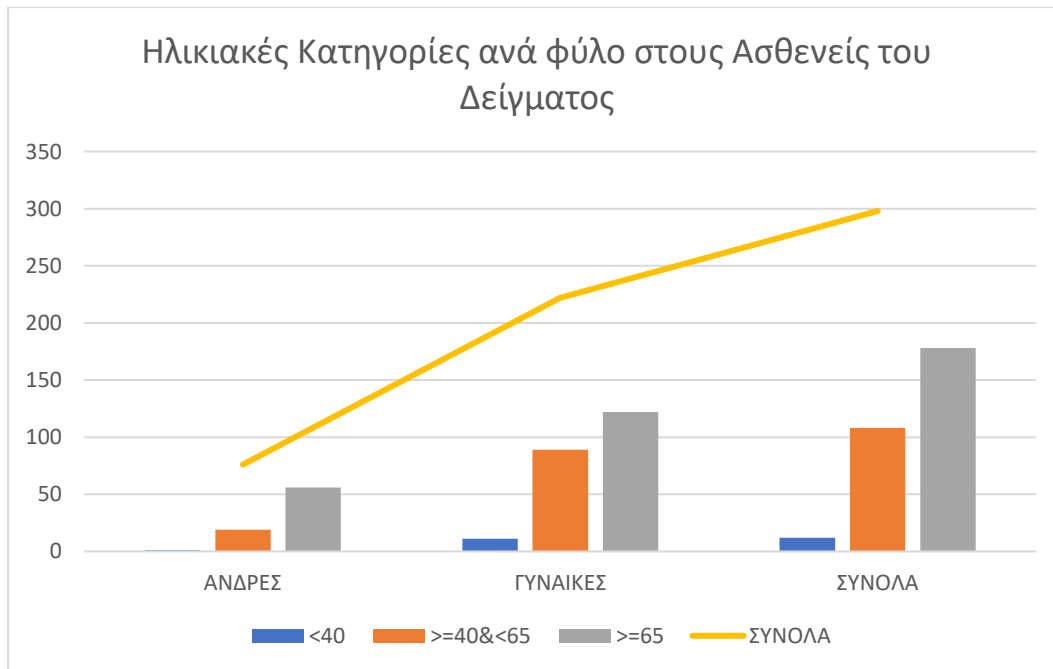
Στο δείγμα που επεξεργάστηκε στα πλαίσια της διπλωματικής η κατανομή συχνότητας για SMOKING (OXI=0 –NAI =1), παρατηρήσεις 1-300, διαμορφώνεται ως εξής:

0	151	50,50%	50,50% *****
1	148	49,50%	100,00% *****

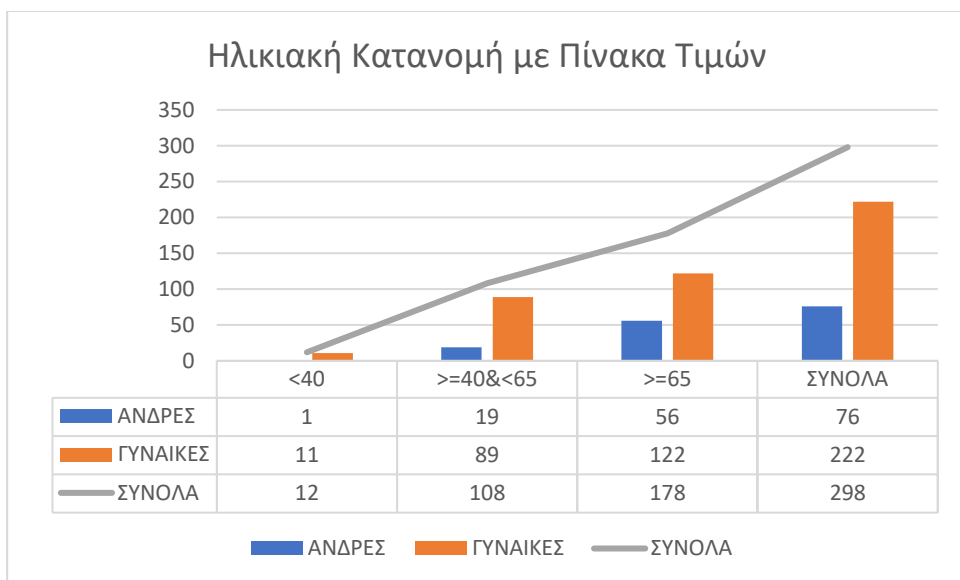
**Απούσες παρατηρήσεις = 1 ( 0,33%)**

Παρατηρείται ότι περίπου το 50% των ασθενών του δείγματος είναι **καπνιστές**, ποσοστό ιδιαίτερα υψηλό. Πράγματι, σύμφωνα με έρευνα του Ινστιτούτου Δημόσιας Υγείας που πραγματοποιήθηκε μέσω της ΚΑΠΑ Research (Φεβρουάριος 2020) στην Ελλάδα, οι καθημερινοί καπνιστές (ένα ή περισσότερα τσιγάρα την ημέρα) έχουν μειωθεί στο 17% του γενικού πληθυσμού.

Σε ότι αφορά την ηλικιακή κατανομή των ασθενών παρατηρείται ότι το 1,3% των ανδρών έχουν ηλικία < 40 ετών, ενώ για τις γυναίκες το αντίστοιχο ποσοστό ανέρχεται σε 5%. Επίσης το 25% έχει ηλικία >=40 ετών & < 65 ετών, ενώ για τις γυναίκες το 40% αντίστοιχα. Τέλος >=65 ετών είναι το 73.7% των ανδρών και το 55% των γυναικών.



**Σχήμα 3.3:** Δημογραφικά στοιχεία δείγματος ασθενών ΡΑ ως προς το φύλο και την ηλικία



**Σχήμα 3.4:** Δημογραφικά στοιχεία δείγματος ασθενών ΡΑ με πίνακα τιμών

Σε ότι αφορά τις συννοσηρότητες, τα στοιχεία του δείγματος επεξεργάστηκαν με το στατιστικό εργαλείο Gretl και προέκυψαν τα ακόλουθα αποτελέσματα:

<b>ΚΑΤΑΝΟΜΗ ΣΥΧΝΟΤΗΤΑΣ ΣΥΝΝΟΣΗΡΟΤΗΤΩΝ ΣΤΟ ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ (Στατιστική επεξεργασία με Gretl)</b>					
<b>1</b>	DM2 (Διαβήτης Τύπου 2)	0	253	84,33%	84,33%
		1	47	<b>15,67%</b>	100,00%
<b>2</b>	ILD (Διάμεσες Πνευμονοπάθειες)	0	257	85,67%	85,67%
		1	43	<b>14,33%</b>	100,00%
<b>3</b>	THYROID (Προσβολή Θυροειδούς)	0	251	83,67%	83,67%
		1	49	<b>16,33%</b>	100,00% *
<b>4</b>	TBC (Φυματίωση)	0	273	91,00%	91,00%
		1	27	<b>9,00%</b>	100,00% *
<b>5</b>	CANCER (Καρκίνος)	0	280	93,65%	93,65%
		1	19	<b>6,35%</b>	100,00%
<b>6</b>	OSTEOPOROSIS (Οστεοπόρωση)	0	149	49,67%	49,67%
		1	151	<b>50,33%</b>	100,00%
<b>7</b>	DEPRESSION (Κατάθλιψη)	0	256	85,62%	85,62%
		1	42	<b>14,05%</b>	99,67%
		2	1	0,33%	100,00%
Απουσίες παρατηρήσεις = 1 ( 0,33%)					
<b>8</b>	HYPERURICEMIA (Υπερουριχαιμία)	0	277	92,33%	92,33%
		1	23	<b>7,67%</b>	100,00%
<b>9</b>	HBV (Ηπατίτιδα Β)	0	291	97,00%	97,00%
		1	9	<b>3,00%</b>	100,00%

<b>10</b>	ΑΥ(ΑρτηριακήΥπέρταση)	0	151	50,33%	50,33%
		1	149	<b>49,67%</b>	100,00%
<b>11</b>	DYSLIPIDAEMIA (Δυσλιπιδαιμία)	0	179	59,67%	59,67%
		1	121	<b>40,33%</b>	100,00% *****
<b>12</b>	ARTHROPLASTY (Αρθροπλαστική)	0	260	86,67%	86,67%
		1	40	<b>13,33%</b>	100,00%
<b>19</b>	H.ZOSTER (Έρπης Ζωστήρ)	0	296	99,00%	99,00%
		1	2	0,67%	99,67%
		2	1	<b>0,33%</b>	100,00%
		Απουσίες παρατηρήσεις = 1 ( 0,33%)			

**Πίνακας 3.5:** OXI=0, NAI=1, KENO=2

Παρατηρείται ότι ορισμένες ασθένειες όπως η οστεοπόρωση, η δυσλιπιδαιμία και η υπέρταση έχουν συχνότητα εμφάνισης σε ποσοστά από 40%-50% στους ασθενείς PA του δείγματος, συμπέρασμα που μπορεί να είναι σημαντικό για έναν κλινικό γιατρό. Επίσης σε στατιστικά σημαντικά ποσοστά από 13% έως και 16% εμφανίζονται ασθένειες όπως κατάθλιψη, διαβήτης, θυροειδής, διάμεσες πνευμονοπάθειες, ενώ ένα ποσοστό 13,3% των ασθενών του δείγματος έχουν προβεί σε αρθροπλαστική.

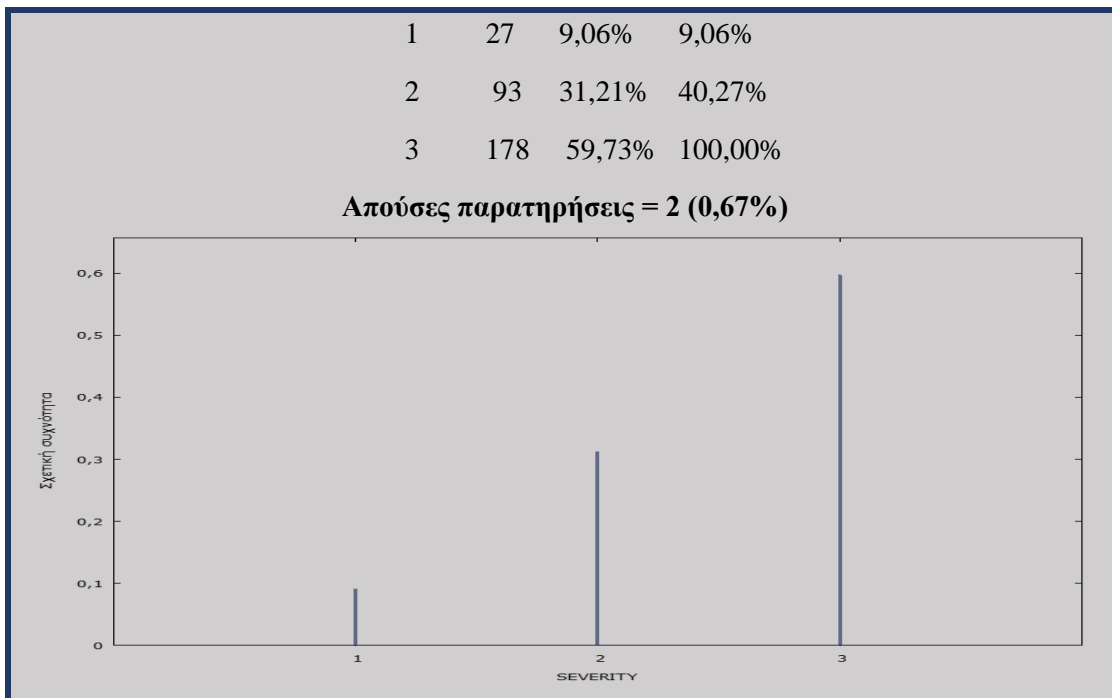
Ασθένειες όπως ο καρκίνος, φυματίωση και υπερουριχαιμία εμφανίζονται σε ποσοστά που κυμαίνονται από 6%-8%,ενώ η ηπατίτιδα Β σε ποσοστό 3% και ο έρπης ζωστήρ στο 0,67% των ασθενών του δείγματος, παρόλες τις βιβλιογραφικές αναφορές για το αντίθετο, ιδιαίτερα ως αποτέλεσμα από τις θεραπείες[97].

### 3.2.2 Δείκτες Σοβαρότητας και Ενεργότητας στην PA

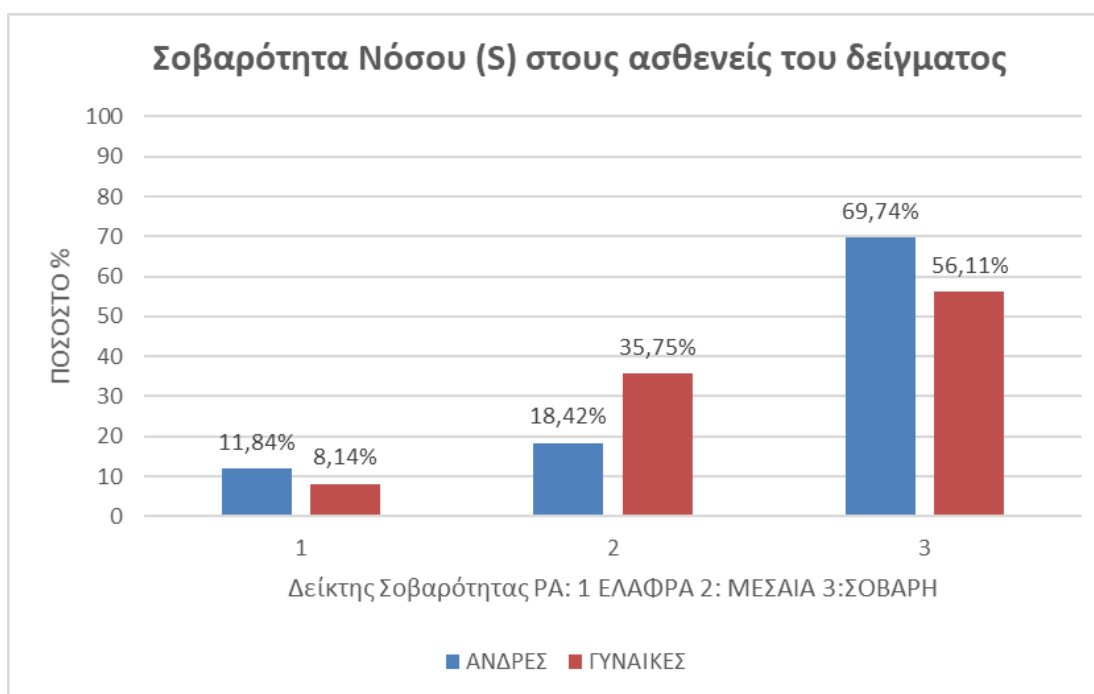
Στη συνέχεια θα συζητηθούν ιδιαίτερα οι δείκτες σοβαρότητας (severity) και ενεργότητας/δραστηριότητας (activity) της PA η συχνότητα εμφάνισης των διακριτών τιμών τους στο σύνολο δεδομένων μας, δεδομένου ότι με αυτούς τους δείκτες θα ασχοληθούμε στο υπόλοιπο της διπλωματικής εργασίας.

Την σοβαρότητα και την ενεργότητα τις έχουμε χαρακτηρίσει ως διακριτές μεταβλητές με τιμές 1, 2 και 3, σύμφωνα με την περιγραφή της πρώτης ενότητας του παρόντος κεφαλαίου και η ταξινόμηση των παρατηρήσεων του συνόλου δεδομένων μας στις ανωτέρω κατηγορίες έγινε από ειδικούς ρευματολόγους. Η ταξινόμηση των ασθενών με βάση τους δείκτες αυτούς εξετάστηκε τόσο για το σύνολο των παρατηρήσεων του δείγματος όσο και για το κάθε φύλο ξεχωριστά.

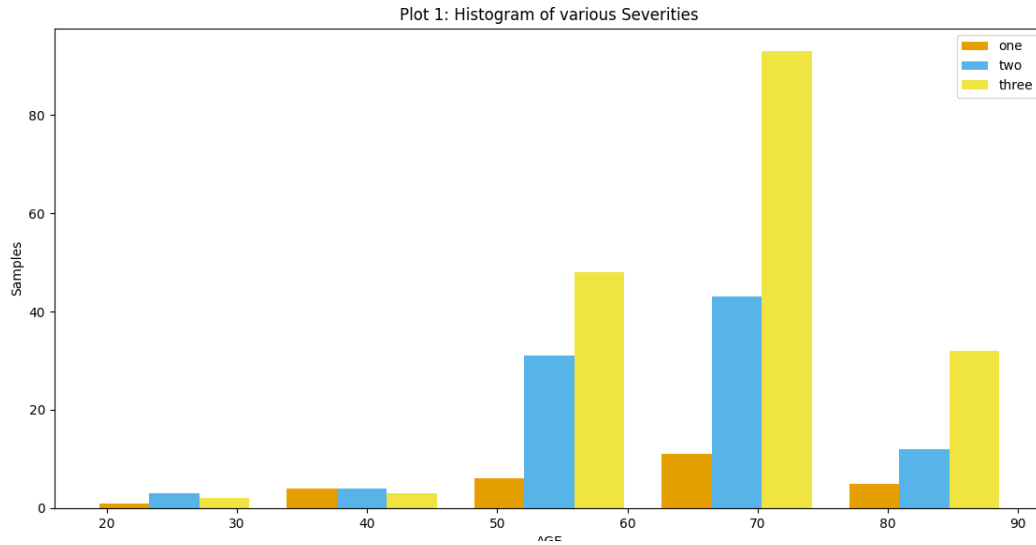
Στο σύνολο παρατηρήσεων του δείγματος (300 παρατηρήσεις) τα ποσοστά για την σοβαρότητα της νόσου διαμορφώνονται ως εξής:



**Σχήμα 3.5:** Σοβαρότητα νόσου για το σύνολο του πληθυσμού (Ανδρες / Γυναίκες)

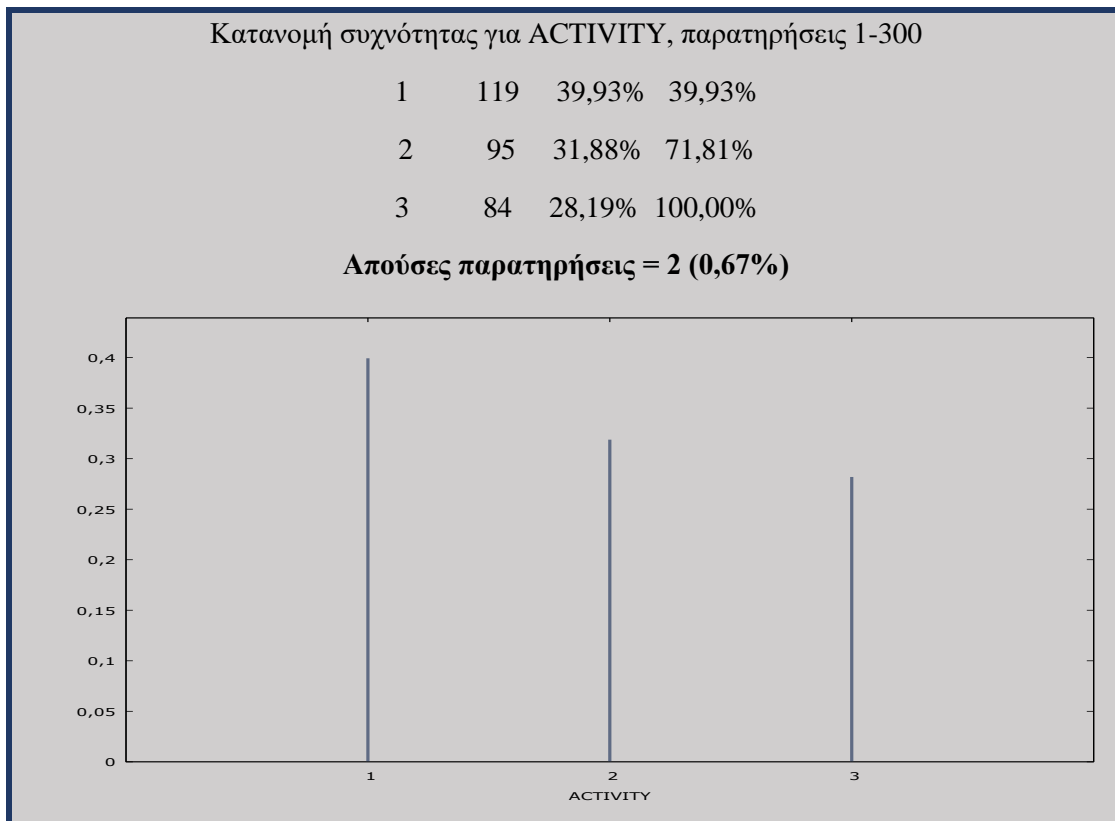


**Σχήμα 3.6:** Σοβαρότητα Νόσου (Ανδρες / Γυναίκες)

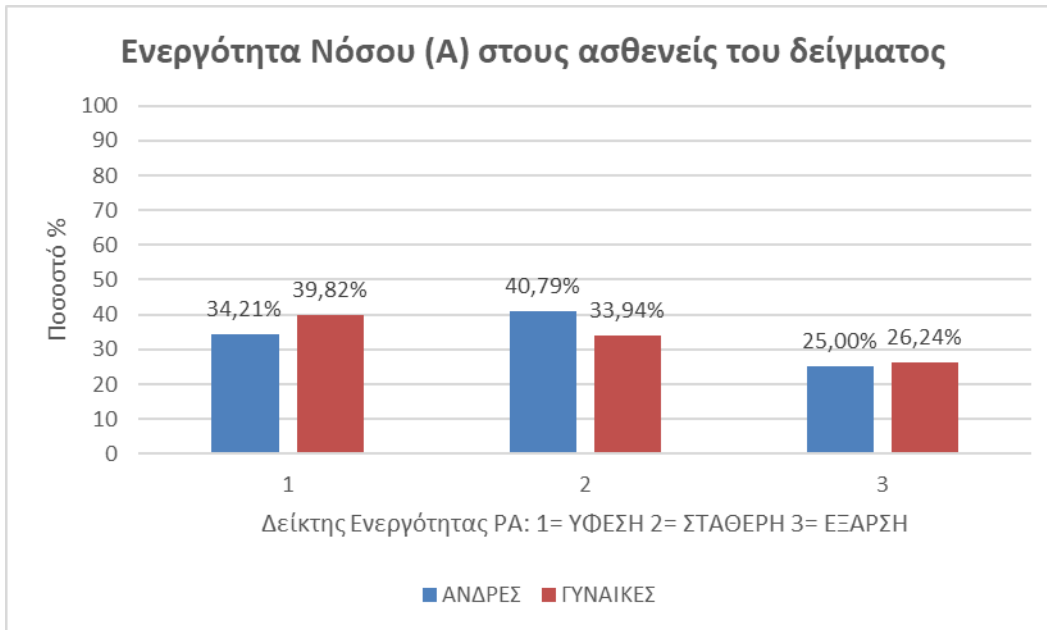


**Σχήμα 3.7:** Ιστόγραμμα επιπέδων σοβαρότητας ανά ηλικία

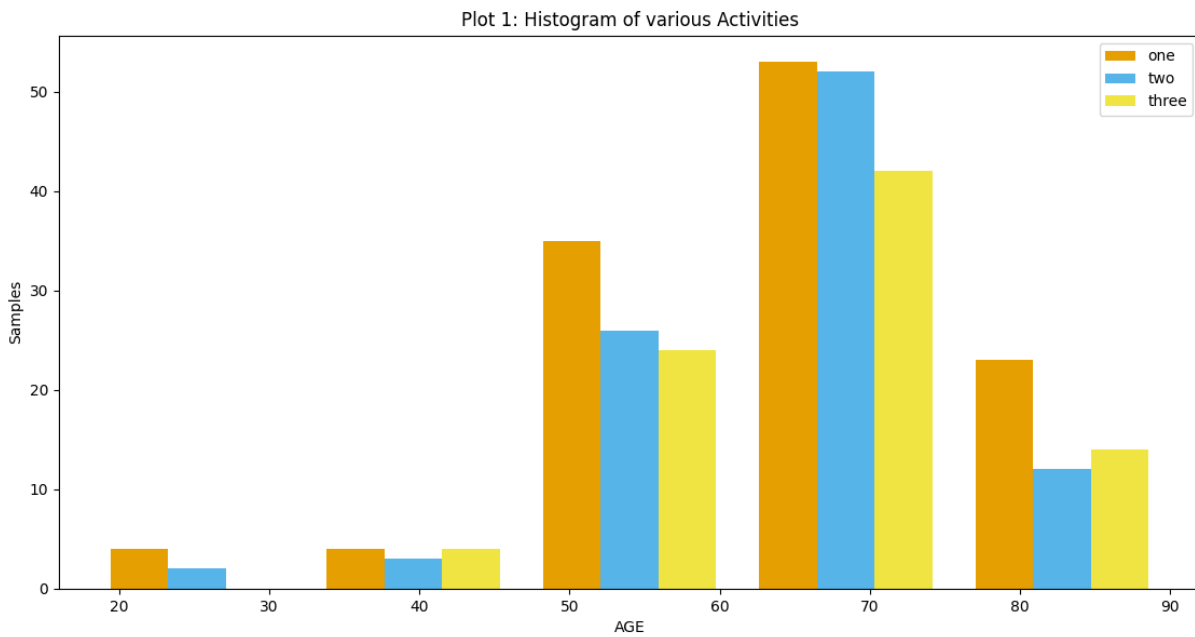
Σε ότι αφορά στον δείκτη ενεργότητας / δραστηριότητας της νόσου στο σύνολο του πληθυσμού του δείγματος (300 παρατηρήσεις), προκύπτουν τα παρακάτω αποτελέσματα:



**Σχήμα 3.8:** Σοβαρότητα νόσου για το σύνολο του πληθυσμού (Ανδρες / Γυναίκες)



**Σχήμα 3.9 :** Ενεργότητα Νόσου ( Άνδρες / Γυναίκες)



**Σχήμα 3.10:** Ιστόγραμμα επιπέδων ενεργότητας ανά ηλικία



Από την παραπάνω γρήγορη στατιστική ανάλυση παρατηρείται ότι η δραστηριότητα και η σοβαρότητα της νόσου δεν ταυτίζονται αναγκαστικά. Για παράδειγμα τη μεγαλύτερη συχνότητα σε ότι αφορά την δραστηριότητα στους ασθενείς του δείγματος έχει η τιμή 1-(ύφεση), ενώ σε ότι αφορά την σοβαρότητα την μεγαλύτερη συχνότητα έχει η τιμή 3- (σοβαρή νόσος). Ως εκ τούτου οι δύο αυτοί δείκτες είναι σημαντικό για τον γιατρό να μελετώνται ξεχωριστά και είναι εξίσου κρίσιμοι για την λήψη εξατομικευμένων ιατρικών αποφάσεων και την εφαρμογή της κατάλληλης θεραπείας.

Σε ότι αφορά την δραστηριότητα (activity) είναι γνωστό ότι εκτιμάται με μεγάλη επιτυχία με τον δείκτη μέτρησης DAS28, ο οποίος συμπεριλαμβάνεται και στα χαρακτηριστικά της βάσης δεδομένων που χρησιμοποιήθηκε. Τα επόμενα κεφάλαια της διπλωματικής εργασίας θα επικεντρωθούν στην επιλογή των κατάλληλων χαρακτηριστικών και την αξιοποίηση των στοιχείων της βάσης δεδομένων ως δεδομένα εκπαίδευσης (training data) και δεδομένα ελέγχου (testing data) σε μοντέλα μηχανικής μάθησης αναφορικά με την εκτίμηση της σοβαρότητας (severity) της PA.

Ειδικότερα, σκοπός είναι η ανάπτυξη μοντέλων μηχανικής μάθησης για την εξαγωγή συμπερασμάτων ως προς τις κατάλληλες μεταβλητές που μπορούν να αξιοποιηθούν για την εκτίμηση της σοβαρότητας (severity) της ρευματοειδούς αρθρίτιδας σε ασθενείς, καθώς και για την δυνατότητα ανάπτυξης ενός αντίστοιχου δείκτη με τον DAS28 για την σοβαρότητα της νόσου σε επόμενη φάση. Τα μοντέλα θα ελεγχθούν ως προς την αξιοπιστία τους (validation) με βάση την κατάταξη ως προς την σοβαρότητα που έδωσαν οι ειδικοί ρευματολόγοι για κάθε παρατήρηση του δείγματος ασθενών.

Τέλος, σε ότι αφορά τον **ποιοτικό έλεγχο των δεδομένων** δεν παρατηρήθηκαν σημαντικά σφάλματα στις τιμές των μεταβλητών, γεγονός που ήταν αναμενόμενο λόγω του ότι το σύνολο δεδομένων είχε ήδη διορθωθεί και καταχωριστεί σε αρχείο excel από ρευματολόγο γιατρό.

Οι απύσες τιμές στις συννοσηρότητες αναφέρονται στον Πίνακα 3.5. Σε ότι αφορά τα εργαστηριακά ευρήματα παρατηρούνται απύσες τιμές σε ορισμένες μεταβλητές όπως HCT (αιματοκρίτης), HGB (αιμοσφαιρίνη), Glucose (ζάχαρο) κοκ, οι οποίες όμως δεν θα χρησιμοποιηθούν στα μοντέλα μηχανικής μάθησης της παρούσας εργασίας, οπότε δεν αλλάζουν το πλήθος των παρατηρήσεων που μπορούν να αξιοποιηθούν.

### **3.3 Περιγραφή Κρίσιμων Μεταβλητών / Χαρακτηριστικών για τα Μοντέλα της Μελέτης**

#### **3.3.1 Δείκτης DAS28**

Ο δείκτης DAS28 (Disease Activity Score-28) δίνει μια αξιόπιστη εκτίμηση της ενεργότητας της νόσου [98]–[100]. Λαμβάνει υπόψη τέσσερις (4) παράγοντες και ειδικότερα:

- Τον αριθμό των διογκωμένων αρθρώσεων (από ένα σύνολο 28 αρθρώσεων).
- Τον αριθμό των ευαίσθητων στην πίεση αρθρώσεων (από ένα σύνολο 28 αρθρώσεων).
- Τα αποτελέσματα των εργαστηριακών εξετάσεων για την C αντιδρώσα πρωτεΐνη (CRP) ή την ΤΚΕ
- Τις απαντήσεις του ασθενούς σε ερωτηματολόγιο αυτοαξιολόγησης (πρωινή δυσκαμψία (N/O), πρωινή κόπωση (N/O) κοκ).

Για τον υπολογισμό της τιμής του δείκτη χρησιμοποιείται ένας μαθηματικός τύπος. Ο DAS28 μπορεί να κυμαίνεται από 0 έως 9.4. Γενικά το σκόρ του DAS28 ερμηνεύεται ως εξής:

- Πάνω από 5.1 υποδεικνύει υψηλή δραστηριότητα νόσου.
- Μεταξύ 3.2 and 5.1 υποδεικνύει μέτρια δραστηριότητα νόσου.
- Μεταξύ 2.6 και 3.2 υποδεικνύει χαμηλή δραστηριότητα νόσου.
- Λιγότερο από 2.6 υποδεικνύει υποχώρηση της νόσου.

Ο DAS28 μεταξύ άλλων μπορεί να χρησιμοποιηθεί για να μετρήσει αν αποδίδει η φαρμακευτική αγωγή σε έναν ασθενή.

#### **3.3.2 Λευκοκυτταρικά Αντιγόνα (*Human Leukocyte Antigens, HLA*)**

Κοινό χαρακτηριστικό στα αυτοάνοσα νοσήματα όπως η ΡΑ είναι η ύπαρξη αυτοδραστικών κλώνων Τ-λεμφοκυττάρων, με αποτέλεσμα η βλάβη στους ιστούς και τα όργανα να προκύπτει από την αντίδραση των κλώνων αυτών προς «εαυτά» αντιγόνα. Τα λευκοκυτταρικά αντιγόνα έχουν περιγραφεί αναλυτικά στο Κεφ.1 της παρούσας.

Η γενετική ανάλυση των γονιδίων HLA σε συνδυασμό με τον έλεγχο αυτοαντισωμάτων σε άτομα με οικογενειακό ιστορικό αυτοανοσίας φαίνεται πως είναι επιτακτική ανάγκη στα

πλαίσια της προσυμπτωματικής διάγνωσης, προκειμένου να αποφευχθεί η εμφάνιση της νόσου [101]–[103].

### **3.3.3 Αντισώματα Ορού έναντι κιτρουλίνης (CCP)**

Τα αντισώματα αντι-CCP (Cyclic Citrullinated Peptide) εμφανίζονται νωρίς στην πορεία της ρευματοειδούς αρθρίτιδας. Τα αντισώματα αυτά έχουν ευαισθησία παρόμοια με αυτήν του ρευματοειδούς παράγοντα (ΡΠ) και πολύ μεγάλη ειδικότητα για τη νόσο. Είναι νέα αντισώματα ειδικά για την ΡΑ και βοηθούν σημαντικά στη διάγνωση. Ανιχνεύονται στον ορό αρρώστων με ΡΑ σε ποσοστό μεγαλύτερο του 80%. Έχουν υψηλή ειδικότητα (98%) και βοηθούν να διακρίνουμε τη ΡΑ από άλλες αρθρίτιδες που την μιμούνται. Προβλέπουν καλύτερα την ΡΑ από τον ΡΠ. Είναι επίσης χρήσιμα για την πρόγνωση της νόσου [104]–[106]. Περιγράφονται αναλυτικά στο Κεφ.1.

### **3.3.4 Ρευματοειδείς Παράγοντες (ΡΠ)**

Στη ΡΑ η ανεύρεση ρευματοειδών παραγόντων είναι ένα από τα διαγνωστικά της κριτήρια. Οι τίτλοι ΡΠ παρουσιάζουν συσχέτιση με τη βαρύτητα της νόσου. Μείωση των τίτλων ΡΠ κατά την πορεία της νόσου αποτελεί ένδειξη αυτόματης ύφεσης ή καλής ανταπόκρισης στη θεραπεία. Ο ρευματοειδής παράγοντας ανιχνεύεται στο 65-85% περίπου των ασθενών με ρευματοειδή αρθρίτιδα. Η εξέταση του ρευματοειδούς παράγοντα ζητείται συνήθως μαζί με την εξέταση ΤΚΕ και CRP αφού και αυτή ως πρωτεΐνη οξείας φάσεως αυξάνεται ραγδαία στα πρώιμα στάδια της νόσου [105]–[107]. Οι ρευματοειδείς παράγοντες έχουν συζητηθεί αναλυτικά στο Κεφ.1.

### **3.3.5 C Αντιδρώσα Πρωτεΐνη (CRP)**

Η CRP αυξάνει σε καταστάσεις που χαρακτηρίζονται από φλεγμονή, λοίμωξη και καταστροφή ιστών. Κατά τη διάρκεια της απάντησης οξείας φάσης μπορεί οι τιμές της CRP μπορεί να αυξηθούν και 1000 φορές. Η CRP χρησιμοποιείται για την διάγνωση ορισμένων ρευματικών νοσημάτων συμπεριλαμβανομένης της ρευματοειδούς αρθρίτιδας. Είναι σημαντικός δείκτης για την παρακολούθηση της πορείας νόσου και του θεραπευτικού αποτελέσματος. Είναι πιο αξιόπιστη από την ΤΚΕ (ταχύτητα καθιζήσεως ερυθρών αίματος) διότι δεν επηρεάζεται από τον αιματοκρίτη, τις ανοσοσφαιρίνες, την ηλικία ή τη λήψη φαρμάκων. Οι

τιμές της μεταβάλλονται ταχύτερα σε σύγκριση με την ΤΚΕ σε περιπτώσεις ύφεσης ή επιδείνωσης της νόσου. Τιμές μεγαλύτερες από 10 mg/L δείχνουν κλινικά σημαντική φλεγμονή [106].

### 3.3.6 Ταχύτητα Καθίζσεως Αίματος (ΤΚΕ)

Η ΤΚΕ εξαρτάται από παράγοντες του πλάσματος όπως πρωτεΐνες οξείας φάσης (CRP, C3, ινωδογόνο, αμυλοειδές Α, φερριτίνη κ.α.) και μεταβολές των σφαιρινών. Τιμές ΤΚΕ > 100 mm/h υποδεικνύουν βακτηριακή λοίμωξη, νεόπλασμα ή συστηματικό ρευματικό νόσημα. Η ΤΚΕ αξιοποιείται για την διάγνωση ορισμένων ρευματικών νοσημάτων όπως ρευματοειδή αρθρίτιδα, ρευματική πολυμυαλγία, κροταφική αρτηρίτιδα κ.ά. Επίσης είναι ιδιαίτερα χρήσιμη στην παρακολούθηση πορείας νόσου και θεραπευτικού αποτελέσματος στην ρευματική πολυμυαλγία, κροταφική αρτηρίτιδα, ρευματικό πυρετό, ΡΑ, οροαρνητικές σπονδυλίτιδες, λοιμώξεις σπονδυλικής στήλης και αρθρώσεων κ.ά [106].

### 3.3.7 Sharp Score

Το Sharp Score είναι ένα σύστημα βαθμολόγησης για την αξιολόγηση των ακτινολογικών ευρημάτων των ασθενών με ΡΑ και αποτελεί δείκτη ακτινολογικής καταστροφής των αρθρώσεων. Η αρχική μέθοδος Sharp περιλάμβανε ακτινογραφίες από χέρια και καρπούς και μεγάλο αριθμό χαρακτηριστικών όπως οστεοπόρωση, επιφανειακές διαβρώσεις, στένωση, αγκύλωση κοκ. Διάφοροι περιορισμοί σε αρκετά από αυτά τα χαρακτηριστικά οδήγησαν στην παράλειψή τους από την τελική βαθμολογία. Επομένως, το Sharp Score διαμορφώθηκε ώστε η μέθοδος να περιλαμβάνει δύο βαθμολογίες, μια για τις διαβρώσεις και μια για στένωση του αρθρικού χώρου, ως εξής:

(1) Διάβρωση: Κλίμακα βαθμολογίας από 0-5 ως εξής: 0 χωρίς αλλοιώσεις, 5 πλήρης καταστροφή της άρθρωσης (τιμές από 0 έως 160). Εξετάζονται 16 περιοχές του χεριού.

(2) Στένωση: Κλίμακα βαθμολογίας από 0-4 ως εξής: 0 φυσιολογική άρθρωση, 1 εστιακή στένωση, 2 ελάττωση κατά <50% του αρθρικού διαστήματος, 3 ελάττωση κατά >50% του αρθρικού διαστήματος, 4 αγκύλωση. (τιμές από 0 έως 144). Εξετάζονται 18 περιοχές του χεριού.

Το συνολικό Sharp Score δημιουργείται με βάση το άθροισμα των βαθμολογιών στένωσης και διάβρωσης των αρθρώσεων που εξετάζονται. Επίσης χρησιμοποιείται το Modified Van Der Heijde erosion and joint narrowing score το οποίο εκτός από χέρια και καρπούς εξετάζει και τα πόδια [108]–[111].

Περιλαμβάνει 16 περιοχές από κάθε χέρι/καρπό και 10 μεταταρσοφαλαγγικές + 2 ΦΦ από τους μεγάλους δακτύλους των ποδιών. Σύνολο (μέγιστο) για το Van Der Heijde score=448 [112], [113].

# 4

## *Μεθοδολογία Μελέτης*

### *4.1 Εισαγωγή*

Πηγή έμπνευσης για την ανάπτυξη των μοντέλων της παρούσας διπλωματικής, αποτέλεσε ο δείκτης **DAS28 (Disease Activity Score-28)**, που δίνει μια αξιόπιστη εκτίμηση για την ενεργότητα της νόσου. Ο DAS28 μεταξύ άλλων μπορεί να χρησιμοποιηθεί για να μετρήσει αν ένα φαρμακευτικό σχήμα δουλεύει σωστά σε έναν ασθενή.

Με βάση τον DAS28 και την ευρύτατη χρησιμοποίησή του την τελευταία δεκαετία στην καθημερινή πρακτική των ρευματολογικών κλινικών, τόσο για την αξιολόγηση της δραστηριότητας της ΡΑ όσο και για την καταλληλότητα της χορηγούμενης θεραπευτικής αγωγής, είχε ενδιαφέρον λοιπόν να εξεταστεί αν υπάρχει δυνατότητα να δημιουργηθεί ένας αντίστοιχος δείκτης που θα δίνει αξιόπιστα αποτελέσματα όσον αφορά την σοβαρότητα της νόσου. Πράγματι, η σοβαρότητα σε συνδυασμό με την δραστηριότητα ή ενεργότητα της νόσου είναι τα κρίσιμα στοιχεία που καθορίζουν την λήψη σημαντικών αποφάσεων από τον κλινικό γιατρό, μεταξύ άλλων την φαρμακευτική αγωγή που θα αποφασίσει ο ειδικός και θα ακολουθήσει ο εκάστοτε ασθενής.

Διάφορες μελέτες έχουν δείξει ότι με την έκβαση και την σοβαρότητα της ΡΑ σχετίζονται ποικίλοι παράγοντες, μεταξύ των οποίων το γενετικό υπόστρωμα. Ειδικότερα, η ύπαρξη λευκοκυτταρικών αντιγόνων HLA αποτελεί σημαντικό προγνωστικό παράγοντα για διάφορα αυτοάνοσα νοσήματα συμπεριλαμβανομένης της ΡΑ.

Επιπροσθέτως, σημαντικοί προγνωστικοί παράγοντες είναι οι δείκτες φλεγμονής CRP, ΤΚΕ όπως και τα αυτοαντισώματα ορού (ΡΠ, αντισώματα αντι-CCP κ.ά.). Ο ρευματοειδής παράγοντας είναι θετικός στο 60% – 70% των ασθενών με ΡΑ, αλλά στερείται ειδικότητα για τη νόσο.

Η αντίχνευση των αντισωμάτων έναντι κυκλικών κιτρουλινοποιημένων πεπτιδίων (anti-CCP) και η εισαγωγή τους στην κλινική πράξη είναι μία σημαντική εξέλιξη της τελευταίας δεκαετίας, μιας και αυτά τα αυτοαντισώματα έχουν την ίδια ευαισθησία με το ρευματοειδή παράγοντα, αλλά έχουν πολύ μεγάλη ειδικότητα για τη Ρ.Α. (93% – 97%). Ας σημειωθεί ότι ποσοστό 10% – 20% των ασθενών με Ρ.Α. έχει θετικά (συνήθως σε χαμηλό τίτλο) αντιπυρηνικά αντισώματα (ANA). Τα αντιπυρηνικά αντισώματα δεν συμπεριλαμβάνονται στα χαρακτηριστικά του δείγματος των ασθενών που αποτελούν την βάση δεδομένων που χρησιμοποιήθηκε.

Ειδικότερα οι παράγοντες που αξιοποιήθηκαν σαν είσοδοι στα μοντέλα που αναπτύχθηκαν είναι οι παρακάτω:

<b>1. HLA-DRB1</b>	<b>5. smoking</b>
<b>2. RF - RFi</b>	<b>6. CRP</b>
<b>3. CCP - CCPi</b>	<b>7. ΤΚΕ</b>
<b>4. Sharp Score</b>	<b>8. duration</b>

Στο πλαίσιο αυτό, δημιουργήθηκαν μοντέλα Μηχανικής Μάθησης με συνδυασμό των παραπάνω παραγόντων σαν εισόδους, ώστε να δούμε ποιοι από αυτούς τους παράγοντες μπορούν να προβλέψουν ή να ομαδοποιήσουν βέλτιστα τα δεδομένα με βάση το διάλυμα του SEVERITY που αποτελεί μια στήλη που έχει προστεθεί και συμπληρωθεί από ειδικούς ρευματολόγους με βάση τα δεδομένα για κάθε ασθενή του δείγματος.

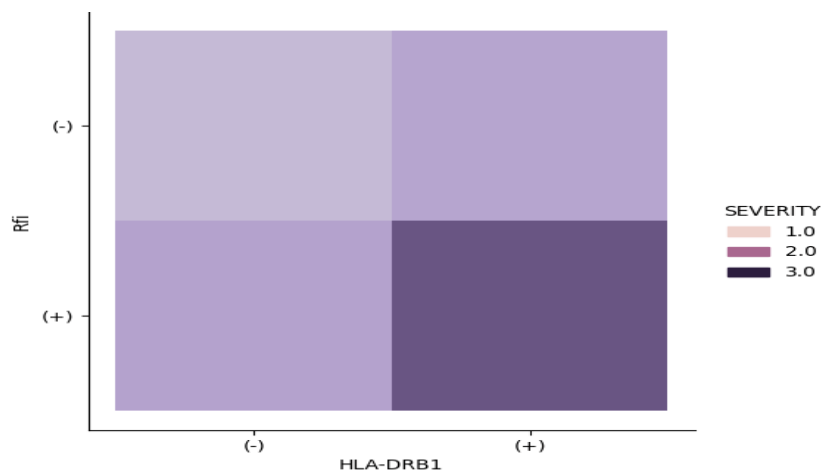
Αναπτύχθηκαν συνολικά πέντε (5) μοντέλα εκ των οποίων τα πρώτα τρία (3) αφορούν μη εποπτευόμενη μάθηση με clustering και τα υπόλοιπα δύο (2) αφορούν μοντέλα εποπτευόμενης μηχανικής μάθησης. Όλα τα μοντέλα που αναπτύχθηκαν βασίζονται σε αλγόριθμους που έχουν περιγραφεί αναλυτικά στο Κεφ. 2 και είναι τα εξής:

- ✓ **Μοντέλο 1:** kmeans
- ✓ **Μοντέλο 2:** GaussianMixture
- ✓ **Μοντέλο 3:** kMedoids
- ✓ **Μοντέλο 4:** k- Nearest Neighbors

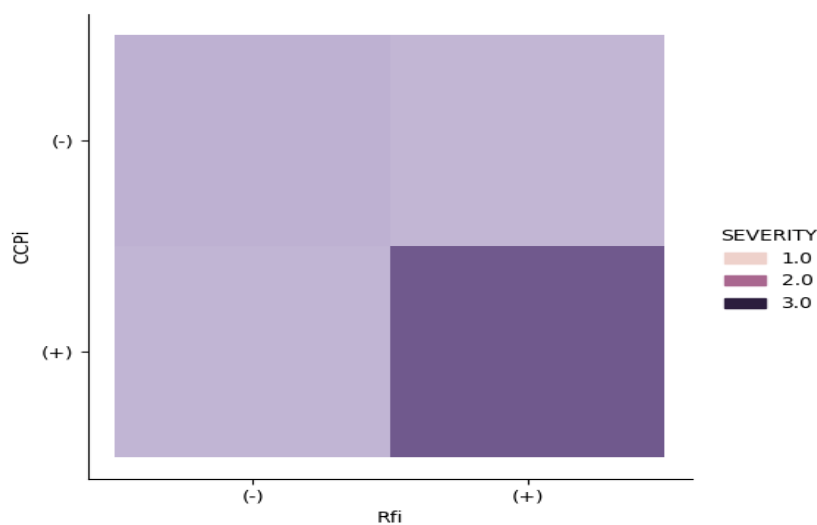
Για την ανάπτυξη των μοντέλων χρησιμοποιήθηκε η **python 3.9** και ενδεικτικά οι παρακάτω βιβλιοθήκες: pandas, numpy, sklearn, seaborn , matplotlib.

## 4.2 Προεπεξεργασία Δεδομένων

Σε πρώτο στάδιο χρησιμοποιώντας γραφικά εργαλεία έπρεπε να αποσαφηνιστεί αν οι παραπάνω παράγοντες που αναφέρθηκαν ότι σχετίζονται με τη σοβαρότητα, έχουν κάποιο συσχετισμό μεταξύ τους (correlation). Όπως φαίνεται από τα παρακάτω διαγράμματα ασθενείς με σοβαρότητα τύπου 3 έχουν όντως σε μεγάλο βαθμό θετικό ρευματοειδή παράγοντα και θετικό CCP.

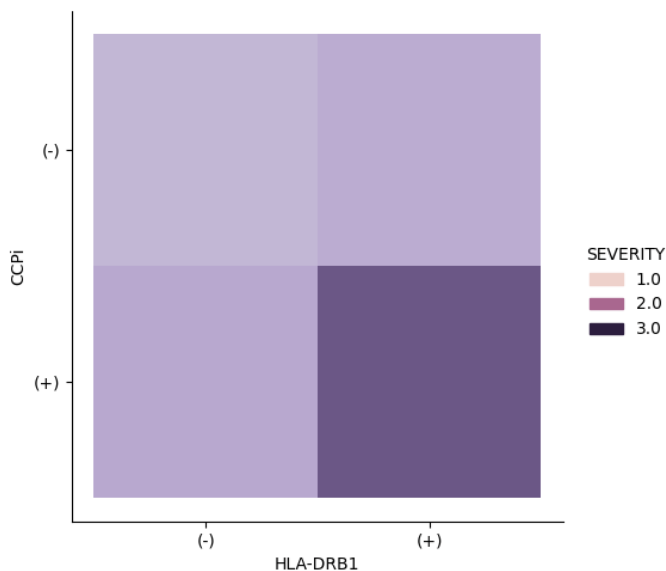


**Σχήμα 4.1.** Density 2 παραμέτρων RFi και HLA-DRB1



**Σχήμα 4.2.** Density 2 παραμέτρων RFi και CCPi





**Σχήμα 4.3.** Density 2 παραμέτρων CCPi και HLA-DRB1

Αρχικά το dataset περιείχε απύσες τιμές οι οποίες θα έπρεπε να αφαιρεθούν ώστε να εισαχθούν στα αντίστοιχα μοντέλα. Αυτό το επιτεύχθηκε χρησιμοποιώντας την εντολή `.dropna()`.

Σε ότι αφορά τα εργαστηριακά ευρήματα, παρατηρούνται απύσες τιμές σε ορισμένες μεταβλητές όπως HCT (αιματοκρίτης), HGB (αιμοσφαιρίνη), Glucose (ζάχαρο) κοκ, οι οποίες όμως δεν θα χρησιμοποιηθούν στα μοντέλα μηχανικής μάθησης της παρούσας εργασίας, οπότε δεν αλλάζουν το πλήθος των παρατηρήσεων που μπορούν να αξιοποιηθούν.

Ειδικότερα για τις μεταβλητές οι οποίες θα αξιοποιήθηκαν στη συνέχεια ήτοι : HLA\_DRB1, CCPi, RFi, TKE και CRP, παρατηρήθηκαν τα εξής:

Μεταβλητή	Λάθος Τιμές	Απούσες Τιμές
HLA_DRB1	Γραμμή 17	Καμία (0)
CCPi	Καμία (0)	Γραμμή 40
RFi	Καμία (0)	Καμία (0)
TKE	Καμία (0)	Καμία (0)
CRP	Καμία (0)	Καμία (0)

**Πίνακας 4.1:** Ποιοτικός έλεγχος δεδομένων

Από τα παραπάνω προκύπτει ότι τα δεδομένα δεν απαιτούν μεγάλο χρόνο για την προεπεξεργασία και τον καθαρισμό τους και το σημαντικότερο θέμα πριν την δημιουργία των μοντέλων είναι η κανονικοποίηση αυτών.

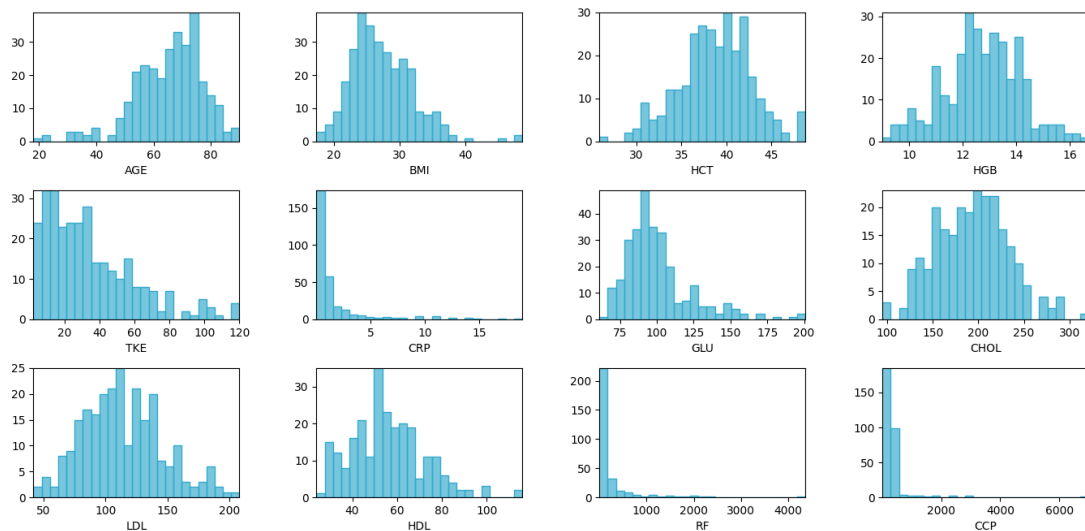
Τέλος τιμές μη αριθμητικής μορφής αντικαταστάθηκαν με αριθμητικές τιμές όπως φαίνεται παρακάτω ώστε να μπορούν να εισαχθούν στα μοντέλα

Αρχική Τιμή	Αντικατάσταση
(+)	1
(-)	0
ΝΑΙ	1
ΟΧΙ	0

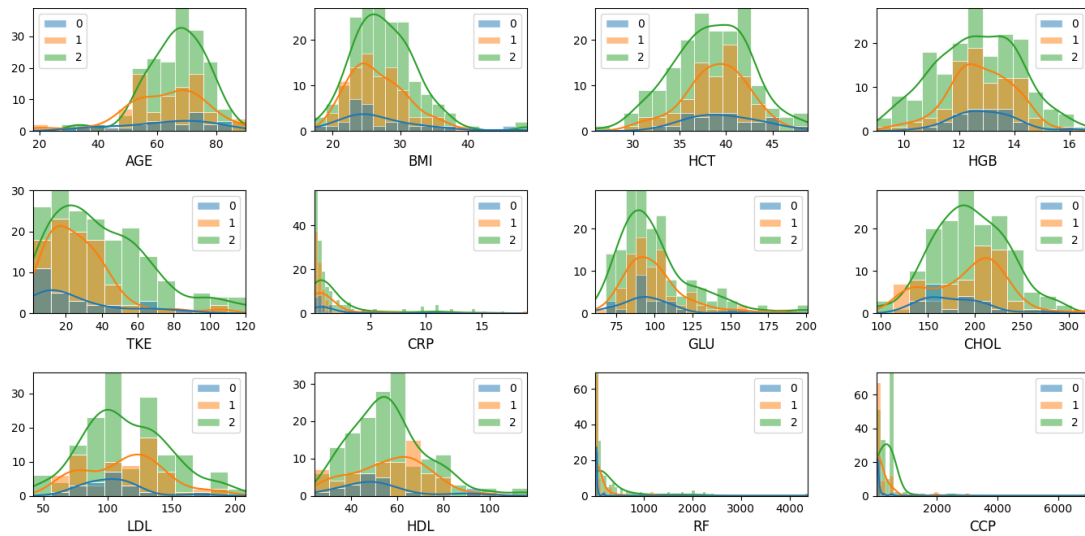
**Πίνακας 4.2:** Ποιοτικός έλεγχος δεδομένων

### Κανονικοποίηση

Παρατηρώντας την κατανομή ορισμένων μεταβλητών αποσαφηνίστηκε η κατεύθυνση που πρέπει να ακολουθηθεί όσον αφορά την κανονικοποίηση. Για παράδειγμα όπως φαίνεται στα παρακάτω διαγράμματα ο RF/CCP που έχουν επιστημονικό ενδιαφέρον έχουν πολύ μεγάλο φάσμα τιμών και μη ομοιόμορφη κατανομή, επομένως ένας απλός μετασχηματισμός μεγίστου-ελαχίστου ώστε να περιοριστούν οι τιμές μεταξύ 0 και 1 δεν θα ήταν αρκετός.



**Σχήμα 4.4:** Κατανομή βασικών μεταβλητών



**Σχήμα 4.5:** Κατανομή βασικών μεταβλητών σε σχέση με τα επίπεδα σοβαρότητας

Συγκεκριμένα χρησιμοποιήθηκαν οι παρακάτω μετασχηματισμοί ανάλογα με τις ανάγκες για κάθε μεταβλητή από την βιβλιοθήκη [sklearn.preprocessing](https://scikit-learn.org/stable/modules/preprocessing.html)

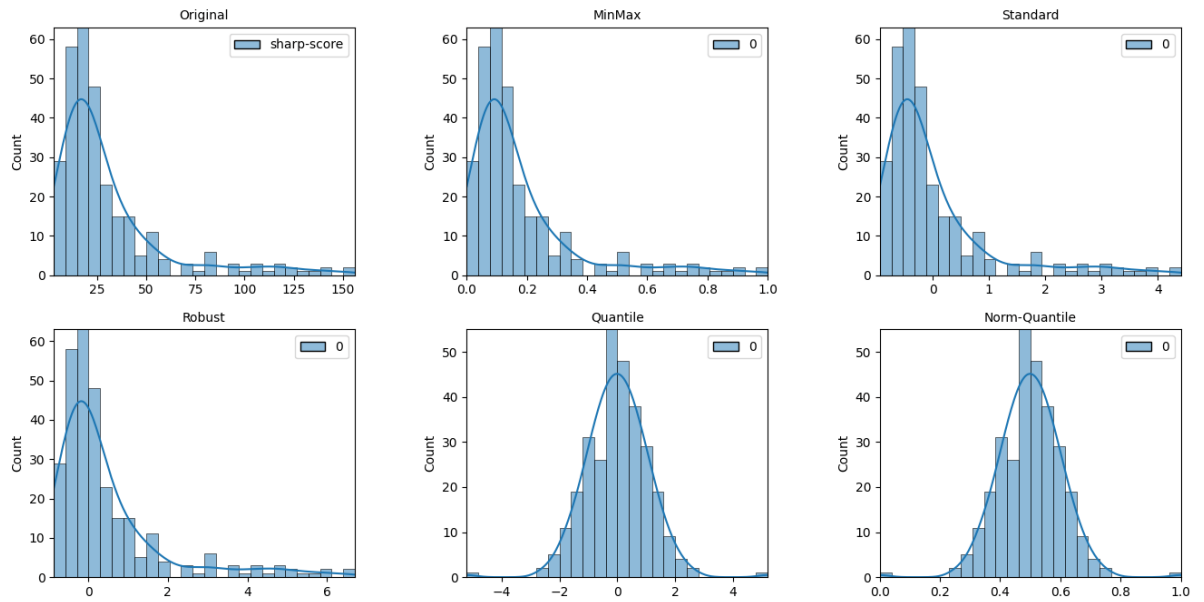
Μετασχηματισμοί
1. MinMaxScaler ()
2. StandardScaler ()
3. RubustScaler ()
4. QuantileTransformer ()

**Πίνακας 4.3:** Ποιοτικός έλεγχος δεδομένων

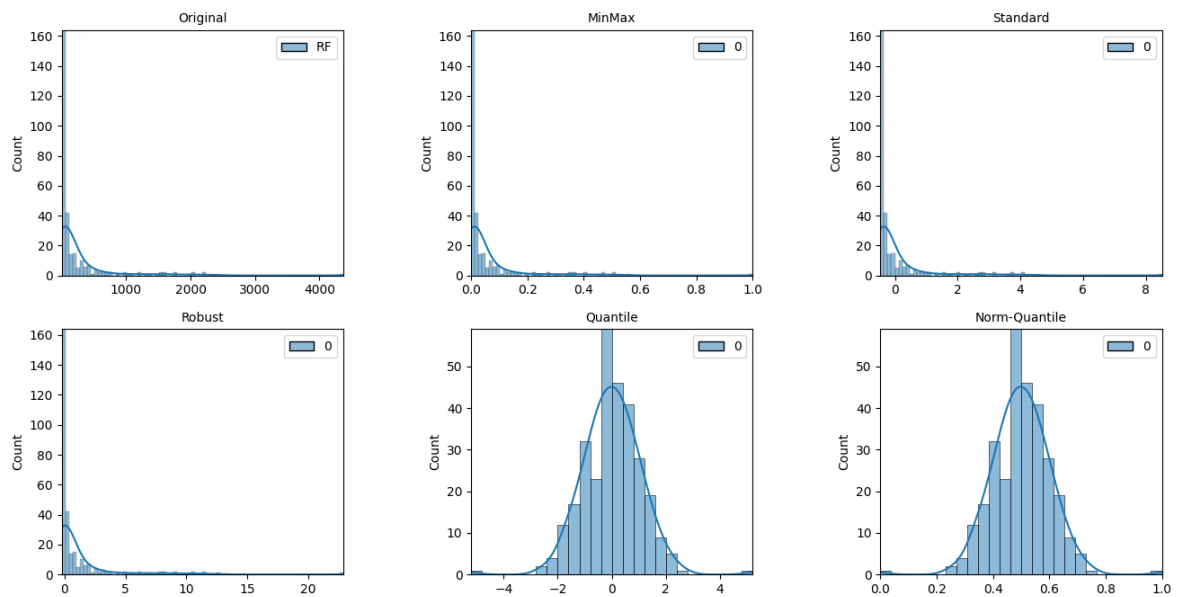
Στα διαγράμματα που ακολουθούν φαίνονται οι μεταβλητές ΡΠ(RF), sharp-score και CCP που αξιοποιήθηκαν κατά κόρον με διάφορους μετασχηματισμούς. Οι τελικοί μετασχηματισμοί που χρησιμοποιήθηκαν για αυτές τις μεταβλητές ήταν QuantileTransformer και στη συνέχεια MinMaxScaler καθώς έπρεπε πρώτα να γίνει η κατανομή πιο ομοιόμορφη και στη συνέχεια να περιοριστεί σε συγκεκριμένα πλαίσια το φάσμα τιμών.

Όσον αφορά μεταβλητές όπως: AGE, DURATION, CRP, DAS28CRP, TKE κλπ έγινε απλός μετασχηματισμός μεγίστου - ελαχίστου ( MinMaxScaler ) καθώς είχαν αρκετά ομοιόμορφη κατανομή και πρόβλημα αποτελούσε μόνο το φάσμα τιμών τους.

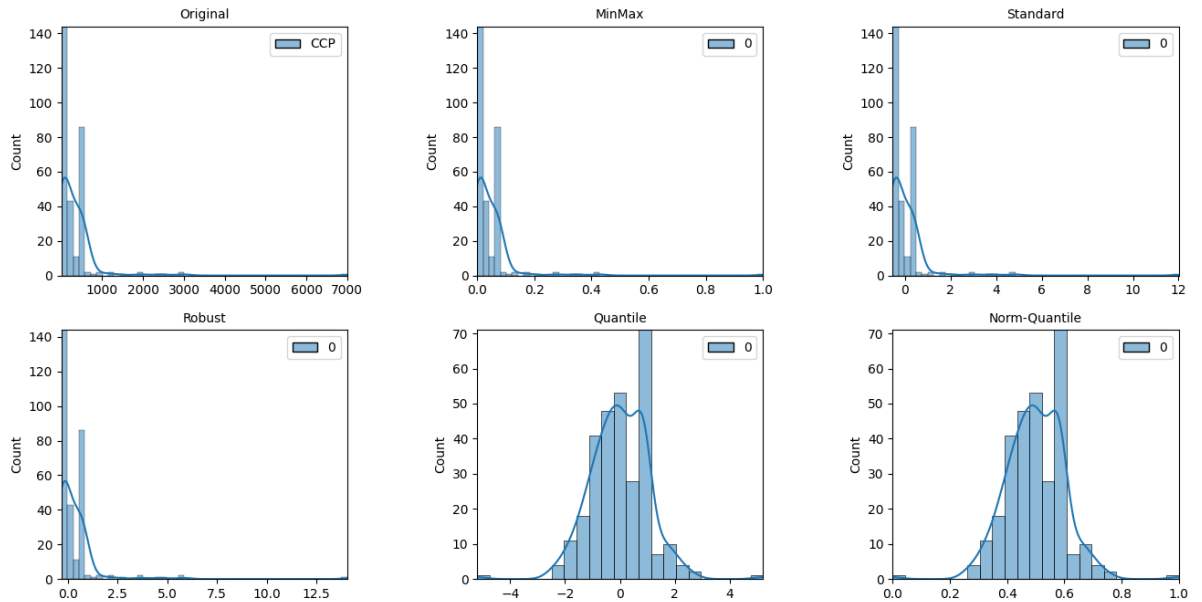
Τέλος για μεταβλητές όπως HLA-DRB1, RFi, CCPi, SMOKING κλπ δεν έγινε κάποιος μετασχηματισμός καθώς έχουν δυαδικές τιμές, έγινε μόνο αντικατάσταση των τιμών με αριθμητικές ( 0 και 1 ).



**Σχήμα 4.6:** Μετασχηματισμοί για τη μεταβλητή sharp-score



**Σχήμα 4.7:** Μετασχηματισμοί για τη μεταβλητή RF



**Σχήμα 4.8:** Μετασχηματισμοί για τη μεταβλητή CCP

### 4.3 Μοντέλα Μηχανικής Μάθησης

Για την ανάπτυξη των μοντέλων εφαρμόστηκε η παρακάτω διαδικασία:

- Βήμα 1: Εισαγωγή απαραίτητων βιβλιοθηκών
- Βήμα 2: καθαρισμός απουσών τιμών
- Βήμα 3: Κανονικοποίηση μεταβλητών που αξιοποιούνται στο εκάστοτε μοντέλο
- Βήμα 4: Εκτέλεση του αντίστοιχου μοντέλου
- Βήμα 5: Καταγραφή και γραφική αναπαράσταση των αποτελεσμάτων

Στο παρόν κεφάλαιο περιγράφονται τα βήματα 1 έως 4. Ειδικότερα για το Βήμα 4 γίνεται αναλυτική παρουσίαση της διαδικασίας. Το Βήμα 5, δηλαδή τα αποτελέσματα του κάθε μοντέλου, θα αναλυθούν στο Κεφάλαιο 5.

#### 4.3.1 Μοντέλο 1 : *kMeans*

Πειραματικά σαν είσοδοι στο μοντέλο *kMeans* δοκιμάστηκαν οι συνδυασμοί μεταβλητών που φαίνονται στον πίνακα που ακολουθεί. Εδώ παρουσιάζονται όλοι οι συνδυασμοί που δοκιμάστηκαν. Στο Κεφάλαιο 5 θα συζητηθούν όσοι από αυτούς παρουσίασαν σχετικά καλή ομαδοποίηση των δεδομένων.

Όπως φαίνεται και από τους παρακάτω συνδυασμούς δοκιμάστηκαν και μεταβλητές που αφορούν την ενεργότητα ώστε να διαπιστωθεί αν διευκολύνουν την ομαδοποίηση όσον αφορά την σοβαρότητα. Τελικά διαπιστώθηκε ότι επιβαρύνουν το μοντέλο και ο καλύτερος συνδυασμός εισόδων είναι ο τελευταίος που φαίνεται στην παρακάτω πίνακα.

Εκ των προτέρων ήταν γνωστό ότι οι πιο σημαντικοί παράγοντες που καθορίζουν την σοβαρότητα ήταν οι εξής: RF, RFi, CCP, CCPi, HLA-DRB1, sharp-score. Γεγονός το οποίο επιβεβαιώνεται από την ομαδοποίηση με kMeans καθώς οι παραπάνω παράγοντες ως είσοδοι οδηγούν στην καλύτερη ομαδοποίηση.

Είσοδοι στο kMeans	
1.	<code>workingDF[["AGE", "TKE", "CRP", "TENDER", "SWOL-LEN", "VAS", "sharp-score", "RF"]]</code>
2.	<code>workingDF[["DAS28TKE", "HDL", "VAS", "sharp-score", "DUPATION"]]</code>
3.	<code>workingDF[["DUPATION", "DAS28TKE", "DAS28CRP", "RF"]]</code>
4.	<code>workingDF[["sharp-score", "RF", "HLA-DRB1", "CCPi"]]</code>
5.	<code>workingDF[["HLA-DRB1", "CCPi", "Rfi", "SMOKING"]]</code>
6.	<code>workingDF[["sharp-score", "DURATION", "ACTIVITY", "HLA-DRB1"]]</code>
7.	<code>workingDF[["sharp-score", "CCP", "CCPi", "RF", "Rfi", "ACTIVITY", "HLA-DRB1"]]</code>
8.	<code>[["sharp-score", "CCP", "CCPi", "RF", "Rfi", "ACTIVITY", "HLA-DRB1", "DAS28TKE"]]</code>
9.	<code>workingDF[["sharp-score", "CCP", "CCPi", "RF", "Rfi", "ACTIVITY", "HLA-DRB1", "DAS28TKE", "DAS28CRP", "HDL", "DURATION", ]]</code>
10.	<code>workingDF[["sharp-score", "DURATION", "CCPi", "CRP", "Rfi", "ACTIVITY", "HLA-DRB1", "DAS28TKE", "DAS28CRP"]]</code>
11.	<code>workingDF[["sharp-score", "CCP", "CCPi", "RF", "Rfi", "ACTIVITY", "HLA-DRB1", "DAS28TKE"]]</code>
12.	<code>workingDF[["sharp-score", "CCP", "CCPi", "RF", "Rfi", "ACTIVITY", "HLA-DRB1"]]</code>
13.	<code>workingDF[["sharp-score", "CCP", "CCPi", "RF", "Rfi", "HLA-DRB1", "DURATION"]]</code>
14.	<code>workingDF[["sharp-score", "CCP", "RF", "HLA-DRB1"]]</code>
15.	<code>workingDF[["CCP", "RF", "HLA-DRB1", "DURATION"]]</code>
16.	<code>workingDF[["CCP", "RF"]]</code>
17.	<code>workingDF[["sharp-score", "CCP", "CCPi", "RF", "Rfi", "HLA-DRB1"]]</code>
18.	<code>workingDF[["<b>sharp-score</b>", "<b>CCP</b>", "<b>RF</b>", "<b>HLA-DRB1</b>"]]</code>

**Πίνακας 4.5:** Δεδομένα εισόδου στο μοντέλο KMeans

Για να μετρηθεί η αποτελεσματικότητα του μοντέλου, συγκρίθηκαν οι συστάδες (clusters) που δημιουργήθηκαν από το Kmeans με τον πίνακα στήλη SEVERITY που περιέχει 3 τιμές

που εκφράζουν την σοβαρότητα και όπως προαναφέρθηκε έχει συμπληρωθεί από κλινικούς ιατρούς ρευματολόγους.

Επομένως η στήλη της σοβαρότητας (severity) αξιολογήθηκε στην κλίμακα από 1 έως 3, όπου όταν έχει τιμή 1 πρόκειται για μια ελαφριά ρευματοειδή, όταν έχει τιμή 2 πρόκειται για μια ρευματοειδή μεσαίας σοβαρότητας και όταν έχει τιμή 3 πρόκειται για μια σοβαρή ρευματοειδή.

Η αξιολόγηση των ασθενών και η συμπλήρωση της σωστής τιμής έγινε από δύο ανεξάρτητους κλινικούς ιατρούς του Νοσοκομείου Ευαγγελισμός, με αποτέλεσμα οι στήλες αυτές να απεικονίζουν τις απόψεις του ειδικού ρευματολόγου για τον κάθε ασθενή.

Επομένως η πρώτη προσπάθεια ομαδοποίησης έγινε με 3 συστάδες (clusters) που όμως δεν παρουσίαζε καλά αποτελέσματα. Διενεργήθηκε λοιπόν έλεγχος ώστε να βρεθεί ο βέλτιστος αριθμός συστάδων (clusters) για τα συγκεκριμένα δεδομένα. Ο έλεγχος αυτός έγινε με 2 τρόπους, αρχικά με Silhouette analysis και στη συνέχεια με Μεθοδολογία Elbow για που παρουσιάζονται αναλυτικά παρακάτω.

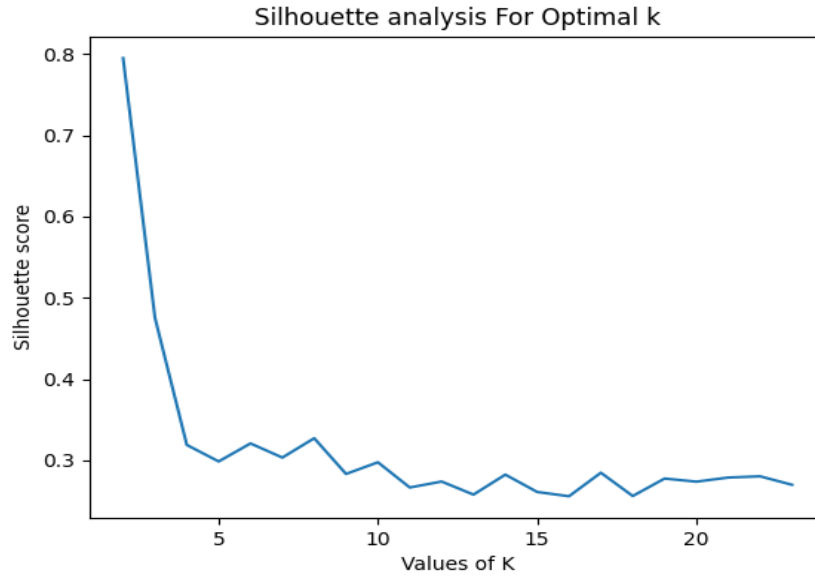
### **Silhouette Analysis**

Η ανάλυση Silhouette μπορεί να χρησιμοποιηθεί για τη μελέτη της απόστασης διαχωρισμού μεταξύ των συστάδων που προκύπτουν.

Η γραφική παράσταση που προκύπτει εμφανίζει ένα μέτρο του πόσο κοντά είναι κάθε σημείο σε ένα σύμπλεγμα (cluster) σε σημεία των γειτονικών συστάδων (clusters) και έτσι παρέχει έναν τρόπο αξιολόγησης παραμέτρων όπως ο αριθμός των συστάδων (clusters) οπτικά. Αυτό το μέτρο έχει εύρος  $[-1, 1]$ .

Οι συντελεστές Silhouette όπως αναφέρονται οι τιμές αυτές που προκύπτουν στον y άξονα του γραφήματος Silhouette κοντά στο +1 υποδεικνύουν ότι το δείγμα απέχει πολύ από τις γειτονικές ομάδες. Η τιμή 0 υποδηλώνει ότι το δείγμα βρίσκεται στο όριο απόφασης μεταξύ δύο γειτονικών συστάδων (clusters) ή πολύ κοντά στο όριο και οι αρνητικές τιμές υποδεικνύουν ότι αυτά τα δείγματα μπορεί να έχουν εκχωρηθεί σε λάθος σύμπλεγμα (cluster).

Όπως φαίνεται από την παρακάτω γραφική παράσταση που προέκυψε από τα δεδομένα που αξιοποιούνται για τα μοντέλα που αναφέρονται στην παρούσα εργασία, ο βέλτιστος αριθμός συστάδων είναι **k=2** καθώς αυτή η τιμή του x άξονα αντιστοιχίζεται σε τιμή κοντά στο +1 του y άξονα.

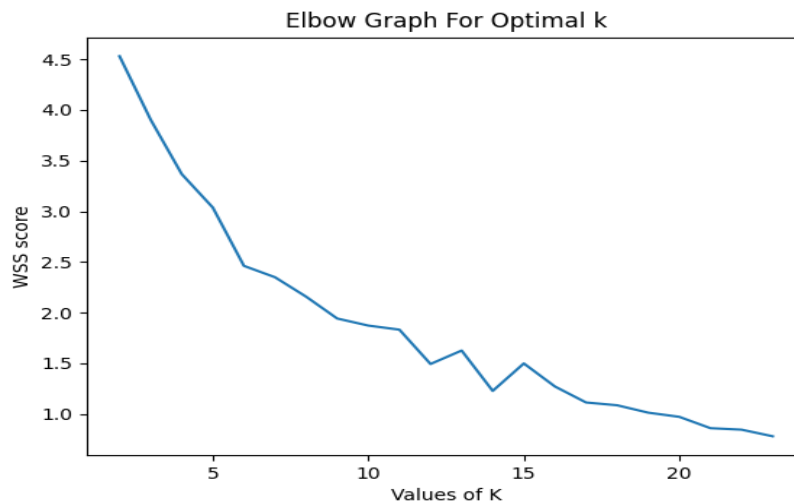


**Σχήμα 4.9:** Γράφημα Silhouette για βέλτιστο αριθμό clusters

### **Μεθοδολογία Elbow**

Η μέθοδος elbow έχει αναλυθεί στο Κεφάλαιο 2 και χρησιμοποιείται για τον προσδιορισμό των clusters σε ένα σύνολο δεδομένων.

Από τους αγκώνες «elbow» ή αλλιώς τα γόνατα της καμπύλης που προκύπτει και φαίνεται παρακάτω μπορεί να προσδιοριστεί και βέλτιστος αριθμός clusters.



**Σχήμα 4.10:** Γράφημα Elbow για βέλτιστο αριθμό clusters



Όπως φαίνεται και από την παραπάνω γραφική, ή μεθοδολογία elbow δεν δίνει ξεκάθαρη απάντηση για τον βέλτιστο αριθμό clusters με βάση το σύνολο δεδομένων που αξιοποιήθηκε. Συγκεκριμένα γόνατο παρουσιάζει για  $k=3$ ,  $k=6$ ,  $k=9$  κλπ.

Μετά από δοκιμές βρέθηκε ότι ο βέλτιστος αριθμός clusters που παρουσιάζει την καλύτερη ομαδοποίηση είναι  **$k=2$** . Επομένως και στα υπόλοιπα μοντέλα που αναπτύχθηκαν ώστε να μπορεί να γίνει σύγκριση των αποτελεσμάτων, έγινε ομαδοποίηση σε 2 clusters.

Επειδή λοιπόν στην πορεία της έρευνας βρέθηκε ότι το σύνολο δεδομένων με βάση τη σοβαρότητα ομαδοποιείται καλύτερα σε 2 συστάδες ζητήθηκε από τους κλινικούς ιατρούς να ξανασυμπληρωθεί η στήλη SEVERITY με δυαδικές τιμές. Δηλαδή στην κλίμακα από 1 έως 2, όπου όταν έχει τιμή 1 πρόκειται για μια ελαφριά ρευματοειδή, όταν έχει τιμή 2 πρόκειται για μια σοβαρή ρευματοειδή.

Θα παρουσιαστούν τα αποτελέσματα και με τις 2 γνωματεύσεις των κλινικών ιατρών και για λόγους διαχωρισμού ονομάστηκαν SEVERITY (σοβαρότητα 1 έως 3) και SEVERITY\_binary (σοβαρότητα 1 και 2).

#### **4.3.2 Μοντέλο 2 : GaussianMixture**

Πειραματικά σαν είσοδοι στο μοντέλο GaussianMixture δοκιμάστηκαν οι συνδυασμοί μεταβλητών που φαίνονται στον πίνακα που παρουσιάστηκε στο Μοντέλο 1.

Χρησιμοποιήθηκαν οι ίδιοι ακριβώς συνδυασμοί σαν είσοδοι με το μοντέλο kMeans και έγινε ομαδοποίηση και εδώ σε 2 clusters, ώστε να μπορεί να γίνει σύγκριση μεταξύ των μοντέλων.

Στην GaussianMixture όπως και στο kMeans, ο συνδυασμός εισόδων με τα καλύτερα αποτελέσματα ήταν ο παρακάτω:

```
teachArr = workingDF[["sharp-score", "CCP", "RF", "HLA-DRB1"]]
```

#### **4.3.3 Μοντέλο 3 : K-Medoids**

Πειραματικά σαν είσοδοι στο μοντέλο K-Medoids δοκιμάστηκαν οι συνδυασμοί μεταβλητών που φαίνονται στον πίνακα που παρουσιάστηκε στο Μοντέλο 1.

Χρησιμοποιήθηκαν οι ίδιοι ακριβώς συνδυασμοί σαν είσοδοι με το μοντέλο kMeans και έγινε ομαδοποίηση και εδώ σε 2 clusters, ώστε να μπορεί να γίνει σύγκριση μεταξύ των μοντέλων.

Στην K-Medoids ο συνδυασμός εισόδων με τα καλύτερα αποτελέσματα ήταν ο παρακάτω:

```
teachArr = workingDF[["sharp-score", "CCP", "CCPi", "RF", "Rfi", "HLA-DRB1"]]
```

#### 4.3.4 Μοντέλο 4 : *k*- Nearest Neighbors

Στο κομμάτι της επιβλεπόμενης Μηχανικής Μάθησης έγινε προσπάθεια πρόβλεψης της σοβαρότητας νόσου κάθε ασθενή με βάση συγκεκριμένα κλινικά χαρακτηριστικά.

Στα μοντέλα επιβλεπόμενης Μηχανικής μάθησης που αναπτύχθηκαν δεν χρειάστηκαν πολλές δοκιμές στον πίνακα εισόδων του μοντέλου καθώς τα χαρακτηριστικά που καθορίζουν την σοβαρότητα της νόσου είναι γνωστά και με βάση αυτά πρέπει να γίνει η εκπαίδευση του μοντέλου.

Συγκεκριμένα δοκιμάστηκαν συνδυασμοί των παρακάτω παραγόντων οι οποίοι σχετίζονται άμεσα με τη σοβαρότητα της νόσου.

Χαρακτηριστικά που σχετίζονται με Σοβαρότητα Νόσου
1. RF – Rfi
2. CCP- CCPi
3. HLA_DRB1
4. Sharp score
5. Smoking
6. Duration
7. CRP
8. TKE

**Πίνακας 4.6**

Στον παρακάτω πίνακα παρουσιάζονται οι συνδυασμοί που χρησιμοποιήθηκαν σαν είσοδοι στο μοντέλο k-NN

Είσοδοι στο kNN
1. <code>workingDF[["sharp-score", "DURATION", "CCP", "RF", "HLA-DRB1"]]</code>
2. <code>workingDF[["sharp-score", "AGE", "CCP", "RF", "HLA-DRB1"]]</code>
3. <code>workingDF[["sharp-score", "DURATION", "CCPi", "Rfi", "HLA-DRB1"]]</code>
4. <code>workingDF[["sharp-score", "DURATION", "CCP", "RF", "HLA-DRB1", "Rfi", "CCPi"]]</code>
5. <code>workingDF[["sharp-score", "CCP", "RF", "HLA-DRB1", "Rfi", "CCPi"]]</code>
6. <code>workingDF[["sharp-score", "SMOKING", "CCP", "RF", "HLA-DRB1"]]</code>

7.	<code>workingDF[["sharp-score", "DURATION", "SMOKING", "CCP", "RF", "HLA-DRB1"]]</code>
8.	<code>workingDF[["DURATION", "CCP", "SMOKING", "RF", "HLA-DRB1"]]</code>
9.	<code>workingDF[["CCP", "RF", "HLA-DRB1"]]</code>
10.	<code>workingDF[["Rfi", "CCPi", "CCP", "RF", "HLA-DRB1"]]</code>
11.	<code>workingDF[["Rfi", "SMOKING", "CCPi", "CCP", "RF", "HLA-DRB1"]]</code>
12.	<code>workingDF[["CCP", "RF", "HLA-DRB1", "SMOKING"]]</code>
13.	<code>workingDF[["sharp-score", "SMOKING", "CCPi", "Rfi", "HLA-DRB1"]]</code>
14.	<code>workingDF[["sharp-score", "CCPi", "Rfi", "HLA-DRB1"]]</code>

**Πίνακας 4.7:** Δεδομένα εισόδου στο μοντέλο KNN

Στην συνέχεια το working dataframe χωρίστηκε σε training data και testing data. Χρησιμοποιήθηκαν διάφοροι ποσοστιαίοι συνδυασμοί αλλά κατά κύριο λόγο χρησιμοποιήθηκε το 10%- 30% του δείγματος για έλεγχο και όλο το υπόλοιπο για εκπαίδευση.

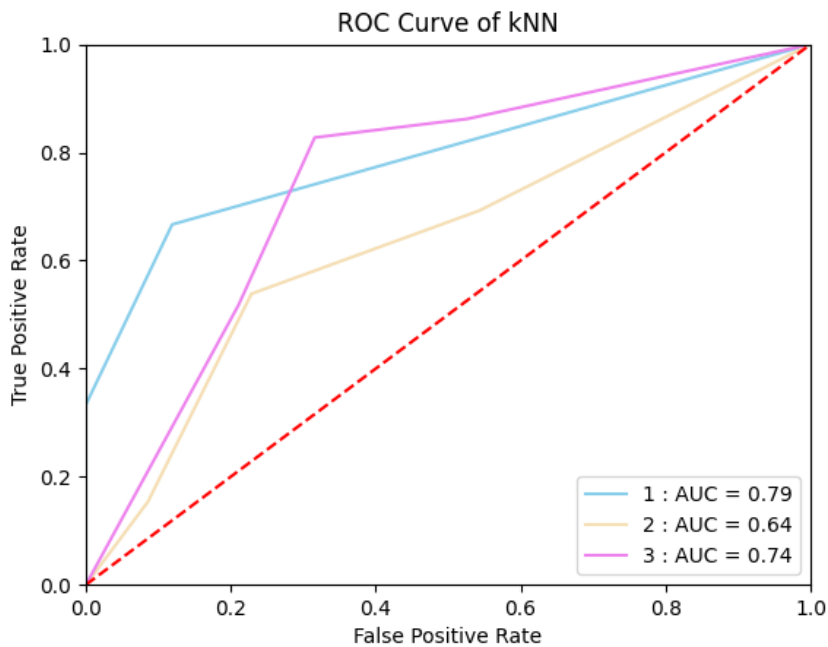
Εδώ επειδή ο πίνακας SEVERITY ο οποίος χρησιμοποιήθηκε για το validation του μοντέλου, εμπεριέχει τρεις διαφορετικές τιμές για την σοβαρότητα το μοντέλο αναπτύχθηκε για k=3 δηλαδή δημιουργήθηκαν 3 ομάδες.

Παρότι τα αποτελέσματα του μοντέλου ήταν αρκετά καλά είχε ενδιαφέρον να ερευνηθεί αν προκύπτουν καλύτερα αποτελέσματα αξιοποιώντας τον πίνακα SEVERITY\_binary για το validation του μοντέλου που δημιουργήθηκε στην πορεία της έρευνας και εμπεριέχει δυαδικές τιμές για την σοβαρότητα. Σε αυτήν την περίπτωση δημιουργήθηκαν 2 ομάδες στην εφαρμογή του μοντέλου (για k=2) και τα αποτελέσματα του μοντέλου ήταν αρκετά καλύτερα.

Ένας τρόπος παρουσίασης των επιδόσεων ανά κλάση ενός κατηγοριοποιητή που αξιοποιήθηκε είναι με τη χρήση του **πίνακα σύγχυσης** (confusion matrix). Ο Πίνακας Σύγχυσης είναι ένας δισδιάστατος πίνακας, όπου οι στήλες αντιστοιχούν στις προβλέψεις και οι γραμμές στις πραγματικές τιμές κλάσης. Στα κελιά του πίνακα αναγράφονται οι αληθινές θετικές (true positive), οι αληθινές αρνητικές (true negative), οι ψευδείς θετικές (false positive) και οι ψευδείς αρνητικές (false negative) προβλέψεις. Οι πίνακες σύγχυσης για τις διάφορες εισόδους που δοκιμάστηκαν θα παρουσιαστούν παρακάτω στο Κεφ.5.

Ένα επιπλέον μέτρο που αξιοποιήθηκε για την εκτίμηση της ανά κλάσης ακρίβειας κάθε κατηγοριοποιητή είναι η **καμπύλη ROC**. Πρόκειται για μια δυσδιάστατη καμπύλη όπου ο οριζόντιος άξονας εκφράζει το μέγεθος specificity ή False Positive Rate και ο κατακόρυφος άξονας εκφράζει το μέγεθος sensitivity ή True Positive Rate.

Ο οριζόντιος άξονας εκφράζει το ποσοστό των αρνητικών παρατηρήσεων, οι οποίες κατηγοριοποιήθηκαν λάθος, και ο κατακόρυφος άξονας εκφράζει το ποσοστό των θετικών παρατηρήσεων, οι οποίες κατηγοριοποιήθηκαν σωστά. Το σημείο 0,0 είναι ένας κατηγοριοποιητής, που δεν προβλέπει ποτέ θετική παρατήρηση. Το σημείο 1,1 είναι ένας κατηγοριοποιητής, που προβλέπει πάντα θετική παρατήρηση. Η διαγώνια γραμμή, από το σημείο 0,0 στο σημείο 1,1 είναι ένας κατηγοριοποιητής που προβλέπει τυχαία την κλάση. Οι κατηγοριοποιητές που βρίσκονται κάτω από τη διαγώνια γραμμή είναι χειρότεροι από την τυχαία πρόβλεψη. Οι κατηγοριοποιητές που βρίσκονται πάνω από τη διαγώνια γραμμή είναι καλύτεροι από την τυχαία πρόβλεψη. Το σημείο 0,1 είναι ο άριστος κατηγοριοποιητής, οι οποίος προβλέπει σωστά όλες τις θετικές και αρνητικές παρατηρήσεις. Γενικώς, όσο πιο μετατοπισμένο είναι προς τα επάνω και προς τα αριστερά ένα σημείο, τόσο καλύτερη θεωρείται η επίδοση. Στο παρακάτω σχήμα φαίνεται ένα παράδειγμα ROC καμπύλης που προέκυψε.



**Σχήμα 4.11:** ROC Curve

# 5

## *Αποτελέσματα*

### *5.1 Μοντέλο 1: KMeans*

Επειδή αναπτύχθηκε πληθώρα μοντέλων με πολλαπλούς συνδυασμούς εισόδων παρακάτω θα παρουσιαστούν τα αποτελέσματα για τις εισόδους που παρουσίασαν την καλύτερη ομαδοποίηση και έχουν μεγαλύτερο επιστημονικό ενδιαφέρον να αναλυθούν.

Επιπλέον θα παρουσιαστούν χωριστά τα αποτελέσματα στην Περίπτωση 1 όπου για την αξιολόγηση του μοντέλου χρησιμοποιήθηκε ο πίνακας SEVERITY ( τιμές σοβαρότητας 1-3) και στην Περίπτωση 2 όπου για την αξιολόγηση του μοντέλου χρησιμοποιήθηκε ο πίνακας SEVERITY\_binary ( δυαδικές τιμές σοβαρότητας 1 και 2 )

#### *5.1.1 Περίπτωση 1 KMeans: SEVERITY*

Στην συγκεκριμένη περίπτωση χρησιμοποιήθηκε ο πίνακας SEVERITY για την επαλήθευση του μοντέλου ο οποίος περιέχει 3 τιμές για την σοβαρότητα:

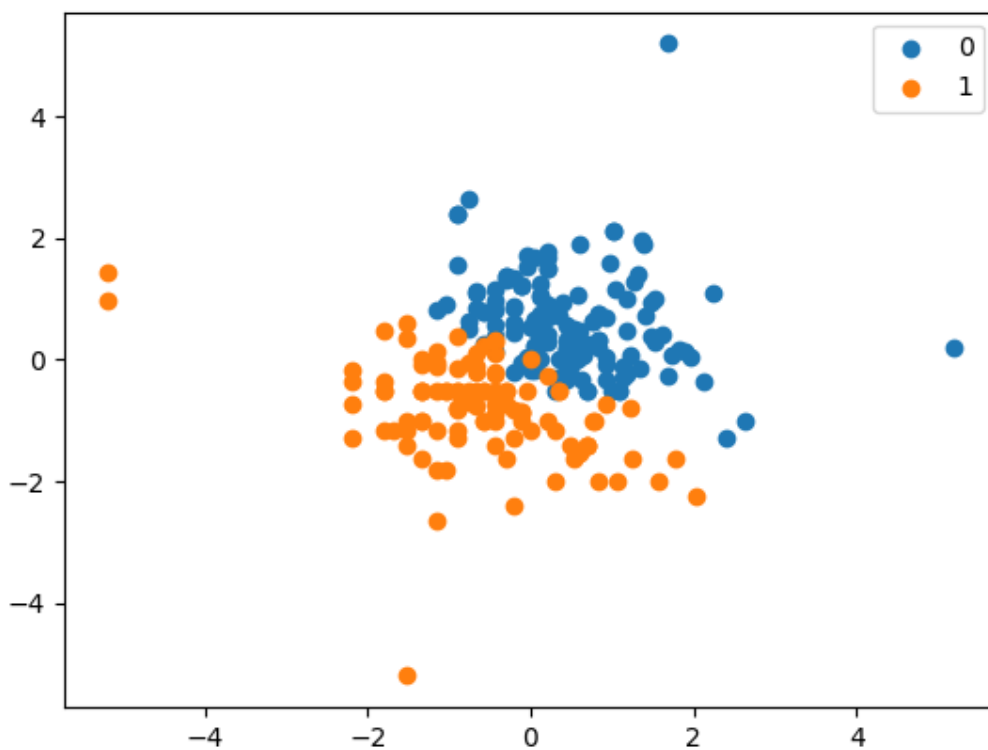
1. Ελαφριά Ρευματοειδής
2. Ρευματοειδής Μέτριας Σοβαρότητας
3. Βαριά Ρευματοειδής

*5.1.1.1 Είσοδος 1 KMeans: `teachArr = workingDF[["sharp-score","RF","HLA-DRB1","CCPi"]].to_numpy()`*

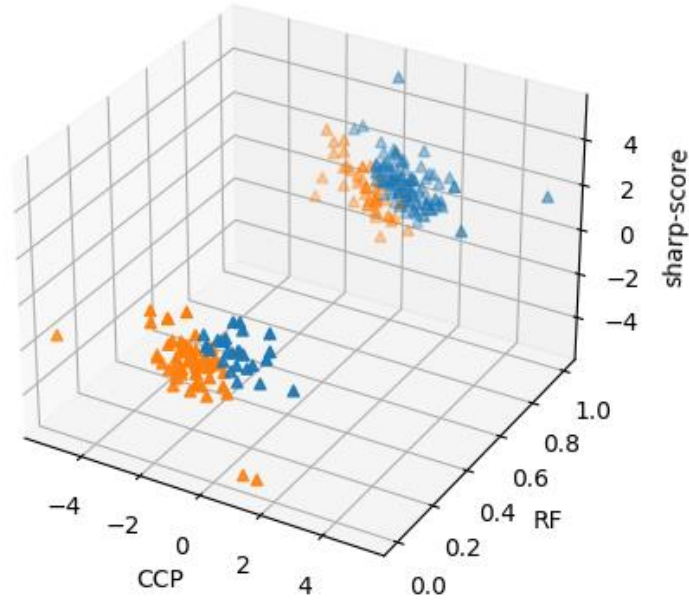
Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.05 of all sev=1	0.95 of all sev=1
0.25 of all sev=2	0.75 of all sev=2
0.77 of all sev=3	0.23 of all sev=3

Όπως φαίνεται στο group 0 περιέχεται το 77% των ασθενών με σοβαρότητα 3 ενώ στο group 1 περιέχεται το 95% των ασθενών με σοβαρότητα 1. Πρόκειται για μια αρκετά καλή ομαδοποίηση. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης



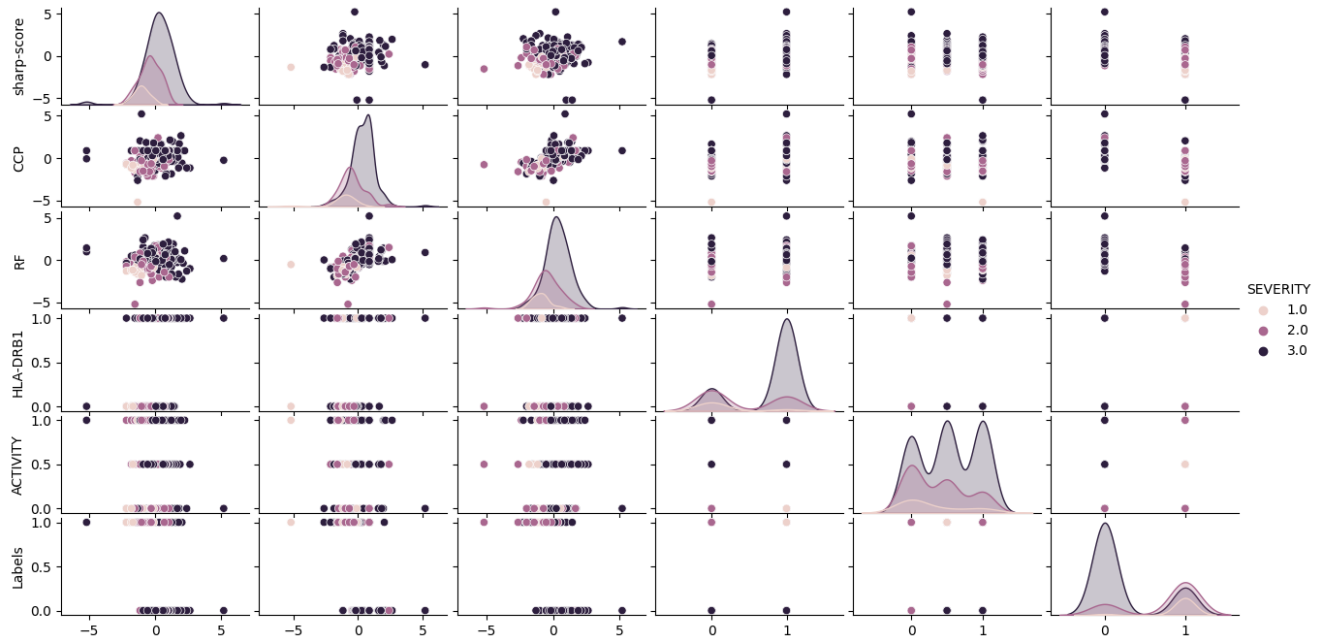
**Σχήμα 5.1 :** Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο kmeans με είσοδο: sharp-score, RF, HLA-DRB1, CCPi



**Σχήμα 5.2 :** Τρισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο kmeans με είσοδο: sharp-score,RF,HLA-DRB1,CCPi



**Σχήμα 5.3 :** Συσγέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο kmeans με είσοδο: sharp-score,RF,HLA-DRB1,CCPi



**Σχήμα 5.4:** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της σοβαρότητας για την κάθε γραφική

Όσον αφορά τα Σχήματα 3 και 4 απεικονίζουν την συσχέτιση των μεταβλητών sharp-score, CCP, RF, HLA-DRB1, ACTIVITY, SEVERITY μεταξύ τους. Στην διαγώνιο είναι η κατανομή καθεμίας από τις παραπάνω μεταβλητές. Το πάνω αριστερά σχήμα είναι η γραφική αναπαράσταση του sharp-score, το 2<sup>ο</sup> κάτω από το sharp-score πάνω αριστερά είναι η γραφική παράσταση CCP και sharp-score, το ακριβώς από κάτω είναι η γραφική παράσταση RF και sharp-score κλπ.

Όσον αφορά το Σχήμα 3 σε κάθε γραφική αναπαράσταση απεικονίζεται η κατανομή στις 2 ομάδες που δημιουργήθηκαν ενώ στο Σχήμα 4 σε κάθε γραφική αναπαράσταση απεικονίζεται η κατανομή σε σχέση με τα 3 επίπεδα σοβαρότητας.

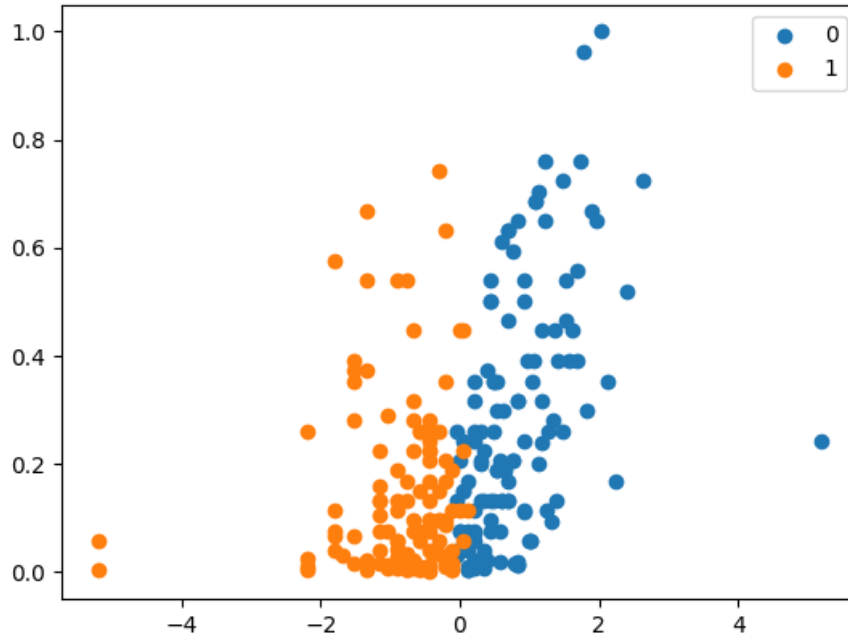
5.1.1.2 **Είσοδος 2 KMeans** : `teachArr = workingDF[["sharp-score", "DURATION", "HLA-DRB1"]].to_numpy()`

Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

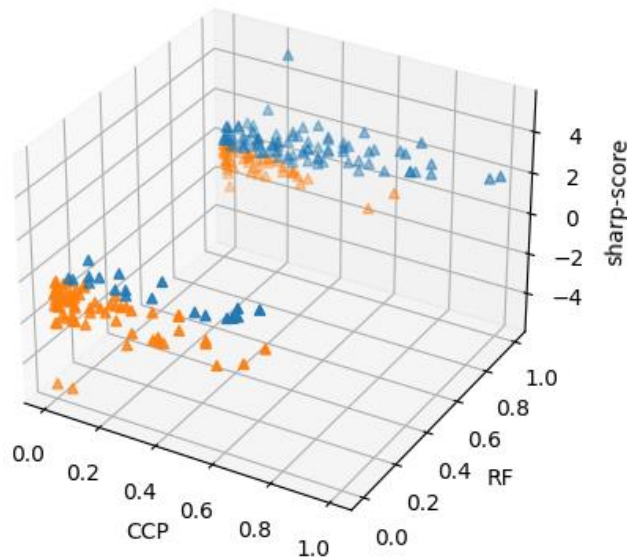
group 0	group 1
0.05 of all sev=1	0.95 of all sev=1
0.32 of all sev=2	0.68 of all sev=2
0.64 of all sev=3	0.36 of all sev=3



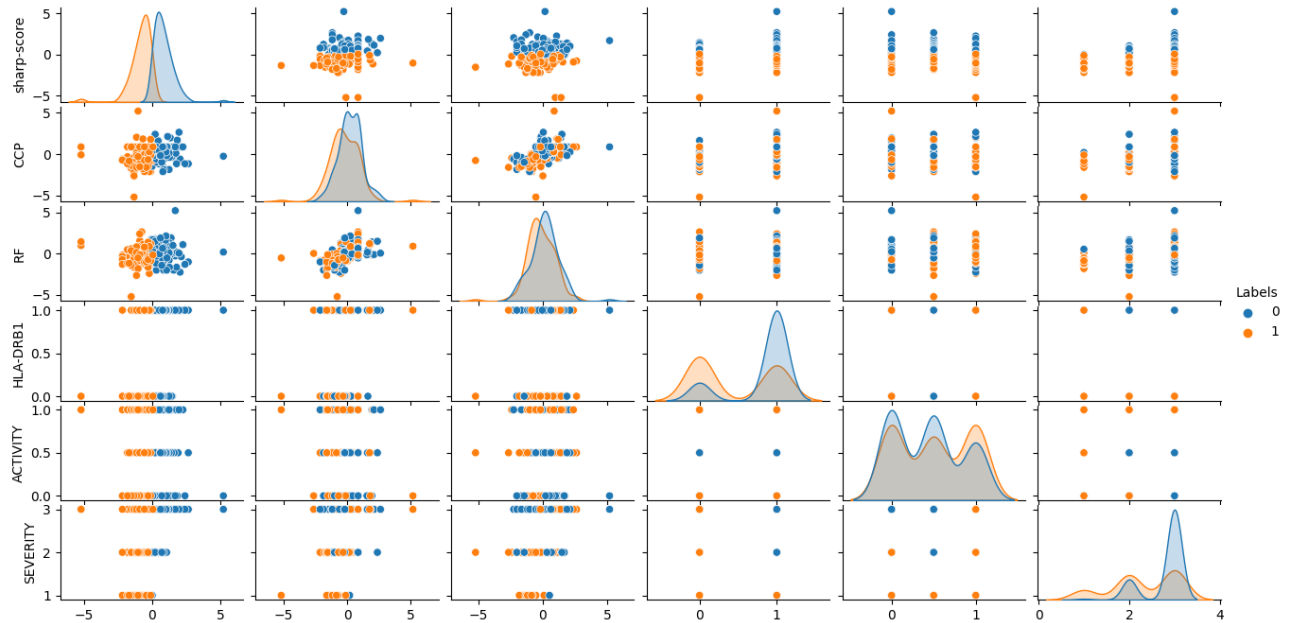
Όπως φαίνεται στο group 0 περιέχεται το 64% των ασθενών με σοβαρότητα 3 ενώ στο group 1 περιέχεται το 95% των ασθενών με σοβαρότητα 1. Επιπλέον η σοβαρότητα 2 βρίσκεται κατά μεγαλύτερο βαθμό στο group 1 (68%). Πρόκειται για μια αρκετά καλή ομαδοποίηση όχι όμως καλύτερη από την είσοδο 1. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης.



**Σχήμα 5.5 :** Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο kmeans με είσοδο: sharp-score,DURATION,HLA-DRB1



**Σχήμα 5.6 :** Τρισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο kmeans με είσοδο: sharp-score,DURATION,HLA-DRB1



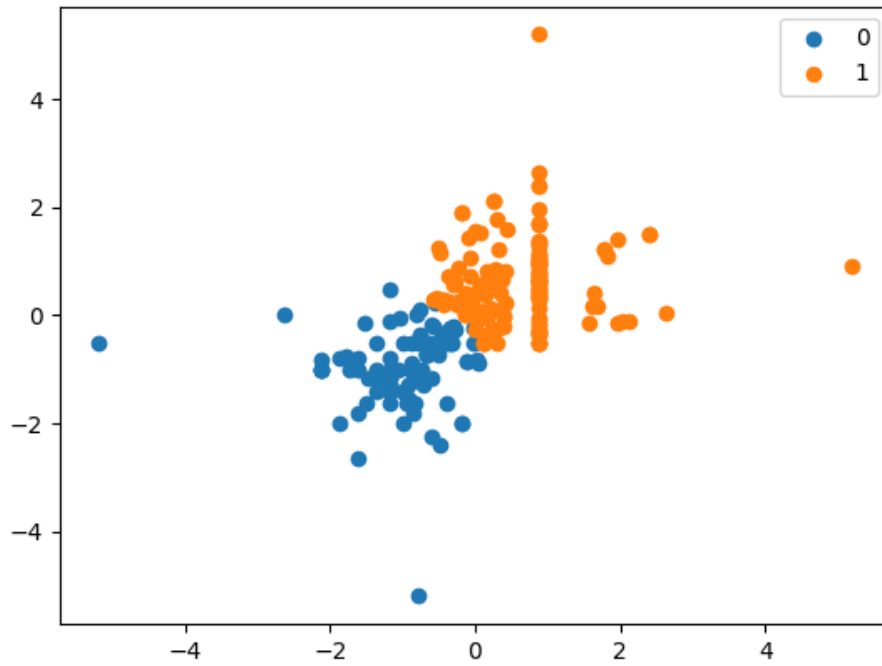
**Σχήμα 5.7 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο kmeans με είσοδο: sharp-score,DURATION,HLA-DRB1

### 5.1.1.3 Είσοδος 3 KMenas: `teachArr = workingDF[["CCP", "RF", "HLA-DRB1"]].to_numpy()`

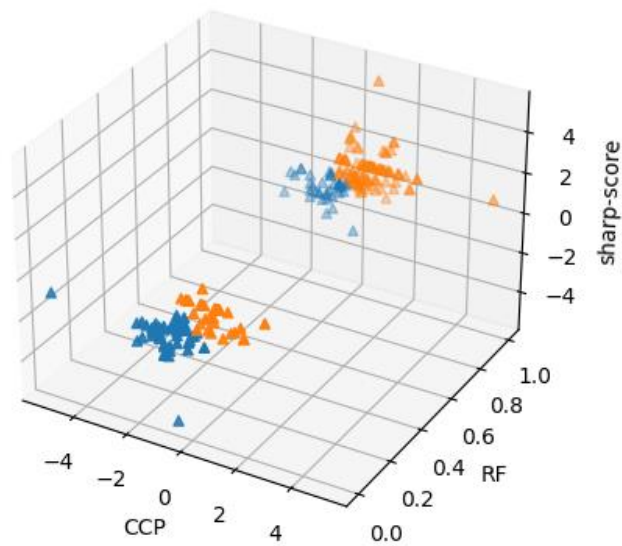
Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.95 of all sev=1	0.05 of all sev=1
0.68 of all sev=2	0.32 of all sev=2
0.19 of all sev=3	0.81 of all sev=3

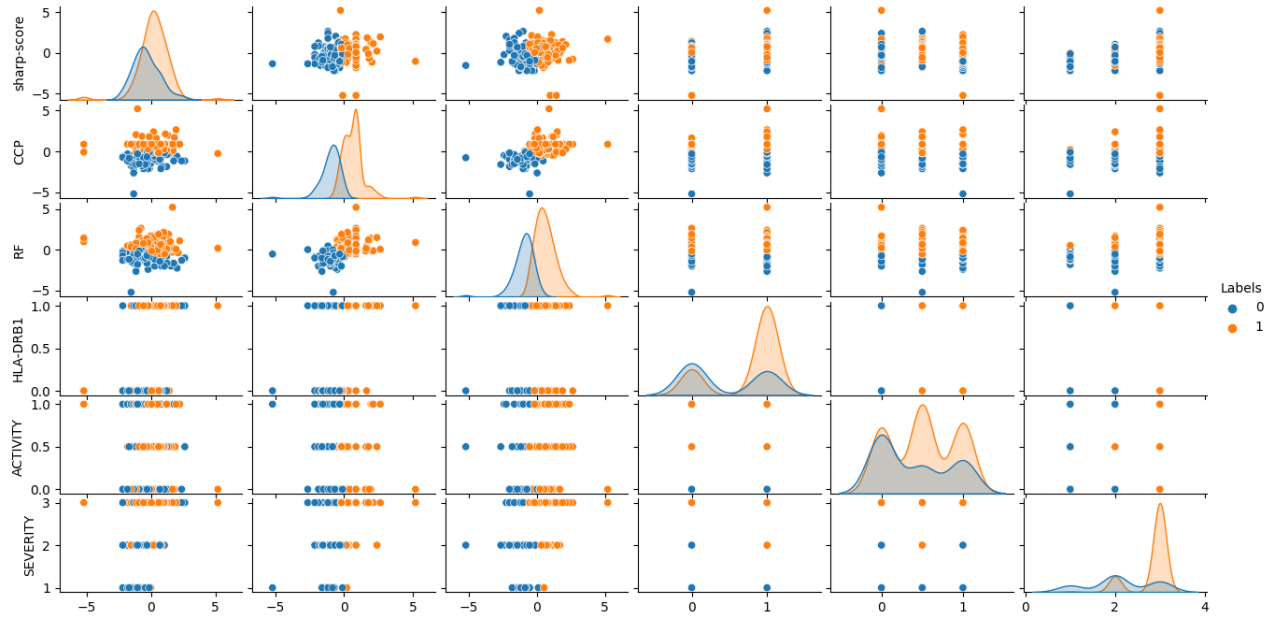
Όπως φαίνεται στο group 0 περιέχεται το 95% των ασθενών με σοβαρότητα 1 ενώ στο group 1 περιέχεται το 81% των ασθενών με σοβαρότητα 1. Επιπλέον η σοβαρότητα 2 βρίσκεται κατά μεγαλύτερο βαθμό στο group 0 (68%). Πρόκειται για μια πολύ καλή ομαδοποίηση, μέχρι στιγμής η καλύτερη από τις τρεις. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης



**Σχήμα 5.8 :** Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο kmeans με είσοδο: CCP,RF,HLA-DRB1



**Σχήμα 5.9 :** Τρισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο kmeans με είσοδο: CCP,RF,HLA-DRB1



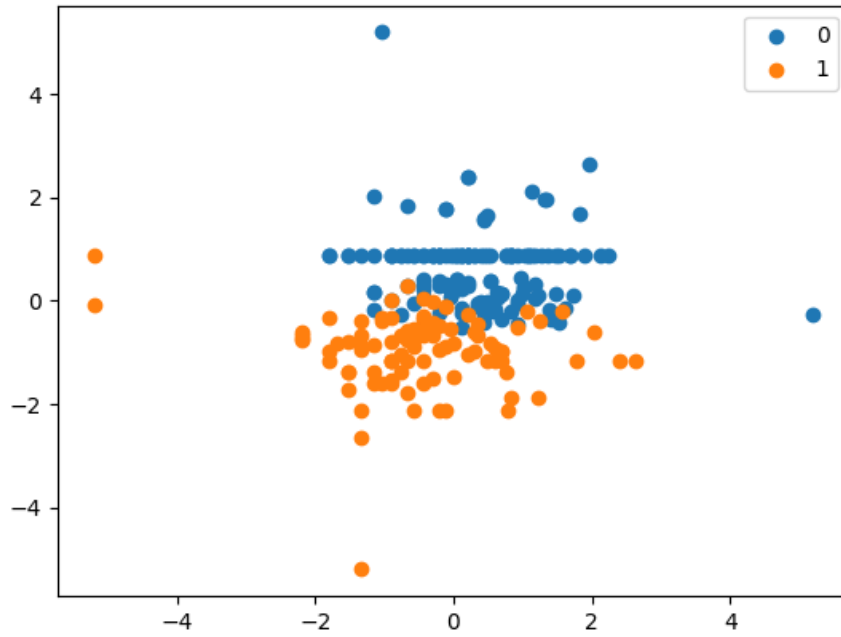
**Σχήμα 5.10 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο kmeans με είσοδο: CCP,RF,HLA-DRB1

5.1.1.4 **Είσοδος 4 KMeans:** `teachArr = workingDF[["sharp-score", "CCP", "CCPi", "RF", "Rfi", "HLA-DRB1"]].to_numpy()`

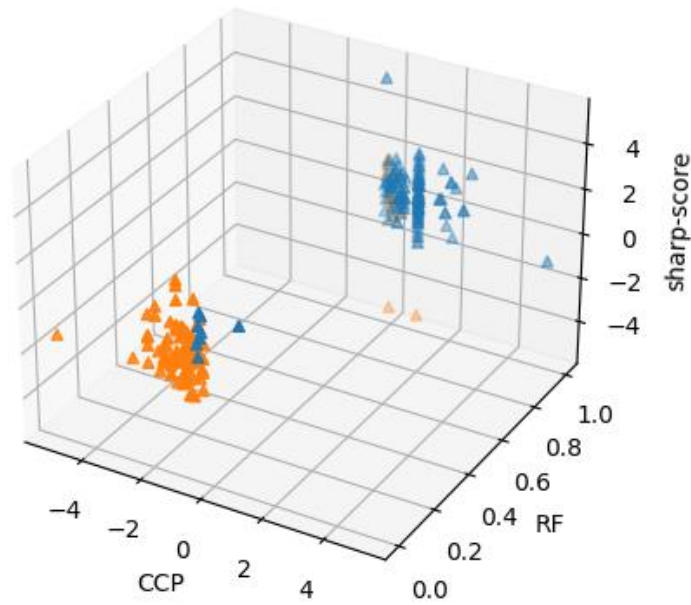
Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.05 of all sev=1	0.95 of all sev=1
0.29 of all sev=2	0.71 of all sev=2
0.80 of all sev=3	0.20 of all sev=3

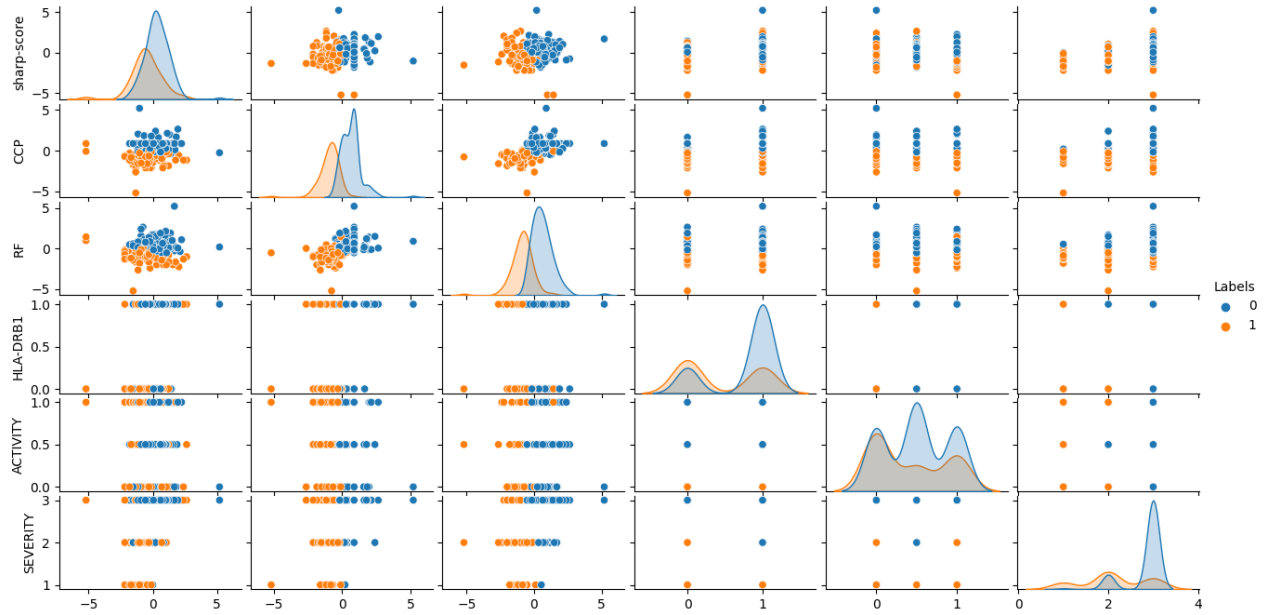
Όπως φαίνεται στο group 0 περιέχεται το 80% των ασθενών με σοβαρότητα 3 ενώ στο group 1 περιέχεται το 95% των ασθενών με σοβαρότητα 1. Επιπλέον η σοβαρότητα 2 βρίσκεται κατά μεγαλύτερο βαθμό στο group 1(71%). Πρόκειται για μια αρκετά καλή ομαδοποίηση, ειδικά σε επόμενους αλγόριθμους έχει ακόμα καλύτερα αποτελέσματα. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης



**Σχήμα 5.11** : Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο kmeans με είσοδο: sharp-score,CCP,CCPi,RF,Rfi,HLA-DRB1



**Σχήμα 5.12** : Τρισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο kmeans με είσοδο: sharp-score,CCP,CCPi,RF,Rfi,HLA-DRB1



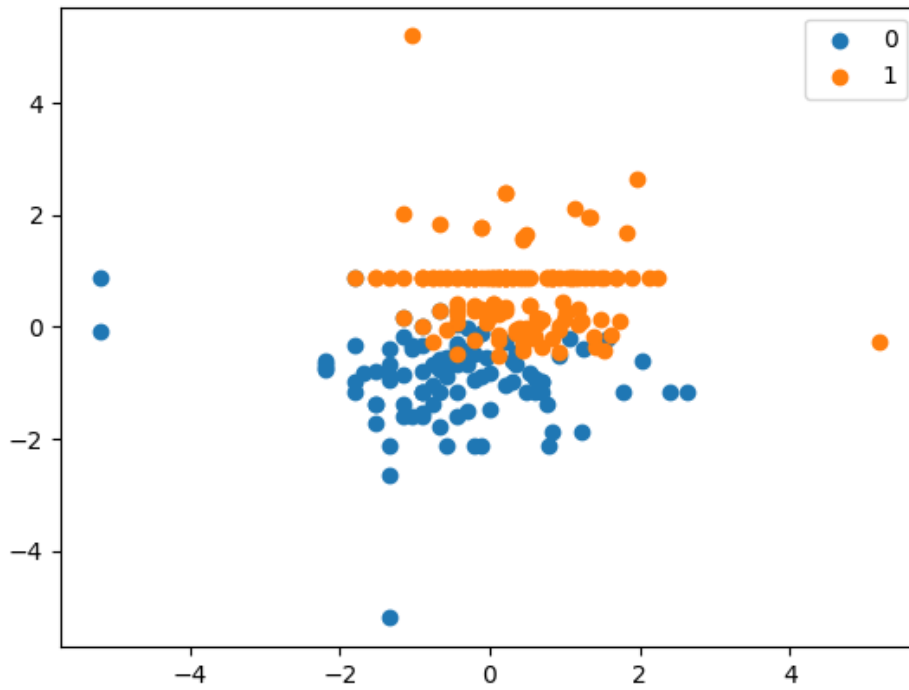
**Σχήμα 5.13 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο kmeans με είσοδο: sharp-score,CCP,CCPi,RF,Rfi,HLA-DRB1

5.1.1.5 **Είσοδος 5 KMeans:** `teachArr = workingDF[["sharp-score","CCP","RF","HLA-DRB1","DURATION"]].to_numpy()`

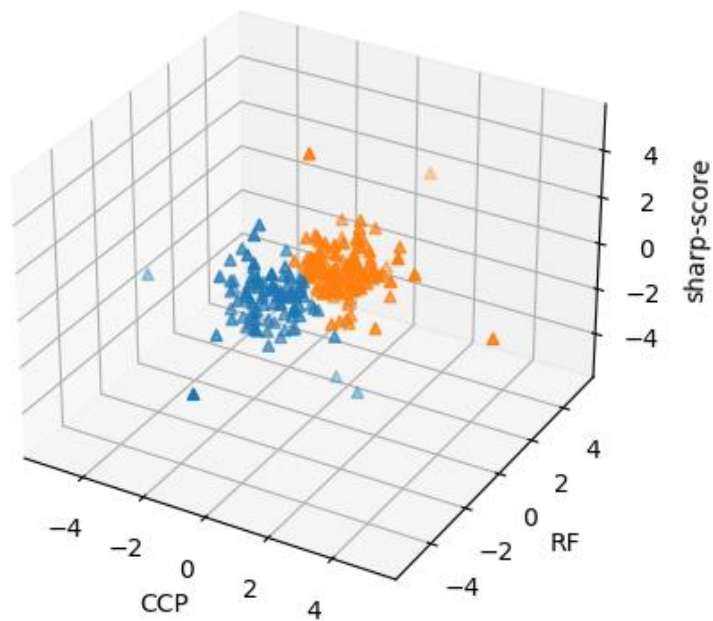
Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.95 of all sev=1	0.05 of all sev=1
0.72 of all sev=2	0.28 of all sev=2
0.21 of all sev=3	0.79 of all sev=3

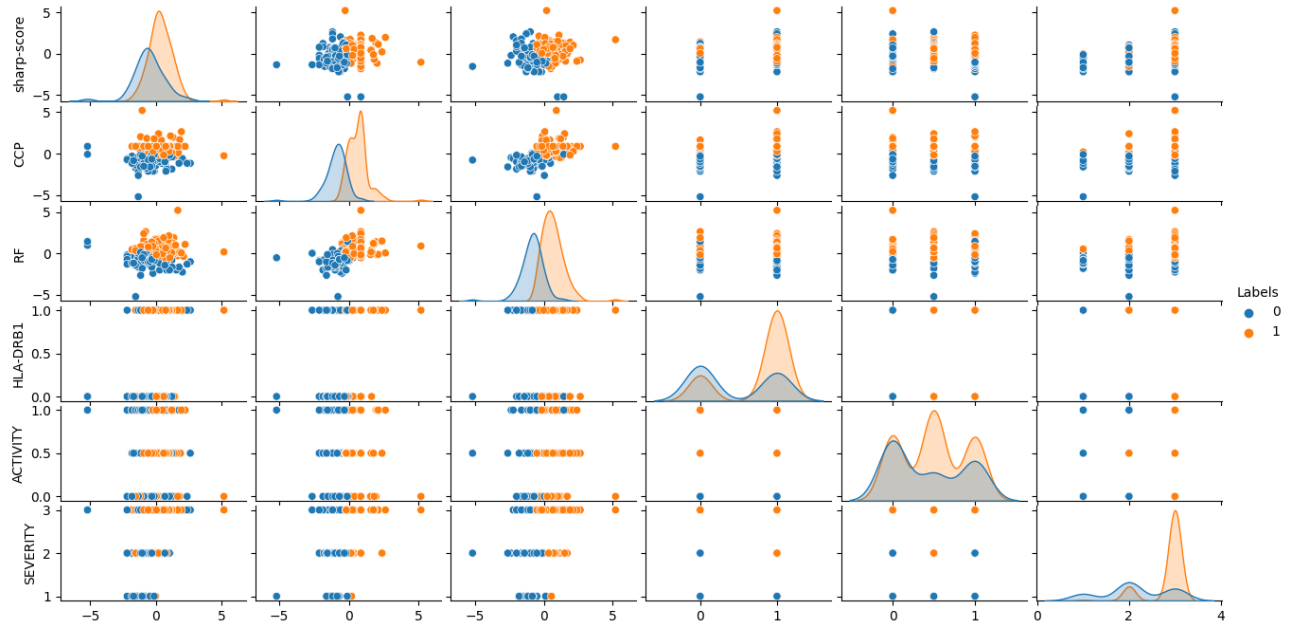
Όπως φαίνεται στο group 0 περιέχεται το 95% των ασθενών με σοβαρότητα 1 ενώ στο group 1 περιέχεται το 79% των ασθενών με σοβαρότητα 3. Επιπλέον η σοβαρότητα 2 βρίσκεται κατά μεγαλύτερο βαθμό στο group 0(72%). Πρόκειται για μια αρκετά καλή ομαδοποίηση. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης



**Σχήμα 5.14** : Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο kmeans με είσοδο: sharp-score,CCP,RF,HLA-DRB1,DURATION



**Σχήμα 5.15** : Τρισδιάστατη απεικόνιση των 2 συστάδων . Μοντέλο kmeans με είσοδο: sharp-score,CCP,RF,HLA-DRB1,DURATION



**Σχήμα 5.16 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο kmeans με είσοδο: sharp-score,CCP,RF,HLA-DRB1,DURATION

**5.1.1.6 Είσοδος 6 KMeans:** `teachArr = workingDF[["sharp-score","CCP","RF","HLA-DRB1"]].to_numpy()`

Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.95 of all sev=1	0.05 of all sev=1
0.72 of all sev=2	0.28 of all sev=2
0.21 of all sev=3	0.79 of all sev=3

Όπως φαίνεται στο group 0 περιέχεται το 95% των ασθενών με σοβαρότητα 1 ενώ στο group 1 περιέχεται το 79% των ασθενών με σοβαρότητα 3. Επιπλέον η σοβαρότητα 2 βρίσκεται κατά μεγαλύτερο βαθμό στο group 0(72%). Πρόκειται για μια αρκετά καλή ομαδοποίηση που δίνει ακριβώς τα ίδια αποτελέσματα με την είσοδο 5. Επομένως συμπεραίνουμε ότι η διάρκεια της νόσου δεν επηρεάζει την ομαδοποίηση με το Kmeans. Οι γραφικές απεικονίσεις της ομαδοποίησης είναι ίδιες με της εισόδου 5 οπότε παραλείπονται.

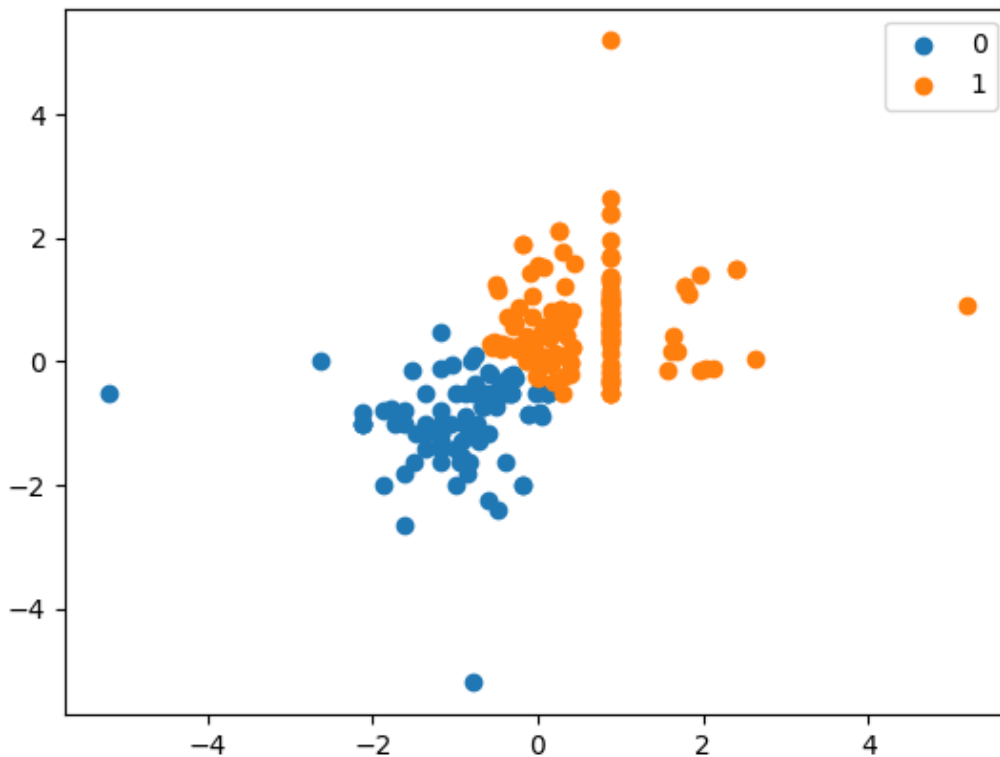
**5.1.1.7 Είσοδος 7 KMeans:** `teachArr = workingDF[["CCP","RF"]].to_numpy()`

Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

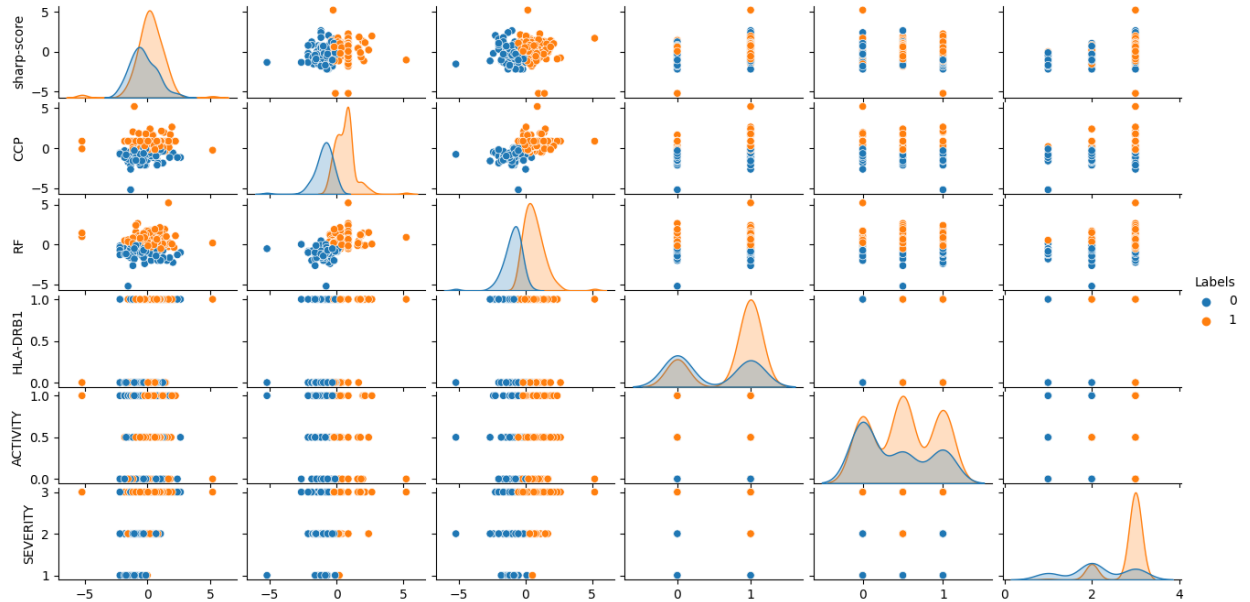


group 0	group 1
0.95 of all sev=1	0.05 of all sev=1
0.68 of all sev=2	0.32 of all sev=2
0.20 of all sev=3	0.80 of all sev=3

Όπως φαίνεται στο group 0 περιέχεται το 95% των ασθενών με σοβαρότητα 1 ενώ στο group 1 περιέχεται το 80% των ασθενών με σοβαρότητα 1. Επιπλέον η σοβαρότητα 2 βρίσκεται κατά μεγαλύτερο βαθμό στο group 0(68%). Πρόκειται για μια αρκετά καλή ομαδοποίηση, έχει πολύ παρόμοια αποτελέσματα με την είσοδο 3 με τις εξής εισόδους [{"CCP", "RF", "HLA-DRB1"}] όπου η είσοδος 3 έχει ελαφρώς καλύτερα αποτελέσματα από την είσοδο 7. Επομένως ο παράγοντας "HLA-DRB1" βοηθάει στην ομαδοποίηση αλλά δεν παίζει ιδιαίτερα σημαντικό ρόλο. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης



**Σχήμα 5.17** : Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο kmeans με είσοδο: sharp-score.CCP.RF



**Σχήμα 5.18 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο kmeans με είσοδο: sharp-score,CCP,RF

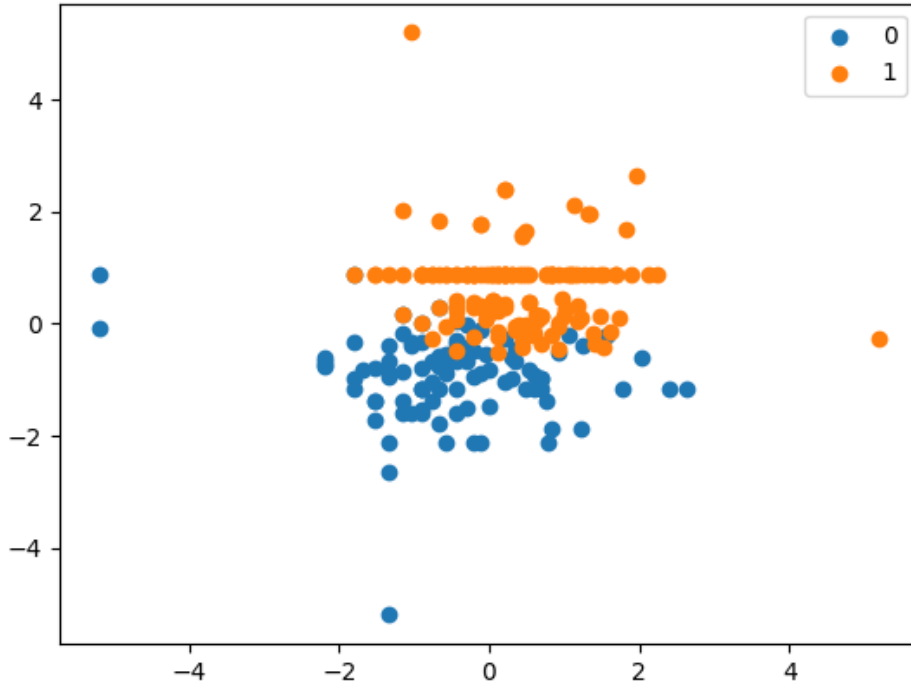
Επειδή έχουμε σαν εισόδους μόνο 2 παράγοντες δεν έχει νόημα η τρισδιάστατη απεικόνιση οπότε και παραλείπεται.

#### 5.1.1.8 **Είσοδος 8 KMeans:** `teachArr = workingDF[["sharp-score", "CCP", "RF"]].to_numpy`

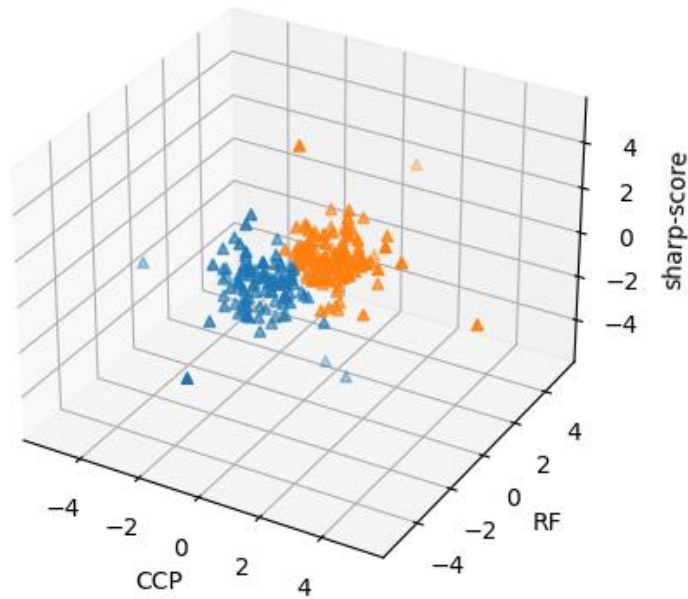
Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.95 of all sev=1	0.05 of all sev=1
0.72 of all sev=2	0.28 of all sev=2
0.21 of all sev=3	0.79 of all sev=3

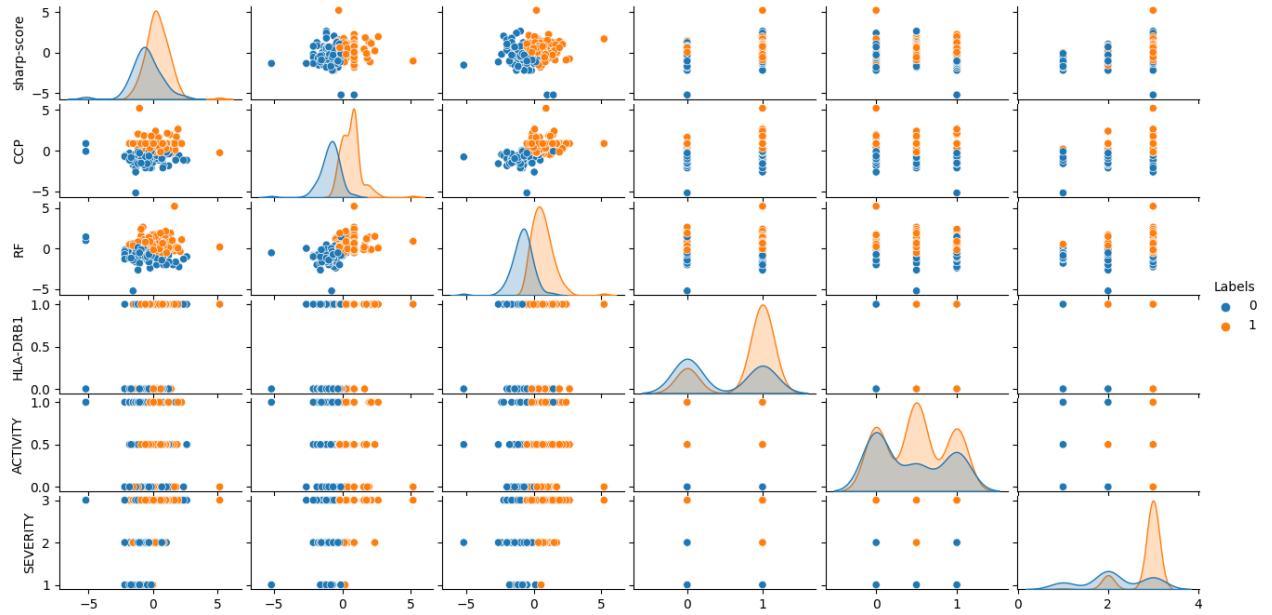
Όπως φαίνεται στο group 0 περιέχεται το 95% των ασθενών με σοβαρότητα 1 ενώ στο group 1 περιέχεται το 79% των ασθενών με σοβαρότητα 3. Επιπλέον η σοβαρότητα 2 βρίσκεται κατά μεγαλύτερο βαθμό στο group 0(72%). Πρόκειται για μια αρκετά καλή ομαδοποίηση, έχει πολύ παρόμοια αποτελέσματα με την είσοδο 7 με την διαφορά ότι εμφανίζει ελαφρώς καλύτερη ομαδοποίηση για την σοβαρότητα 2 και ελαφρώς χειρότερη για την σοβαρότητα 3. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης.



**Σχήμα 5.19** : Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο kmeans με είσοδο: sharp-score,CCP,RF



**Σχήμα 5.20** : Τρισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο kmeans με είσοδο: sharp-score,CCP,RF



**Σχήμα 5.21:** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο kmeans με είσοδο: sharp-score,CCP,RF

Τέλος όσον αφορά το μοντέλο kmeans ακολουθεί ένας συγκεντρωτικός πίνακας με τις εισόδους που παρουσιάστηκαν παραπάνω και την ομαδοποίηση που προέκυψε από το μοντέλο.

ΕΙΣΟΔΟΣ	ΟΜΑΔΑ 1	ΟΜΑΔΑ 2
1. [ ["sharp-score", "CCPi", "RF", "HLA-DRB1" ] ]	0.05 of all sev=1 0.25 of all sev=2 0.77 of all sev=3	0.95 of all sev=1 0.75 of all sev=2 0.23 of all sev=3
2. [ ["sharp-score", "DURATION", "HLA-DRB1" ] ]	0.05 of all sev=1 0.32 of all sev=2 0.64 of all sev=3	0.95 of all sev=1 0.68 of all sev=2 0.36 of all sev=3
3. [ ["CCP", "RF", "HLA-DRB1" ] ]	0.95 of all sev=1 0.68 of all sev=2 0.19 of all sev=3	0.05 of all sev=1 0.32 of all sev=2 0.81 of all sev=3

4. [ ["sharp-score", "CCP", "CCPi", "RF", "Rfi", "HLA-DRB1" ] ]	0.05 of all sev=1 0.29 of all sev=2 0.80 of all sev=3	0.95 of all sev=1 0.71 of all sev=2 0.20 of all sev=3
5. [ ["sharp-score", "CCP", "RF", "HLA-DRB1", "DURATION" ] ]	0.95 of all sev=1 0.72 of all sev=2 0.21 of all sev=3	0.05 of all sev=1 0.28 of all sev=2 0.79 of all sev=3
6. [ ["sharp-score", "CCP", "RF", "HLA-DRB1" ] ]	0.95 of all sev=1 0.72 of all sev=2 0.21 of all sev=3	0.05 of all sev=1 0.28 of all sev=2 0.79 of all sev=3
7. [ ["CCP", "RF" ] ]	0.95 of all sev=1 0.68 of all sev=2 0.20 of all sev=3	0.05 of all sev=1 0.32 of all sev=2 0.80 of all sev=3
8. [ ["sharp-score", "CCP", "RF" ] ]	0.95 of all sev=1 0.72 of all sev=2 0.21 of all sev=3	0.05 of all sev=1 0.28 of all sev=2 0.79 of all sev=3

**Πίνακας 5.1:** Αποτελέσματα ομαδοποίησης KMeans με βάση τον πίνακα SEVERITY

Όπως φαίνεται από τα συγκεντρωτικά αποτελέσματα οι καλύτερες ομαδοποιήσεις με το KMeans, έχοντας σαν πίνακα επαλήθευσης SEVERITY με τρεις τιμές για την σοβαρότητα της ρευματοειδούς ,είναι με τις παρακάτω εισόδους:

- [ ["sharp-score", "CCP", "CCPi", "RF", "Rfi", "HLA-DRB1" ] ] (ΕΙΣΟΔΟΣ 4)
  - Πέτυχε 95% ομαδοποίηση για την σοβαρότητα 1 , 71% ομαδοποίηση για την σοβαρότητα 2 και 80% ομαδοποίηση για την σοβαρότητα 3.
- [ ["CCP", "RF" ] ] (ΕΙΣΟΔΟΣ 7)
  - Πέτυχε 95% ομαδοποίηση για την σοβαρότητα 1 , 68% ομαδοποίηση για την σοβαρότητα 2 και 80% ομαδοποίηση για την σοβαρότητα 3.

Όλες οι εισοδοί όμως εμφανίζουν αντίστοιχα επίπεδα ομαδοποίησης οπότε δεν μπορεί να διακριθεί ξεκάθαρα η καλύτερη.

### 5.1.2 Περίπτωση 2 KMeans: SEVERITY\_binary

Όπως αναφέρθηκε στο Κεφ.4 στην πορεία της έρευνας δημιουργήθηκε ένας επιπλέον πίνακας σοβαρότητας με δυαδικές τιμές ώστε να ελεγχθεί αν τα μοντέλα παρουσιάζουν καλύτερη ομαδοποίηση και κατηγοριοποίηση με βάση δυαδικές τιμές σοβαρότητας.

Στον πίνακα που ακολουθεί παρουσιάζεται η ομαδοποίηση που προέκυψε με το μοντέλο KMeans έχοντας σαν πίνακα ελέγχου τον SEVERITY\_binary. Ο πίνακας SEVERITY\_binary περιέχει 2 τιμές για την σοβαρότητα:

1. Ελαφριά Ρευματοειδής
2. Βαριά Ρευματοειδής

Όπως φαίνεται παρακάτω προκύπτει ακόμα καλύτερη ομαδοποίηση καθώς εξαλείφεται η σοβαρότητα τύπου 2 η οποία στα παραπάνω μοντέλα βρισκόταν είτε εξ ημισείας είτε κατά κάποιο ποσοστό περίπου 60-70% σε μια εκ των δύο ομάδων. Για όλες τις εισόδους που δοκιμάστηκαν και οι δύο ομάδες που δημιουργήθηκαν περιέχουν το 65% και πάνω της σοβαρότητας 1 ή 2 αντίστοιχα.

Στα συγκεκριμένα μοντέλα και πάλι έχουμε καλύτερη ομαδοποίηση για την σοβαρότητα 1 η οποία βρίσκεται στο 86%-91% στις αντίστοιχες ομάδες ενώ η σοβαρότητα τύπου 2 βρίσκεται κατά 65%-79% στις αντίστοιχες ομάδες.

Η καλύτερη ομαδοποίηση όσον αφορά την σοβαρότητα τύπου 1 προκύπτει για την είσοδο 1: `[["sharp-score","CCP1","RF","HLA-DRB1"]]` όπου το 91% των ασθενών με σοβαρότητα 1 ομαδοποιούνται στο group 1. Αντίστοιχα η καλύτερη ομαδοποίηση όσον αφορά την σοβαρότητα τύπου 2 προκύπτει για την είσοδο 3: `[["CCP","RF","HLA-DRB1"]]` όπου το 79% των ασθενών με σοβαρότητα 2 ομαδοποιούνται στο group 1.

Τέλος για λόγους συντομίας παραλείπονται οι γραφικές απεικονίσεις της ομαδοποίησης για τις 8 εισόδους και ακολουθεί ένας συγκεντρωτικός πίνακας με τα αποτελέσματα που προέκυψαν από κάθε είσοδο.

INPUT	GROUP 0	GROUP 1
1. [["sharp-score","CCPi","RF","HLA-DRB1"]]	0.09 of all sev=1	<b>0.91 of all sev=1</b>
	<b>0.74 of all sev=2</b>	0.26 of all sev=2
2. [["sharp-score","DURATION","HLA-DRB1"]]	0.12 of all sev=1	<b>0.88 of all sev=1</b>
	<b>0.65 of all sev=2</b>	0.35 of all sev=2
3. [["CCP","RF","HLA-DRB1"]]	<b>0.86 of all sev=1</b>	0.14 of all sev=1
	0.21 of all sev=2	<b>0.79 of all sev=2</b>
4. [["sharp-score","CCP","CCPi","RF","Rfi","HLA-DRB1"]]	0.12 of all sev=1	<b>0.88 of all sev=1</b>
	<b>0.78 of all sev=2</b>	0.22 of all sev=2
5. [["sharp-score","CCP","RF","HLA-DRB1","DURATION"]]	<b>0.89 of all sev=1</b>	0.11 of all sev=1
	0.23 of all sev=2	<b>0.77 of all sev=2</b>
6. [["sharp-score","CCP","RF","HLA-DRB1"]]	<b>0.89 of all sev=1</b>	0.11 of all sev=1
	0.23 of all sev=2	<b>0.77 of all sev=2</b>
7. [["CCP","RF"]]	<b>0.86 of all sev=1</b>	0.14 of all sev=1
	0.22 of all sev=2	<b>0.78 of all sev=2</b>
8. [["sharp-score","CCP","RF"]]	<b>0.89 of all sev=1</b>	0.11 of all sev=1
	0.23 of all sev=2	<b>0.77 of all sev=2</b>

**Πίνακας 5.2:** Αποτελέσματα ομαδοποίησης KMeans με βάση τον πίνακα SEVERITY\_binary

## 5.2 Μοντέλο 2: GaussianMixture

Επειδή αναπτύχθηκε πληθώρα μοντέλων με πολλαπλούς συνδυασμούς εισόδων παρακάτω θα παρουσιαστούν τα αποτελέσματα για τις εισόδους που παρουσίασαν την καλύτερη ομαδοποίηση με βάση το πρώτο μοντέλο kMeans ώστε να μπορεί να υπάρξει σύγκριση των αποτελεσμάτων και των μεθόδων

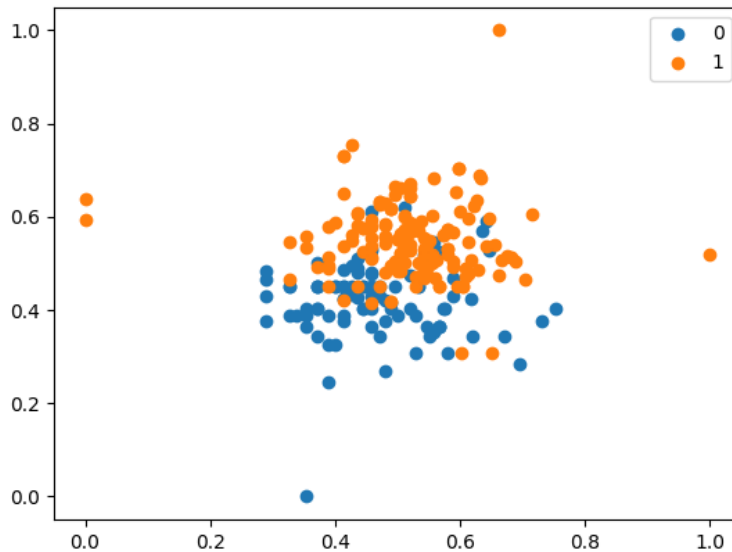
### 5.2.1 Περίπτωση 1 GaussianMixture: SEVERITY

5.2.1.1 Είσοδος 1 GaussianMixture : `teachArr = workingDF[["sharp-score","RF","HLA-DRB1","CCPi"]].to_numpy()`

Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

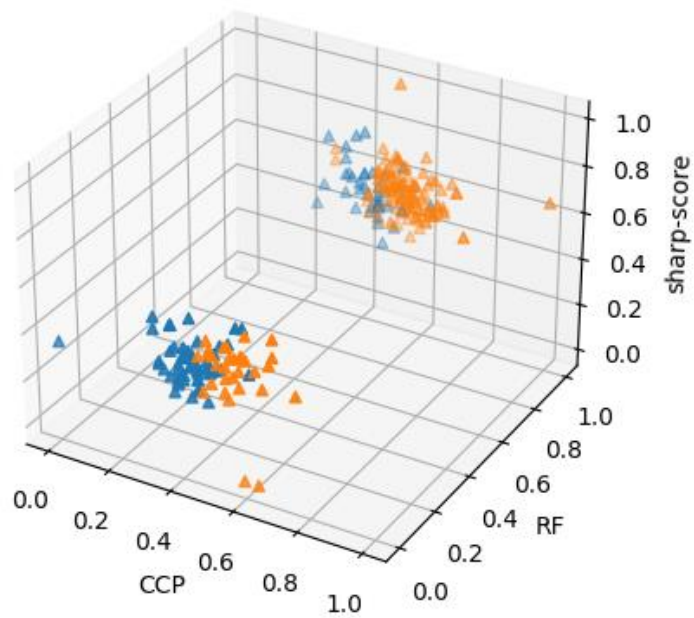
group 0	group 1
0.89 of all sev=1	0.11 of all sev=1
0.71 of all sev=2	0.29 of all sev=2
0.19 of all sev=3	0.81 of all sev=3

Όπως φαίνεται στο group 0 περιέχεται το 89% των ασθενών με σοβαρότητα 1 ενώ στο group 1 περιέχεται το 81% των ασθενών με σοβαρότητα 3. Επιπλέον η σοβαρότητα 2 βρίσκεται κατά μεγαλύτερο βαθμό στο group 0 (71%). Πρόκειται για μια αρκετά καλή ομαδοποίηση. Σε σχέση με το Kmeans εμφανίζει καλύτερη ομαδοποίηση για την σοβαρότητα 3 ενώ το kmeans εμφανίζει καλύτερη ομαδοποίηση για τη σοβαρότητα 3. Και τα 2 μοντέλα ομαδοποιούν παρόμοια την σοβαρότητα 2 στην ομάδα που περιέχει το μεγαλύτερο ποσοστό της σοβαρότητας 1. Συγκριτικά το Kmeans εμφανίζει καλύτερη ομαδοποίηση. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης.

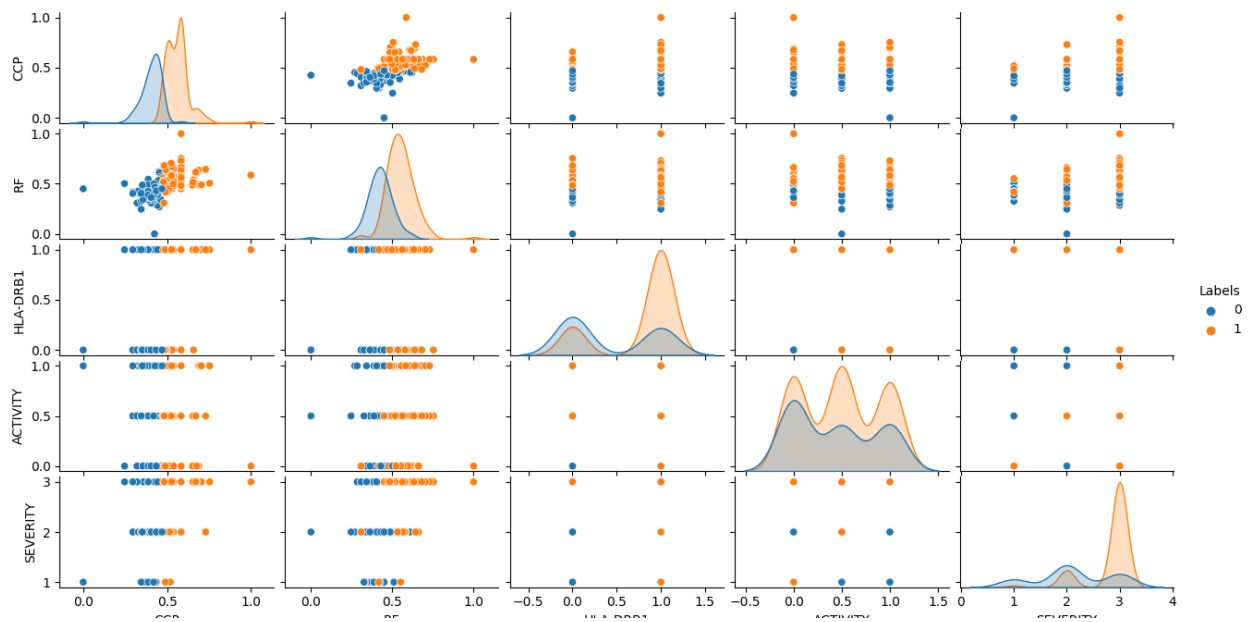


**Σχήμα 5.22** : Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο GaussianMixture με είσοδο: sharp-score,RF,HLA-DRB1,CCPi





**Σχήμα 5.23 :** Τρισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο GaussianMixture με είσοδο: sharp-score,RF,HLA-DRB1,CCPi



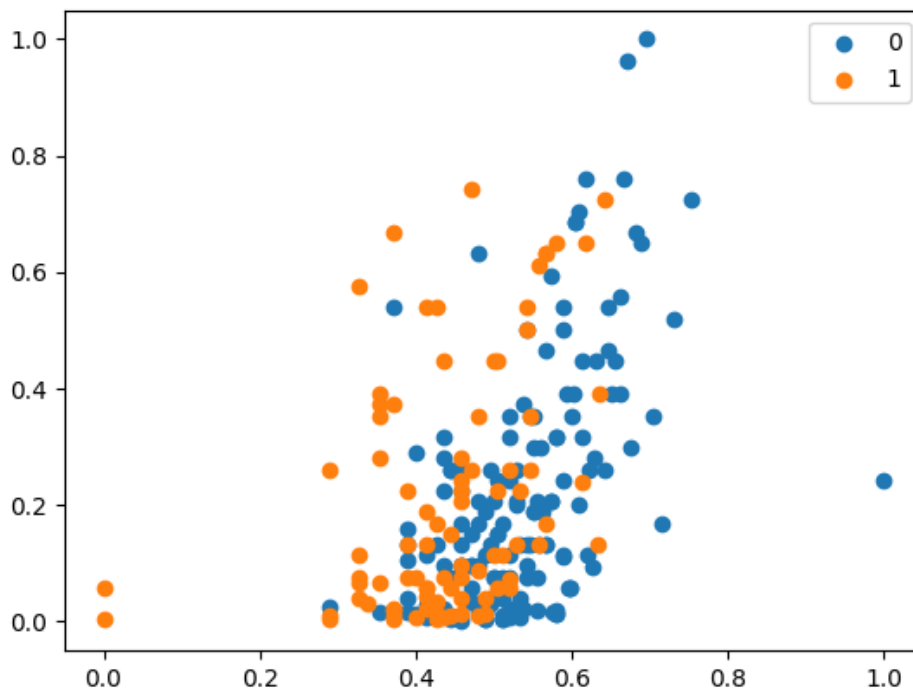
**Σχήμα 5.24 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο GaussianMixture με είσοδο: sharp-score,RF,HLA-DRB1,CCPi

5.2.1.2 Είσοδος 2 GaussianMixture : `teachArr = workingDF[["sharp-score","DURATION","HLA-DRB1"]].to_numpy()`

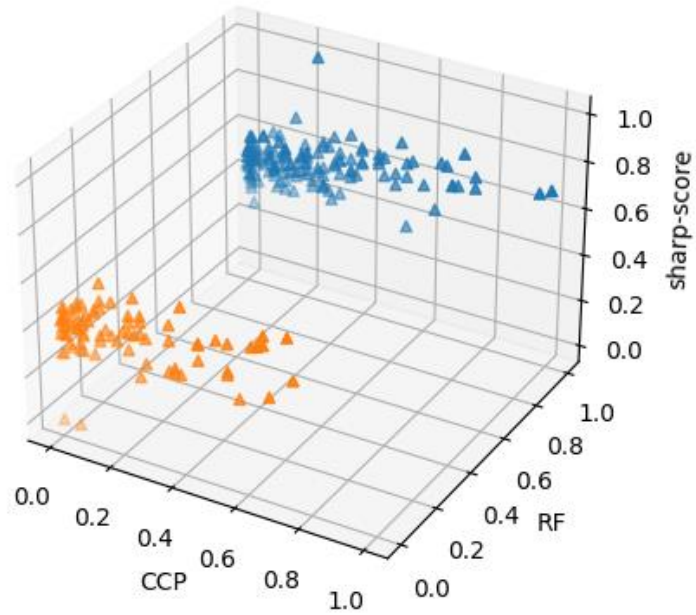
Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.16 of all sev=1	0.84 of all sev=1
0.41 of all sev=2	0.59 of all sev=2
0.80 of all sev=3	0.20 of all sev=3

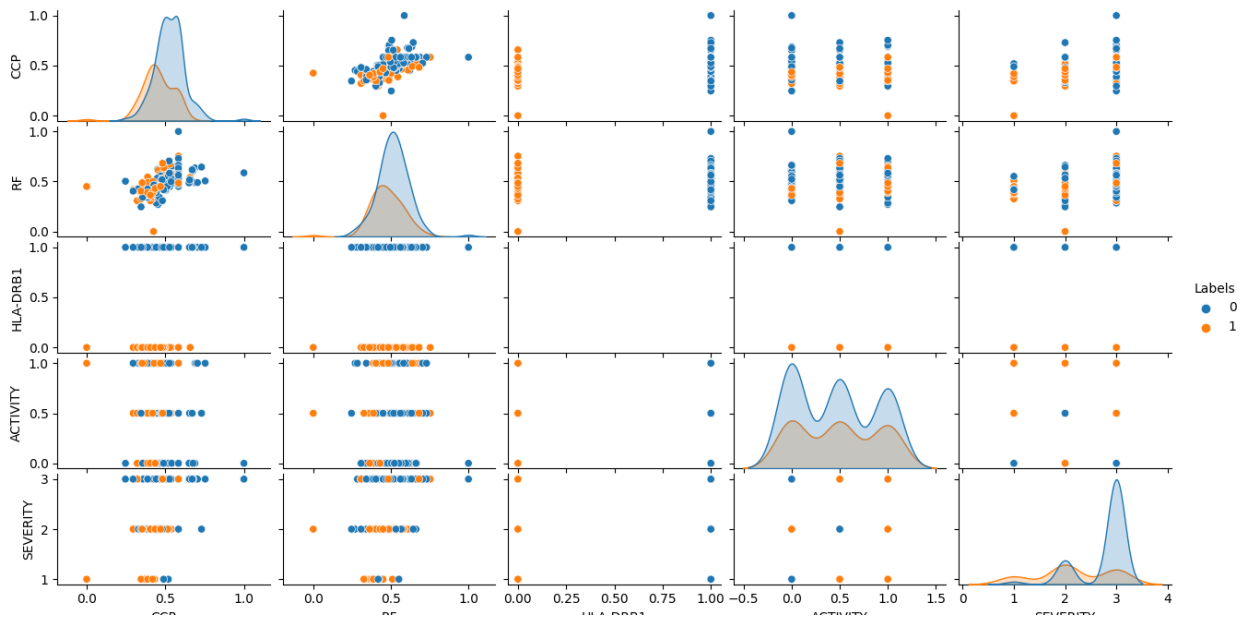
Όπως φαίνεται στο group 0 περιέχεται το 80% των ασθενών με σοβαρότητα 3 ενώ στο group 1 περιέχεται το 84% των ασθενών με σοβαρότητα 1. Επιπλέον η σοβαρότητα 2 βρίσκεται εξίσου και στις 2 ομάδες .Πρόκειται για μια αρκετά καλή ομαδοποίηση. Σε σχέση με το Kmeans εμφανίζει σαφώς καλύτερη ομαδοποίηση για την σοβαρότητα 3 ενώ το kmeans εμφανίζει καλύτερη ομαδοποίηση τόσο για τη σοβαρότητα 1 όσο και για τη σοβαρότητα 2. Συγκριτικά το Kmeans εμφανίζει καλύτερη ομαδοποίηση. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης.



Σχήμα 5.25 : Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο GaussianMixture με είσοδο: `sharp-score,DURATION,HLA-DRB1`



**Σχήμα 5.26 :** Τρισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο GaussianMixture με είσοδο: sharp-score,DURATION,HLA-DRB1



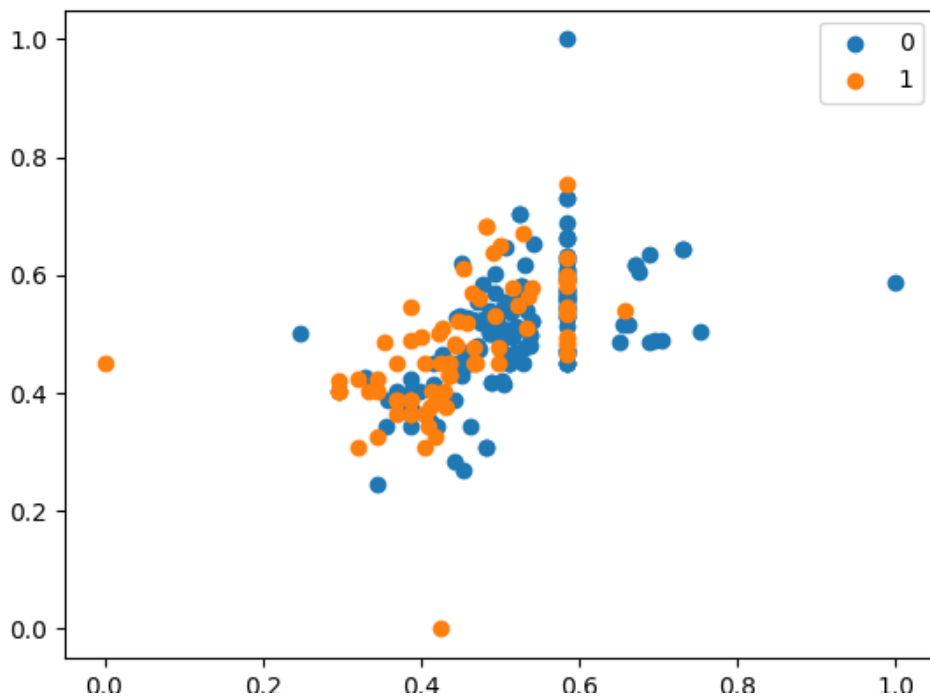
**Σχήμα 5.27 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο GaussianMixture με είσοδο: sharp-score,DURATION,HLA-DRB1

5.2.1.3 Είσοδος 3 GaussianMixture: `teachArr = workingDF[["CCP", "RF", "HLA-DRB1"]].to_numpy()`

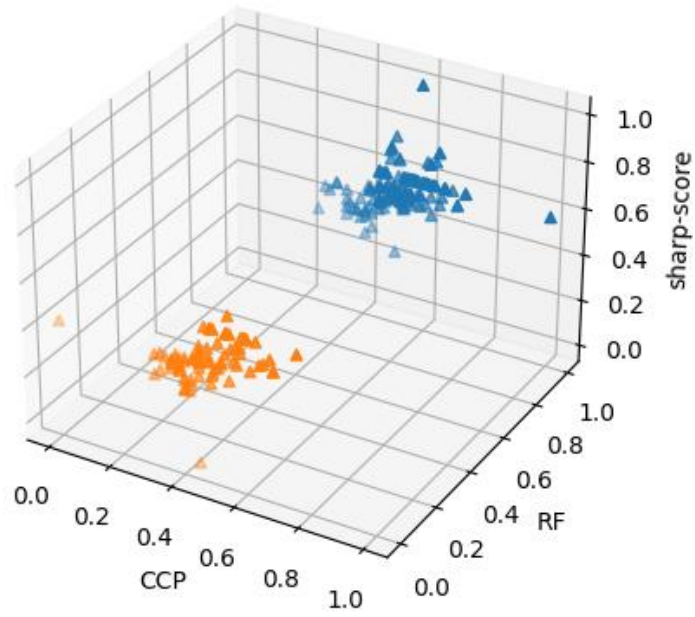
Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.16 of all sev=1	0.84 of all sev=1
0.41 of all sev=2	0.59 of all sev=2
0.80 of all sev=3	0.20 of all sev=3

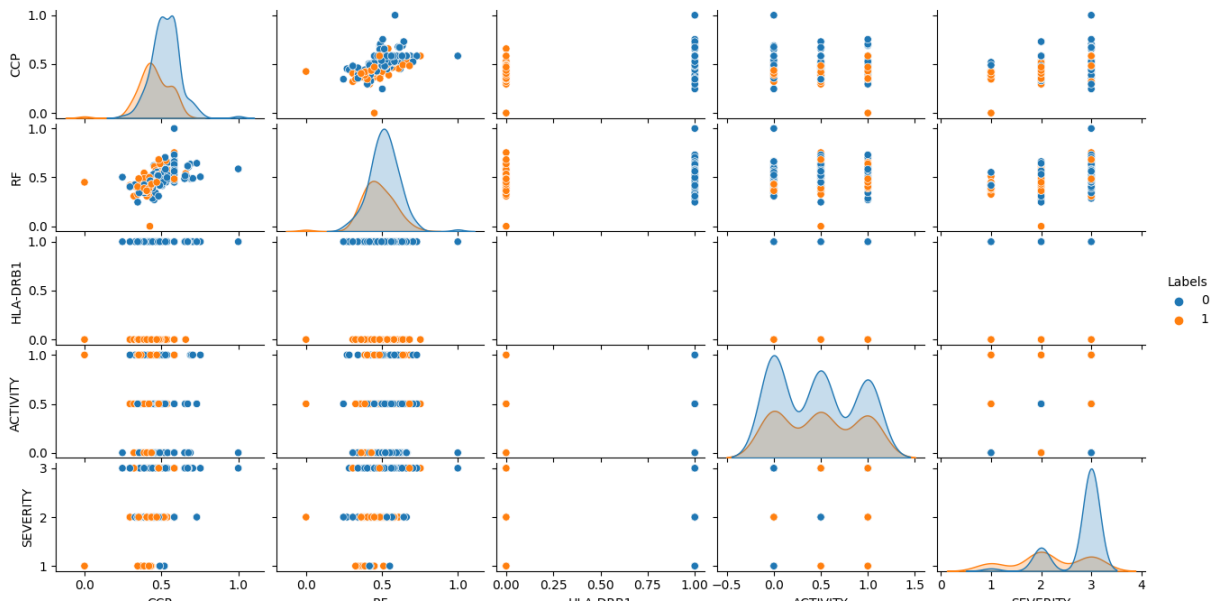
Όπως φαίνεται στο group 0 περιέχεται το 80% των ασθενών με σοβαρότητα 3 ενώ στο group 1 περιέχεται το 84% των ασθενών με σοβαρότητα 1. Επιπλέον η σοβαρότητα 2 βρίσκεται εξίσου και στις 2 ομάδες. Παρατηρείται η ίδια ακριβώς ομαδοποίηση με την είσοδο 2. Συγκριτικά το Kmeans εμφανίζει σαφώς καλύτερη ομαδοποίηση. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης.



Σχήμα 5.28 : Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο GaussianMixture με είσοδο: CCP,RF,HLA-DRB1



**Σχήμα 5.29 :** Τρισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο GaussianMixture με είσοδο: CCP,RF,HLA-DRB1



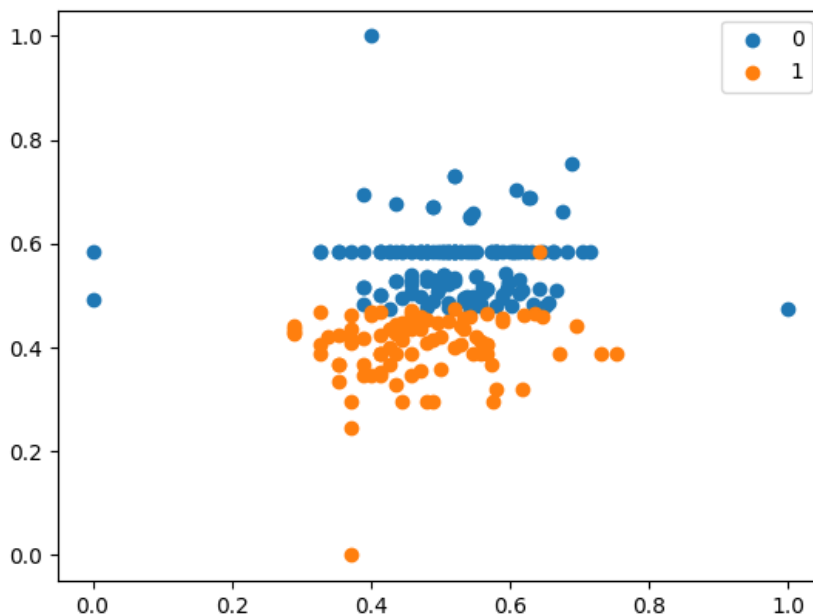
**Σχήμα 5.30 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο GaussianMixture με είσοδο: CCP,RF,HLA-DRB1

5.2.1.4 ***Είσοδος 4 GaussianMixture:*** `teachArr = workingDF[["sharp-score","CCP","CCPi","RF","Rfi","HLA-DRB1"]].to_numpy()`

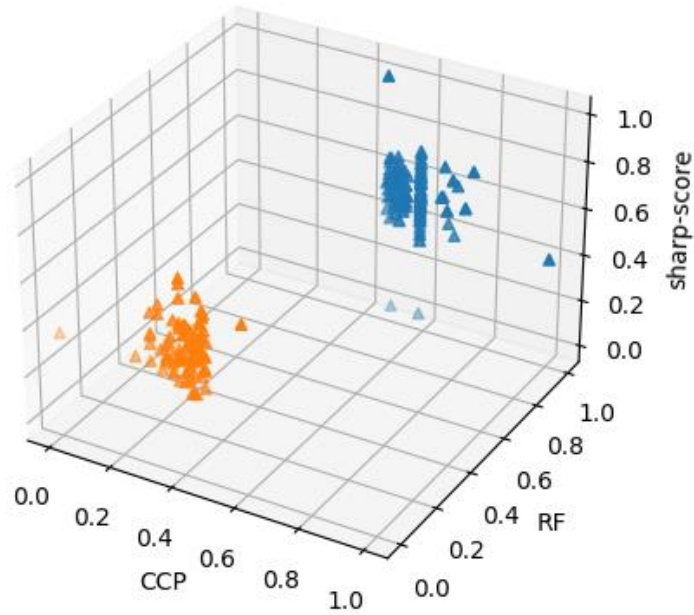
Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.11 of all sev=1	0.89 of all sev=1
0.29 of all sev=2	0.71 of all sev=2
0.81 of all sev=3	0.19 of all sev=3

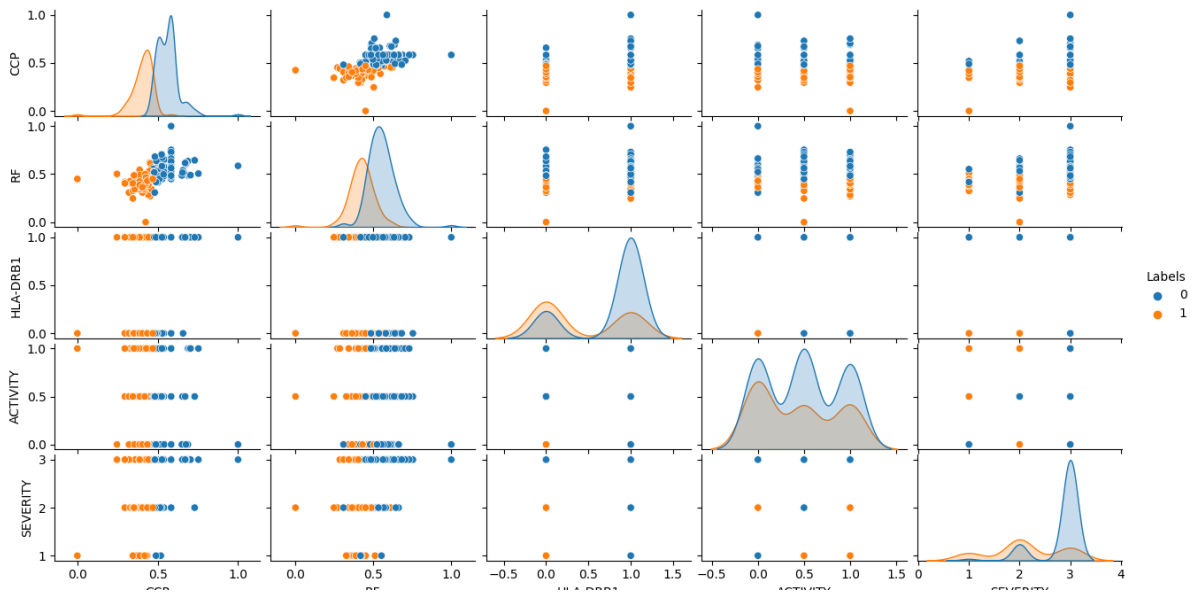
Όπως φαίνεται στο group 0 περιέχεται το 81% των ασθενών με σοβαρότητα 3 ενώ στο group 1 περιέχεται το 89% των ασθενών με σοβαρότητα 1. Επιπλέον η σοβαρότητα 2 βρίσκεται κατά μεγαλύτερο βαθμό στο group 1 (71%). Πρόκειται για μια αρκετά καλή ομαδοποίηση. Πρόκειται για το καλύτερο μοντέλο που προέκυψε με Gaussian Mixture. Σε σχέση με το Kmeans εμφανίζει την ίδια ομαδοποίηση για την σοβαρότητα 3 (περίπου 80%) ενώ το kmeans εμφανίζει καλύτερη ομαδοποίηση για τη σοβαρότητα 1 ενώ παρουσιάζουν την ίδια ομαδοποίηση για τη σοβαρότητα 2 η οποία κατά 71% βρίσκεται στην ομάδα σοβαρότητας 1. Συγκριτικά τα δυο μοντέλα παρουσιάζουν παρόμοια ομαδοποίηση. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης.



**Σχήμα 5.31 :** Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο GaussianMixture με είσοδο: sharp-score,CCP,CCPi,RF,Rfi,HLA-DRB1



**Σχήμα 5.32 :** Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο GaussianMixture με είσοδο: sharp-score,CCP,CCPi,RF,Rfi,HLA-DRB1



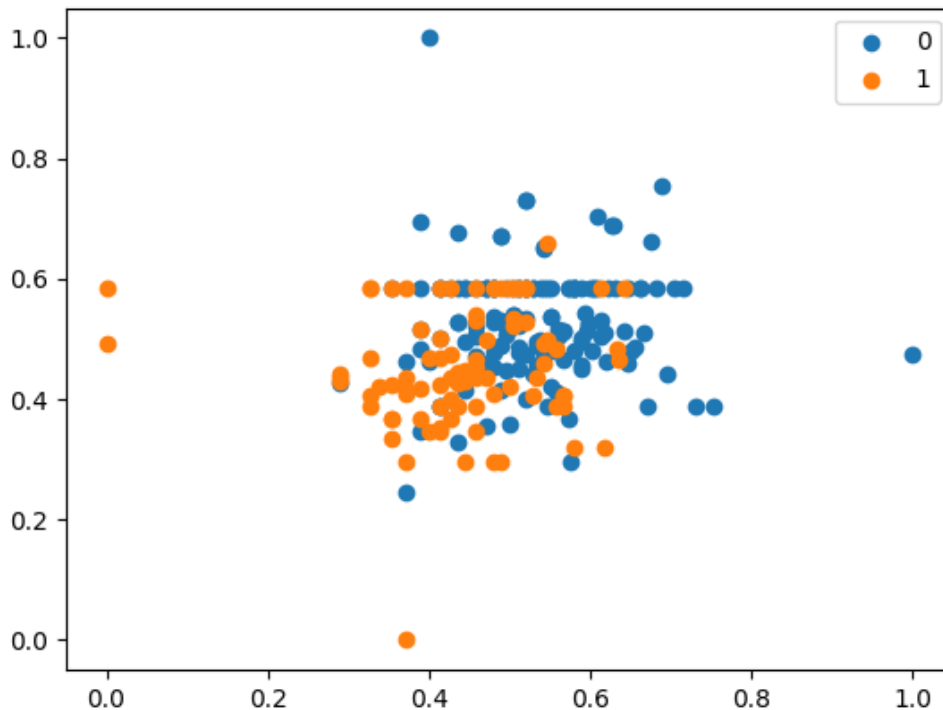
**Σχήμα 5.33 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο GaussianMixture με είσοδο: sharp-score,CCP,CCPi,RF,Rfi,HLA-DRB1

5.2.1.5 Είσοδος 5 GaussianMixture: `teachArr = workingDF[["sharp-score","CCP","RF","HLA-DRB1","DURATION"]].to_numpy()`

Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

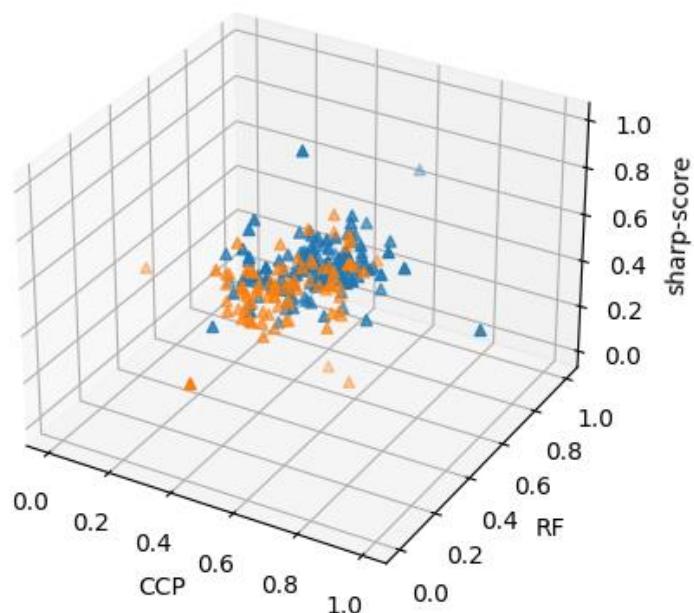
group 0	group 1
0.16 of all sev=1	0.84 of all sev=1
0.41 of all sev=2	0.59 of all sev=2
0.80 of all sev=3	0.20 of all sev=3

Όπως φαίνεται στο group 0 περιέχεται το 80% των ασθενών με σοβαρότητα 3 ενώ στο group 1 περιέχεται το 84% των ασθενών με σοβαρότητα 1. Επιπλέον η σοβαρότητα 2 βρίσκεται εξίσου και στις 2 ομάδες. Παρατηρείται η ίδια ακριβώς ομαδοποίηση με την είσοδο 2 και 3. Κοινός παραγοντας των τριών εισόδων είναι η παράμετρος HLA\_DRB1, επομένως παίζει πολύ μεγάλο ρόλο στην ομαδοποίηση με GaussianMixture και κατευθύνει σχεδόν εξολοκλήρου τα μοντέλα. Συγκριτικά το Kmeans με την ίδια είσοδο εμφανίζει σαφώς καλύτερη ομαδοποίηση. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης.

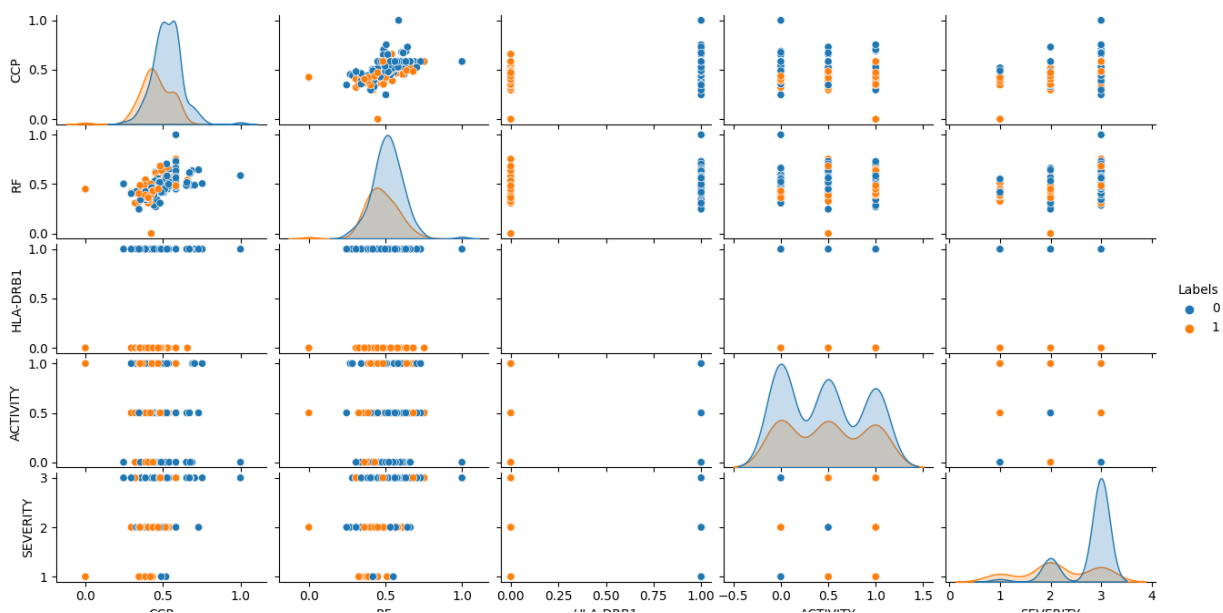


**Σχήμα 5.34** : Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο GaussianMixture με είσοδο: sharp-score,CCP,RF,HLA-DRB1,DURATION





**Σχήμα 5.35 :** Τρισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο GaussianMixture με είσοδο: sharp-score,CCP,RF,HLA-DRB1,DURATION



**Σχήμα 5.36 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο GaussianMixture με είσοδο: sharp-score,CCP,RF,HLA-DRB1,DURATION

5.2.1.6 Είσοδος 6 GaussianMixture: `teachArr = workingDF[["sharp-score", "CCP", "RF", "HLA-DRB1"]].to_numpy()`

Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.16 of all sev=1	0.84 of all sev=1
0.41 of all sev=2	0.59 of all sev=2
0.80 of all sev=3	0.20 of all sev=3

Όπως φαίνεται στο group 0 περιέχεται το 80% των ασθενών με σοβαρότητα 3 ενώ στο group 1 περιέχεται το 84% των ασθενών με σοβαρότητα 1. Επιπλέον η σοβαρότητα 2 βρίσκεται εξίσου και στις 2 ομάδες. Παρατηρείται η ίδια ακριβώς ομαδοποίηση με την είσοδο 2, 3 και 5. Κοινός παράγοντας των τριών εισόδων είναι η παράμετρος HLA\_DRB1, επομένως παίζει πολύ μεγάλο ρόλο στην ομαδοποίηση με GaussianMixture και κατευθύνει σχεδόν εξολοκλήρου τα μοντέλα.

Συγκριτικά το Kmeans με την ίδια είσοδο εμφανίζει σαφώς καλύτερη ομαδοποίηση. Οι γραφικές απεικονίσεις της ομαδοποίησης είναι ίδιες με της εισόδου 5 οπότε παραλείπονται.

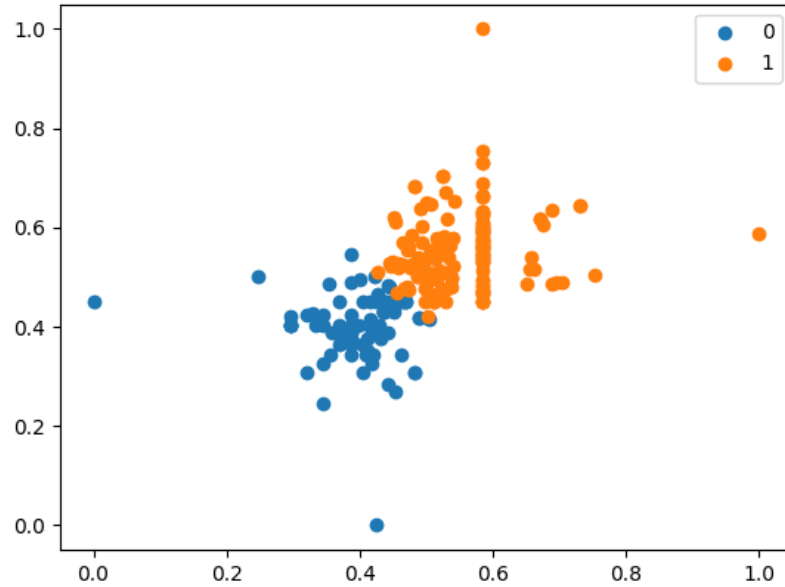
5.2.1.7 Είσοδος 7 GaussianMixture: `teachArr = workingDF[["CCP", "RF"]].to_numpy()`

Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

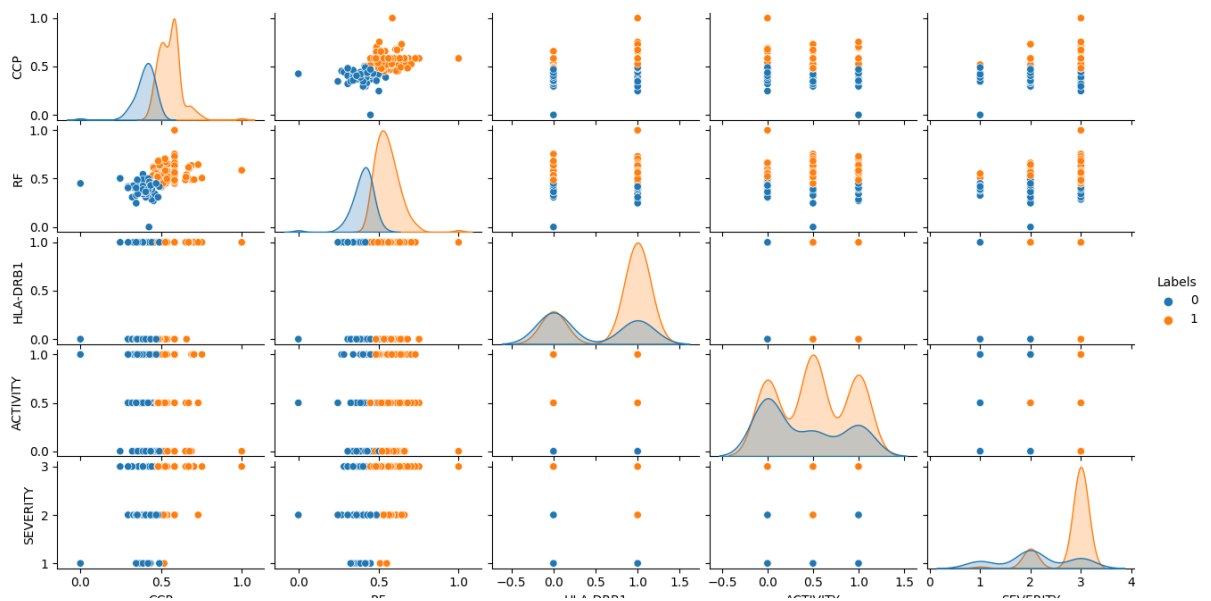
group 0	group 1
0.89 of all sev=1	0.11 of all sev=1
0.63 of all sev=2	0.37 of all sev=2
0.16 of all sev=3	0.84 of all sev=3

Όπως φαίνεται στο group 0 περιέχεται το 89% των ασθενών με σοβαρότητα 1 ενώ στο group 1 περιέχεται το 84% των ασθενών με σοβαρότητα 3. Επιπλέον η σοβαρότητα 2 βρίσκεται κατά μεγαλύτερο βαθμό στο group 0 (63%). Πρόκειται για μια αρκετά καλή ομαδοποίηση, έχει πολύ παρόμοια αποτελέσματα με την είσοδο 4 με τις εξής εισόδους [["sharp-score", "CCP", "CCPi", "RF", "Rfi", "HLA-DRB1"]] όπου η είσοδος 4 εμφανίζει ελαφρώς καλύτερη ομαδοποίηση στη σοβαρότητα 2 και ελαφρώς χειρότερη στην σοβαρότητα 3. Όσον αφορά την αντίστοιχη είσοδο στο Kmeans και πάλι εμφανίζει συνολικά καλύτερη ομαδοποίηση αλλά κινούνται σε αντίστοιχα επίπεδα. Ακολουθούν γραφικές

απεικονίσεις της ομαδοποίησης, επειδή έχουμε σαν εισόδους μόνο 2 παράγοντες δεν έχει νόημα η τρισδιάστατη απεικόνιση οπότε και παραλείπεται.



**Σχήμα 5.37 :** Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο GaussianMixture με είσοδο: CCP,RF



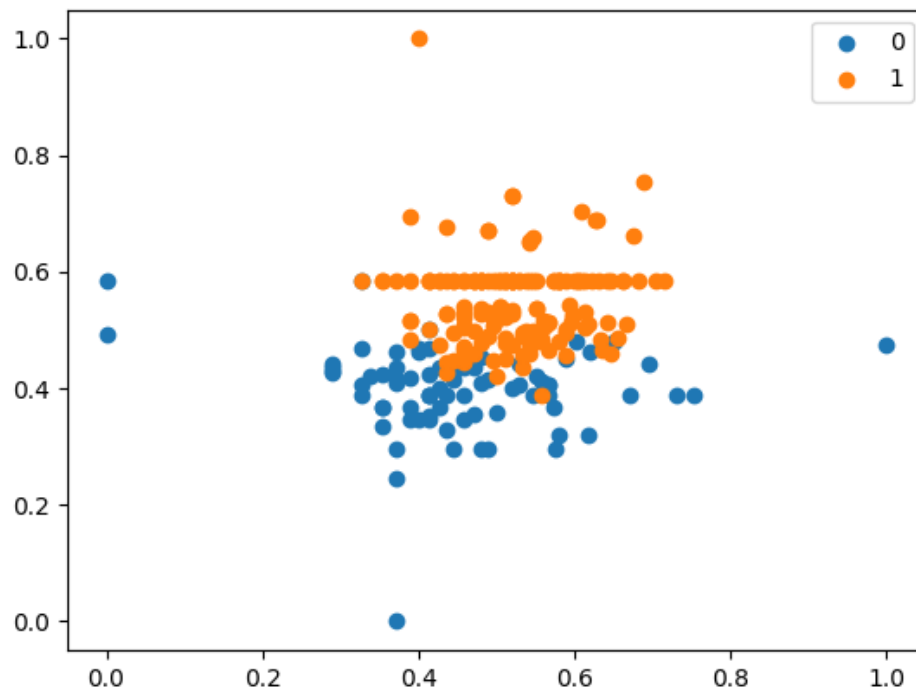
**Σχήμα 5.38 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο GaussianMixture με είσοδο: CCP,RF

5.2.1.8 **Είσοδος 8 GaussianMixture:** `teachArr = workingDF[["sharp-score", "CCP", "RF"]].to_numpy`

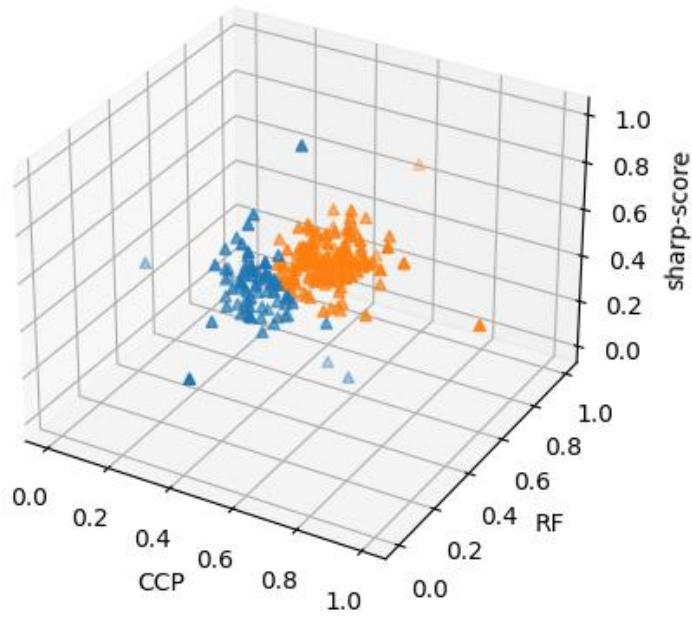
Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.84 of all sev=1	0.16 of all sev=1
0.57 of all sev=2	0.43 of all sev=2
0.15 of all sev=3	0.85 of all sev=3

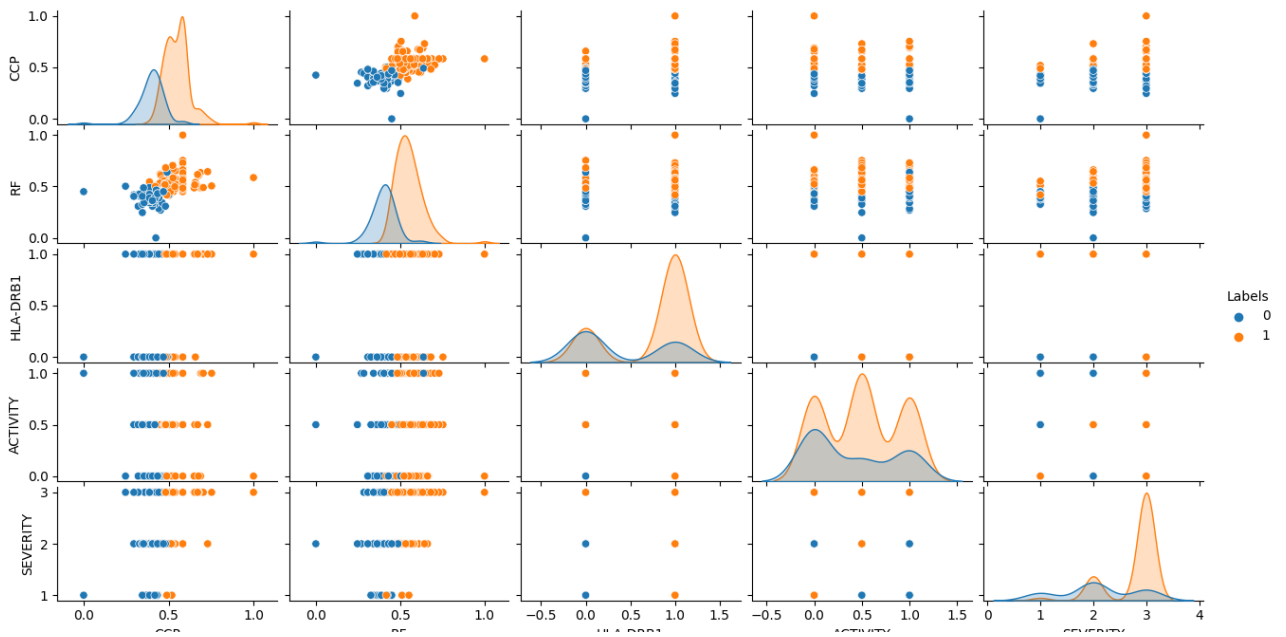
Όπως φαίνεται στο group 0 περιέχεται το 84% των ασθενών με σοβαρότητα 1 ενώ στο group 1 περιέχεται το 85% των ασθενών με σοβαρότητα 3. Επιπλέον η σοβαρότητα 2 βρίσκεται εξίσου διαχωρισμένη στις δυο ομάδες. Πρόκειται για μια πολύ καλή ομαδοποίηση όσον αφορά την σοβαρότητα 1 και 3. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης.



**Σχήμα 5.39 :** Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο GaussianMixture με είσοδο: `sharp-score, CCP, RF`



**Σχήμα 5.40 :** Τρισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο GaussianMixture με είσοδο: sharp-score, CCP,RF



**Σχήμα 5.41 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο GaussianMixture με είσοδο: sharp-score, CCP,RF

Τέλος όσον αφορά το μοντέλο GaussianMixture ακολουθεί ένας συγκεντρωτικός πίνακας με τις εισόδους που παρουσιάστηκαν παραπάνω και την ομαδοποίηση που προέκυψε από το μοντέλο.

ΕΙΣΟΔΟΣ	ΟΜΑΔΑ 1	ΟΜΑΔΑ 2
1. [ ["sharp-score", "CCPi", "RF", "HLA-DRB1"] ]	0.89 of all sev=1 0.71 of all sev=2 0.19 of all sev=3	0.11 of all sev=1 0.29 of all sev=2 0.81 of all sev=3
2. [ ["sharp-score", "DURATION", "HLA-DRB1"] ]	0.16 of all sev=1 0.41 of all sev=2 0.80 of all sev=3	0.84 of all sev=1 0.59 of all sev=2 0.20 of all sev=3
3. [ ["CCP", "RF", "HLA-DRB1"] ]	0.16 of all sev=1 0.41 of all sev=2 0.80 of all sev=3	0.84 of all sev=1 0.59 of all sev=2 0.20 of all sev=3
4. [ ["sharp-score", "CCP", "CCPi", "RF", "Rfi", "HLA-DRB1"] ]	0.11 of all sev=1 0.29 of all sev=2 0.81 of all sev=3	0.89 of all sev=1 0.71 of all sev=2 0.19 of all sev=3
5. [ ["sharp-score", "CCP", "RF", "HLA-DRB1", "DURATION"] ]	0.16 of all sev=1 0.41 of all sev=2 0.80 of all sev=3	0.84 of all sev=1 0.59 of all sev=2 0.20 of all sev=3
6. [ ["sharp-score", "CCP", "RF", "HLA-DRB1"] ]	0.16 of all sev=1 0.41 of all sev=2 0.80 of all sev=3	0.84 of all sev=1 0.59 of all sev=2 0.20 of all sev=3
7. [ ["CCP", "RF"] ]	0.89 of all sev=1 0.63 of all sev=2 0.16 of all sev=3	0.11 of all sev=1 0.37 of all sev=2 0.84 of all sev=3
8. [ ["sharp-score", "CCP", "RF"] ]	0.84 of all sev=1 0.57 of all sev=2 0.15 of all sev=3	0.16 of all sev=1 0.43 of all sev=2 0.85 of all sev=3

**Πίνακας 5.3:** Αποτελέσματα ομαδοποίησης GaussianMixture με βάση τον πίνακα SEVERITY

Όπως φαίνεται από τα συγκεντρωτικά αποτελέσματα οι καλύτερες ομαδοποιήσεις με το GaussianMixture, έχοντας σαν πίνακα επαλήθευσης SEVERITY με τρεις τιμές για την σοβαρότητα της ρευματοειδούς ,είναι με τις παρακάτω εισόδους:

- `[["sharp-score", "CCPi", "RF", "HLA-DRB1"]]` (ΕΙΣΟΔΟΣ 1)
  - Πέτυχε 89% ομαδοποίηση για την σοβαρότητα 1 , 71% ομαδοποίηση για την σοβαρότητα 2 και 81% ομαδοποίηση για την σοβαρότητα 3.
- `[["CCP", "RF"]]` (ΕΙΣΟΔΟΣ 7)
  - Πέτυχε 89% ομαδοποίηση για την σοβαρότητα 1 , 63% ομαδοποίηση για την σοβαρότητα 2 και 84% ομαδοποίηση για την σοβαρότητα 3. Δεν εμφανίζει καλή ομαδοποίηση για την σοβαρότητα 1 αλλά εμφανίζει εξαιρετική για τις άλλες δύο.

Όλες οι εισόδοι όμως εμφανίζουν αντίστοιχα επίπεδα ομαδοποίησης. Αξίζει να σημειωθεί όμως ότι οι εισόδοι 2,3,5,6 και 8 δεν μπορούν να ομαδοποιήσουν την σοβαρότητα 2 σε μία εκ των 2 συστάδων και την χωρίζουν σχεδόν κατά ήμισυ στις 2 ομάδες. Αυτός είναι ένας από τους λόγους που υποδεικνύουν ότι η σοβαρότητα τύπου 2 «μπερδεύει» τα μοντέλα.

### 5.2.2 Περίπτωση 2 GaussianMixture: SEVERITY\_binary

Στον παρακάτω πίνακα παρουσιάζεται η ομαδοποίηση που προέκυψε με το μοντέλο GaussianMixture έχοντας σαν πίνακα ελέγχου τον SEVERITY\_binary που περιέχει δυαδικές τιμές για την σοβαρότητα. Πάλι προκύπτει ακόμα καλύτερη ομαδοποίηση από ότι με τον πίνακα ελέγχου τον SEVERITY.

Για όλες τις εισόδους που δοκιμάστηκαν και οι δυο ομάδες που δημιουργήθηκαν περιέχουν το 77% και πάνω της σοβαρότητας 1 ή 2 αντίστοιχα. Συνολικά προκύπτει καλύτερη ομαδοποίηση με την GaussianMixture από ότι με την KMeans.

Στα συγκεκριμένα μοντέλα δεν προκύπτει καλύτερη ομαδοποίηση για την σοβαρότητα 1 αλλά αντίστοιχη με της ομαδοποίηση της σοβαρότητας 2.

Η σοβαρότητα 1 βρίσκεται στο 74%-88% στις αντίστοιχες ομάδες ενώ η σοβαρότητα τύπου 2 βρίσκεται κατά 79%-83% στις αντίστοιχες ομάδες.

Η καλύτερη ομαδοποίηση όσον αφορά την σοβαρότητα τύπου 1 προκύπτει για την είσοδο 3: `[["sharp-score", "CCP", "CCPi", "RF", "RF", "HLA-DRB1"]]` όπου το 88% των ασθενών με σοβαρότητα 1 ομαδοποιούνται στο group 1. Αντίστοιχα η καλύτερη ομαδοποίηση όσον αφορά την σοβαρότητα τύπου 2 προκύπτει για την είσοδο 8: `[["sharp-score", "CCP", "RF"]]` όπου το 79% των ασθενών με σοβαρότητα 2 ομαδοποιούνται στο group 0.

Τέλος για λόγους συντομίας παραλείπονται οι γραφικές απεικονίσεις της ομαδοποίησης για τις 8 εισόδους και ακολουθεί ένας συγκεντρωτικός πίνακας με τα αποτελέσματα που προέκυψαν από κάθε είσοδο.

INPUT	GROUP 0	GROUP 1
1. [["sharp-score","CCPi","RF","HLA-DRB1"]]	0.23 of all sev=1 <b>0.79 of all sev=2</b>	<b>0.77 of all sev=1</b> 0.21 of all sev=2
2. [["sharp-score","DURATION","HLA-DRB1"]]	0.23 of all sev=1 <b>0.79 of all sev=2</b>	<b>0.77 of all sev=1</b> 0.21 of all sev=2
3. [["CCP","RF","HLA-DRB1"]]	0.23 of all sev=1 <b>0.79 of all sev=2</b>	<b>0.77 of all sev=1</b> 0.21 of all sev=2
4. [["sharp-score","CCP","CCPi","RF","Rfi","HLA-DRB1"]]	0.12 of all sev=1 <b>0.79 of all sev=2</b>	<b>0.88 of all sev=1</b> 0.21 of all sev=2
5. [["sharp-score","CCP","RF","HLA-DRB1","DURATION"]]	0.23 of all sev=1 <b>0.79 of all sev=2</b>	<b>0.77 of all sev=1</b> 0.21 of all sev=2
6. [["sharp-score","CCP","RF","HLA-DRB1"]]	0.23 of all sev=1 <b>0.79 of all sev=2</b>	<b>0.77 of all sev=1</b> 0.21 of all sev=2
7. [["CCP","RF"]]	0.18 of all sev=1 <b>0.82 of all sev=2</b>	<b>0.82 of all sev=1</b> 0.18 of all sev=2
8. [["sharp-score","CCP","RF"]]	0.26 of all sev=1 <b>0.83 of all sev=2</b>	<b>0.74 of all sev=1</b> 0.17 of all sev=2

**Πίνακας 5.4:** Αποτελέσματα ομαδοποίησης GaussianMixture με βάση τον πίνακα SEVERITY binary



## 5.3 Μοντέλο 3: K-Medoids

Επειδή αναπτύχθηκε πληθώρα μοντέλων με πολλαπλούς συνδυασμούς εισόδων παρακάτω θα παρουσιαστούν τα αποτελέσματα για τις εισόδους που παρουσίασαν την καλύτερη ομαδοποίηση με βάση το πρώτο μοντέλο kMeans ώστε να μπορεί να υπάρξει σύγκριση των αποτελεσμάτων και των μεθόδων.

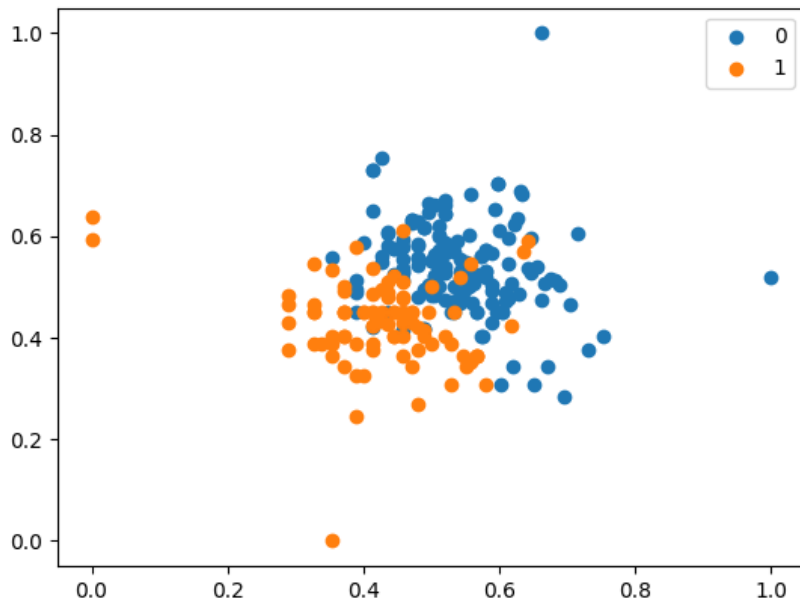
### 5.3.1 Περίπτωση 1 K-Medoids: SEVERITY

5.3.1.1 Είσοδος 1 K-Medoids: `teachArr = workingDF[["sharp-score","RF","HLA-DRB1","CCPi"]].to_numpy()`

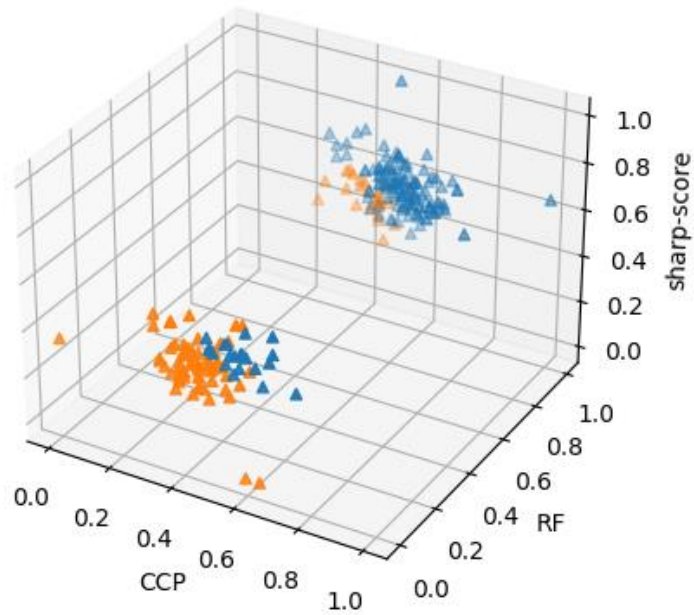
Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.11 of all sev=1	0.89 of all sev=1
0.31 of all sev=2	0.69 of all sev=2
0.84 of all sev=3	0.16 of all sev=3

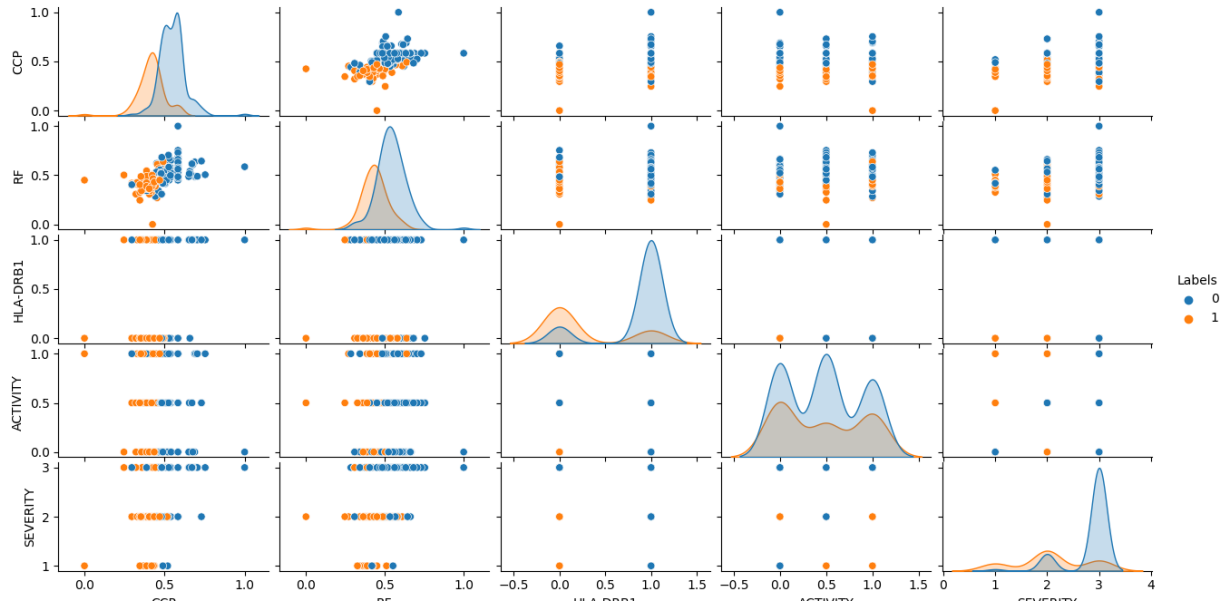
Όπως φαίνεται στο group 0 περιέχεται το 84% των ασθενών με σοβαρότητα 3 ενώ στο group 1 περιέχεται το 89% των ασθενών με σοβαρότητα 3. Επιπλέον η σοβαρότητα 2 βρίσκεται κατά μεγαλύτερο βαθμό στο group 1 (69%). Πρόκειται για μια αρκετά καλή ομαδοποίηση. Σε σχέση με το Kmeans εμφανίζει καλύτερη ομαδοποίηση για την σοβαρότητα 3 ενώ το kmeans εμφανίζει καλύτερη ομαδοποίηση για τη σοβαρότητα 1. Σε σχέση με το μοντέλο GaussianMixture εμφανίζει την ίδια ομαδοποίηση για τη σοβαρότητα 1, ελαφρώς καλύτερη για τη σοβαρότητα 3 και ελαφρώς χειρότερη για τη σοβαρότητα 2. Συγκριτικά το Kmeans εμφανίζει καλύτερη ομαδοποίηση. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης.



**Σχήμα 5.42 :** Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο K-Medoids με είσοδο: sharp-score,RF,HLA-DRB1,CPi



**Σχήμα 5.43 :** Τρισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο K-Medoids με είσοδο: sharp-score,RF,HLA-DRB1,CPi



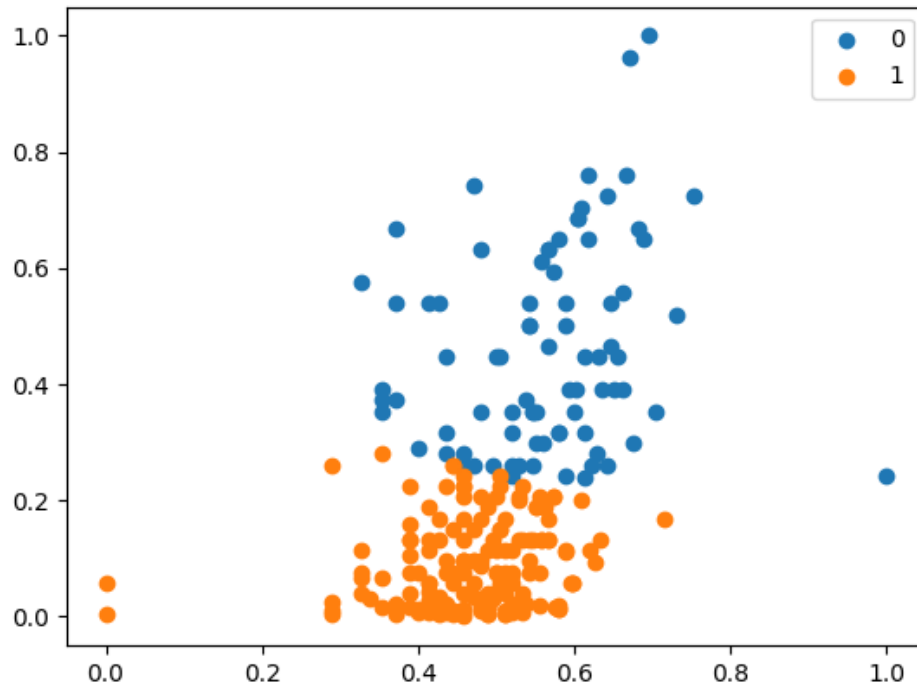
**Σχήμα 5.44 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο K-Medoids με είσοδο: sharp-score,RF,HLA-DRB1,CPi

5.3.1.2 **Είσοδος 2 K-Medoids :** `teachArr = workingDF[["sharp-score","DURATION","HLA-DRB1"]].to_numpy()`

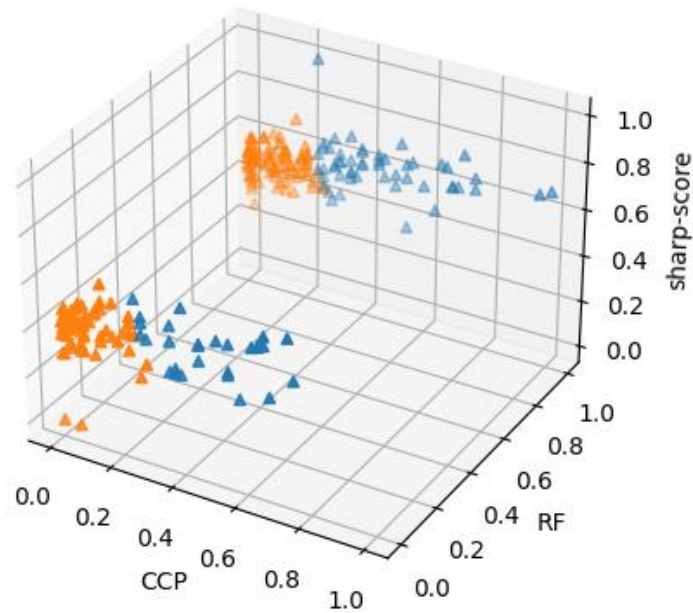
Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.05 of all sev=1	0.95 of all sev=1
0.32 of all sev=2	0.68 of all sev=2
0.38 of all sev=3	0.62 of all sev=3

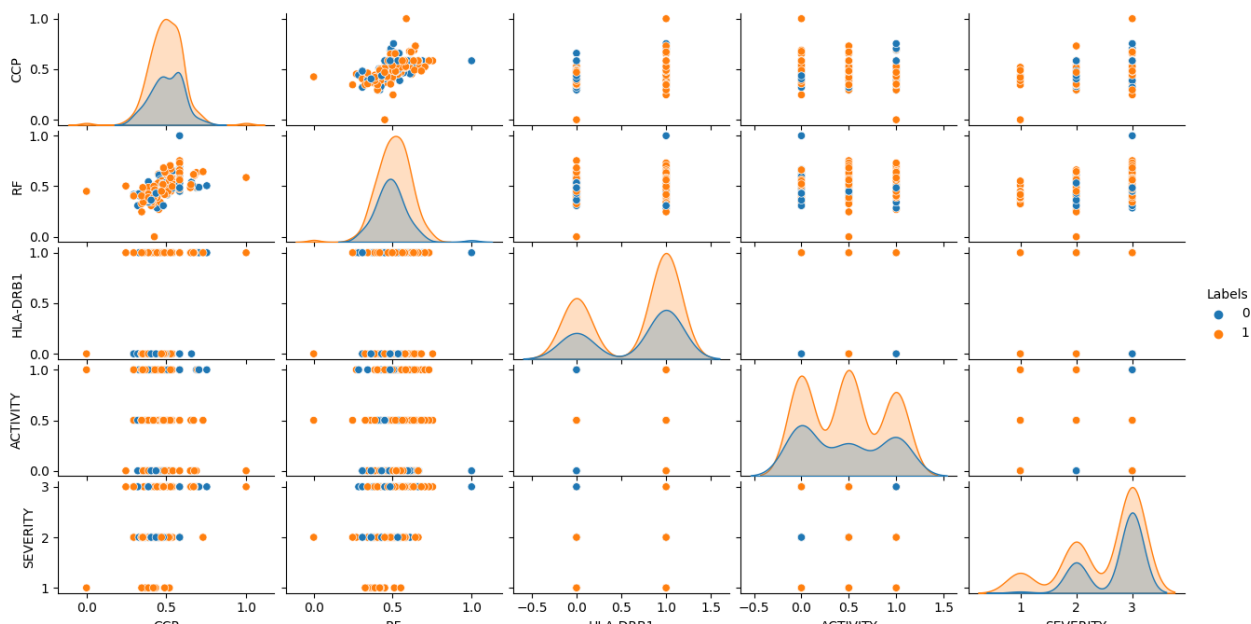
Όπως φαίνεται στο group 0 περιέχεται το 38% των ασθενών με σοβαρότητα 3 ενώ στο group 1 περιέχεται το 95% των ασθενών με σοβαρότητα 1. Επιπλέον η σοβαρότητα 2 βρίσκεται σε μεγαλύτερο βαθμό στο group 1 (68%). Πρόκειται για την χειρότερη ομαδοποίηση μέχρι στιγμής. Τόσο το KMeans όσο και το GaussianMixture παρουσιάζουν καλύτερη ομαδοποίηση. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης.



**Σχήμα 5.45 :** Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο K-Medoids με είσοδο: sharp-score,DURATION,HLA-DRB1



**Σχήμα 5.46 :** Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο K-Medoids με είσοδο: sharp-score,DURATION,HLA-DRB1



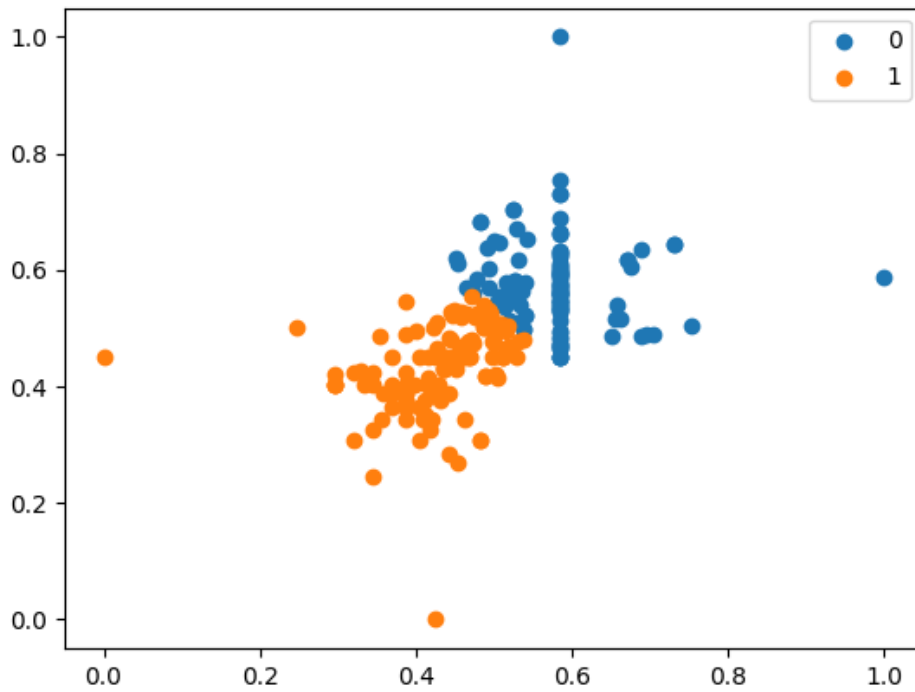
**Σχήμα 5.47 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο K-Medoids με είσοδο: sharp-score,DURATION,HLA-DRB1

5.3.1.3 **Είσοδος 3 K-Medoids:** `teachArr = workingDF[["CCP", "RF", "HLA-DRB1"]].to_numpy()`

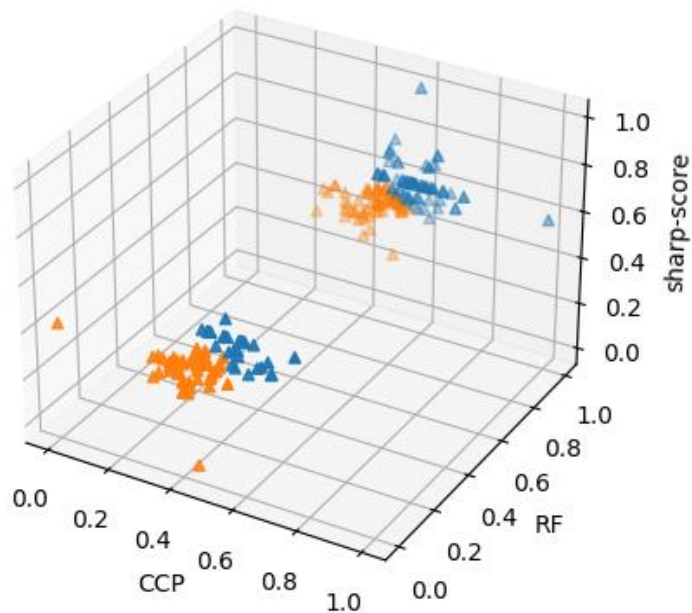
Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.05 of all sev=1	0.95 of all sev=1
0.25 of all sev=2	0.75 of all sev=2
0.63 of all sev=3	0.37 of all sev=3

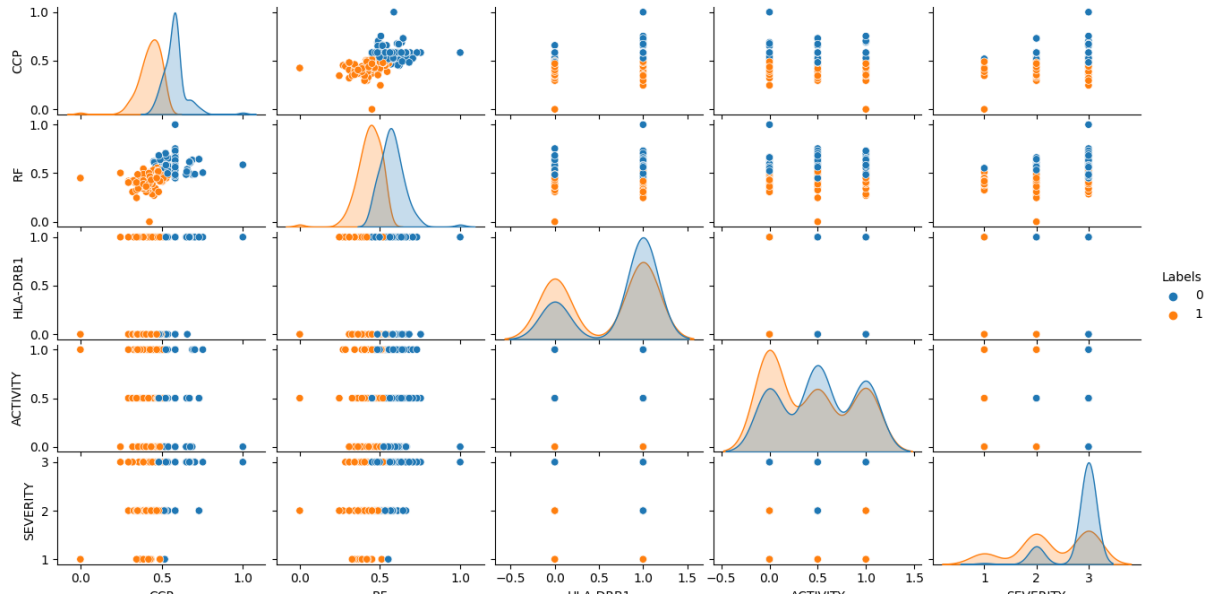
Όπως φαίνεται στο group 0 περιέχεται το 63% των ασθενών με σοβαρότητα 3 ενώ στο group 1 περιέχεται το 95% των ασθενών με σοβαρότητα 1. Επιπλέον η σοβαρότητα 2 βρίσκεται σε μεγαλύτερο βαθμό στο group 1 (75%) Συγκριτικά το Kmeans εμφανίζει την ίδια ομαδοποίηση όσον αφορά τη σοβαρότητα 1 και 3 και καλύτερη ομαδοποίηση όσον αφορά την σοβαρότητα 2. Η GaussianMinture εμφανίζει την καλύτερη ομαδοποίηση από όλα τα μοντέλα για την σοβαρότητα 3 και χειρότερη για τις άλλες δύο. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης.



**Σχήμα 5.48 :** Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο K-Medoids με είσοδο:  
CCP,RF,HLA-DRB1



**Σχήμα 5.49 :** Τρισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο K-Medoids με είσοδο:  
CCP,RF,HLA-DRB1



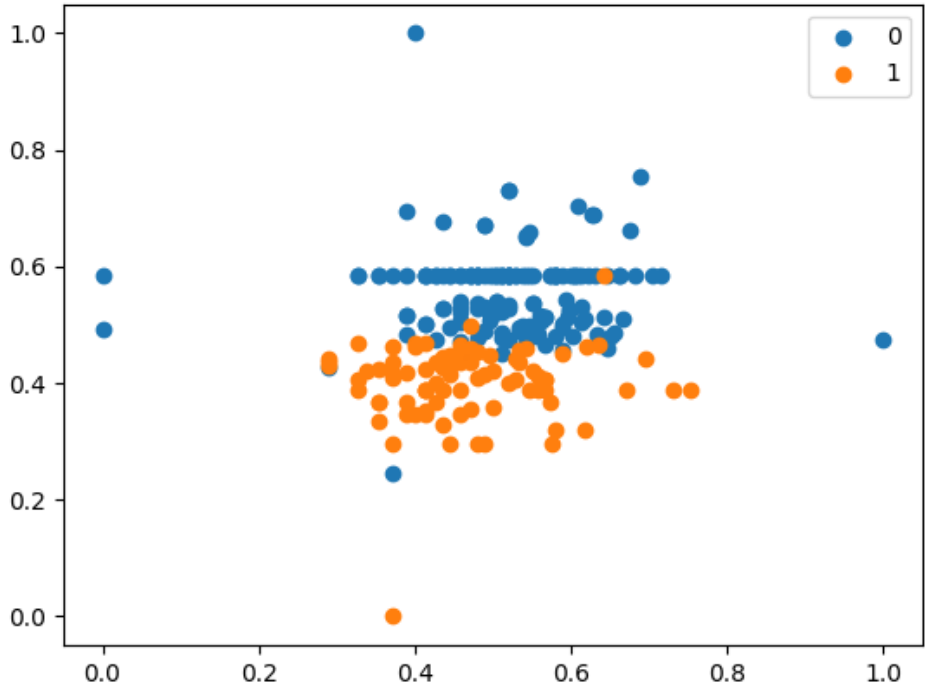
**Σχήμα 5.50 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο K-Medoids με είσοδο: CCP,RF,HLA-DRB1

5.3.1.4 **Είσοδος 4 K-Medoids:** `teachArr = workingDF[["sharp-score", "CCP", "CCPi", "RF", "Rfi", "HLA-DRB1"]].to_numpy()`

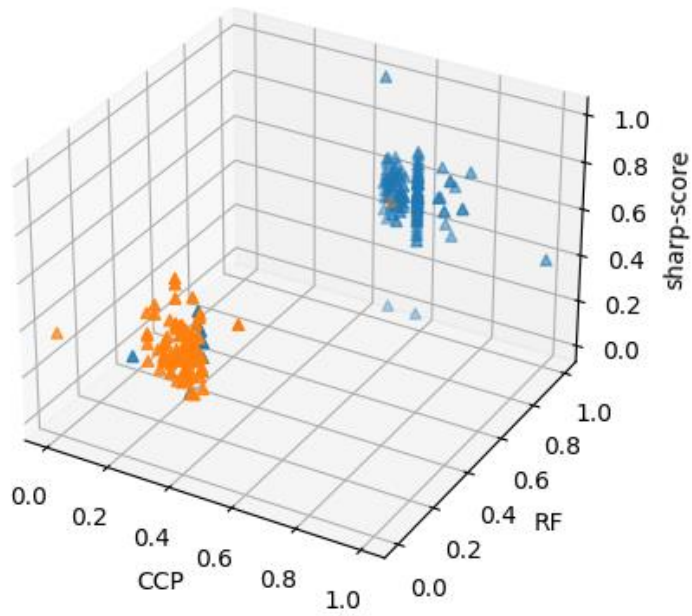
Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.11 of all sev=1	0.89 of all sev=1
0.34 of all sev=2	0.66 of all sev=2
0.85 of all sev=3	0.15 of all sev=3

Όπως φαίνεται στο group 0 περιέχεται το 85% των ασθενών με σοβαρότητα 3 ενώ στο group 1 περιέχεται το 89% των ασθενών με σοβαρότητα 1. Επιπλέον η σοβαρότητα 2 βρίσκεται κατά μεγαλύτερο βαθμό στο group 1 (66%). Πρόκειται για μια αρκετά καλή ομαδοποίηση. Το συγκεκριμένο μοντέλο έχει παρόμοια αποτελέσματα με το μοντέλο Gaussian Mixture για την ίδια είσοδο. Σε σχέση με το Kmeans εμφανίζει την ίδια ομαδοποίηση για την σοβαρότητα 3 (περίπου 80%) ενώ το kmeans εμφανίζει καλύτερη ομαδοποίηση για τη σοβαρότητα 1 ενώ παρουσιάζουν παρόμοια ομαδοποίηση για τη σοβαρότητα. Συγκριτικά τα τρία μοντέλα παρουσιάζουν παρόμοια ομαδοποίηση. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης.

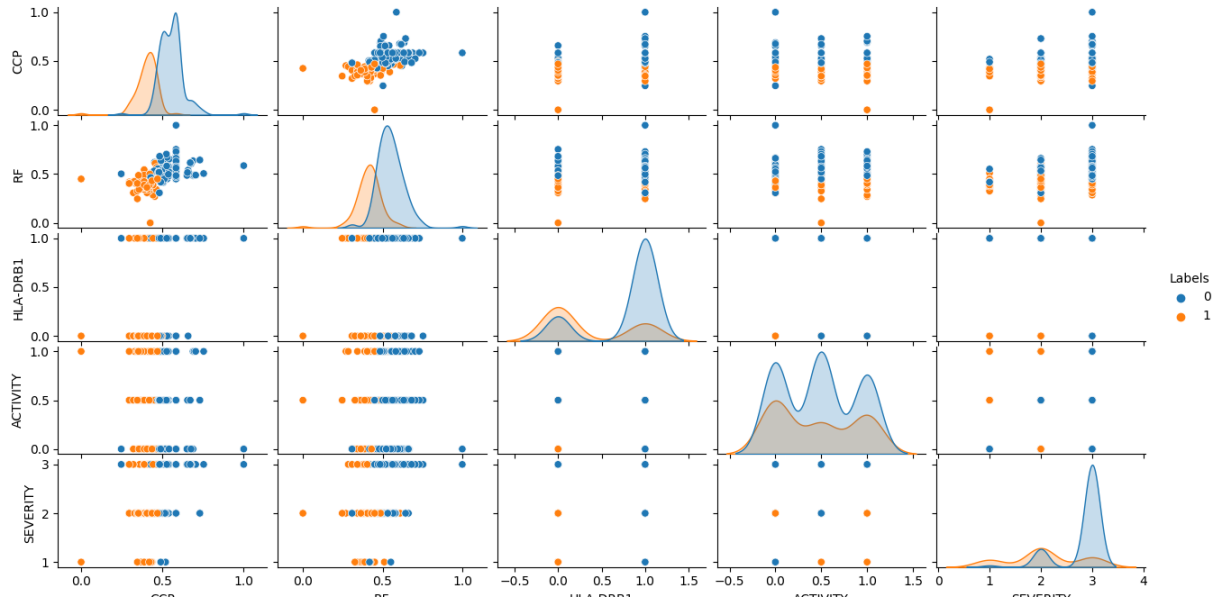


**Σχήμα 5.51:** Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο K-Medoids με είσοδο: sharp-score,CCP,CCPi,RF,Rfi,HLA-DRB1



**Σχήμα 5.52:** Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο K-Medoids με είσοδο: sharp-score,CCP,CCPi,RF,Rfi,HLA-DRB1





**Σχήμα 5.53 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο K-Medoids με είσοδο: sharp-score,CCP,CCPi,RF,Rfi,HLA-DRB1

5.3.1.5 **Είσοδος 5 K-Medoids:** `teachArr = workingDF[["sharp-score","CCP","RF","HLA-DRB1","DURATION"]].to_numpy()`

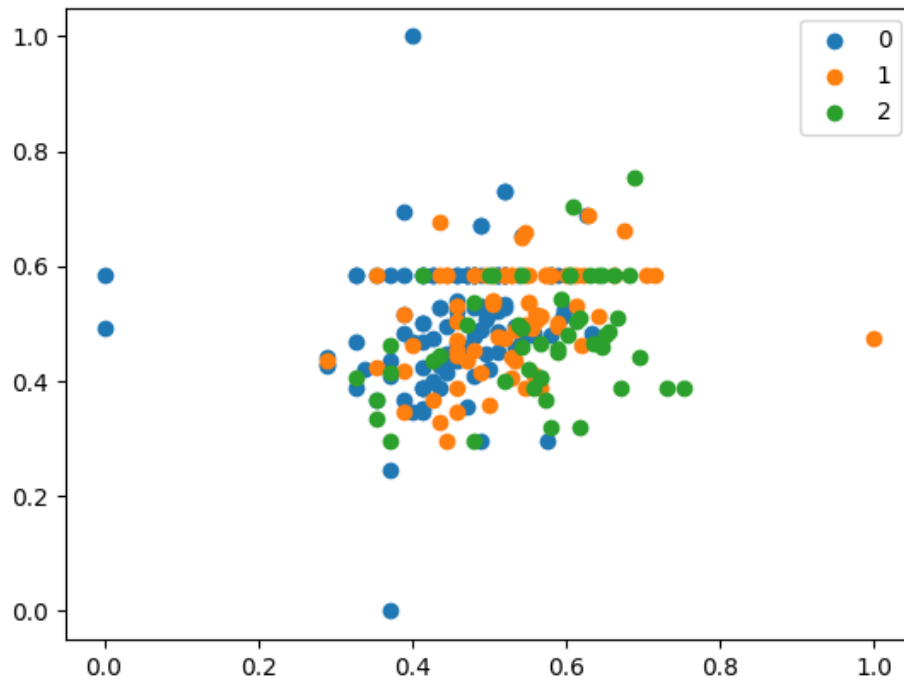
Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
1.00 of all sev=1	
1.00 of all sev=2	<b>Cluster 1 is empty!</b>
1.00 of all sev=3	

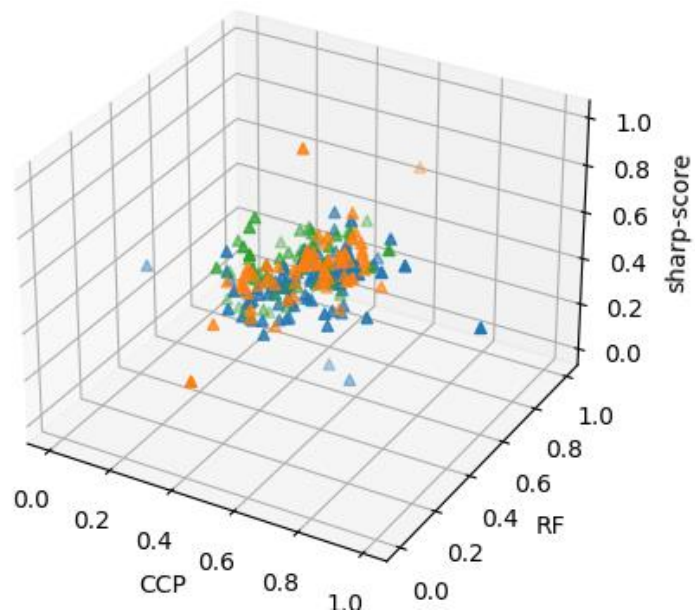
Το μοντέλο k-Medoids δεν μπορεί να δημιουργήσει 2 ομάδες για την είσοδο 5. Για k=3 παράγει τα παρακάτω αποτελέσματα

group 0	group 1	group 2
0.84 of all sev=1	0.11 of all sev=1	0.05 of all sev=1
0.46 of all sev=2	0.32 of all sev=2	0.22 of all sev=2
0.44 of all sev=3	0.33 of all sev=3	0.23 of all sev=3

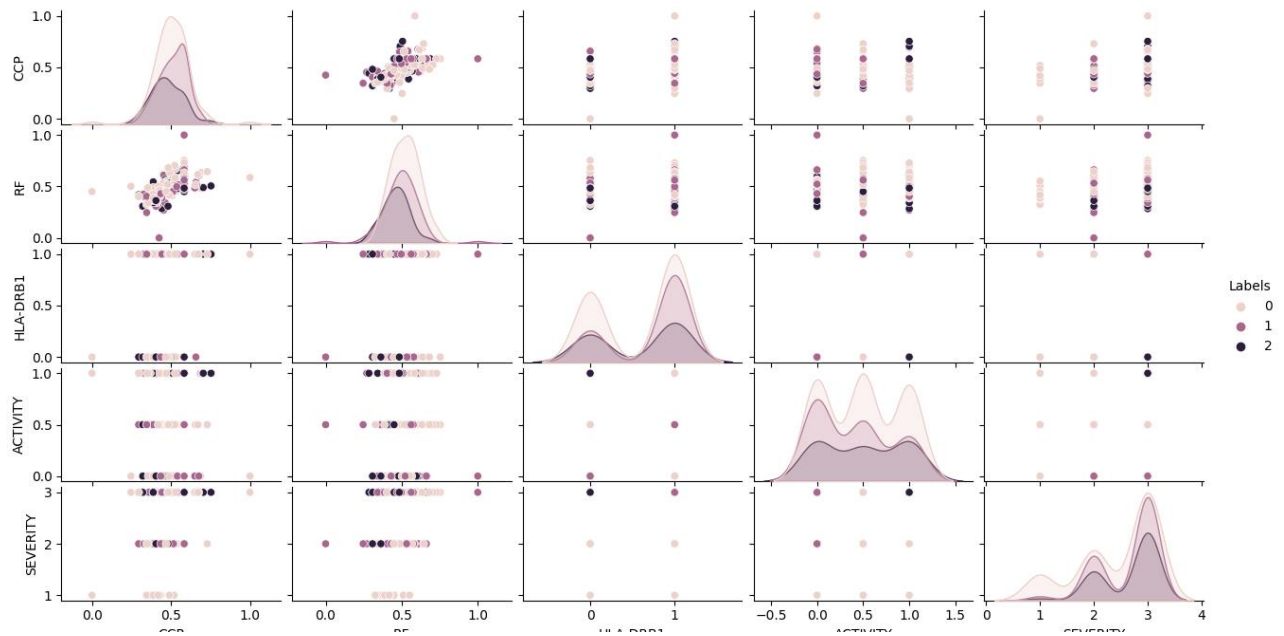
Για k=3 ομαδοποιεί σωστά μόνο τη σοβαρότητα 1. Πρόκειται για μια κακή ομαδοποίηση επομένως δεν θα αναφερθεί στη συνέχεια στα συγκεντρωτικά αποτελέσματα του μοντέλου. Ενδεικτικά οι γραφικές παραστάσεις φαίνονται παρακάτω.



**Σχήμα 5.54 :** Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο K-Medoids με είσοδο: sharp-score,CCP,RF,HLA-DRB1,DURATION



**Σχήμα 5.55 :** Τρισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο K-Medoids με είσοδο: sharp-score,CCP,RF,HLA-DRB1,DURATION



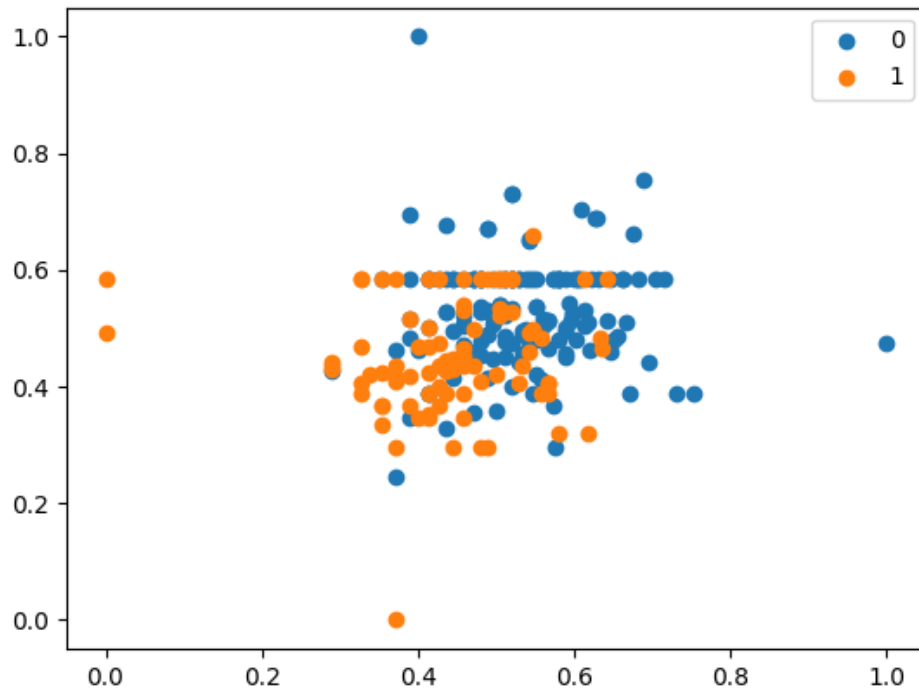
**Σχήμα 5.56 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο K-Medoids με είσοδο: sharp-score,CCP,RF,HLA-DRB1,DURATION

5.3.1.6 **Είσοδος 6 K-Medoids:** `teachArr = workingDF[["sharp-score","CCP","RF","HLA-DRB1"]].to_numpy()`

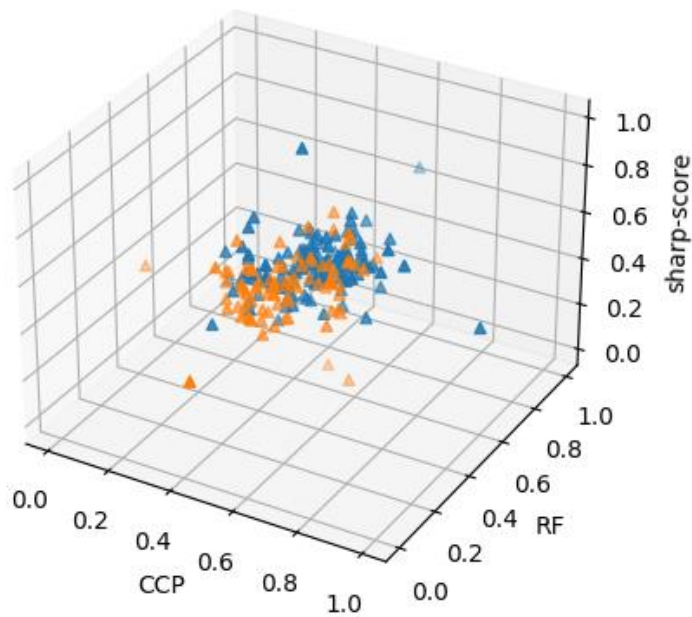
Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.16 of all sev=1	0.84 of all sev=1
0.41 of all sev=2	0.59 of all sev=2
0.80 of all sev=3	0.20 of all sev=3

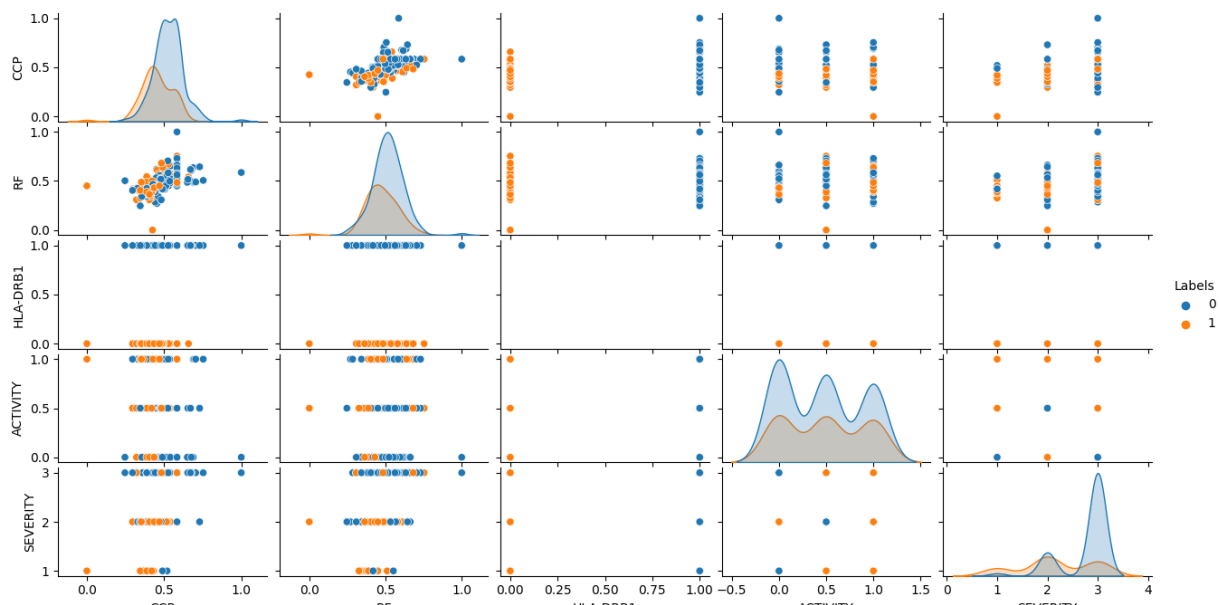
Όπως φαίνεται στο group 0 περιέχεται το 80% των ασθενών με σοβαρότητα 3 ενώ στο group 1 περιέχεται το 84% των ασθενών με σοβαρότητα 1. Επιπλέον η σοβαρότητα 2 βρίσκεται εξίσου και στις 2 ομάδες .Παρατηρείται παρόμοια ομαδοποίηση με το μοντέλο GaussianMixture. Συγκριτικά το Kmeans με την ίδια είσοδο εμφανίζει σαφώς καλύτερη ομαδοποίηση. Οι γραφικές απεικονίσεις της ομαδοποίησης φαίνονται παρακάτω



**Σχήμα 5.57 :** Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο K-Medoids με είσοδο: sharp-score,CCP,RF,HLA-DRB1



**Σχήμα 5.58 :** Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο K-Medoids με είσοδο: sharp-score,CCP,RF,HLA-DRB1



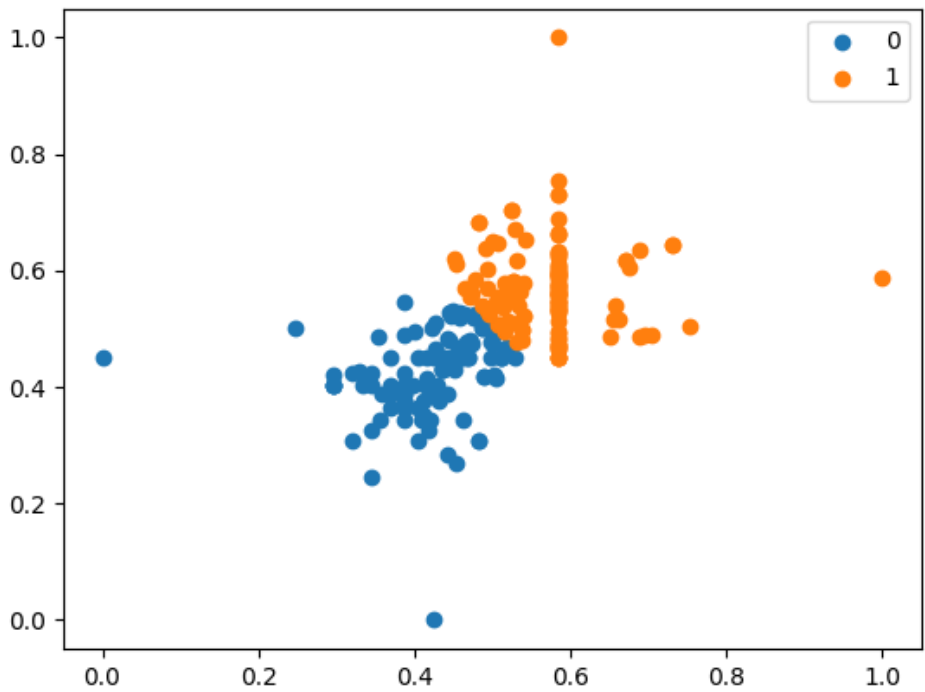
**Σχήμα 5.59 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο K-Medoids με είσοδο: sharp-score,CCP,RF,HLA-DRB1

### 5.3.1.7 Είσοδος 7 K-Medoids: `teachArr = workingDF[["CCP", "RF"]].to_numpy()`

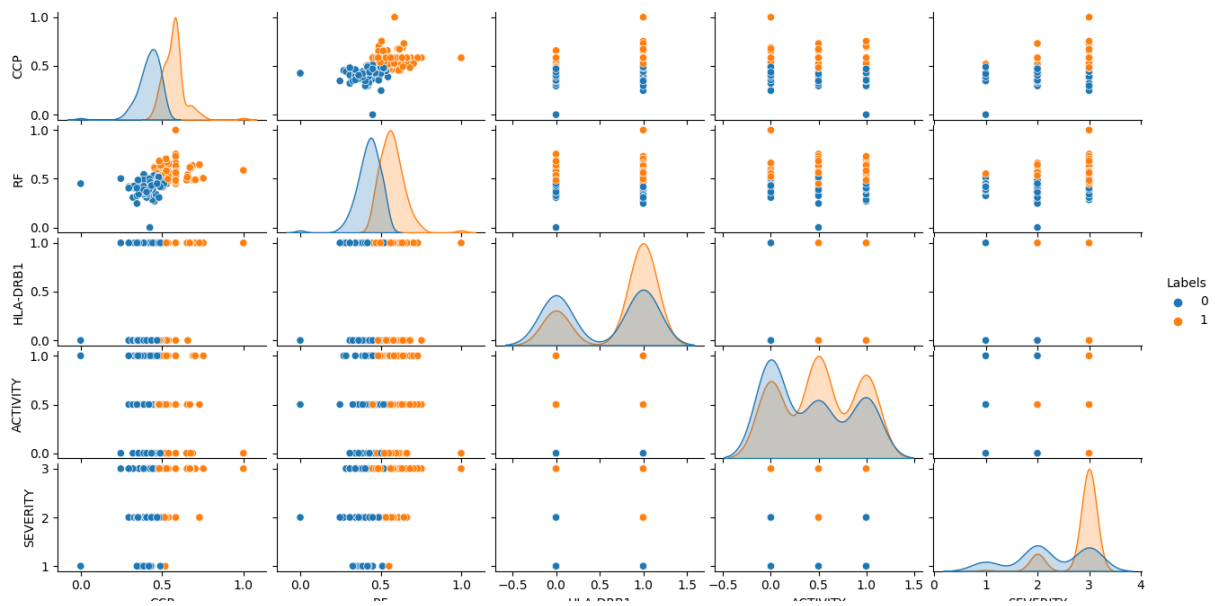
Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.95 of all sev=1	0.05 of all sev=1
0.74 of all sev=2	0.26 of all sev=2
0.30 of all sev=3	0.70 of all sev=3

Όπως φαίνεται στο group 0 περιέχεται το 95% των ασθενών με σοβαρότητα 1 ενώ στο group 1 περιέχεται το 70% των ασθενών με σοβαρότητα 3. Επιπλέον η σοβαρότητα 2 βρίσκεται κατά μεγαλύτερο βαθμό στο group 0 (74%). Πρόκειται για μια πολύ καλή ομαδοποίηση. Όσον αφορά την αντίστοιχη είσοδο στο Kmeans εμφανίζει την ίδια ομαδοποίηση για σοβαρότητα 1,ελαφρώς καλύτερη ομαδοποίηση για σοβαρότητα 3 και χειρότερη ομαδοποίηση για σοβαρότητα 2. Όσον αφορά την αντίστοιχη είσοδο στο GaussianMixture εμφανίζει καλύτερη ομαδοποίηση μόνο για τη σοβαρότητα 2.Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης , επειδή έχουμε σαν εισόδους μόνο 2 παράγοντες δεν έχει νόημα η τρισδιάστατη απεικόνιση οπότε και παραλείπεται.



**Σχήμα 5.60 :** Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο K-Medoids με είσοδο: CCP,RF



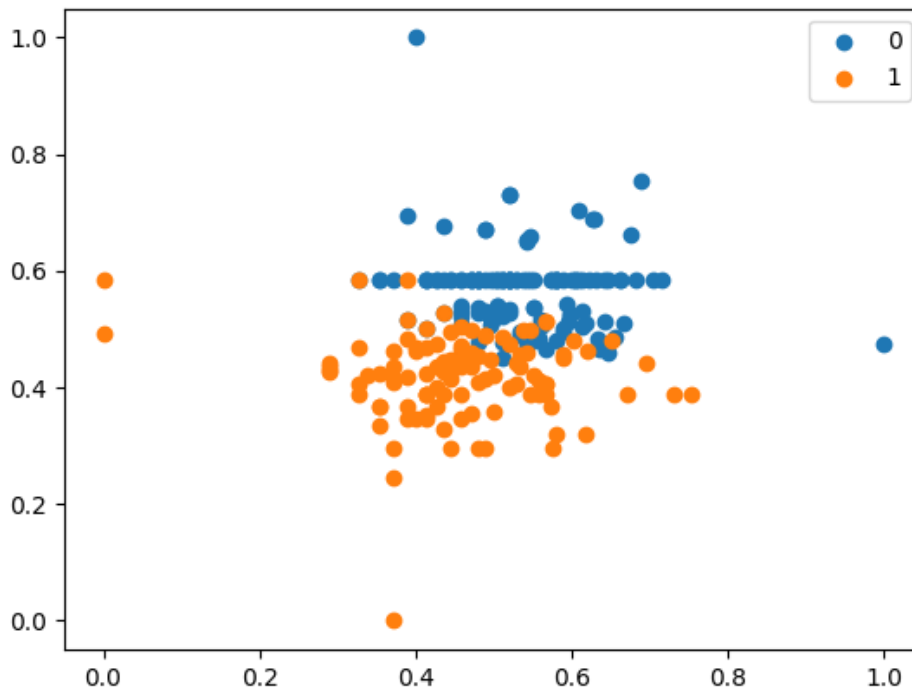
**Σχήμα 5.61 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο K-Medoids με είσοδο: CCP,RF

5.3.1.8 Είσοδος 8 K-Medoids: `teachArr = workingDF[["sharp-score","CCP","RF"]].to_numpy`

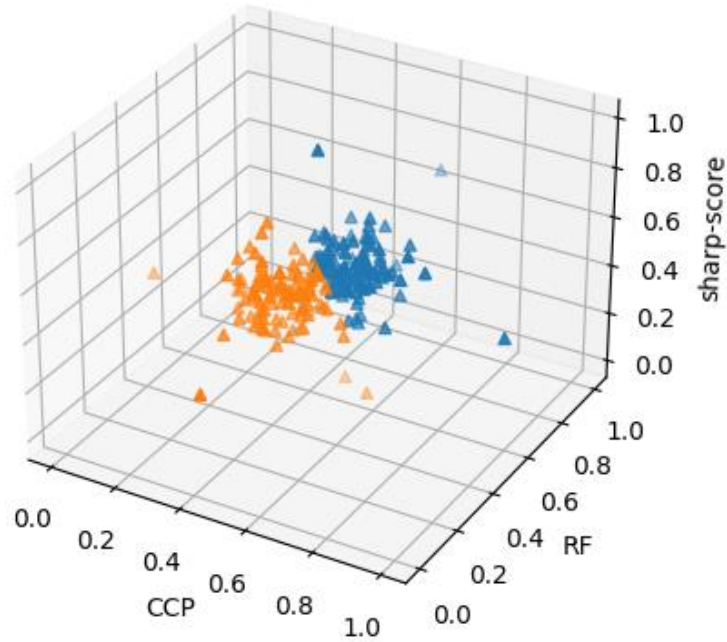
Με βάση την παραπάνω είσοδο προέκυψε η ομαδοποίηση που παρουσιάζεται παρακάτω

group 0	group 1
0.05 of all sev=1	0.95 of all sev=1
0.24 of all sev=2	0.76 of all sev=2
0.75 of all sev=3	0.25 of all sev=3

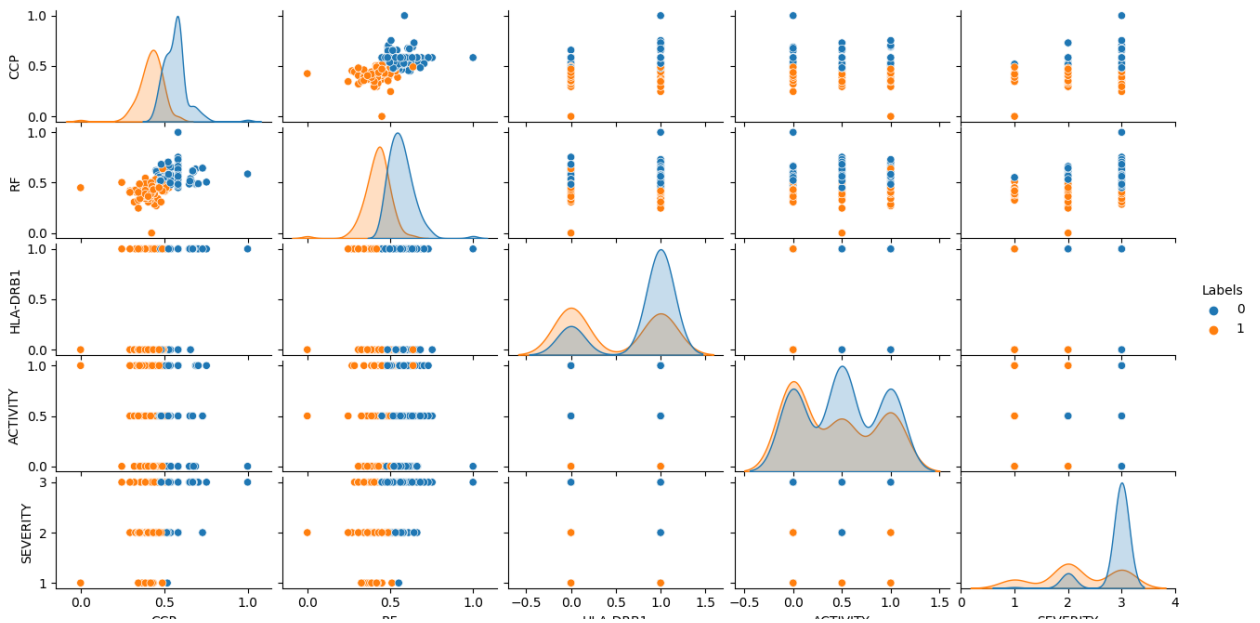
Όπως φαίνεται στο group 0 περιέχεται το 75% των ασθενών με σοβαρότητα 3 ενώ στο group 1 περιέχεται το 95% των ασθενών με σοβαρότητα 1. Επιπλέον η σοβαρότητα 2 βρίσκεται σε μεγαλύτερο βαθμό στο group 1 (76%). Πρόκειται για μια πολύ καλή ομαδοποίηση όσον αφορά την σοβαρότητα 1 και 2. Σε σχέση με το Kmeans έχει παρόμοια αποτελέσματα. Τέλος το μοντέλο GaussianMixture παρουσιάζει καλύτερη ομαδοποίηση από τα άλλα 2 μόνο για την σοβαρότητα 3 και χειρότερη για τις άλλες δύο. Ακολουθούν γραφικές απεικονίσεις της ομαδοποίησης.



Σχήμα 5.62 : Δισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο K-Medoids με είσοδο: `sharp-score,CCP,RF`



**Σχήμα 5.63 :** Τρισδιάστατη απεικόνιση των 2 συστάδων. Μοντέλο K-Medoids με είσοδο: sharp-score, CCP, RF



**Σχήμα 5.64 :** Συσχέτιση κρίσιμων παραγόντων και απεικόνιση της ομαδοποίησης για την κάθε γραφική. Μοντέλο K-Medoids με είσοδο: sharp-score, CCP, RF



Τέλος όσον αφορά το μοντέλο K-Medoids ακολουθεί ένας συγκεντρωτικός πίνακας με τις εισόδους που παρουσιάστηκαν παραπάνω και την ομαδοποίηση που προέκυψε από το μοντέλο.

ΕΙΣΟΔΟΣ	ΟΜΑΔΑ 1	ΟΜΑΔΑ 2
1. [ ["sharp-score", "CCPi", "RF", "HLA-DRB1"] ]	0.11 of all sev=1 0.31 of all sev=2 0.84 of all sev=3	0.89 of all sev=1 0.69 of all sev=2 0.16 of all sev=3
2. [ ["sharp-score", "DURATION", "HLA-DRB1"] ]	0.05 of all sev=1 0.32 of all sev=2 0.38 of all sev=3	0.95 of all sev=1 0.68 of all sev=2 0.62 of all sev=3
3. [ ["CCP", "RF", "HLA-DRB1"] ]	0.05 of all sev=1 0.25 of all sev=2 0.63 of all sev=3	0.95 of all sev=1 0.75 of all sev=2 0.37 of all sev=3
4. [ ["sharp-score", "CCP", "CCPi", "RF", "Rfi", "HLA-DRB1"] ]	0.11 of all sev=1 0.34 of all sev=2 0.85 of all sev=3	0.89 of all sev=1 0.66 of all sev=2 0.15 of all sev=3
5. [ ["sharp-score", "CCP", "RF", "HLA-DRB1", "DURATION"] ]	---	---
6. [ ["sharp-score", "CCP", "RF", "HLA-DRB1"] ]	0.16 of all sev=1 0.41 of all sev=2 0.80 of all sev=3	0.84 of all sev=1 0.59 of all sev=2 0.20 of all sev=3
7. [ ["CCP", "RF"] ]	0.95 of all sev=1 0.74 of all sev=2 0.30 of all sev=3	0.05 of all sev=1 0.26 of all sev=2 0.70 of all sev=3
8. [ ["sharp-score", "CCP", "RF"] ]	0.05 of all sev=1 0.24 of all sev=2 0.75 of all sev=3	0.95 of all sev=1 0.76 of all sev=2 0.25 of all sev=3

**Πίνακας 5.5:** Αποτελέσματα ομαδοποίησης με βάση τον πίνακα SEVERITY

Όπως φαίνεται από τα συγκεντρωτικά αποτελέσματα οι καλύτερες ομαδοποιήσεις με το K-Medoids, έχοντας σαν πίνακα επαλήθευσης SEVERITY με τρεις τιμές για την σοβαρότητα της ρευματοειδούς ,είναι με τις παρακάτω εισόδους:

- [ ["sharp-score", "CCP", "CCPi", "RF", "Rfi", "HLA-DRB1"] ] (ΕΙΣΟΔΟΣ 4)
  - Πέτυχε 89% ομαδοποίηση για την σοβαρότητα 1 , 66% ομαδοποίηση για την σοβαρότητα 2 και 85% ομαδοποίηση για την σοβαρότητα 3.
- [ ["sharp-score", "CCP", "RF"] ] (ΕΙΣΟΔΟΣ 8)
  - Πέτυχε 95% ομαδοποίηση για την σοβαρότητα 1 , 76% ομαδοποίηση για την σοβαρότητα 2 και 75% ομαδοποίηση για την σοβαρότητα 3.

### 5.3.2 Περίπτωση 2 K-Medoids: SEVERITY\_binary

Στον παρακάτω πίνακα παρουσιάζεται η ομαδοποίηση που προέκυψε με το μοντέλο GaussianMixture έχοντας σαν πίνακα ελέγχου τον SEVERITY\_binary που περιέχει δυαδικές τιμές για την σοβαρότητα. Πάλι προκύπτει ακόμα καλύτερη ομαδοποίηση από ότι με τον πίνακα ελέγχου τον SEVERITY.

Για όλες τις εισόδους που δοκιμάστηκαν εκτός της εισόδου 2 και 5 και οι δυο ομάδες που δημιουργήθηκαν περιέχουν το 62% και πάνω της σοβαρότητας 1 ή 2 αντίστοιχα. Συνολικά προκύπτει χειρότερη ομαδοποίηση με την K-Medoids από ότι με την KMeans και την GaussianMixture.

Στα συγκεκριμένα μοντέλα έχουμε καλύτερη ομαδοποίηση για την σοβαρότητα 1 όπως με την K-Means. Η σοβαρότητα 1 βρίσκεται στο 77%-94% στις αντίστοιχες ομάδες ενώ η σοβαρότητα τύπου 2 βρίσκεται κατά 62%-83% στις αντίστοιχες ομάδες.

Η καλύτερη ομαδοποίηση όσον αφορά την σοβαρότητα τύπου 1 προκύπτει για την είσοδο 8: [ ["sharp-score","CCP","RF"] ] όπου το 94% των ασθενών με σοβαρότητα 1 ομαδοποιούνται στο group 1. Αντίστοιχα η καλύτερη ομαδοποίηση όσον αφορά την σοβαρότητα τύπου 2 προκύπτει για την είσοδο 1: [ ["sharp-score","CCPi","RF","HLA-DRB1"] ] και 4: [ ["sharp-score","CCP","CCPi","RF","Rfi","HLA-DRB1"] ] όπου το 83% των ασθενών με σοβαρότητα 2 ομαδοποιούνται στο group 0.

Συνολικά το K-Medoids παρουσιάζει χειρότερη ομαδοποίηση από τα προηγούμενα δύο μοντέλα αλλά η ομαδοποίηση που προκύπτει από την είσοδο 1 είναι από τις πιο ισορροπημένες ομαδοποιήσεις για τις δύο σοβαρότητες.

Για λόγους συντομίας δεν παρουσιάζονται οι γραφικές παραστάσεις που προέκυψαν.

INPUT	GROUP 0	GROUP 1
1. [{"sharp-score", "CCPi", "RF", "HLA-DRB1"}]	0.11 of all sev=1 <b>0.83 of all sev=2</b>	<b>0.89 of all sev=1</b> 0.17 of all sev=2
2. [{"sharp-score", "DURATION", "HLA-DRB1"}]	0.23 of all sev=1 0.38 of all sev=2	0.77 of all sev=1 0.62 of all sev=2
3. [{"CCP", "RF", "HLA-DRB1"}]	0.11 of all sev=1 <b>0.62 of all sev=2</b>	<b>0.89 of all sev=1</b> 0.38 of all sev=2
4. [{"sharp-score", "CCP", "CCPi", "RF", "Rfi", "HLA-DRB1"}]	0.15 of all sev=1 <b>0.83 of all sev=2</b>	<b>0.85 of all sev=1</b> 0.17 of all sev=2
5. [{"sharp-score", "CCP", "RF", "HLA-DRB1", "DURATION"}]	---	---
6. [{"sharp-score", "CCP", "RF", "HLA-DRB1"}]	0.23 of all sev=1 <b>0.79 of all sev=2</b>	<b>0.77 of all sev=1</b> 0.21 of all sev=2
7. [{"CCP", "RF"}]	<b>0.89 of all sev=1</b> 0.32 of all sev=2	0.11 of all sev=1 <b>0.68 of all sev=2</b>
8. [{"sharp-score", "CCP", "RF"}]	0.06 of all sev=1 <b>0.73 of all sev=2</b>	<b>0.94 of all sev=1</b> 0.27 of all sev=2

**Πίνακας 5.6:** Αποτελέσματα ομαδοποίησης K-Medoids με βάση τον πίνακα SEVERITY binary

Τέλος παρουσιάζεται ένας συγκεντρωτικός πίνακας με τα αποτελέσματα που προέκυψαν για κάθε είσοδο και με τις τρεις μεθόδους.

INPUT
1. [{"sharp-score", "CCPi", "RF", "HLA-DRB1"}]
2. [{"sharp-score", "DURATION", "HLA-DRB1"}]
3. [{"CCP", "RF", "HLA-DRB1"}]
4. [{"sharp-score", "CCP", "CCPi", "RF", "Rfi", "HLA-DRB1"}]
5. [{"sharp-score", "CCP", "RF", "HLA-DRB1", "DURATION"}]
6. [{"sharp-score", "CCP", "RF", "HLA-DRB1"}]
7. [{"CCP", "RF"}]
8. [{"sharp-score", "CCP", "RF"}]

**Πίνακας 5.7:** Σύνολο εισόδων που παρουσιάστηκαν

	K-Means		GaussianMixture		K-Medoids	
	GROUP 0	GROUP 1	GROUP 0	GROUP 1	GROUP 0	GROUP 1
1.	0.09 of all sev=1	<b>0.91</b> of all sev=1	0.23 of all sev=1	<b>0.77</b> of all sev=1	0.11 of all sev=1	<b>0.89</b> of all sev=1
	<b>0.74</b> of all sev=2	0.26 of all sev=2	<b>0.79</b> of all sev=2	0.21 of all sev=2	<b>0.83</b> of all sev=2	0.17 of all sev=2
2.	0.12 of all sev=1	<b>0.88</b> of all sev=1	0.23 of all sev=1	<b>0.77</b> of all sev=1	0.23 of all sev=1	<b>0.77</b> of all sev=1
	<b>0.65</b> of all sev=2	0.35 of all sev=2	<b>0.79</b> of all sev=2	0.21 of all sev=2	<b>0.38</b> of all sev=2	0.62 of all sev=2
3.	<b>0.86</b> of all sev=1	0.14 of all sev=1	0.23 of all sev=1	<b>0.77</b> of all sev=1	0.11 of all sev=1	<b>0.89</b> of all sev=1
	0.21 of all sev=2	<b>0.79</b> of all sev=2	<b>0.79</b> of all sev=2	0.21 of all sev=2	<b>0.62</b> of all sev=2	0.38 of all sev=2
4.	0.12 of all sev=1	<b>0.88</b> of all sev=1	0.12 of all sev=1	<b>0.88</b> of all sev=1	0.15 of all sev=1	<b>0.85</b> of all sev=1
	<b>0.78</b> of all sev=2	0.22 of all sev=2	<b>0.79</b> of all sev=2	0.21 of all sev=2	<b>0.83</b> of all sev=2	0.17 of all sev=2
5.	<b>0.89</b> of all sev=1	0.11 of all sev=1	0.23 of all sev=1	<b>0.77</b> of all sev=1		
	0.23 of all sev=2	<b>0.77</b> of all sev=2	<b>0.79</b> of all sev=2	0.21 of all sev=2	—	—
6.	<b>0.89</b> of all sev=1	0.11 of all sev=1	0.23 of all sev=1	<b>0.77</b> of all sev=1	0.23 of all sev=1	<b>0.77</b> of all sev=1
	0.23 of all sev=2	<b>0.77</b> of all sev=2	<b>0.79</b> of all sev=2	0.21 of all sev=2	<b>0.79</b> of all sev=2	0.21 of all sev=2
7.	<b>0.86</b> of all sev=1	0.14 of all sev=1	0.18 of all sev=1	<b>0.82</b> of all sev=1	<b>0.89</b> of all sev=1	0.11 of all sev=1
	0.22 of all sev=2	<b>0.78</b> of all sev=2	<b>0.82</b> of all sev=2	0.18 of all sev=2	0.32 of all sev=2	<b>0.68</b> of all sev=2
8.	<b>0.89</b> of all sev=1	0.11 of all sev=1	0.26 of all sev=1	<b>0.74</b> of all sev=1	0.06 of all sev=1	<b>0.94</b> of all sev=1
	0.23 of all sev=2	<b>0.77</b> of all sev=2	<b>0.83</b> of all sev=2	0.17 of all sev=2	<b>0.73</b> of all sev=2	0.27 of all sev=2

**Πίνακας 5.7:** Συνολικά αποτελέσματα και για τις τρεις μεθόδους με βάση τον πίνακα

SEVERITY\_binary

Severity	K-Means	GaussianMixture	K-Medoids
Sev 1	86%-91%	74%-88%	77%-94%
Sev 2	65%-79%	79%-83%	62%-83%

**Πίνακας 5.8:** Ποσοστά ομαδοποίησης κάθε σοβαρότητας για τις τρεις μεθόδους με βάση τον πίνακα

SEVERITY\_binary

## 5.4 Μοντέλο 4: K-Nearest Neighbors

Για τα μοντέλα επιβλεπόμενης μάθησης που αναπτύχθηκαν θα παρουσιαστούν όλες οι εισοδοί που αξιοποιήθηκαν καθώς όλες οι εισοδοί παρουσιάζουν ενδιαφέρον.

Παρόλα αυτά θα αναλυθούν περαιτέρω οι εισοδοί που έδωσαν το υψηλότερο accuracy. Επιπλέον θα παρουσιαστούν χωριστά τα αποτελέσματα στην Περίπτωση 1 όπου για validation του μοντέλου χρησιμοποιήθηκε ο πίνακας SEVERITY (τιμές σοβαρότητας 1-3) και στην Περίπτωση 2 όπου για validation του μοντέλου χρησιμοποιήθηκε ο πίνακας SEVERITY\_binary (δυαδικές τιμές σοβαρότητας 1 και 2)

#### 5.4.1 Περίπτωση 1: SEVERITY

Όπως προαναφέρθηκε στην συγκεκριμένη περίπτωση χρησιμοποιήθηκε ο πίνακας SEVERITY για την επαλήθευση του μοντέλου ο οποίος περιέχει 3 τιμές για την σοβαρότητα:

4. Ελαφριά Ρευματοειδής
5. Ρευματοειδής Μέτριας Σοβαρότητας
6. Βαριά Ρευματοειδής

Στον παρακάτω πίνακα παρουσιάζονται οι εισοδοί του μοντέλου και το accuracy του μοντέλου για κάθε test size που δοκιμάστηκε. Πρέπει να σημειωθεί εδώ ότι δημιουργήθηκαν 3 ομάδες καθώς ο πίνακας SEVERITY περιέχει 3 τιμές για την σοβαρότητα.

Στην συνέχεια παρουσιάζονται χωριστά τα αποτελέσματα για ορισμένες εισόδους με καλή ακρίβεια.

INPUT	ACCURACY test_size=0.15	ACCURACY test_size=0.2	ACCURACY test_size=0.25
1. [{"sharp-score","DURATION","CCP","RF","HLA-DRB1"}]	80.5%	70.8%	70%
2. [{"sharp-score","AGE","CCP","RF","HLA-DRB1"}]	66.7%	62.5%	65%
3. [{"sharp-score","DURATION","CCPi","Rfi","HLA-DRB1"}]	58.3%	62.5%	65%
4. [{"sharp-score","DURATION","CCP","RF","HLA-DRB1","Rfi","CCPi"}]	75%	70.8%	73.3%
5. [{"sharp-score","CCP","RF","HLA-DRB1","Rfi","CCPi"}]	63.9%	64.6%	65%
6. [{"sharp-score","SMOKING","CCP","RF","HLA-DRB1"}]	72.2%	72.9%	73.3%
7. [{"sharp-score","DURATION","SMOKING","CCP","RF","HLA-DRB1"}]	72.2%	66.7%	70%
8. [{"DURATION","CCP","SMOKING","RF","HLA-DRB1"}]	66.7%	58.3%	68.3%
9. [{"CCP","RF","HLA-DRB1"}]	63.9%	60.4%	61.7%
10. [{"Rfi","CCPi","CCP","RF","HLA-DRB1"}]	63.9%	62.5%	58.3%
11. [{"Rfi","SMOKING","CCPi","CCP","RF","HLA-DRB1"}]	66.7%	64.6%	65%
12. [{"CCP","RF","HLA-DRB1","SMOKING"}]	66.7%	64.6%	63.3%
13. [{"sharp-score","SMOKING","CCPi","Rfi","HLA-DRB1"}]	58.3%	60.4%	61.7%
14. [{"sharp-score","CCPi","Rfi","HLA-DRB1"}]	61.1%	58.3%	61.7%

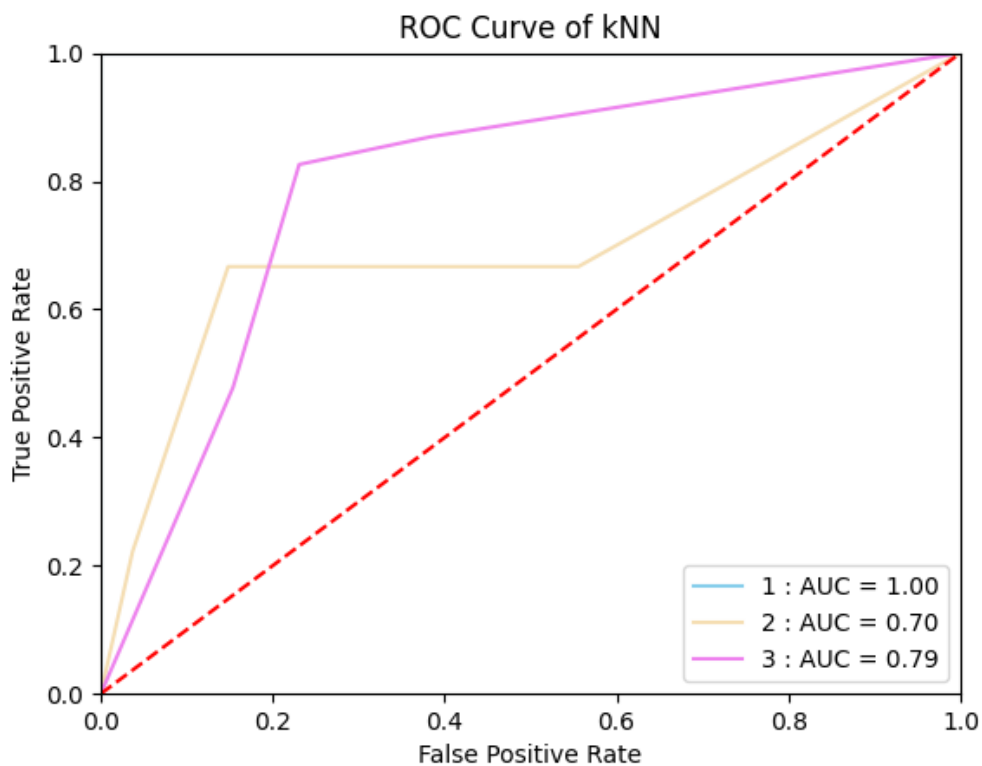
**Πίνακας 5.9:** Accuracy για k=3

5.4.1.1 Είσοδος 1 KNN :  $X = \text{workingDF}[["sharp-score", "DURATION", "CCP", "RF", "HLA-DRB1"]].\text{to\_numpy}()$

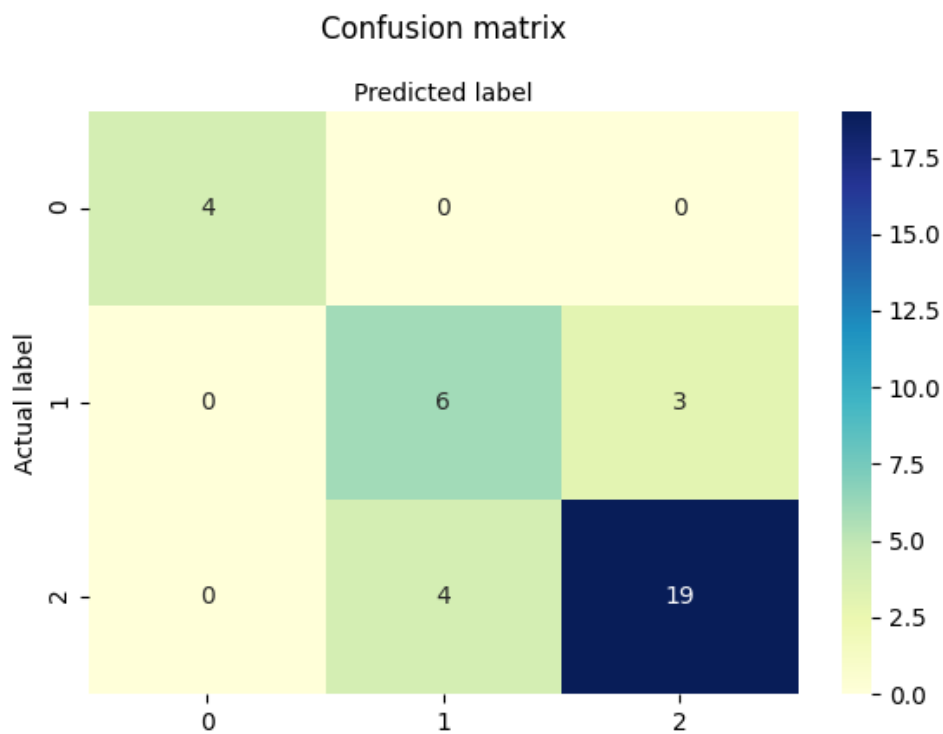
Παρακάτω παρουσιάζονται τα αποτελέσματα που προέκυψαν για  $\text{test\_size}=0.15$  που σημαίνει ότι το 85% του dataset χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου και το υπόλοιπο 15% για την επαλήθευση της κατηγοριοποίησης.

Δεδομένα Εκπαίδευσης	204 ασθενείς
Δεδομένα Επαλήθευσης	36 ασθενείς
Ακρίβεια	<b>0.8055555555555556</b>

Ακολουθούν οι γραφικές απεικονίσεις που προέκυψαν από την κατηγοριοποίηση.



Σχήμα 5.65 : ROC Curve για την Είσοδο 1 με  $\text{test\_size}=0.15$



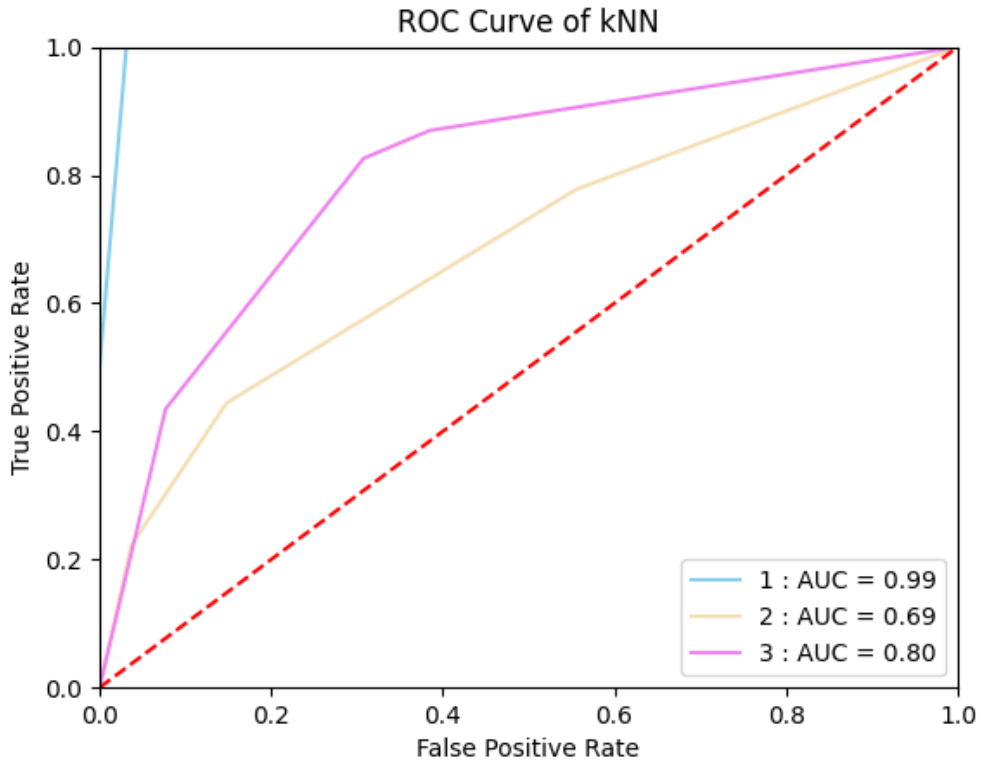
Σχήμα 5.66 : Πίνακας Σύγκρισης για την Είσοδο 1 με test\_size=0.15

5.4.1.2 **Είσοδος 4 KNN** :  $X = \text{workingDF}[["sharp-score", "DURATION", "CCP", "RF", "HLA-DRB1", "Rfi", "CCPi"]].to\_numpy()$

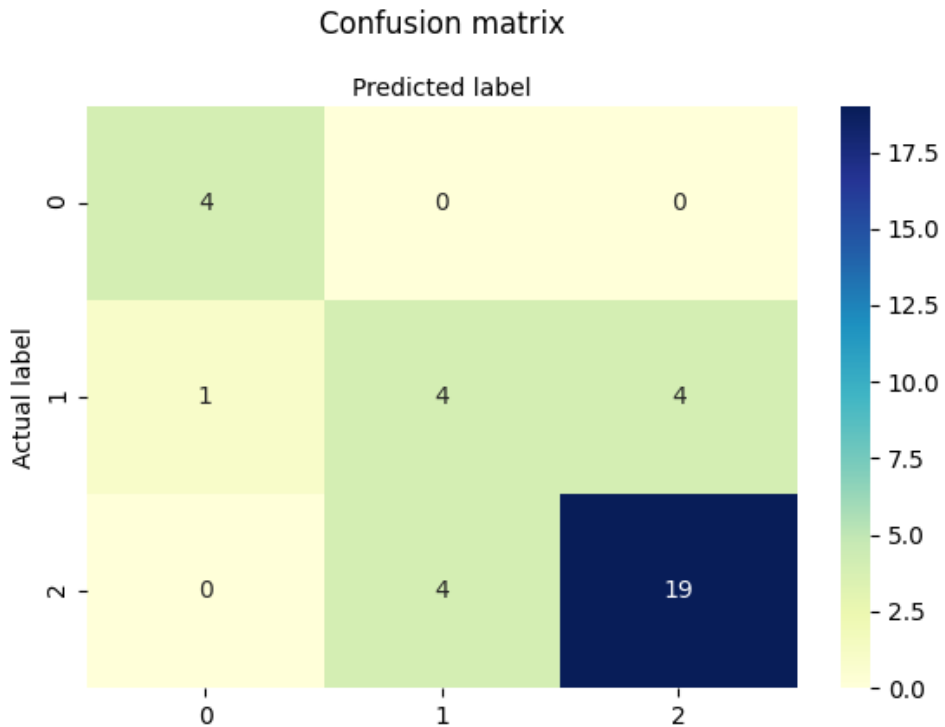
Παρακάτω παρουσιάζονται τα αποτελέσματα που προέκυψαν για test\_size=0.15

Δεδομένα Εκπαίδευσης	204 ασθενείς
Δεδομένα Επαλήθευσης	36 ασθενείς
Ακρίβεια	<b>0.75</b>

Ακολουθούν οι γραφικές απεικονίσεις που προέκυψαν από την κατηγοριοποίηση.



Σχήμα 5.67 : ROC Curve για την Είσοδο 4 με test\_size=0.15



Σχήμα 5.68 : Πίνακας Σύγκρισης για την Είσοδο 4 με test\_size=0.15

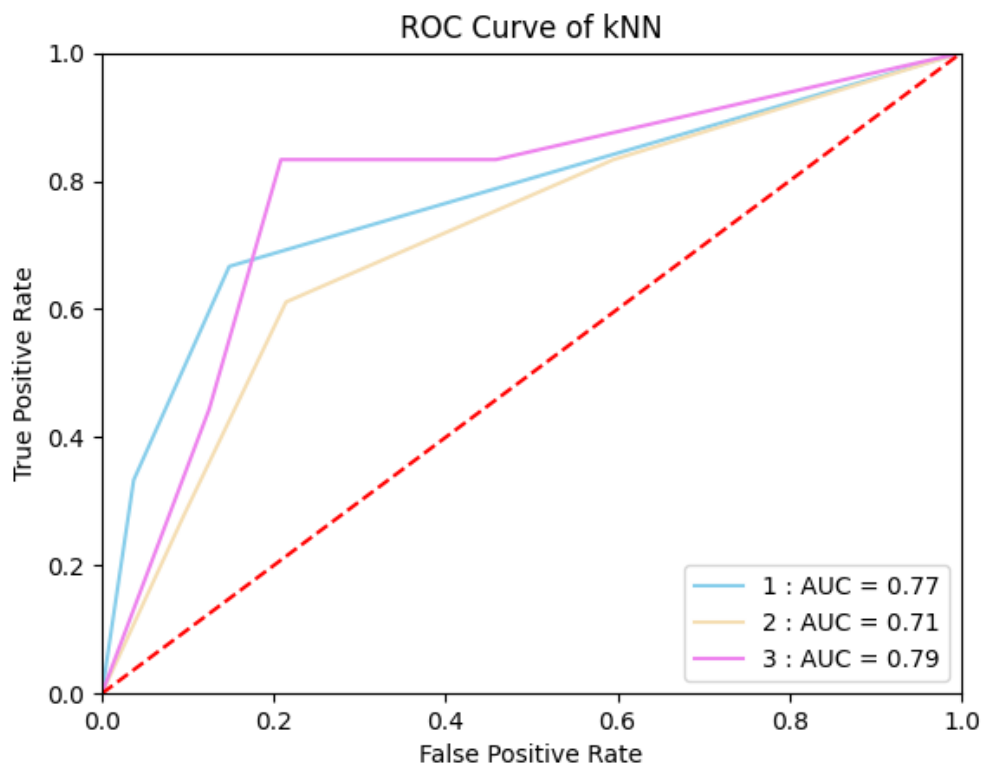


5.4.1.3 Είσοδος 6 KNN :  $X = \text{workingDF}[["sharp-score", "SMOKING", "CCP", "RF", "HLA-DRB1"]].\text{to\_numpy}()$

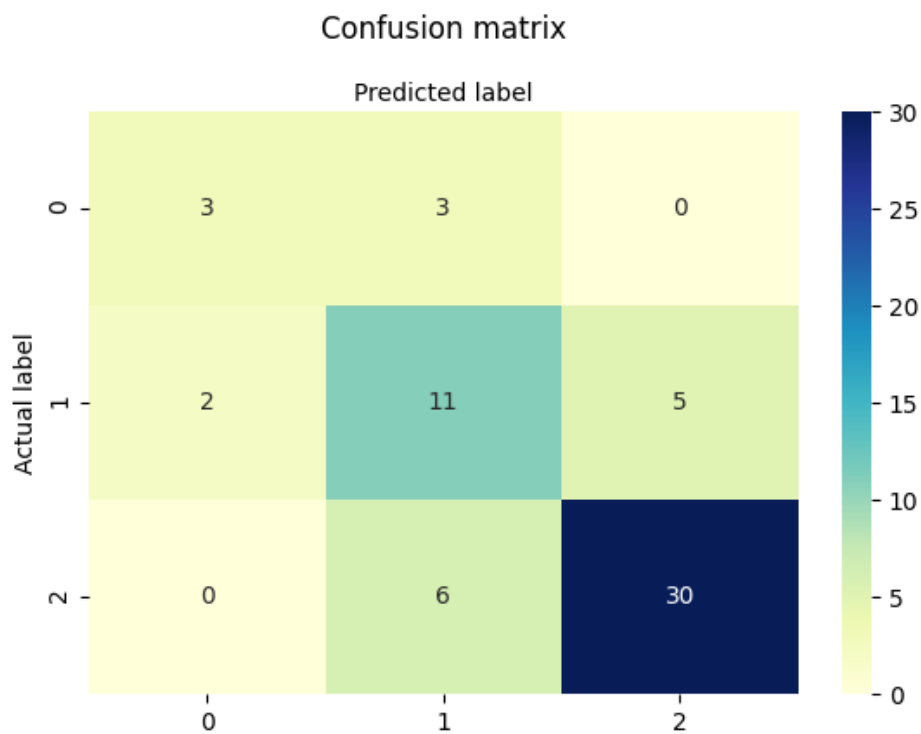
Παρακάτω παρουσιάζονται τα αποτελέσματα που προέκυψαν για  $\text{test\_size}=0.25$  που σημαίνει ότι το 75% του dataset χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου και το υπόλοιπο 25% για την επαλήθευση της κατηγοριοποίησης.

Δεδομένα Εκπαίδευσης	180 ασθενείς
Δεδομένα Επαλήθευσης	60 ασθενείς
Ακρίβεια	<b>0.7333333333333333</b>

Ακολουθούν οι γραφικές απεικονίσεις που προέκυψαν από την κατηγοριοποίηση.



Σχήμα 5.69 : ROC Curve για την Είσοδο 6 με  $\text{test\_size}=0.25$



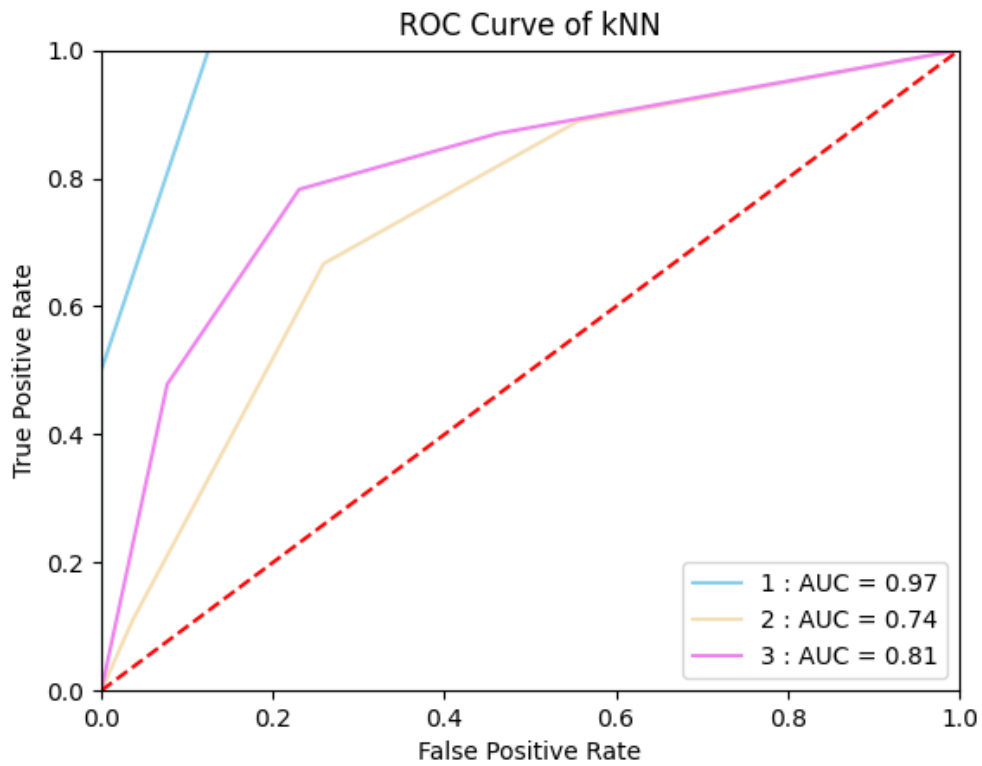
Σχήμα 70 : Πίνακας Σύγκρισης για την Είσοδο 6 με test\_size=0.25

5.4.1.4 Είσοδος 7 KNN :  $X = \text{workingDF}[["sharp-score", "DURATION", "SMOKING", "CCP", "RF", "HLA-DRB1"]].to\_numpy()$

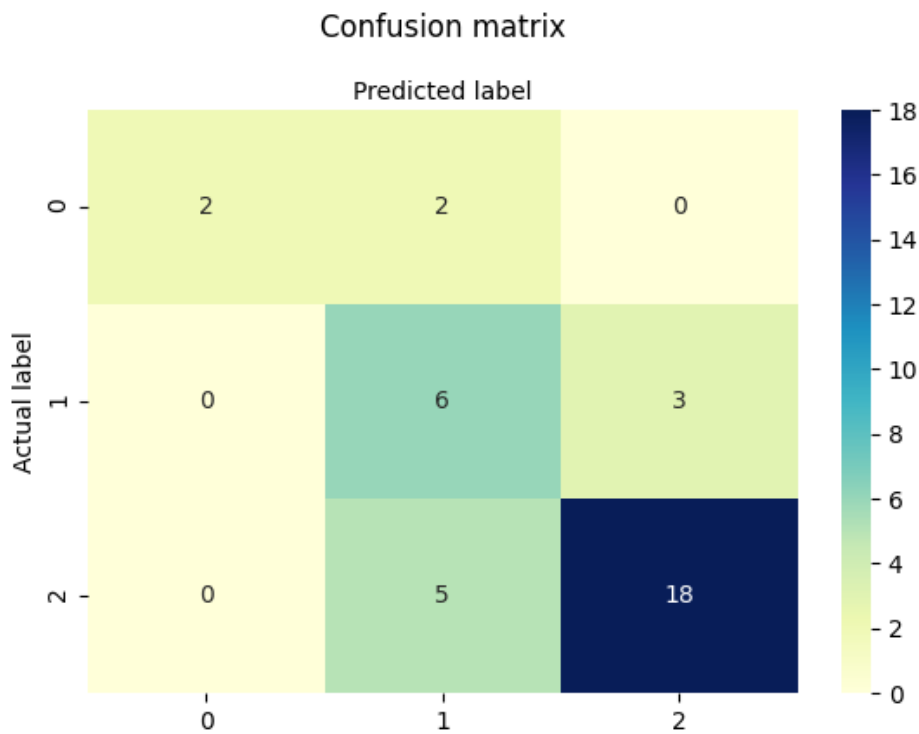
Παρακάτω παρουσιάζονται τα αποτελέσματα που προέκυψαν για test\_size=0.15.

Δεδομένα Εκπαίδευσης	204 ασθενείς
Δεδομένα Επαλήθευσης	36 ασθενείς
Ακρίβεια	<b>0.7222222222222222</b>

Ακολουθούν οι γραφικές απεικονίσεις που προέκυψαν από την κατηγοριοποίηση.



Σχήμα 5.71 : ROC Curve για την Είσοδο 7 με test\_size=0.15



Σχήμα 5.72 : Πίνακας Σύγκρισης για την Είσοδο 7 με test\_size=0.15

### 5.4.2 Περίπτωση 2: SEVERITY\_binary

Όπως προαναφέρθηκε στην συγκεκριμένη περίπτωση χρησιμοποιήθηκε ο πίνακας SEVERITY\_binary για την επαλήθευση του μοντέλου ο οποίος περιέχει 2 τιμές για την σοβαρότητα:

3. Ελαφριά Ρευματοειδής
4. Βαριά Ρευματοειδής

Στον παρακάτω πίνακα παρουσιάζονται οι εισοδοί του μοντέλου και το accuracy του μοντέλου για κάθε test size που δοκιμάστηκε. Πρέπει να σημειωθεί εδώ ότι δημιουργήθηκαν 2 ομάδες καθώς ο πίνακας SEVERITY περιέχει 2 τιμές για την σοβαρότητα.

Στην συνέχεια παρουσιάζονται χωριστά τα αποτελέσματα για ορισμένες εισόδους με καλή ακρίβεια.

INPUT	ACCURACY test_size=0.15	ACCURACY test_size=0.2	ACCURACY test_size=0.25
1. [{"sharp-score", "DURATION", "CCP", "RF", "HLA-DRB1"}]	82.2%	86.4%	86.5%
2. [{"sharp-score", "AGE", "CCP", "RF", "HLA-DRB1"}]	80%	84.7%	85.1%
3. [{"sharp-score", "DURATION", "CCPi", "Rfi", "HLA-DRB1"}]	86.7%	89.8%	87.8%
4. [{"sharp-score", "DURATION", "CCP", "RF", "HLA-DRB1", "Rfi", "CCPi"}]	80%	86.4%	83.7%
5. [{"sharp-score", "CCP", "RF", "HLA-DRB1", "Rfi", "CCPi"}]	84.4%	86.4%	86.5%
6. [{"sharp-score", "SMOKING", "CCP", "RF", "HLA-DRB1"}]	77.7%	81.3%	79.7%
7. [{"sharp-score", "DURATION", "SMOKING", "CCP", "RF", "HLA-DRB1"}]	80%	84.8%	85.1%
8. [{"DURATION", "CCP", "SMOKING", "RF", "HLA-DRB1"}]	75.5%	81.3%	81%
9. [{"CCP", "RF", "HLA-DRB1"}]	80%	83%	82.4%
10. [{"Rfi", "CCPi", "CCP", "RF", "HLA-DRB1"}]	80%	83%	80%
11. [{"Rfi", "SMOKING", "CCPi", "CCP", "RF", "HLA-DRB1"}]	75.5%	81.3%	79.7%
12. [{"CCP", "RF", "HLA-DRB1", "SMOKING"}]	75.5%	79.7%	82.4%
13. [{"sharp-score", "SMOKING", "CCPi", "Rfi", "HLA-DRB1"}]	82.2%	84.7%	83.7%
14. [{"sharp-score", "CCPi", "Rfi", "HLA-DRB1"}]	84.4%	86.4%	86.5%

**Πίνακας 5.10:** Accuracy για k=2

Όπως φαίνεται από τα αποτελέσματα στην περίπτωση της σοβαρότητας με δυαδικές τιμές προέκυψε καλύτερη κατηγοριοποίηση. Παρόλα αυτά οι ίδιες εισοδοί που παρουσίαζαν τα καλύτερα αποτελέσματα στα μοντέλα της προηγούμενης παραγράφου παρουσιάζουν τα καλύτερα αποτελέσματα και στην δυαδική κατηγοριοποίηση.

Εδώ προέκυψε και το βέλτιστο μοντέλο για την είσοδο [{"sharp-score", "DURATION", "CCPi", "Rfi", "HLA-DRB1"}] που παρουσίασε ακρίβεια 90% το αντίστοιχο μοντέλο για τρεις διαβαθμίσεις σοβαρότητας είχε δώσει αποτέλεσμα 62.5 %.

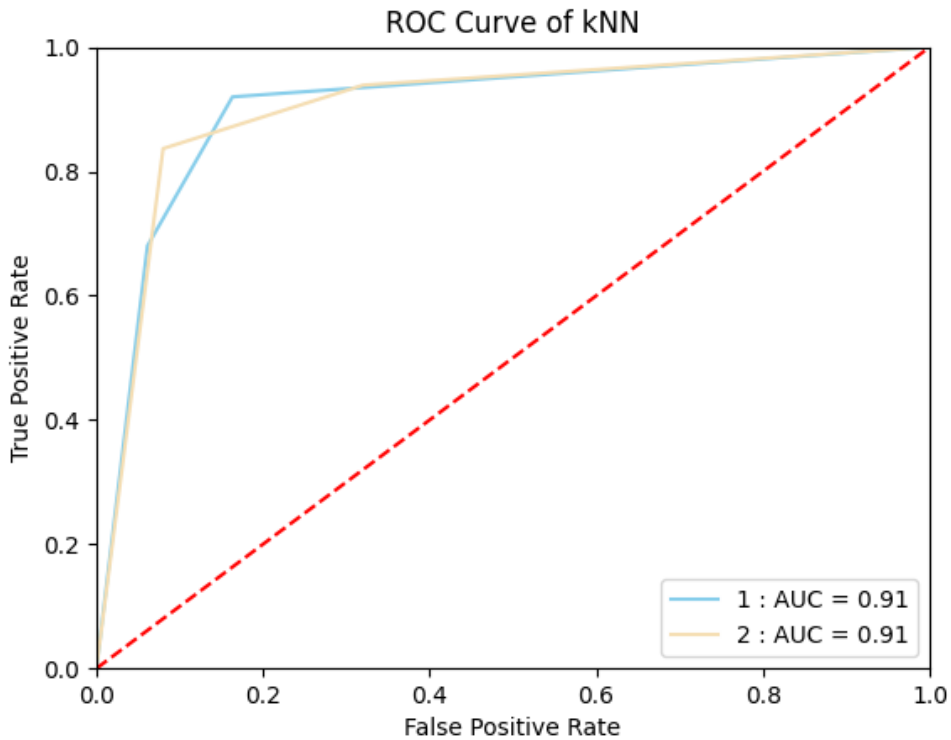
Επομένως δεν υπάρχει απόλυτη ταύτιση των μοντέλων που αναπτύχθηκαν για τρεις διαβαθμίσεις της σοβαρότητας και για δύο διαβαθμίσεις της σοβαρότητας.

5.4.2.1 **Είσοδος 1 KNN** :  $X = \text{workingDF}[["sharp-score", "DURATION", "CCP", "RF", "HLA-DRB1"]].\text{to\_numpy}()$

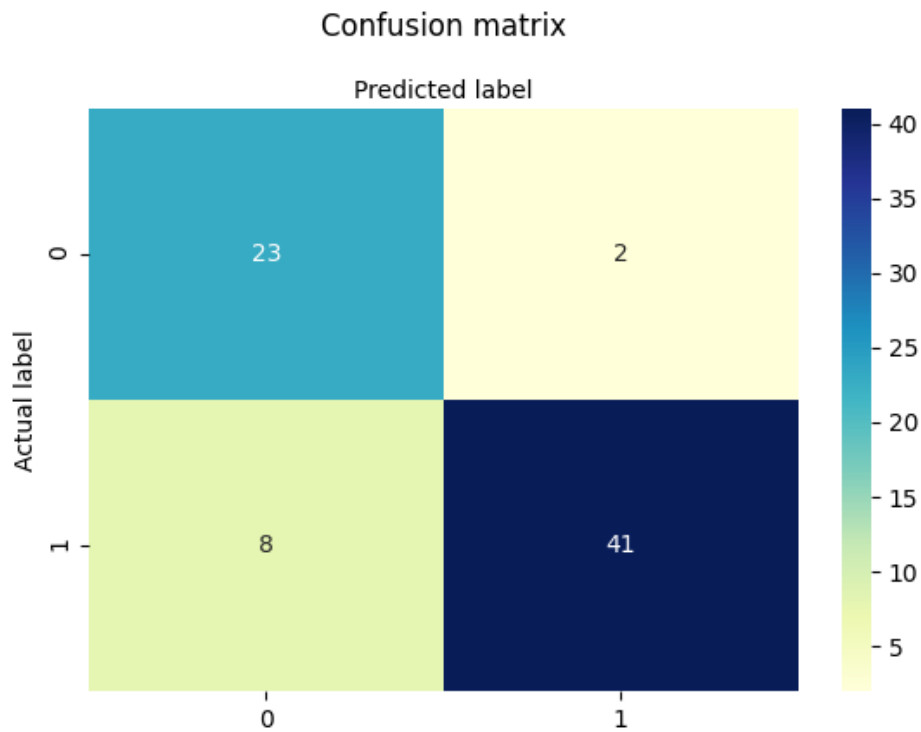
Παρακάτω παρουσιάζονται τα αποτελέσματα που προέκυψαν για  $\text{test\_size}=0.25$

Δεδομένα Εκπαίδευσης	221 ασθενείς
Δεδομένα Επαλήθευσης	74 ασθενείς
Ακρίβεια	<b>0.8648648648648649</b>

Ακολουθούν οι γραφικές απεικονίσεις που προέκυψαν από την κατηγοριοποίηση



**Σχήμα 5.73** : ROC Curve για την Είσοδο 1 με  $\text{test\_size}=0.25$



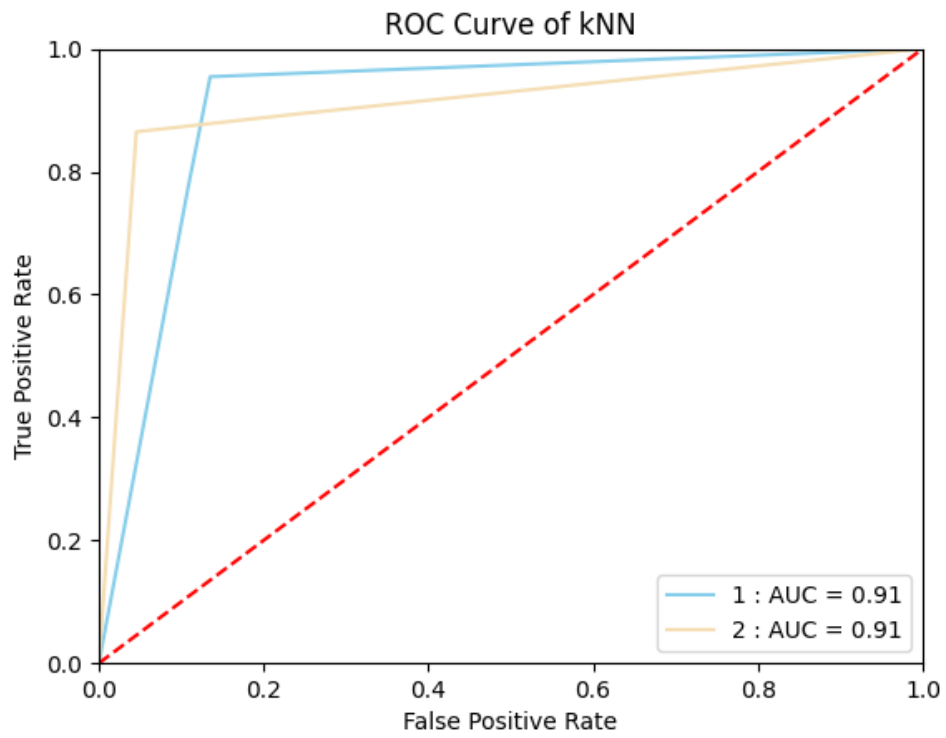
**Σχήμα 5.74** : Πίνακας Σύγκρισης για την Είσοδο 1 με test\_size=0.25

5.4.2.2 **Είσοδος 3 KNN** :  $X = \text{workingDF}[["sharp-score", "DURATION", "CCPi", "Rfi", "HLA-DRB1"]].to\_numpy()$

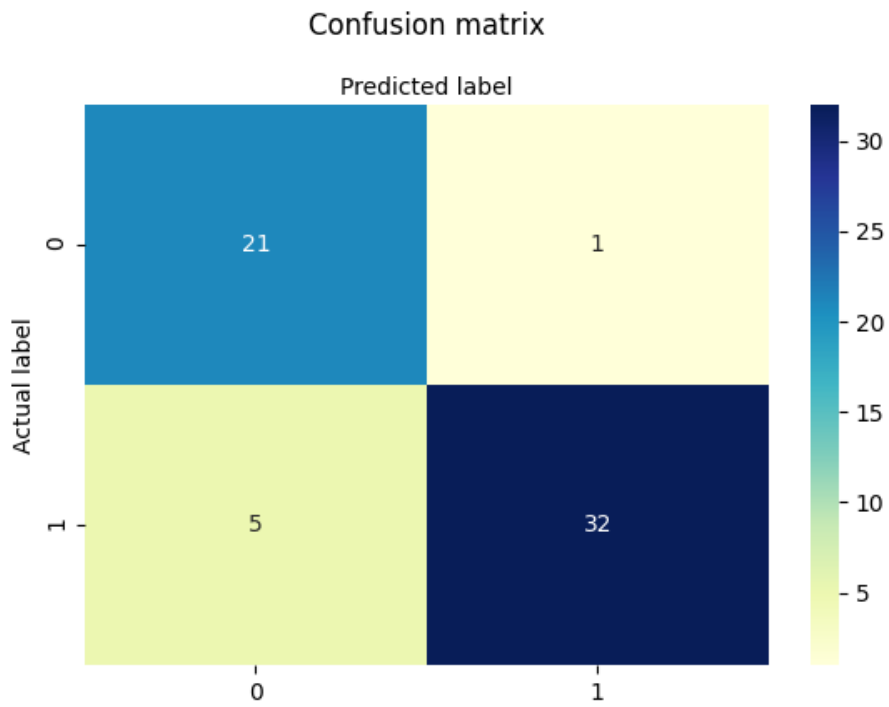
Παρακάτω παρουσιάζονται τα αποτελέσματα που προέκυψαν για test\_size=0.2

Δεδομένα Εκπαίδευσης	236 ασθενείς
Δεδομένα Επαλήθευσης	59 ασθενείς
Ακρίβεια	<b>0.8983050847457628</b>

Ακολουθούν οι γραφικές απεικονίσεις που προέκυψαν από την κατηγοριοποίηση



Σχήμα 5.75 : ROC Curve για την Είσοδο 3 με test\_size=0.2



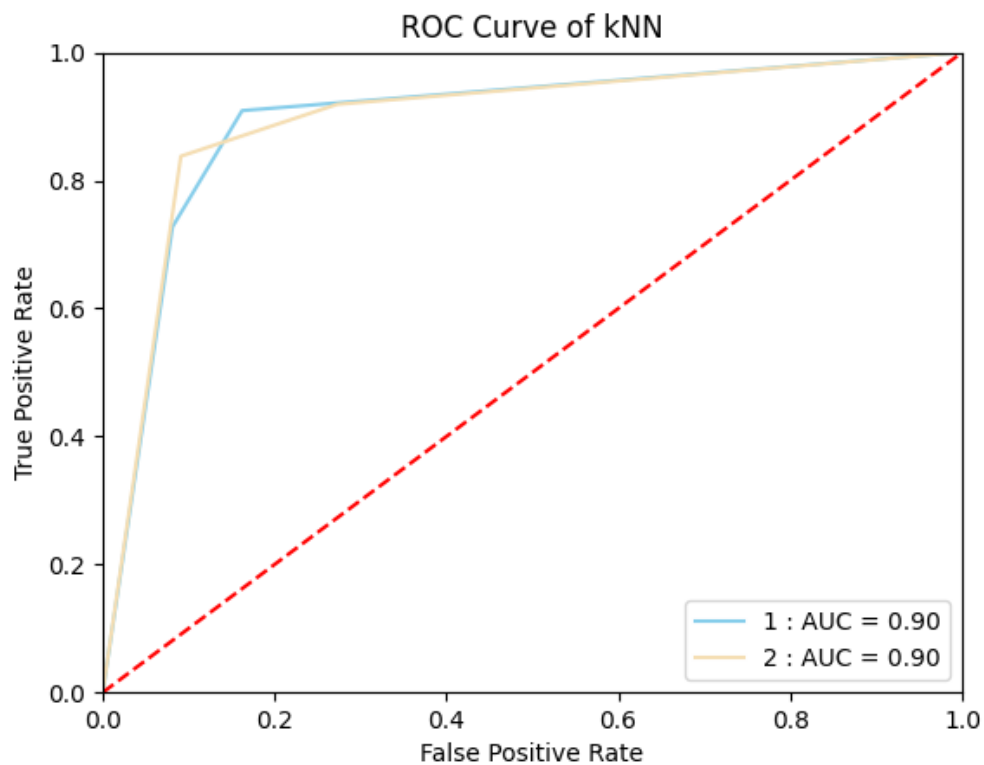
Σχήμα 5.76 : Πίνακας Σύγκρισης για την Είσοδο 3 με test\_size=0.2

5.4.2.3 Είσοδος 4 KNN :  $X = \text{workingDF}[["sharp-score", "DURATION", "CCP", "RF", "HLA-DRB1", "Rfi", "CCPi"]].to\_numpy()$

Παρακάτω παρουσιάζονται τα αποτελέσματα που προέκυψαν για  $\text{test\_size}=0.2$

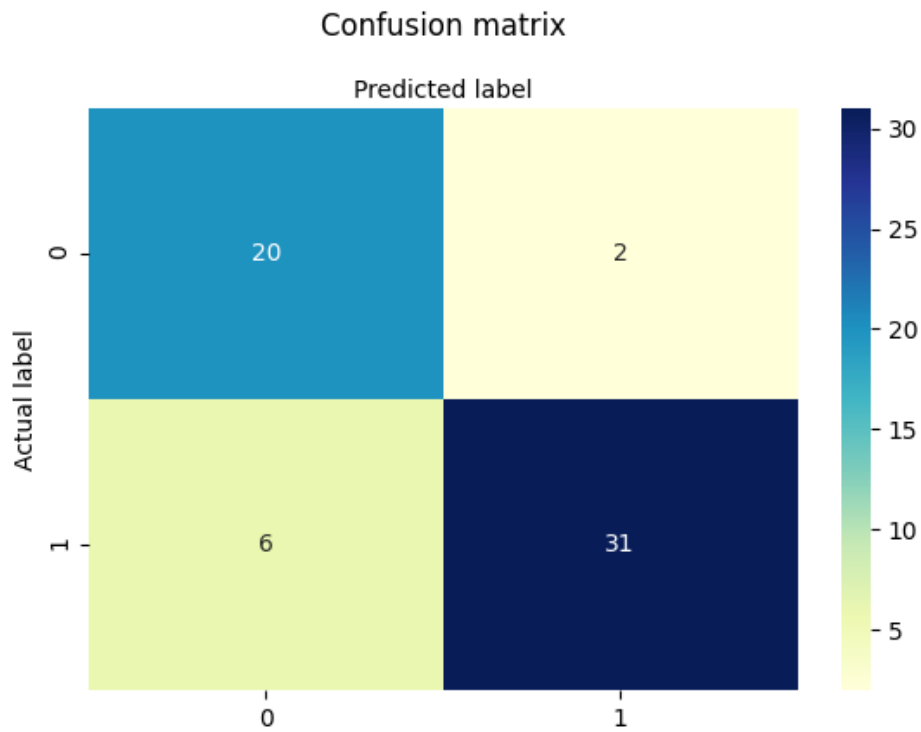
Δεδομένα Εκπαίδευσης	236 ασθενείς
Δεδομένα Επαλήθευσης	59 ασθενείς
Ακρίβεια	<b>0.864406779661017</b>

Ακολουθούν οι γραφικές απεικονίσεις που προέκυψαν από την κατηγοριοποίηση



Σχήμα 5.77 : ROC Curve για την Είσοδο 4 με  $\text{test\_size}=0.2$





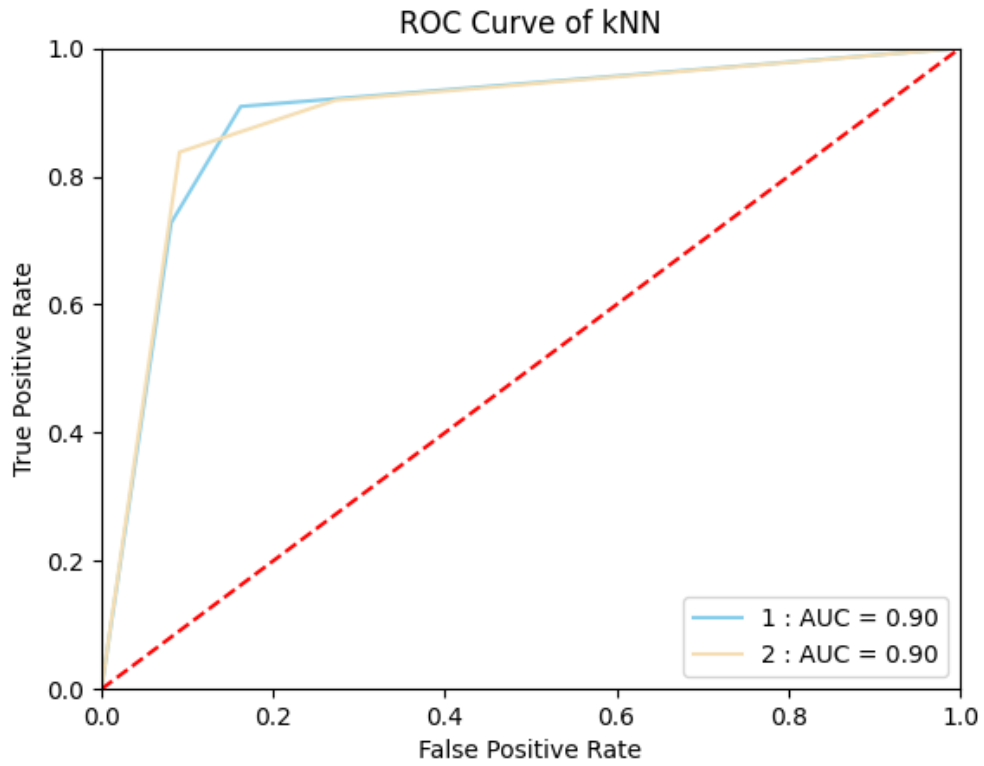
**Σχήμα 5.78** : Πίνακας Σύγκρισης για την Είσοδο 4 με `test_size=0.2`

5.4.2.4 **Είσοδος 7 KNN** : `X = workingDF[["sharp-score", "DURATION", "CCP", "RF", "HLA-DRB1", "Rfi", "CCPi"]].to_numpy()`

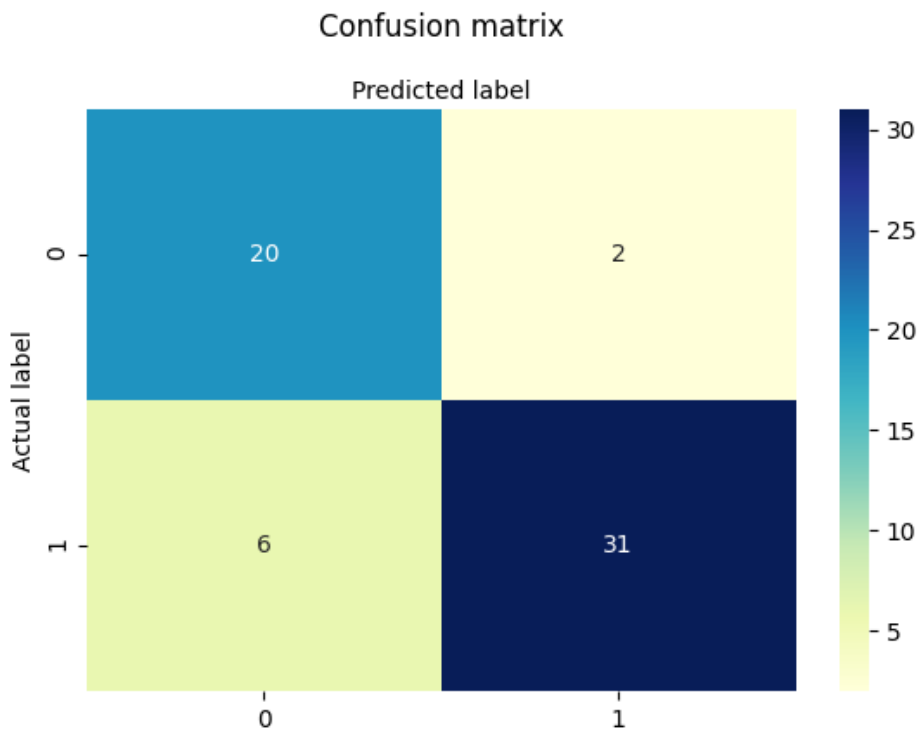
Παρακάτω παρουσιάζονται τα αποτελέσματα που προέκυψαν για `test_size=0.2`

Δεδομένα Εκπαίδευσης	236 ασθενείς
Δεδομένα Επαλήθευσης	59 ασθενείς
<b>Ακρίβεια</b>	<b>0.864406779661017</b>

Ακολουθούν οι γραφικές απεικονίσεις που προέκυψαν από την κατηγοριοποίηση



Σχήμα 5.79 : ROC Curve για την Είσοδο 7 με test\_size=0.2



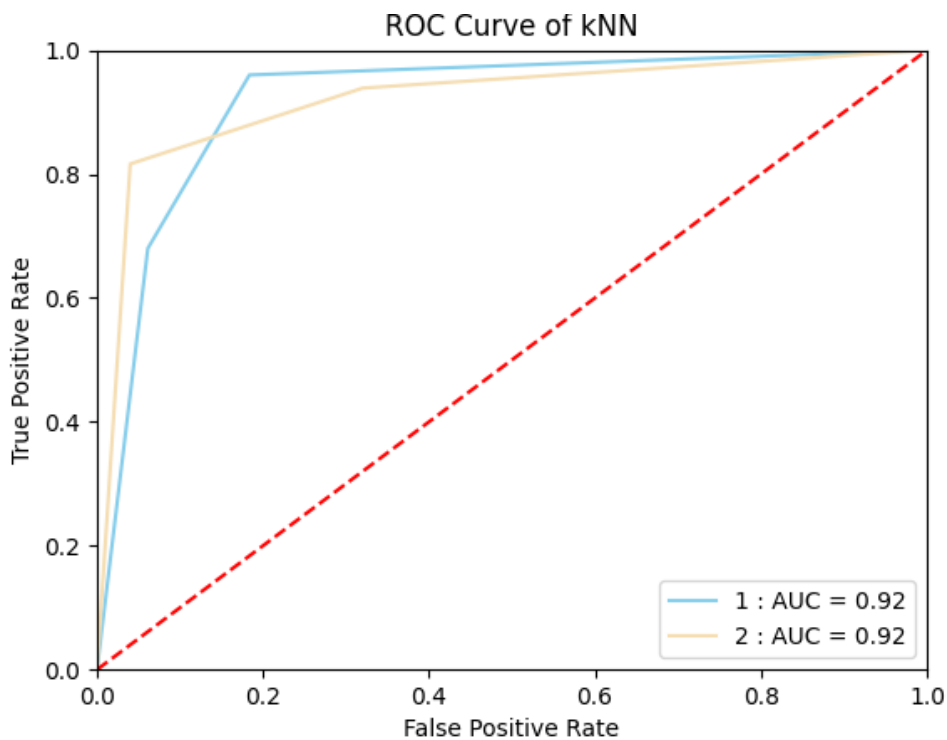
Σχήμα 5.80 : Πίνακας Σύγκρισης για την Είσοδο 7 με test\_size=0.2

5.4.2.5 **Είσοδος 14 KNN:**  $X = \text{workingDF}[["sharp-score", "CCPi", "Rfi", "HLA-DRB1"]].\text{to\_numpy}()$

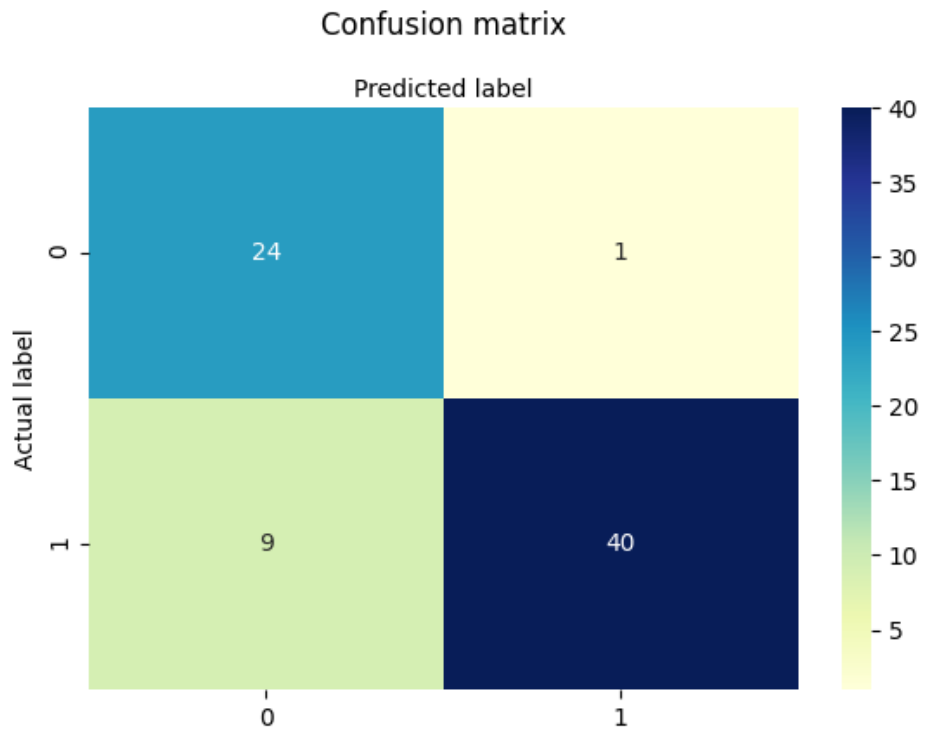
Παρακάτω παρουσιάζονται τα αποτελέσματα που προέκυψαν για  $\text{test\_size}=0.25$

Δεδομένα Εκπαίδευσης	221 ασθενείς
Δεδομένα Επαλήθευσης	74 ασθενείς
Ακρίβεια	<b>0.8648648648648649</b>

Ακολουθούν οι γραφικές απεικονίσεις που προέκυψαν από την κατηγοριοποίηση



**Σχήμα 5.81 :** ROC Curve για την Είσοδο 14 με  $\text{test\_size}=0.25$



Σχήμα 5.82 : Πίνακας Σύγκρισης για την Είσοδο 14 με test\_size=0.25

# 6

## *Συμπεράσματα & Επίλογος*

### *6.1 Συμπεράσματα*

Στην παρούσα διπλωματική εργασία επιχειρήθηκε η δημιουργία μοντέλων με σκοπό την εξαγωγή συμπερασμάτων όσον αφορά το επίπεδο σοβαρότητας της Ρευματοειδούς Αρθρίτιδας. Πηγή έμπνευσης αποτέλεσε το διαγνωστικό εργαλείο DAS28 που χρησιμοποιείται για την ταξινόμηση της ενεργότητας. Έχει μάλιστα δημιουργηθεί εφαρμογή στην οποία οι ιατροί μπορούν να δώσουν τιμές για 4 χαρακτηριστικά και να λάβουν σαν αποτέλεσμα μια τιμή που δείχνει αν ο ασθενής έχει ρευματοειδή σε ύφεση ή σε έξαρση. Θα ήταν χρήσιμο λοιπόν να αναπτυχθεί ένα αντίστοιχο εργαλείο που να μπορεί να δώσει συμπεράσματα όσον αφορά την σοβαρότητα της νόσου. Άλλωστε το φαρμακευτικό σχήμα που θα χορηγηθεί στον ασθενή προκύπτει τόσο από την ενεργότητα όσο και από την σοβαρότητα της νόσου. Επιπλέον, όπως φαίνεται και από τα δεδομένα που επεξεργάστηκαν για την δημιουργία των μοντέλων, δεν υπάρχει ταύτιση των τιμών της ενεργότητας και της σοβαρότητας για κάθε ασθενή. Για παράδειγμα ένας ασθενής μπορεί να έχει Ρευματοειδή Αρθρίτιδα σε έξαρση αλλά να μην πρόκειται για μια σοβαρή ρευματοειδή και αντίστοιχα κάποιος ασθενής να έχει σοβαρή ρευματοειδή και η νόσος παρόλα αυτά να βρίσκεται σε ύφεση. Συνδυασμός των δύο αυτών παραγόντων είναι που θα καθορίσει το φαρμακευτικό σχήμα που θα ακολουθήσει ο ιατρός.

Η έρευνα που έγινε στα πλαίσια της διπλωματικής καταδεικνύει ότι με χρήση συγκεκριμένων χαρακτηριστικών μπορεί να προκύψουν αρκετά αξιόπιστα αποτελέσματα όσον αφορά την σοβαρότητα. Προέκυψε άλλωστε μοντέλο που λαμβάνοντας σαν είσοδο 5

χαρακτηριστικά μπορεί με 90% ακρίβεια να υποδείξει αν ο ασθενής έχει ελαφριά ή σοβαρή ρευματοειδή.

Τα μοντέλα που παρουσιάστηκαν στο Κεφάλαιο 5 είναι αυτά που είχαν τα καλύτερα αποτελέσματα, συνολικά δοκιμάστηκαν εκατοντάδες συνδυασμοί μεταβλητών εισόδου για όλα τα μοντέλα. Τελικά όμως οι εισοδοί που έδιναν καλή ομαδοποίηση ή κατηγοριοποίηση στα μοντέλα που αναπτύχθηκαν ήταν συνδυασμός των παρακάτω βασικών μεταβλητών

Κρίσιμες Μεταβλητές
1. sharp-score
2. RF - Rfi
3. CCP- CCPi
4. HLA_DRB1
5. SMOKING
6. DURATION

Όπως παρατηρήθηκε από τα μοντέλα που αναπτύχθηκαν, την μεγαλύτερη βαρύτητα είχαν οι πρώτες 4 μεταβλητές. Το παραπάνω συμπέρασμα που προέκυψε από την ανάπτυξη των μοντέλων συμφωνεί με τις γνωματεύσεις των γιατρών καθώς κατά κύριο λόγο οι παραπάνω παράγοντες εξετάζονται ώστε να καθοριστεί η σοβαρότητα της νόσου κάθε ασθενή. Επομένως τα μοντέλα επιβεβαιώνουν την πραγματικότητα το οποίο σημαίνει ότι μπορεί στο μέλλον να αναπτυχθεί ένα εργαλείο που να μπορεί με 4-5 βασικά χαρακτηριστικά να προβλέψει την σοβαρότητα και να αποτελέσει πηγή επαλήθευσης της γνωμάτευσης κάθε γιατρού.

Ακολουθεί σύντομος σχολιασμός των μοντέλων που προέκυψαν.

### **6.1.1 Μοντέλα Μη- Εποπτευόμενης Μηχανικής Μάθησης**

Αναπτύχθηκαν τρεις διαφορετικοί αλγόριθμοι με χρήση των ίδιων ακριβώς εισόδων για μπορέσει να υπάρξει σύγκριση των μοντέλων που προέκυψαν.

Αρχικά όπως προέκυψε από την μέθοδο Elbow και την Silhouette ανάλυση και έπειτα από αρκετές δοκιμές ο βέλτιστος αριθμός ομάδων ώστε να επιτευχθεί η καλύτερη δυνατή ομαδοποίηση ήταν δύο. Σε πρώτη φάση χρησιμοποιήθηκε ως πίνακας επαλήθευσης-επικύρωσης της ομαδοποίησης που προέκυψε ο πίνακας SEVERITY που περιέχει τρεις τιμές για την σοβαρότητα (1 , 2 και 3 ).

Τα μοντέλα που προέκυψαν όπως φαίνεται από τους πίνακες 5.1 , 5.3 και 5.5 παρουσίασαν αρκετά καλή ομαδοποίηση.

Συγκεκριμένα με βάση των πίνακα SEVERITY η καλύτερη ομαδοποίηση με το **K-Means** προέκυψε για τις παρακάτω 2 εισόδους:

ΕΙΣΟΔΟΣ	ΟΜΑΔΑ 1	ΟΜΑΔΑ 2
4. [ ["sharp-score", "CCP", "CCPi", "RF", "Rfi", "HLA-DRB1" ] ]	0.05 of all sev=1 0.29 of all sev=2 <b>0.80</b> of all sev=3	<b>0.95</b> of all sev=1 <b>0.71</b> of all sev=2 0.20 of all sev=3
7. [ ["CCP", "RF" ] ]	<b>0.95</b> of all sev=1 <b>0.68</b> of all sev=2 0.20 of all sev=3	0.05 of all sev=1 0.32 of all sev=2 <b>0.80</b> of all sev=3

Το μοντέλο που προέκυψε από την είσοδο 4 εμφανίζει καλύτερη ομαδοποίηση για την σοβαρότητα 2 σε σχέση με το μοντέλο που προέκυψε από την είσοδο 7. Όσον αφορά την σοβαρότητα 1 και 3, παρουσιάζουν την ίδια ακριβώς ομαδοποίηση.

Με βάση των πίνακα SEVERITY η καλύτερη ομαδοποίηση με το **GaussianMixture** προέκυψε για τις παρακάτω 2 εισόδους:

ΕΙΣΟΔΟΣ	ΟΜΑΔΑ 1	ΟΜΑΔΑ 2
1. [ ["sharp-score", "CCPi", "RF", "HLA-DRB1" ] ]	<b>0.89</b> of all sev=1 <b>0.71</b> of all sev=2 0.19 of all sev=3	0.11 of all sev=1 0.29 of all sev=2 <b>0.81</b> of all sev=3
7. [ ["CCP", "RF" ] ]	<b>0.89</b> of all sev=1 <b>0.63</b> of all sev=2 0.16 of all sev=3	0.11 of all sev=1 0.37 of all sev=2 <b>0.84</b> of all sev=3

Το μοντέλο που προέκυψε από την είσοδο 1 εμφανίζει καλύτερη ομαδοποίηση για την σοβαρότητα 2 σε σχέση με το μοντέλο που προέκυψε από την είσοδο 7. Όσον αφορά την σοβαρότητα 1 και 3 η είσοδος 7 παρουσιάζει ελαφρώς καλύτερη ομαδοποίηση. Συνολικά εμφανίζουν αντίστοιχα αποτελέσματα. Σε σχέση με τα μοντέλα του K-Means η GaussianMixture εμφανίζει πιο ισορροπημένα αποτελέσματα. Συγκεκριμένα ομαδοποιεί εξίσου καλά την σοβαρότητα 1 και 3 σε αντίθεση με το K-Means που ομαδοποιεί καλύτερα την σοβαρότητα 1.

Με βάση των πίνακα SEVERITY η καλύτερη ομαδοποίηση με το **K-Medoids** προέκυψε για τις παρακάτω 2 εισόδους:

ΕΙΣΟΔΟΣ	ΟΜΑΔΑ 1	ΟΜΑΔΑ 2
4. [ ["sharp-score", "CCP", "CCPi", "RF", "Rfi", "HLA-DRB1"] ]	0.11 of all sev=1 0.34 of all sev=2 <b>0.85</b> of all sev=3	<b>0.89</b> of all sev=1 <b>0.66</b> of all sev=2 0.15 of all sev=3
8. [ ["sharp-score", "CCP", "RF"] ]	0.05 of all sev=1 0.24 of all sev=2 <b>0.75</b> of all sev=3	<b>0.95</b> of all sev=1 <b>0.76</b> of all sev=2 0.25 of all sev=3

Το μοντέλο που προέκυψε από την είσοδο 8 εμφανίζει καλύτερη ομαδοποίηση για την σοβαρότητα 1 και 2 σε σχέση με το μοντέλο που προέκυψε από την είσοδο 4. Όσον αφορά την σοβαρότητα 3, η είσοδος 4 παρουσιάζει σαφώς καλύτερη ομαδοποίηση. Συνολικά καλύτερη ομαδοποίηση προκύπτει με την είσοδο 8 και σε σχέση με τα μοντέλα K-Means και GaussianMixture, η μέθοδος K-Medoids εμφανίζει ελαφρώς χειρότερα αποτελέσματα.

Με βάση τα παραπάνω βασικοί παράμετροι που συμβάλουν στην καλή ομαδοποίηση είναι οι εξής: sharp-score, RF, CCP, HLA\_DRB1.

Σε δεύτερη φάση χρησιμοποιήθηκε ως πίνακα επαλήθευσης- επικύρωσης της ομαδοποίησης που προέκυψε ο πίνακας SEVERITY\_binary που περιέχει δύο τιμές για την σοβαρότητα (1 και 2).

Τα μοντέλα που προέκυψαν όπως φαίνεται από τους πίνακες 5.2, 5.4 και 5.6 παρουσίασαν ακόμα καλύτερη ομαδοποίηση από τα αποτελέσματα που προέκυψαν με βάση τον πίνακα SEVERITY.

Παρακάτω παρουσιάζονται τα καλύτερα αποτελέσματα που εντοπίστηκαν από τον πίνακα 5.7. Συγκεκριμένα με πράσινο είναι για κάθε είσοδο που έδωσε καλά αποτελέσματα η μέθοδος μη-επιβλεπόμενης μηχανικής μάθησης που χρησιμοποιήθηκε.

Όπως φαίνεται παρακάτω η είσοδος 1 έδωσε καλά αποτελέσματα τόσο με χρήση του K-Means όσο και με χρήση του K-Medoids. Το μοντέλο που προέκυψε με K-Means παρουσιάζει ελαφρώς καλύτερη ομαδοποίηση για την σοβαρότητα 1. Συνολικά όμως το καλύτερο μοντέλο για την είσοδο 1 προέκυψε με χρήση της K-Medoids.

Μια επιπλέον είσοδος που έδωσε καλά αποτελέσματα ήταν η είσοδος 4. Συνολικά το καλύτερο μοντέλο για την είσοδο 4 προέκυψε με χρήση της K-Medoids καθώς ήταν το μόνο



μοντέλο που παρουσίασε εξίσου καλή ομαδοποίηση τόσο για τη σοβαρότητα 1 όσο και για τη σοβαρότητα 2.

Τέλος για λόγους πληρότητας αναφέρονται ως αρκετά καλά τα αποτελέσματα που προέκυψαν με την είσοδο 7 με χρήση της GaussianMixture που έδωσε απόλυτα ισορροπημένη ομαδοποίηση για τις δύο σοβαρότητες. Τέλος τα αποτελέσματα της είσοδου 8 με χρήση της K-medoids που πέτυχαν την καλύτερη ομαδοποίηση από όλα τα μοντέλα για την σοβαρότητα 1 στο 94%

	K-Means		GaussianMixture		K-Medoids	
	GROUP 0	GROUP 1	GROUP 0	GROUP 1	GROUP 0	GROUP 1
<b>1.</b>	0.09 of all sev=1	<b>0.91</b> of all sev=1	0.23 of all sev=1	<b>0.77</b> of all sev=1	0.11 of all sev=1	<b>0.89</b> of all sev=1
	<b>0.74</b> of all sev=2	0.26 of all sev=2	<b>0.79</b> of all sev=2	0.21 of all sev=2	<b>0.83</b> of all sev=2	0.17 of all sev=2
<b>4.</b>	0.12 of all sev=1	<b>0.88</b> of all sev=1	0.12 of all sev=1	<b>0.88</b> of all sev=1	0.15 of all sev=1	<b>0.85</b> of all sev=1
	<b>0.78</b> of all sev=2	0.22 of all sev=2	<b>0.79</b> of all sev=2	0.21 of all sev=2	<b>0.83</b> of all sev=2	0.17 of all sev=2
<b>7.</b>	<b>0.86</b> of all sev=1	0.14 of all sev=1	0.18 of all sev=1	<b>0.82</b> of all sev=1	<b>0.89</b> of all sev=1	0.11 of all sev=1
	0.22 of all sev=2	<b>0.78</b> of all sev=2	<b>0.82</b> of all sev=2	0.18 of all sev=2	0.32 of all sev=2	<b>0.68</b> of all sev=2
<b>8.</b>	<b>0.89</b> of all sev=1	0.11 of all sev=1	0.26 of all sev=1	<b>0.74</b> of all sev=1	0.06 of all sev=1	<b>0.94</b> of all sev=1
	0.23 of all sev=2	<b>0.77</b> of all sev=2	<b>0.83</b> of all sev=2	0.17 of all sev=2	<b>0.73</b> of all sev=2	0.27 of all sev=2

Συνολικά τόσο με χρήση του πίνακα SEVERITY όσο και με χρήση του πίνακα SEVERITY\_binary η καλύτερη ομαδοποίηση προέκυψε με τις εισόδους που φαίνονται στον παρακάτω πίνακα. Ανάλογα με την είσοδο κάποιες μέθοδοι όπως το K-Means παρουσίαζε καλύτερα αποτελέσματα αλλά κατά κύριο λόγο όλα τα μοντέλα κινούνταν σε παρόμοια επίπεδα.

ΕΙΣΟΔΟΙ	
<b>1.</b>	[["sharp-score", "CCPi", "RF", "HLA-DRB1"]]
<b>4.</b>	[["sharp-score", "CCP", "CCPi", "RF", "Rfi", "HLA-DRB1"]]
<b>7.</b>	[["CCP", "RF"]]
<b>8.</b>	[["sharp-score", "CCP", "RF"]]

### 6.1.2 Μοντέλα Μη- Εποπτευόμενης Μηχανικής Μάθησης

Όσον αφορά τα μοντέλα εποπτευόμενης μηχανικής μάθησης που αναπτύχθηκαν τα καλύτερα αποτελέσματα παρουσιάζονται παρακάτω.

Σε πρώτη φάση χρησιμοποιήθηκε ως πίνακας επαλήθευσης- επικύρωσης της κατηγοριοποίησης ο πίνακας SEVERITY που περιέχει τρεις τιμές για την σοβαρότητα ( 1 , 2 και 3 ) και δημιουργήθηκαν 3 ομάδες για τη σοβαρότητα.

Όπως φαίνεται παρακάτω το καλύτερο μοντέλο προέκυψε με την είσοδο 1 που έδωσε accuracy 80.5% αξιοποιώντας το 85% του δείγματος για εκπαίδευση του μοντέλου και το 15% για επαλήθευση.

INPUT	ACCURACY test_size=0.15
<b>1.</b> [{"sharp-score","DURATION","CCP","RF","HLA-DRB1"}]	<b>80.5%</b>

Σε δεύτερη φάση χρησιμοποιήθηκε ως πίνακας επαλήθευσης- επικύρωσης της κατηγοριοποίησης ο πίνακας SEVERITY\_binary που περιέχει δύο τιμές για την σοβαρότητα (1 και 2 ) και δημιουργήθηκαν 2 ομάδες για τη σοβαρότητα.

Όπως φαίνεται παρακάτω το καλύτερο μοντέλο προέκυψε με την είσοδο 3 που έδωσε accuracy 89.8% αξιοποιώντας το 80% του δείγματος για εκπαίδευση του μοντέλου και το 20% για επαλήθευση.

INPUT	ACCURACY test_size=0.2
<b>3.</b> [{"sharp-score","DURATION","CCPi","Rfi","HLA-DRB1"}]	<b>89.8%</b>

Η διαφορά της εισόδου 1 και 3 είναι ότι η είσοδος 1 αξιοποιεί την τιμή του RF και CCP ενώ η είσοδος 3 αξιοποιεί τον αν ο ασθενής έχει θετικό ή αρνητικό RF και CCP. Κατά τα άλλα πρόκειται για παρόμοιες εισόδους επομένως προκύπτει ξεκάθαρα ποιες μεταβλητές παράγουν τα μοντέλα με τα καλύτερα αποτελέσματα.

Με βάση τα μοντέλα εποπτευόμενης μηχανικής μάθησης που αναπτύχθηκαν πέρα από τις βασικές μεταβλητές sharp-score,CCP-CCPi,RF-RFi,HLA\_DRB1 που είχαν εντοπιστεί από τα μοντέλα μη-επιβλεπόμενης μηχανικής μάθησης ως βασικές μεταβλητές για πρόβλεψη της σοβαρότητας προέκυψε επιπλέον ότι βασικό ρόλο παίζει και η διάρκεια της νόσου όσον αφορά την σοβαρότητα. Γεγονός που και πάλι επιβεβαιώνεται από τους ιατρούς.

Όλα τα μοντέλα που παρουσιάστηκαν εμφανίζουν πολύ καλά αποτελέσματα πράγμα το οποίο υποδεικνύει ότι η έρευνα όσον αφορά την σοβαρότητα της Ρευματοειδούς με ένα ακόμη πιο ποικιλόμορφο δείγμα μπορεί να δώσει εξαιρετικά αποτελέσματα, ειδικά αν συνυπολογιστούν και επιπλέον παράγοντες όπως αναφέρεται παρακάτω στο 6.3

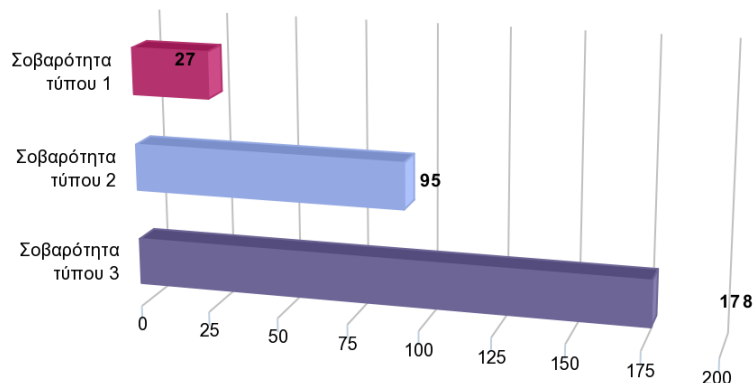
## 6.2 Περιορισμοί της έρευνας

Παρότι προέκυψαν πολύ καλά αποτελέσματα και από την ομαδοποίηση των δεδομένων με βάση τη σοβαρότητα και αντίστοιχα με την κατηγοριοποίηση αυτών, το σύνολο δεδομένων που αξιοποιήθηκε έθεσε αρκετούς περιορισμούς. Καταρχάς πρόκειται για ένα αρκετά μικρό δείγμα ώστε να εκπαιδευτούν τα μοντέλα Μηχανικής Μάθησης. Σίγουρα ένα δείγμα 10.000 ασθενών θα έδινε πολύ πιο αξιόπιστα αποτελέσματα από το δείγμα 300 ασθενών που χρησιμοποιήθηκε.

Επιπλέον, την έρευνα περιόρισε το γεγονός ότι τα δεδομένα προέρχονται από δύο μεγάλα νοσοκομεία, με ειδικά ρευματολογικά τμήματα και κλινικές, στα οποία καταφεύγουν ασθενείς που έχουν ήδη ενεργή ή/και σοβαρή ρευματοειδή αρθρίτιδα και στους οποίους η νόσος συνήθως δεν βρίσκεται στα αρχικά της στάδια.

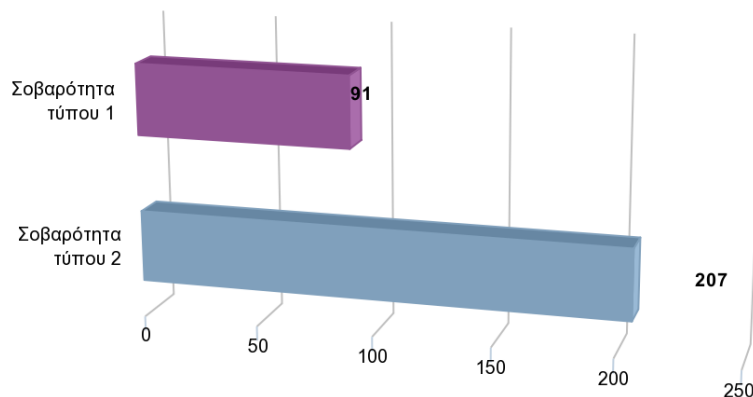
Όπως φαίνεται από τα παρακάτω γραφήματα οι περισσότεροι ασθενείς είχαν σοβαρότητα 3 στο δείγμα με 3 τιμές σοβαρότητας και αντίστοιχα σοβαρότητα 2 στο δείγμα με δυαδικές τιμές της σοβαρότητας.

### Σοβαρότητα



**Σχήμα 6.1:** Σύνολο ασθενών για κάθε επίπεδο σοβαρότητας

## Σοβαρότητα



Σχήμα 6.2: Σύνολο ασθενών για κάθε επίπεδο σοβαρότητας

Το γεγονός ότι τα δεδομένα ήταν μη-ισορροπημένα είναι από τους λόγους που τα μοντέλα που αναπτύχθηκαν δεν είχαν βέλτιστα αποτελέσματα. Επομένως μια σημαντική βελτίωση στα μοντέλα που αναπτύχθηκαν θα προέκυπτε από την χρήση τεχνικών που χειρίζονται τα μη ισορροπημένα δεδομένα ώστε να περιοριστεί αυτή η αδυναμία των μοντέλων χωρίς να είναι απαραίτητο να βρεθούν πιο ισορροπημένοι πληθυσμοί. Πρόκειται για μεθόδους που δίνουν στο σύστημα την «εντύπωση» της επεξεργασίας ισορροπημένων δεδομένων.

Παρόλα αυτά, το ιδανικό σενάριο είναι στο μέλλον να συλλεχθούν δεδομένα από μικρότερες κλινικές, διαγνωστικά κέντρα, κέντρα υγείας και ιδιωτικά ιατρεία, στα οποία είναι περισσότερο που καταφεύγουν ασθενείς με λιγότερο σοβαρή νόσο. Επίσης να αξιοποιηθούν και άλλες διαθέσιμες βάσεις δεδομένων για τον έλεγχο των μοντέλων.

### **6.3 Προτάσεις για μελλοντική έρευνα**

Αρχικά το σύνολο δεδομένων που χρησιμοποιήθηκε για την παρούσα διπλωματική δίνει ακόμα πολλές δυνατότητες που χρονικά δεν μπορούσαν να ερευνηθούν στα πλαίσια μιας

διπλωματικής εργασίας. Παρακάτω ακολουθούν ορισμένες προτάσεις για το πως θα μπορούσε να αξιοποιηθεί το εν λόγω σύνολο δεδομένων.

- Θα μπορούσε να ερευνηθεί η συσχέτιση των συννοσηροτήτων με την φαρμακευτική αγωγή που ακολουθεί ο κάθε ασθενής. Άλλωστε πολλά από τα φάρμακα που χορηγούνται για την αντιμετώπιση της ρευματοειδούς επηρεάζουν το ανοσοποιητικό σύστημα και δίνουν χώρο στην ανάπτυξη άλλων ασθενειών.
- Θα μπορούσε να ερευνηθεί η συσχέτιση της Ρευματοειδούς Αρθρίτιδας που πρόκειται για αυτοάνοσο νόσημα, με την ανάπτυξη Αρθρίτιδας σε ασθενείς ανάλογα με την διάρκεια της νόσου, την ηλικία και τον βαθμό της αλλοίωσης των αρθρώσεων του ασθενή που προκύπτει από το sharp score.
- Θα μπορούσε να ερευνηθεί αν υπάρχει συσχέτιση μεταξύ της σοβαρότητας και της ενεργότητας της νόσου όπως και κατά πόσο το φαρμακευτικό σχήμα που αξιοποιείται σχετίζεται με τους παραπάνω δύο παράγοντες.
- Θα μπορούσε μελλοντικά να αναπτυχθεί μια εφαρμογή που να προσφέρει προτάσεις για το βέλτιστο φαρμακευτικό σχήμα λαμβάνοντας υπόψη την κλινική εικόνα του ασθενούς, τα επίπεδα ενεργότητας και σοβαρότητας της νόσου, τυχόν συννοσηρότητες όπως και εργαστηριακά αποτελέσματα.

Όσον αφορά την έρευνα που ξεκίνησε με την παρούσα διπλωματική εργασία θα μπορούσε να εξελιχθεί με τους παρακάτω τρόπους:

- Αρχικά θα μπορούσε να συλλεχθεί ένα μεγαλύτερο δείγμα ασθενών από ποικίλα διαγνωστικά κέντρα και διαφορετικές πηγές ώστε να δημιουργηθεί ένα πιο ισορροπημένο σύνολο δεδομένων, το οποίο θα μπορούσε να επιβεβαιώσει τα αποτελέσματα της παρούσας διπλωματικής χωρίς να προκύπτουν λόγοι αμφισβήτησης των αποτελεσμάτων.
- Τέλος, θα μπορούσαν να αναπτυχθούν μοντέλα πρόβλεψης όσον αφορά την σοβαρότητα και να δημιουργηθεί μια εφαρμογή που θα έχει πρόσβαση κάθε ιατρός. Η εφαρμογή θα έχει τον ίδιο σκοπό με την αντίστοιχη εφαρμογή που έχει αναπτυχθεί για το DAS28. Να λαμβάνει σαν είσοδο συγκεκριμένα χαρακτηριστικά και να προσφέρει μια γνωμάτευση όσον αφορά την σοβαρότητα της νόσου του ασθενή.

## 6.4 Επίλογος

Οι ρευματικές νόσοι συμπεριλαμβανομένης της ΡΑ δεν είναι ακόμα θεραπεύσιμες. Στις περισσότερες περιπτώσεις η ΡΑ είναι ετερογενής στην εκδήλωσή της. Η εξέλιξη της νόσου, τα επίπεδα ανταπόκρισης στις θεραπείες και η ανάπτυξη συννοσηροτήτων συχνά διαφέρουν μεταξύ των ασθενών. Η προσέγγιση για εξατομικευμένες θεραπείες δεν χρησιμοποιείται ακόμα ευρέως στους ασθενείς. Επί του παρόντος, οι κλινικές αποφάσεις βασίζονται σε επιστημονικά στοιχεία, την προσωπική εμπειρία του ρευματολόγου και τις προτιμήσεις του ασθενούς. Σε αυτό το πλαίσιο, η χρήση της μηχανικής μάθησης μπορεί να αποτελέσει τον καταλυτικό παράγοντα αλλαγής και την κινητήρια δύναμη πίσω από την αύξηση της γνώσης σε διάφορους τομείς της ρευματολογίας.

Δεδομένα του ιστορικού των ασθενών από ψηφιακούς φακέλους, δεδομένα γενετικής έκφρασης, δεδομένα από φορητές διαγνωστικές συσκευές και δεδομένα από επεξεργασία εικόνας μπορούν στο μέλλον να παρέχουν υποστήριξη σε μια ιατρική ακριβείας στην ρευματολογία και ειδικότερα στην ρευματοειδή αρθρίτιδα.

Ένας επίσης πολύ σημαντικός τομέας για έρευνα είναι η ανάπτυξη μοντέλων για τον προσδιορισμό και τον εντοπισμό παραγόντων κινδύνου. Η έγκαιρη διάγνωση με την υποστήριξη της μηχανικής μάθησης ή η πρόγνωση των εξάρσεων της νόσου μπορεί να βοηθήσει στην πρόληψη των επιπτώσεων σε διάφορα όργανα του ασθενούς και να βοηθήσει στην διαχείριση της μακροχρόνιας νόσου. Η ταξινόμηση ασθενών σε υποτύπους γονιδιακής έκφρασης μπορεί να επιτρέψει την επιλογή της πιο κατάλληλης θεραπείας για κάθε ασθενή.

Όπως είδαμε μεγάλος αριθμός πρόσφατων μελετών εστιάζουν στην εξατομικευμένη πρόγνωση της ρευματοειδούς αρθρίτιδας και θα ακολουθήσουν σίγουρα περισσότερα μοντέλα. Η εποπτευόμενη μάθηση είναι επί του παρόντος η μέθοδος που εφαρμόζεται συχνότερα. Η μηχανική μάθηση μπορεί να υποστηρίξει την διαδικασία λήψης κλινικών αποφάσεων και στην ανίχνευση και εκτίμηση κινδύνου της νόσου.

Από τις πρόσφατες μελέτες διαφαίνεται ότι στο μέλλον, η ML θα βοηθήσει τους ρευματολόγους να επιλέγουν την καταλληλότερη θεραπεία για τον ασθενή με ρευματοειδή αρθρίτιδα, για παράδειγμα χρησιμοποιώντας την ενισχυτική μάθηση. Όσες περισσότερες πληροφορίες είναι διαθέσιμες, τόσο καλύτερη θα είναι η απόδοση της ML για την υποστήριξη του ρευματολόγου στην πρόβλεψη της νόσου και στην επιλογή θεραπείας.

Η σημαντικότερη πρόκληση για την επιτυχή εφαρμογή της μηχανικής μάθησης είναι η ποσότητα και η ποιότητα των δεδομένων. Κακής ποιότητας δεδομένα εκπαίδευσης (π.χ.

θορυβώδη δεδομένα, τιμές που λείπουν, ακανόνιστες επισκέψεις) από τους ηλεκτρονικούς φακέλους ασθενών μειώνουν τη συνολική ποιότητα του μοντέλου. Τα θέματα της ασφάλειας δεδομένων, της προστασίας ευαίσθητων προσωπικών δεδομένων αποτελούν και αυτά επίσης βασικές προκλήσεις, που θα πρέπει να συζητηθούν περισσότερο στο μέλλον.

# 7

## Βιβλιογραφία

- [1] P. E. Poubelle, A. Chakravarti, M. J. Fernandes, K. Doiron, and A.-A. Marceau, “Open Access Differential expression of RANK, RANK-L, and osteoprotegerin by synovial fluid neutrophils from patients with rheumatoid arthritis and by healthy human blood neutrophils,” 2007, doi: 10.1186/ar2137.
- [2] E. D. Harris Jr and G. S. Firestein, “Clinical features of rheumatoid arthritis,” in *Kelley’s Textbook of Rheumatology*, Eighth., 2009.
- [3] G. S. Firestein, “Etiology and pathogenesis of rheumatoid arthritis,” in *Kelley’s Textbook of Rheumatology*, Eighth., 2009.
- [4] E. H. Noss and M. B. Brenner, “The role and therapeutic implications of fibroblast-like synoviocytes in inflammation and cartilage erosion in rheumatoid arthritis,” *Immunological Reviews*, vol. 223, no. 1, pp. 252–270, Jun. 2008, doi: 10.1111/J.1600-065X.2008.00648.X.
- [5] E. M. Gravallese, “Bone destruction in arthritis,” *Annals of the Rheumatic Diseases*, vol. 61, no. suppl 2, pp. ii84–ii86, Nov. 2002, doi: 10.1136/ARD.61.SUPPL\_2.II84.
- [6] A. R. Pettit *et al.*, “TRANCE/RANKL Knockout Mice Are Protected from Bone Erosion in a Serum Transfer Model of Arthritis,” *The American Journal of Pathology*, vol. 159, no. 5, pp. 1689–1699, Nov. 2001, doi: 10.1016/S0002-9440(10)63016-7.
- [7] P. Stastny, “Association of the B-Cell Alloantigen DRw4 with Rheumatoid Arthritis,” <http://dx.doi.org/10.1056/NEJM197804202981602>, vol. 298, no. 16, pp. 869–871, Jan. 2010, doi: 10.1056/NEJM197804202981602.



- [8] K. A. Boki, A. A. Drosos, A. G. Tzioufas, J. S. Lanchbury, G. S. Panayi, and H. M. Moutsopoulos, "Examination of HLA-DR4 as a severity marker for rheumatoid arthritis in Greek patients.," *Annals of the Rheumatic Diseases*, vol. 52, no. 7, pp. 517–519, Jul. 1993, doi: 10.1136/ARD.52.7.517.
- [9] A. H. M. van der Helm-Van Mil, K. N. Verpoort, F. C. Breedveld, T. W. J. Huizinga, R. E. M. Toes, and R. R. P. de Vries, "The HLA–DRB1 shared epitope alleles are primarily a risk factor for anti–cyclic citrullinated peptide antibodies and are not an independent risk factor for development of rheumatoid arthritis," *Arthritis & Rheumatism*, vol. 54, no. 4, pp. 1117–1121, Apr. 2006, doi: 10.1002/ART.21739.
- [10] C. M. Weyand, K. C. Hicok, D. L. Conn, and J. J. Goronzy, "The influence of HLA-DRB1 genes on disease severity in rheumatoid arthritis," *Annals of Internal Medicine*, vol. 117, no. 10, pp. 801–806, Nov. 1992, doi: 10.7326/0003-4819-117-10-801.
- [11] H. C. Gérard *et al.*, "Chromosomal DNA From a Variety of Bacterial Species Is Present in Synovial Tissue From Patients With Various Forms of Arthritis," *ARTHRITIS & RHEUMATISM*, vol. 44, no. 7, pp. 1689–1697, 2001, doi: 10.1002/1529-0131.
- [12] E. M. Gravallese and P. A. Moach, "The rheumatoid joint: synovitis and tissue destruction," in *Rheumatology*, Fifth., 2011.
- [13] R. D. Brasington JR, "Clinical features of rheumatoid arthritis," in *Rheumatology*, Fifth., 2011.
- [14] L. McGuigan, D. Burke, and A. Fleming, "Tarsal tunnel syndrome and peripheral neuropathy in rheumatoid disease.," *Annals of the Rheumatic Diseases*, vol. 42, no. 2, pp. 128–131, Apr. 1983, doi: 10.1136/ARD.42.2.128.
- [15] W. N. Roberts, L. H. Daltroy, and R. J. Anderson, "Stability of normal joint findings in persistent classic rheumatoid arthritis," *Arthritis & Rheumatism*, vol. 31, no. 2, pp. 267–271, 1988, doi: 10.1002/ART.1780310215.
- [16] D. N. Collins, C. L. Barnes, and R. L. FitzRandolph, "Cervical spine instability in rheumatoid patients having total hip or knee arthroplasty.," *Clinical Orthopaedics and Related Research*, vol. 272, no. 272, pp. 127–135, Nov. 1991, doi: 10.1097/00003086-199111000-00019.

- [17] X. Puéchal *et al.*, “Peripheral neuropathy with necrotizing vasculitis in rheumatoid arthritis,” *Arthritis & Rheumatism*, vol. 38, no. 11, pp. 1618–1629, Nov. 1995, doi: 10.1002/ART.1780381114.
- [18] C. Turesson and E. L. Matteson, “Extraarticular features of rheumatoid arthritis and systemic involvement,” *Rheumatology: Sixth Edition*, vol. 1–2, pp. 712–719, Jun. 2014, doi: 10.1016/B978-0-323-09138-1.00086-3.
- [19] C. S. Foster, S. L. Forstot, and L. A. Wilson, “Mortality Rate in Rheumatoid Arthritis Patients Developing Necrotizing Scleritis or Peripheral Ulcerative Keratitis: Effects of Systemic Immunosuppression,” *Ophthalmology*, vol. 91, no. 10, pp. 1253–1263, Oct. 1984, doi: 10.1016/S0161-6420(84)34160-4.
- [20] C. Turesson, W. M. O’Fallon, C. S. Crowson, S. E. Gabriel, and E. L. Matteson, “Occurrence of extraarticular disease manifestations is associated with excess mortality in a community based cohort of patients with rheumatoid arthritis,” *The Journal of Rheumatology*, vol. 29, no. 1, 2002.
- [21] H. J. Dinant, W. Hissink Muller, E. M. D. van Berg-Loonen, L. E. Nijenhuis, and C. P. Engelfriet, “HLA-DRw4 in Felty’s syndrome,” *Arthritis & Rheumatism*, vol. 23, no. 11, pp. 1336–1336, Nov. 1980, doi: 10.1002/ART.1780231126.
- [22] G. Gridley, J. H. Klippel, R. N. Hoover, and J. F. Fraumeni, “Incidence of cancer among men with the Felty syndrome,” *Annals of Internal Medicine*, vol. 120, no. 1, pp. 35–39, Jan. 1994, doi: 10.7326/0003-4819-120-1-199401010-00006.
- [23] B. Sutton, A. Corper, V. Bonagura, and M. Taussig, “The structure and origin of rheumatoid factors,” *Immunology Today*, vol. 21, no. 4, pp. 177–183, Apr. 2000, doi: 10.1016/S0167-5699(00)01589-9.
- [24] L. Klareskog, J. Rönnelid, K. Lundberg, L. Padyukov, and L. Alfredsson, “Immunity to Citrullinated Proteins in Rheumatoid Arthritis,” <http://dx.doi.org/10.1146/annurev.immunol.26.021607.090244>, vol. 26, pp. 651–675, Mar. 2008, doi: 10.1146/ANNUREV.IMMUNOL.26.021607.090244.
- [25] C. Vincent, L. Nogueire, C. Clavel, M. Sebbag, and G. Serre, “Autoantibodies to citrullinated proteins: ACPA,” <https://doi.org/10.1080/08916930400022582>, vol. 38, no. 1, pp. 17–24, Feb. 2009, doi: 10.1080/08916930400022582.
- [26] S. W. Syversen *et al.*, “High anti-cyclic citrullinated peptide levels and an algorithm of four variables predict radiographic progression in patients with rheumatoid

- arthritis: results from a 10-year longitudinal study,” *Annals of the Rheumatic Diseases*, vol. 67, no. 2, pp. 212–217, Feb. 2008, doi: 10.1136/ARD.2006.068247.
- [27] D. Aletaha *et al.*, “2010 Rheumatoid arthritis classification criteria: An American College of Rheumatology/European League Against Rheumatism collaborative initiative,” *Arthritis & Rheumatism*, vol. 62, no. 9, pp. 2569–2581, Sep. 2010, doi: 10.1002/ART.27584.
- [28] M. G. Feely and T. R. Mivuls, “Targeted Treatment of the Rheumatic Diseases,” in *Management of Established Rheumatoid Arthritis*, Sanders-Elsevier, 2010.
- [29] M. L. Hetland, “Management of Rheumatoid Arthritis Principles and Strategies for Antirheumatic Pharmacotherapy,” in *Targeted Treatment of the Rheumatic Diseases*, 2010. Accessed: May 03, 2022. [Online]. Available: [https://books.google.gr/books?hl=en&lr=&id=1yOLTwoEr4kC&oi=fnd&pg=PA1&dq=55.%09Merete+Lond+Hetland+2010.+Management+of+Rheumatoid+Arthritis+Principles+and+Strategies+for+Antirheumatic+Pharmacotherapy.+Chapter+1.+Targeted+Treatment+of+the+Rheumatic+Diseases+2010.&ots=oFwaeCQBxx&sig=Kcw5PKTWuicZXqH1gfQkmmmgBjU&redir\\_esc=y#v=onepage&q&f=false](https://books.google.gr/books?hl=en&lr=&id=1yOLTwoEr4kC&oi=fnd&pg=PA1&dq=55.%09Merete+Lond+Hetland+2010.+Management+of+Rheumatoid+Arthritis+Principles+and+Strategies+for+Antirheumatic+Pharmacotherapy.+Chapter+1.+Targeted+Treatment+of+the+Rheumatic+Diseases+2010.&ots=oFwaeCQBxx&sig=Kcw5PKTWuicZXqH1gfQkmmmgBjU&redir_esc=y#v=onepage&q&f=false)
- [30] S. Carette, “All Patients with Rheumatoid Arthritis Should Receive Corticosteroids as Part of Their Management,” *The Journal of Rheumatology*, 2007, Accessed: May 03, 2022. [Online]. Available: <https://www.jrheum.org/content/jrheum/34/4/656.full.pdf>
- [31] J. R. O’dell, “77 - Treatment of Rheumatoid Arthritis,” *Firestein & Kelley’s Textbook of Rheumatology’ 2-Volume Set*, pp. 1258-1283.e6, 2021, doi: 10.1016/B978-0-323-63920-0.00077-1.
- [32] A. Alonso-Ruiz, J. I. Pijoan, E. Ansuategui, A. Urkaregi, M. Calabozo, and A. Quintana, “Tumor necrosis factor alpha drugs in rheumatoid arthritis: Systematic review and metaanalysis of efficacy and safety,” *BMC Musculoskeletal Disorders*, vol. 9, no. 1, pp. 1–27, Apr. 2008, doi: 10.1186/1471-2474-9-52/TABLES/6.
- [33] R. Axmann *et al.*, “CTLA-4 directly inhibits osteoclast formation,” *Annals of the Rheumatic Diseases*, vol. 67, no. 11, pp. 1603–1609, Nov. 2008, doi: 10.1136/ARD.2007.080713.
- [34] A. Finckh *et al.*, “B cell depletion may be more effective than switching to an alternative anti-tumor necrosis factor agent in rheumatoid arthritis patients with

- inadequate response to anti-tumor necrosis factor agents,” *Arthritis & Rheumatism*, vol. 56, no. 5, pp. 1417–1423, May 2007, doi: 10.1002/ART.22520.
- [35] K. G. Saag *et al.*, “American College of Rheumatology 2008 recommendations for the use of nonbiologic and biologic disease-modifying antirheumatic drugs in rheumatoid arthritis,” *Arthritis Rheum*, vol. 59, no. 6, pp. 762–784, Jun. 2008, doi: 10.1002/ART.23721.
- [36] L. Klareskog *et al.*, “Therapeutic effect of the combination of etanercept and methotrexate compared with each treatment alone in patients with rheumatoid arthritis: double-blind randomised controlled trial,” *Lancet*, vol. 363, no. 9410, pp. 675–681, Feb. 2004, doi: 10.1016/S0140-6736(04)15640-7.
- [37] M. Hügler, P. Omoumi, J. M. van Laar, J. Boedecker, and T. Hügler, “Applied machine learning and artificial intelligence in rheumatology,” *Rheumatology Advances in Practice*, vol. 4, no. 1, Jan. 2020, doi: 10.1093/RAP/RKAA005.
- [38] K. Zarkogianni, M. Athanasiou, and A. C. Thanopoulou, “Comparison of Machine Learning Approaches Toward Assessing the Risk of Developing Cardiovascular Disease as a Long-Term Diabetes Complication,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1637–1647, Sep. 2018, doi: 10.1109/JBHI.2017.2765639.
- [39] E. S. Adamidi, K. Mitsis, and K. S. Nikita, “Artificial intelligence in clinical care amidst COVID-19 pandemic: A systematic review,” *Computational and Structural Biotechnology Journal*, vol. 19, pp. 2833–2850, Jan. 2021, doi: 10.1016/J.CSBJ.2021.05.010.
- [40] M. Athanasiou, K. Sfrintzeri, K. Zarkogianni, A. C. Thanopoulou, and K. S. Nikita, “An explainable XGBoost-based approach towards assessing the risk of cardiovascular disease in patients with Type 2 Diabetes Mellitus,” *Proceedings - IEEE 20th International Conference on Bioinformatics and Bioengineering, BIBE 2020*, pp. 859–864, Oct. 2020, doi: 10.1109/BIBE50027.2020.00146.
- [41] K. M. Kingsmore, C. E. Puglisi, A. C. Grammer, and P. E. Lipsky, “An introduction to machine learning and analysis of its use in rheumatic diseases,” *Nat Rev Rheumatol*, vol. 17, no. 12, pp. 710–730, Dec. 2021, doi: 10.1038/S41584-021-00708-W.

- [42] M. Krusche and G. Burmester, “The use of machine learning to optimize treatment of rheumatic diseases,” *rheumatology.medicinematters.com*, 2020, Accessed: May 04, 2022. [Online]. Available: <https://rheumatology.medicinematters.com/artificial-intelligence/devices-and-technology/machine-learning-treatment-rheumatic-disease/17540860>
- [43] D. J. Odgers, N. Tellis, H. Hall, and M. Dumontier, “Using LASSO Regression to Predict Rheumatoid Arthritis Treatment Efficacy,” *AMIA Summits on Translational Science Proceedings*, vol. 2016, p. 176, 2016, Accessed: May 04, 2022. [Online]. Available: [/pmc/articles/PMC5001752/](https://pubmed.ncbi.nlm.nih.gov/35001752/)
- [44] K. J. Kim and I. Tagkopoulos, “Application of machine learning in rheumatic disease research,” *The Korean Journal of Internal Medicine*, vol. 34, no. 4, p. 708, 2019, doi: 10.3904/KJIM.2018.349.
- [45] N. Foulquier and A. Saraux, “Application of machine learning in predicting early diagnosis of rheumatic diseases,” *rheumatology.medicinematters.com*, 2019, Accessed: May 04, 2022. [Online]. Available: <https://rheumatology.medicinematters.com/diagnosis-and-screening/artificial-intelligence/machine-learning-diagnosis-rheumatic-disease/17437330>
- [46] I. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας, and Η. Σακελλαρίου, *Τεχνητή Νοημοσύνη*, 3η. Εκδόσεις Πανεπιστημίου, 2011.
- [47] S. Kotsiantis, “Ομάδες ταξινομητών για την αύξηση της ακρίβειας των μεθόδων μηχανικής μάθησης και εξόρυξης γνώσης,” Doctoral Dissertation, University of Patra, Patra, 2005.
- [48] E. Kyrkos, *Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων*. Open Academic Editions, 2015. Accessed: May 04, 2022. [Online]. Available: <https://repository.kallipos.gr/handle/11419/1226>
- [49] “Gaussian Mixture Models Explained | by Oscar Contreras Carrasco | Towards Data Science.” <https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95> (accessed May 04, 2022).
- [50] “Clustering in Machine Learning - GeeksforGeeks.” <https://www.geeksforgeeks.org/clustering-in-machine-learning/> (accessed May 04, 2022).

- [51] Z. Abdullah and A. , R. Hamdan, “Hierarchical Clustering Algorithms in Data Mining,” *International Journal of Computer and Information Engineering*, vol. 9, no. 10, 2015, Accessed: May 04, 2022. [Online]. Available: <https://www.semanticscholar.org/paper/Hierarchical-Clustering-Algorithms-in-Data-Mining-Abdullah-Hamdan/83772350afc3fac5ec13f9d04bc1cda29ba7c8c9>
- [52] “Chapter 4 Unsupervised Learning | An Introduction to Machine Learning with R.” <https://lgatto.github.io/IntroMachineLearningWithR/unsupervised-learning.html> (accessed May 04, 2022).
- [53] “K means Clustering - Introduction - GeeksforGeeks.” <https://www.geeksforgeeks.org/k-means-clustering-introduction/> (accessed May 04, 2022).
- [54] “2.1. Gaussian mixture models — scikit-learn 1.0.2 documentation.” <https://scikit-learn.org/stable/modules/mixture.html> (accessed May 04, 2022).
- [55] D. J. Ketchen and C. L. Shook, “The application of cluster analysis in Strategic Management Research: An analysis and critique,” *Strategic Management Journal*, vol. 17, no. 6, pp. 441–458, Jun. 1996, doi: [https://doi.org/10.1002/\(SICI\)1097-0266\(199606\)17:6<441::AID-SMJ819>3.0.CO;2-G](https://doi.org/10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G).
- [56] “Elbow method (clustering) - Wikipedia.” [https://en.wikipedia.org/wiki/Elbow\\_method\\_\(clustering\)](https://en.wikipedia.org/wiki/Elbow_method_(clustering)) (accessed May 04, 2022).
- [57] “Elbow Method for optimal value of k in KMeans - GeeksforGeeks.” <https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/> (accessed May 04, 2022).
- [58] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, 3rd ed. 2011. Accessed: May 05, 2022. [Online]. Available: [https://books.google.gr/books?hl=en&lr=&id=pQws07tdpjoC&oi=fnd&pg=PP1&dq=Han,+Pei,+%26+Kamber,+2011&ots=tAIr2UqEVX&sig=rZSpbrFG5T1ErIldDO7XNWtj2YrI&redir\\_esc=y#v=onepage&q=Han%2C%20Pei%2C%20%26%20Kamber%2C%202011&f=false](https://books.google.gr/books?hl=en&lr=&id=pQws07tdpjoC&oi=fnd&pg=PP1&dq=Han,+Pei,+%26+Kamber,+2011&ots=tAIr2UqEVX&sig=rZSpbrFG5T1ErIldDO7XNWtj2YrI&redir_esc=y#v=onepage&q=Han%2C%20Pei%2C%20%26%20Kamber%2C%202011&f=false)
- [59] “K Nearest Neighbor - an overview | ScienceDirect Topics.” <https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/k-nearest-neighbor> (accessed May 05, 2022).

- [60] A. Subasi, K. Khateeb, T. Brahim, and A. Sarirete, "Human activity recognition using machine learning methods in a smart healthcare environment," *Innovation in Health Informatics: A Smart Healthcare Primer*, pp. 123–144, Jan. 2020, doi: 10.1016/B978-0-12-819043-2.00005-8.
- [61] W. Tao *et al.*, "Multiomics and Machine Learning Accurately Predict Clinical Response to Adalimumab and Etanercept Therapy in Patients With Rheumatoid Arthritis," *Arthritis and Rheumatology*, vol. 73, no. 2, pp. 212–222, Feb. 2021, doi: 10.1002/ART.41516/ABSTRACT.
- [62] B. Norgeot *et al.*, "Assessment of a Deep Learning Model Based on Electronic Health Record Data to Forecast Clinical Outcomes in Patients With Rheumatoid Arthritis," *JAMA Network Open*, vol. 2, no. 3, pp. e190606–e190606, Mar. 2019, doi: 10.1001/JAMANETWORKOPEN.2019.0606.
- [63] Y. Guan *et al.*, "Machine Learning to Predict Anti-Tumor Necrosis Factor Drug Responses of Rheumatoid Arthritis Patients by Integrating Clinical and Genetic Markers," *Arthritis & Rheumatology*, vol. 71, no. 12, pp. 1987–1996, Dec. 2019, doi: 10.1002/ART.41056.
- [64] S. M. Zhou *et al.*, "Defining Disease Phenotypes in Primary Care Electronic Health Records by a Machine Learning Approach: A Case Study in Identifying Rheumatoid Arthritis," *PLOS ONE*, vol. 11, no. 5, p. e0154515, May 2016, doi: 10.1371/JOURNAL.PONE.0154515.
- [65] K. P. Liao *et al.*, "Electronic medical records for discovery research in rheumatoid arthritis," *Arthritis Care & Research*, vol. 62, no. 8, pp. 1120–1127, Aug. 2010, doi: 10.1002/ACR.20184.
- [66] R. J. Carroll *et al.*, "Portability of an algorithm to identify rheumatoid arthritis in electronic health records," *Journal of the American Medical Informatics Association*, vol. 19, no. e1, pp. e162–e169, Jun. 2012, doi: 10.1136/AMIAJNL-2011-000583.
- [67] J. K. H. Andersen *et al.*, "Neural networks for automatic scoring of arthritis disease activity on ultrasound images," *RMD Open*, vol. 5, no. 1, p. e000891, Mar. 2019, doi: 10.1136/RMDOPEN-2018-000891.
- [68] J. Tang *et al.*, "Grading of rheumatoid arthritis on ultrasound images with deep convolutional neural network," *IEEE International Ultrasonics Symposium, IUS*, vol. 2018-January, 2018, doi: 10.1109/ULTSYM.2018.8579871.

- [69] J. Tang *et al.*, “Enhancing convolutional neural network scheme for rheumatoid arthritis grading with limited clinical data\*,” *Chinese Physics B*, vol. 28, no. 3, p. 038701, Mar. 2019, doi: 10.1088/1674-1056/28/3/038701.
- [70] K. Üreten, H. Erbay, and H. H. Maraş, “Detection of rheumatoid arthritis from hand radiographs using a convolutional neural network,” *Clinical Rheumatology 2019 39:4*, vol. 39, no. 4, pp. 969–974, Mar. 2019, doi: 10.1007/S10067-019-04487-4.
- [71] S. Murakami, K. Hatano, J. Tan, H. Kim, and T. Aoki, “Automatic identification of bone erosions in rheumatoid arthritis from hand radiographs based on deep convolutional neural network,” *Multimedia Tools and Applications 2017 77:9*, vol. 77, no. 9, pp. 10921–10937, Dec. 2017, doi: 10.1007/S11042-017-5449-4.
- [72] J. Rohrbach, T. Reinhard, B. Sick, and O. Dürr, “Bone erosion scoring for rheumatoid arthritis with deep convolutional neural networks,” *Computers & Electrical Engineering*, vol. 78, pp. 472–481, Sep. 2019, doi: 10.1016/J.COMPELECENG.2019.08.003.
- [73] M. Betancourt-Hernández, G. Viera-López, and Serrano-Muñoz, “Automatic Diagnosis of Rheumatoid Arthritis from Hand Radiographs using Convolutional Neural Networks,” *Revista Cubana de Física*, 2018, Accessed: May 04, 2022. [Online]. Available: [http://revistacubanadefisica.org/index.php/rcf/article/view/RCF\\_35\\_1\\_39](http://revistacubanadefisica.org/index.php/rcf/article/view/RCF_35_1_39)
- [74] R. J. Hemalatha, V. Vijaybaskar, and T. R. Thamizhvani, “Automatic localization of anatomical regions in medical ultrasound images of rheumatoid arthritis using deep learning,” *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 233, no. 6, pp. 657–667, Jun. 2019, doi: 10.1177/0954411919845747.
- [75] H. Dehghani, Y. Feng, D. Lighter, L. Zhang, and Y. Wang, “Deep Neural Networks Improve Diagnostic Accuracy of Rheumatoid Arthritis using Diffuse Optical Tomography,” *Clinical and Preclinical Optical Diagnostics II (2019)*, paper 11074\_52, vol. Part F142-ECBO 2019, p. 11074\_52, Jun. 2019, doi: 10.1117/12.2526787.
- [76] F. Liu *et al.*, “Deep learning approach for evaluating knee MR images: Achieving high diagnostic performance for cartilage lesion detection,” *Radiology*, vol. 289, no. 1, pp. 160–169, Oct. 2018, doi:



10.1148/RADIOL.2018172986/ASSET/IMAGES/LARGE/RADIOL.2018172986.TBL5.JPEG.

- [77] A. Vodencarevic *et al.*, “Advanced machine learning for predicting individual risk of flares in rheumatoid arthritis patients tapering biologic drugs,” *Arthritis Research and Therapy*, vol. 23, no. 1, pp. 1–8, Dec. 2021, doi: 10.1186/S13075-021-02439-5/FIGURES/2.
- [78] M. Kalweit *et al.*, “Personalized prediction of disease activity in patients with rheumatoid arthritis using an adaptive deep neural network,” *PLOS ONE*, vol. 16, no. 6, p. e0252289, Jun. 2021, doi: 10.1371/JOURNAL.PONE.0252289.
- [79] L. Gossec *et al.*, “Detection of Flares by Decrease in Physical Activity, Collected Using Wearable Activity Trackers in Rheumatoid Arthritis or Axial Spondyloarthritis: An Application of Machine Learning Analyses in Rheumatology,” *Arthritis Care & Research*, vol. 71, no. 10, pp. 1336–1343, Oct. 2019, doi: 10.1002/ACR.23768.
- [80] A. Madrid-García *et al.*, “Outpatient Readmission in Rheumatology: A Machine Learning Predictive Model of Patient’s Return to the Clinic,” *Journal of Clinical Medicine 2019, Vol. 8, Page 1156*, vol. 8, no. 8, p. 1156, Aug. 2019, doi: 10.3390/JCM8081156.
- [81] D. E. Orange *et al.*, “Identification of Three Rheumatoid Arthritis Disease Subtypes by Machine Learning Integration of Synovial Histologic Features and RNA Sequencing Data,” *Arthritis Rheumatol*, vol. 70, no. 5, pp. 690–701, May 2018, doi: 10.1002/ART.40428.
- [82] M. Luque-Tévar *et al.*, “Integrative Clinical, Molecular, and Computational Analysis Identify Novel Biomarkers and Differential Profiles of Anti-TNF Response in Rheumatoid Arthritis,” *Frontiers in Immunology*, vol. 12, p. 592, Mar. 2021, doi: 10.3389/FIMMU.2021.631662/BIBTEX.
- [83] F. B. S. Briggs *et al.*, “Supervised machine learning and logistic regression identifies novel epistatic risk factors with PTPN22 for rheumatoid arthritis,” *Genes & Immunity 2010 11:3*, vol. 11, no. 3, pp. 199–208, Jan. 2010, doi: 10.1038/gene.2009.110.
- [84] N. Nair and A. G. Wilson, “Can machine learning predict responses to TNF inhibitors?,” *Nature Reviews Rheumatology 2019 15:12*, vol. 15, no. 12, pp. 702–704, Oct. 2019, doi: 10.1038/s41584-019-0320-9.

- [85] R. M. Plenge *et al.*, “Crowdsourcing genetic prediction of clinical utility in the Rheumatoid Arthritis Responder Challenge,” *Nature Genetics* 2013 45:5, vol. 45, no. 5, pp. 468–469, Apr. 2013, doi: 10.1038/ng.2623.
- [86] S. van Looy, B. vander Cruyssen, B. Wyns, J. Meeus, F. de Keyser, and L. Boullart, “Comparing Statistics with Machine Learning Models to Predict Dose Increase of Infliximab for Rheumatoid Arthritis Patients,” *Acta Press*, 2005. Accessed: May 04, 2022. [Online]. Available: <https://www.actapress.com/Abstract.aspx?paperId=21798>
- [87] C. Lin *et al.*, “Automatic Prediction of Rheumatoid Arthritis Disease Activity from the Electronic Medical Records,” *PLOS ONE*, vol. 8, no. 8, p. e69932, Aug. 2013, doi: 10.1371/JOURNAL.PONE.0069932.
- [88] Z. Shiezadeh, H. Sajedi, and E. Aflakie, “DIAGNOSIS OF RHEUMATOID ARTHRITIS USING AN ENSEMBLE LEARNING APPROACH,” pp. 139–148, 2015, doi: 10.5121/csit.2015.51512.
- [89] J. M. Lezcano-Valverde *et al.*, “Development and validation of a multivariate predictive model for rheumatoid arthritis mortality using a machine learning approach,” *Scientific Reports* 2017 7:1, vol. 7, no. 1, pp. 1–10, Aug. 2017, doi: 10.1038/s41598-017-10558-w.
- [90] C. H. Feldman *et al.*, “Supplementing Claims Data with Electronic Medical Records to Improve Estimation and Classification of Rheumatoid Arthritis Disease Activity: A Machine Learning Approach,” *ACR Open Rheumatology*, vol. 1, no. 9, pp. 552–559, Nov. 2019, doi: 10.1002/ACR2.11068.
- [91] H. Sharon, I. Elamvazuthi, C. K. Lu, S. Parasuraman, and E. Natarajan, “Development of Rheumatoid Arthritis Classification from Electronic Image Sensor Using Ensemble Method,” *Sensors* 2020, Vol. 20, Page 167, vol. 20, no. 1, p. 167, Dec. 2019, doi: 10.3390/S20010167.
- [92] A. Pandit and T. R. D. J. Radstake, “Machine learning in rheumatology approaches the clinic,” *Nature Reviews Rheumatology* 2020 16:2, vol. 16, no. 2, pp. 69–70, Jan. 2020, doi: 10.1038/s41584-019-0361-0.
- [93] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, “Machine learning in medicine: a practical introduction,” *BMC Medical Research Methodology*, vol. 19, no. 1, pp. 1–18, Mar. 2019, doi: 10.1186/S12874-019-0681-4/TABLES/5.

- [94] C. Y. Chin, S. Y. Hsieh, and V. S. Tseng, “eDRAM: Effective early disease risk assessment with matrix factorization on a large-scale medical database: A case study on rheumatoid arthritis,” *PLOS ONE*, vol. 13, no. 11, p. e0207579, Nov. 2018, doi: 10.1371/JOURNAL.PONE.0207579.
- [95] Π. Κατσιμπρή, Ρ. Δπιμελήτρια Β' Γ' Παθολογική, Κ. Πανεπιστημιακό, and Ν. <<αττικον>>, “Ρευματοειδής Αρθρίτιδα”.
- [96] K. Chang, S. M. Yang, S. H. Kim, K. H. Han, S. J. Park, and J. il Shin, “Smoking and Rheumatoid Arthritis,” *International Journal of Molecular Sciences 2014*, Vol. 15, Pages 22279-22295, vol. 15, no. 12, pp. 22279–22295, Dec. 2014, doi: 10.3390/IJMS151222279.
- [97] T. D. Wilsdon and C. L. Hill, “Managing the drug treatment of rheumatoid arthritis,” *Australian Prescriber*, vol. 40, no. 2, p. 51, 2017, doi: 10.18773/AUSTPRESCR.2017.012.
- [98] B. F. Leeb *et al.*, “Disease activity measurement of rheumatoid arthritis: Comparison of the simplified disease activity index (SDAI) and the disease activity score including 28 joints (DAS28) in daily routine,” *Arthritis Care & Research*, vol. 53, no. 1, pp. 56–60, Feb. 2005, doi: 10.1002/ART.20923.
- [99] “Disease Activity Scores (DAS and DAS28) Definition | Defined by Arthritis-Health.” <https://www.arthritis-health.com/glossary/disease-activity-scores-das-and-das28> (accessed May 05, 2022).
- [100] “Misreading Disease Activity With DAS28.” <https://www.rheumatologynetwork.com/view/misreading-disease-activity-das28> (accessed May 05, 2022).
- [101] T. Markatseli and Θ. Μαγκατσέλη, “Καταστροφή των περιφερικών αρθρώσεων στη ρευματοειδή αρθρίτιδα και προγνωστικοί δείκτες,” Jan. 2010, doi: 10.12681/EADD/18628.
- [102] G. Kinikli, A. Ateş, M. Turgay, G. Akay, S. Kinikli, and G. Tokgöz, “HLA-DRB1 genes and disease severity in rheumatoid arthritis in Turkey,” *Scand J Rheumatol*, vol. 32, no. 5, pp. 277–280, 2003, doi: 10.1080/03009740310003901.
- [103] Γ. Κοκαράκη and G. Κοκαράκη, “Υπότυποι των αντιγόνων ιστοσυμβατότητας HLA-DRB1\*04 σε έλληνες ασθενείς και συσχέτισή τους με αυτοάνοσα νοσήματα,” Jan. 2010, doi: 10.12681/EADD/28304.

- [104] R. Wu, O. Shovman, Y. Zhang, B. Gilburd, G. Zandman-Goddard, and Y. Shoenfeld, “Increased prevalence of anti-third generation cyclic citrullinated peptide antibodies in patients with rheumatoid arthritis and CREST syndrome,” *Clin Rev Allergy Immunol*, vol. 32, no. 1, pp. 47–55, 2007, doi: 10.1007/BF02686081.
- [105] S. Rantapää-Dahlqvist *et al.*, “Antibodies against cyclic citrullinated peptide and IgA rheumatoid factor predict the development of rheumatoid arthritis,” *Arthritis Rheum*, vol. 48, no. 10, pp. 2741–2749, 2003, doi: 10.1002/ART.11223.
- [106] “ΣΥΜΒΟΛΗ ΕΡΓΑΣΤΗΡΙΑΚΩΝ ΕΥΡΗΜΑΤΩΝ ΣΤΗ ΔΙΑΓΝΩΣΗ & ΠΑΡΑΚΟΛΟΥΘΗΣΗ ΡΕΥΜΑΤΙΚΩΝ ΝΟΣΗΜΑΤΩΝ”.
- [107] D. Wilson, “Rheumatoid factors in patients with rheumatoid arthritis,” *Canadian Family Physician*, vol. 52, no. 2, p. 180, Feb. 2006, Accessed: May 05, 2022. [Online]. Available: /pmc/articles/PMC1479719/
- [108] J. T. Sharp *et al.*, “How many joints in the hands and wrists should be included in a score of radiologic abnormalities used to assess rheumatoid arthritis?,” *Arthritis & Rheumatism*, vol. 28, no. 12, pp. 1326–1335, 1985, doi: 10.1002/ART.1780281203.
- [109] J. T. Sharp, M. D. Lidsky, L. C. Collins, and J. Moreland, “Methods of scoring the progression of radiologic changes in rheumatoid arthritis. Correlation of radiologic, clinical and laboratory abnormalities,” *Arthritis & Rheumatism*, vol. 14, no. 6, pp. 706–720, Nov. 1971, doi: 10.1002/ART.1780140605.
- [110] P. A. Ory, “Interpreting radiographic data in rheumatoid arthritis,” *Annals of the Rheumatic Diseases*, vol. 62, no. 7, pp. 597–604, Jul. 2003, doi: 10.1136/ARD.62.7.597.
- [111] V. Ravindran and S. Rachapalli, “An overview of commonly used radiographic scoring methods in rheumatoid arthritis clinical trials,” *Clinical Rheumatology 2010 30:1*, vol. 30, no. 1, pp. 1–6, Aug. 2010, doi: 10.1007/S10067-010-1554-8.
- [112] D. van der Heijde, T. Dankert, F. Nieman, R. Rau, and M. Boers, “Reliability and sensitivity to change of a simplification of the Sharp/van der Heijde radiological assessment in rheumatoid arthritis,” *Rheumatology*, vol. 38, no. 10, pp. 941–947, Oct. 1999, doi: 10.1093/RHEUMATOLOGY/38.10.941.
- [113] “ΚΛΙΝΙΚΗ ΕΚΤΙΜΗΣΗ ΑΣΘΕΝΟΥΣ ΜΕ ΡΑ”.

