Εθνικό Μετσόβιο
Πολυτεχνείο -
Διατμηματικό Πρόγραμμα
Μεταπτυχιακών Σπουδών

Μαθηματική Προτυποποίηση σε Σύγχρονες Τεχνολογίες και τη Χρηματοοικονομική

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τεχνικές διαχείρισης μη ισορροπημένων σύνολων δεδομένων δυαδικής κατηγοριοποίησης στη μηχανική μάθηση

Handling techniques for imbalanced binary classification datasets in machine learning

Σταματάκης Αργύρης

Αθήνα 2022

Σύνοψη

Στο σύγχρονο κόσμο, η χρήση της Επιστήμης των Δεδομένων είναι διάχυτη σε όλα τα επιστημονικά πεδία. Η προσπάθεια εύρεσης του πιο ικανού αλγόριθμου που θα μπορεί να εκτιμάει σχεδόν τέλεια τα αποτελέσματα για κάθε καινούργιο σύνολο δεδομένων είναι ένας από τους πιο σημαντικούς και ενδιαφέροντες στόχους των επιστημόνων. Ο σκοπός της παρούσας διπλωματικής εργασίας είναι να εμβαθύνει στο πρόβλημα που προκύπτει όταν σε ένα σύνολο δεδομένων δυαδικής κατηγοριοποίησης η μια από τις δυο κλάσεις υπερτερεί σημαντικά σε αριθμό εγγραφών σε σχέση με την άλλη. Αυτό συνιστά πρόβλημα, γιατί συνήθως οι αλγόριθμοι τείνουν να εστιάζουν στην επικρατούσα κλάση, αγνοώντας τη μικρότερη κλάση.

Η εφαρμογή με την οποία ερευνούμε τα θεωρητικά δεδομένα, που παρουσιάζουμε στα πρώτα κεφάλαια, αναφέρεται σε μια εταιρεία που ενδιαφέρεται να γνωρίζει ποιοι από τους υπαλλήλους που θα επιλέξει για να τους εκπαιδεύσει, θα παραμείνουν στην εταιρεία ως υπάλληλοι και δε θα χρησιμοποιήσουν την εκπαίδευση για να βρουν μια άλλη εργασία. Η επικρατούσα κλάση που αφορά τους υπαλλήλους που παραμένουν στην εταιρεία έχει ένα ποσοστό 75% από το σύνολο δεδομένων και η περίπτωση αυτή θεωρείται ως ένας ήπιος βαθμός ανισορροπίας του συνόλου. Η έρευνα στην οποία θα προβούμε θα εστιάσει στις μεθόδους διαχείρισης μη ισορροπημένων δεδομένων κατηγοριοποίησης.

Abstract

In the modern world, the use of Data Science is wide spread in all the scientific fields. Trying to find the most capable algorithm that can almost perfectly evaluate the results for each new dataset is one of the most important and interesting goals of scientists. The purpose of this thesis is to delve into the problem that arises when in a dataset of binary categorization one of the two classes is significantly bigger in number of records compared to the other. This is a problem, because algorithms usually tend to focus on the dominant class, ignoring the smaller class.

The application with which we research the theory that is presented in the first chapters, refers to a company that is interested in knowing which of the employees that will be chosen to be educated, will remain in the company as employees and will not use the training courses to find another job. The dominant class, concerning the employees who remain in the company, has a percentage of 75% of the dataset and this case is considered as a mild degree of total imbalance. Our research will be focused on methods for managing imbalanced categorization dataset.

Περιεχόμενα

1 Εισαγωγή

Η ανάλυση δεδομένων είναι ένας επιστημονικός κλάδος που μας επιτρέπει να εκτιμήσουμε τις εξαρτημένες μεταβλητές τις οποίες μελετάμε από ένα σύνολο δεδομένων που αποτελείται από ορισμένες ανεξάρτητες μεταβλητές. Σημαντικότατος παράγοντας στην ανάλυση είναι η γλώσσα προγραμματισμού που θα μας βοηθήσει με τη χρήση βιβλιοθηκών της, να επεξεργαστούμε τα δεδομένα μας.

Το αρχικό και πλέον χρονοβόρο βήμα για την ανάλυση δεδομένων είναι η προεπεξεργασία των δεδομένων που αποτελείται από τρία στάδια, όπως τα έχουμε ορίσει στην εργασία μας. Το πρώτο αποτελεί τη διαχείριση δεδομένων που έχει ως κύριο στόχο τα δεδομένα μας να είναι τακτοποιημένα και οριοθετημένα, ώστε να μπορούν να χρησιμοποιηθούν χωρίς προβλήματα από τα επόμενα στάδια της προεπεξεργασίας. Το δεύτερο στάδιο είναι η διερευνητική ανάλυση δεδομένων, ΔΑΔ, η οποία μας επιτρέπει να έχουμε μια οπτικοποιημένη αντίληψη των δεδομένων που έχουμε συλλέξει και να μπορούμε να εστιάσουμε την έρευνα μας όχι σε όλες τις ανεξάρτητες μεταβλητές του σύνολου δεδομένων, αλλά μόνο σε εκείνες που φαίνεται να επηρεάζουν περισσότερο την εξαρτημένη μεταβλητή. Το τρίτο στάδιο, αφορά τη μηχανική χαρακτηριστικών και αναφέρεται στη χρήση μετασχηματισμών και τεχνικών που μπορεί να βοηθήσουν το υπολογιστικό σύστημα να γίνει πιο λειτουργικό και να εξάγει πιο αντικειμενικά και ακριβή αποτελέσματα, μειώνοντας παράλληλα το θόρυβο που υπάρχει στα δεδομένα.

Η μηχανική μάθηση παίζει σημαντικό ρόλο, αφού βοηθάει τους υπολογιστές να μαθαίνουν από το σύνολο δεδομένων, χωρίς να υπάρχει κάποιο πρόγραμμα που να τους καθοδηγεί για τον ακριβή τρόπο με τον οποίο θα εξάγουν τα συμπεράσματα. Ουσιαστικά, δίνει στον υπολογιστή τη λογική με βάση την οποία θα οικοδομήσει τον αλγόριθμο που θα ακολουθήσει. Στηριζόμενοι σε αυτόν τον αλγόριθμο, ο υπολογιστής θα εκτιμήσει τις τιμές της εξαρτημένης μεταβλητής του συνόλου δεδομένων που του δίνουμε και μετά, θα ελέγξουμε την απόδοση του κάθε αλγόριθμου με κάποια από τις μετρικές αξιολόγησης που διαθέτουμε και η οποία θα ταιριάζει στο αντίστοιχο σύνολο δεδομένων. Με αυτό τον τρόπο μπορούμε να καταλήξουμε στον καλύτερο αλγόριθμο από όσους θα δοκιμάσουμε.

Τέλος όσον αφορά την εργασία μας, θα εστιάσουμε την έρευνας μας στις μεθόδους διαχείρισης μη ισορροπημένων δεδομένων κατηγοριοποίησης. Έτσι, θα εξισορροπήσουμε το πλήθος των εγγραφών μεταξύ των δυο κλάσεων της εξαρτημένης μεταβλητής, αφού όπως έχει αποδειχθεί η συγκεκριμένη ανισορροπία επηρεάζει αρνητικά τόσο τους αλγόριθμους που προσπαθούν να εκτιμήσουν τις εξαρτημένες μεταβλητές, όσο και ορισμένες μετρικές αξιολόγησης. Στην έρευνα μας θα εφαρμόσουμε και τις τρεις κατηγορίες των συγκεκριμένων μεθόδων διαχείρισης μη ισορροπημένων δεδομένων, δηλαδή υπο-δειγματοληψίας, υπερ-δειγματοληψίας και μεικτές.

2 Ανάλυση Δεδομένων (Data Analysis)

Η ανάλυση δεδομένων είναι μια διαδικασία συλλογής, καθαρισμού, επεξεργασίας και μοντελοποίησης δεδομένων με στόχο την ανακάλυψη χρήσιμων πληροφοριών, την εξαγωγή συμπερασμάτων και την καθοδήγηση ως προς τη λήψη αποφάσεων. Η ανάλυση δεδομένων έχει πολλαπλές πτυχές και προσεγγίσεις, που περιλαμβάνουν ποικίλες τεχνικές και χρησιμοποιείται σε διάφορους εργασιακούς και επιστημονικούς τομείς. Στον σημερινό κόσμο, παίζει ρόλο στη λήψη αποφάσεων χρησιμοποιώντας πιο επιστημονικά κριτήρια και βοηθώντας τους **επιστήμονες δεδομένων** (**data scientists**) να καταλήγουν σε ασφαλέστερα συμπεράσματα.

Η **εξόρυξη δεδομένων** (**data mining**) είναι μια συγκεκριμένη τεχνική ανάλυσης δεδομένων που ο στόχος της δεν είναι η καθαρά περιγραφική στατιστική, αλλά εστιάζει στη στατιστική μοντελοποίηση και στην εύρεση εκτιμήσεων, ώστε να προβλέπει αποτελέσματα. Στις στατιστικές εφαρμογές, η ανάλυση δεδομένων μπορεί να χωριστεί σε περιγραφικές στατιστικές, δηλαδή τη **διερευνητική ανάλυση δεδομένων, ΔΑΔ** (**Exploratory Data Analysis, EDA**) και την **επιβεβαιωτική ανάλυση δεδομένων** (**Confirmatory Data Analysis, CDA**). ***[1]***

Η ΔΑΔ επικεντρώνεται στην ανακάλυψη νέων **χαρακτηριστικών** (**features**) στα δεδομένα, ενώ η επιβεβαιωτική ανάλυση δεδομένων επικεντρώνεται στην επιβεβαίωση ή απόρριψη υπαρχόντων υποθέσεων. Ακόμα, η **προγνωστική αναλυτική** (**predictive analytics**) επικεντρώνεται στην εφαρμογή στατιστικών μοντέλων για προβλέψεις ή **κατηγοριοποιήσεις** (**classifications**), ενώ η **αναλυτική κειμένου** (**text analytics**) εφαρμόζει στατιστικές, γλωσσικές και δομικές τεχνικές για την εξαγωγή και την κατηγοριοποίηση πληροφοριών από πηγές κειμένου και αφορά μη δομημένα δεδομένα.

2.1 Συλλογή δεδομένων (Gathering data)

Η συλλογή των ακατέργαστων δεδομένων, που αποτελεί την πρώτη φάση της διαχείρισης δεδομένων, γίνεται με ποικίλους τρόπους.

Οι πιο συνήθεις είναι:

1) **Με τη χρήση του API από κάποιο site.**

Εάν τα απαιτούμενα δεδομένα είναι διαθέσιμα σε συγκεκριμένους ιστότοπους, τότε μπορούμε να χρησιμοποιήσουμε, σε περίπτωση που υπάρχουν, τα API των ιστοτόπων, έτσι ώστε να αποθηκεύσουμε τα δεδομένα στη δική μας τοπική βάση δεδομένων ή όπου αλλού θέλουμε για να μπορέσουμε στη συνέχεια να τα επεξεργαστούμε. Συχνά, με αυτό τον τρόπο τα δεδομένα που συλλέγουμε από το διαδίκτυο μας παρέχονται σε μορφή JSON και απαιτείται περαιτέρω επεξεργασία για τη μετατροπή του JSON στη μορφή ".csv" που χρησιμοποιείται συνήθως στην ανάλυση δεδομένων.

2) **Βάσεις δεδομένων**

Εάν τα απαιτούμενα δεδομένα είναι διαθέσιμα στις βάσεις δεδομένων κάποιων εταιρειών ή οργανισμών που έχουμε πρόσβαση, τότε μπορούμε εύκολα, χρησιμοποιώντας ερωτήματα SQL, να εξαγάγουμε τα δεδομένα που θέλουμε από αυτές.

3) **Ιστότοποι που έχουν ως αντικείμενο τους την Επιστήμη των Δεδομένων (Data Science)**

Μπορούμε να πάρουμε δεδομένα για να τα αναλύσουμε από ιστότοπους που έχουν ως αντικείμενο τους την **επιστήμη των δεδομένων** (**data science**), όπως για παράδειγμα είναι το kaggle.com, που εξειδικεύεται σε αντίστοιχους διαγωνισμούς ή και πρακτικές εξοικείωσης με το αντικείμενο. Σε τέτοιους ιστότοπους μπορούμε να βρούμε τεράστια πληθώρα δεδομένων, τα οποία μάλιστα βρίσκονται έτοιμα σε αρχεία της μορφής ".csv", προκειμένου να ξεκινήσουμε άμεσα την ανάλυση τους.

2.2 Γλώσσες προγραμματισμού για την Ανάλυση Δεδομένων

Υπάρχουν δυο εργαλεία για να μπορέσει να χρησιμοποιηθεί εκτενώς ο υπολογιστής στην ανάλυση δεδομένων. Πρόκειται για δυο γλώσσες προγραμματισμού που παρέχουν τη δυνατότητα να διερευνηθεί σε βάθος μια ανάλυση δεδομένων. Και οι δυο γλώσσες προγραμματισμού ενσωματώνουν σχεδόν στο σύνολο τους όλες τις γνώσεις της επιστήμης δεδομένων, παρέχοντας υπερβολικά μεγάλες δυνατότητες στους χρήστες τους.

1) **Python**

Είναι ίσως η πιο πλήρης, αντικειμενοστρεφής γλώσσα προγραμματισμού με υψηλότατες προγραμματιστικές και υπολογιστικές δυνατότητες, καθώς και πολλά πακέτα και εργαλεία που εξυπηρετούν την ανάλυση δεδομένων. Οι ενσωματωμένες δομές δεδομένων υψηλού επιπέδου που παρέχει, σε συνδυασμό με τη δυναμική δέσμευση για τις μεταβλητές, την καθιστούν πολύ ελκυστική για απευθείας χρήση σε προβλήματα μηχανικής μάθησης. Επίσης, μπορεί να χρησιμοποιηθεί ως συνδετικός κρίκος μεταξύ εφαρμογών από διαφορετικές γλώσσες προγραμματισμού.

2) **R**

Πρόκειται για μια γλώσσα προγραμματισμού ανοιχτού κώδικα που έχει ως κυριότερο αντικείμενο τα μαθηματικά και κατ’ επέκταση την υπολογιστική στατιστική και τα γραφικά. Η γλώσσα R χρησιμοποιείται ευρέως στην επιστήμη των δεδομένων μεταξύ των στατιστικολόγων, για την ανάπτυξη στατιστικών παρατηρήσεων και ανάλυσης δεδομένων.

2.3 Προεπεξεργασία δεδομένων (Data preprocessing)

Η προεπεξεργασία δεδομένων είναι ένα σημαντικό βήμα πριν τη διαδικασία της **εξόρυξης δεδομένων** (**data mining**). Η συλλογή δεδομένων που είδαμε έχει το μειονέκτημα πως τα δεδομένα που αποκτούμε έχουν πολλά προβλήματα και δεν μπορούν να χρησιμοποιηθούν αυτούσια, αφού πρόκειται για **ακατέργαστα δεδομένα** (**raw data**). Έτσι, μπορούμε να βρούμε κάποιες τιμές που δεν έχουν καμία λογική εξήγηση, όπως για παράδειγμα μια αρνητική τιμή για ένα **χαρακτηριστικό** (**feature**) με τίτλο "Βάρος" ή ένα μη ρεαλιστικό συνδυασμό, όπως το φύλο να είναι άντρας και η τιμή για το χαρακτηριστικό "Εγκυμοσύνη" να είναι θετική. Ακόμα, μπορεί σε πολλά χαρακτηριστικά - πεδία να υπάρχουν κενές τιμές.

Αν οι συγκεκριμένες τιμές παραμείνουν χωρίς να ληφθεί καμία μέριμνα για αυτές, όταν η ανάλυση δεδομένων φτάσει στην εξόρυξη δεδομένων και συγκεκριμένα στην εύρεση αλγορίθμου, τότε θα παραχθούν εντελώς παραπλανητικά αποτελέσματα. Επομένως, η ορθότητα και η ποιότητα των δεδομένων είναι πρωταρχικής σημασίας και πρέπει να διευθετηθεί πριν φτάσουμε σε οποιοδήποτε συμπέρασμα.

Συχνά, η προεπεξεργασία δεδομένων είναι η πιο σημαντική φάση ενός προγράμματος **μηχανικής μάθησης** (**machine learning**). Εάν υπάρχουν άσχετες και λάθος πληροφορίες, θόρυβοι που εμποδίζουν τη σωστή εκτίμηση των τιμών και αναξιόπιστα δεδομένα, τότε η ανακάλυψη αποτελεσματικών προτύπων και αλγόριθμων κατά τη διάρκεια της εκπαίδευσης μπορεί να είναι αδύνατη. Το προϊόν της προεπεξεργασίας δεδομένων είναι το τελικό **σύνολο εκπαίδευσης** (**training set**).

Όσον αφορά τις επιμέρους διεργασίες, υπάρχουν αμφισβητήσεις σχετικά με την κατηγορία της προεπεξεργασίας στην οποία ανήκουν ορισμένες από αυτές, αφού διαφορετικοί επιστήμονες τις καταχωρούν σε άλλες κατηγορίες. Επειδή λοιπόν, τα όρια ανάμεσα στις κατηγορίες και στο τι περιλαμβάνουν είναι κάποιες φορές δυσδιάκριτα, ερευνώντας το ζήτημα καταλήξαμε πως μια αποδεκτή λίστα για την προεπεξεργασία δεδομένων είναι η παρακάτω:

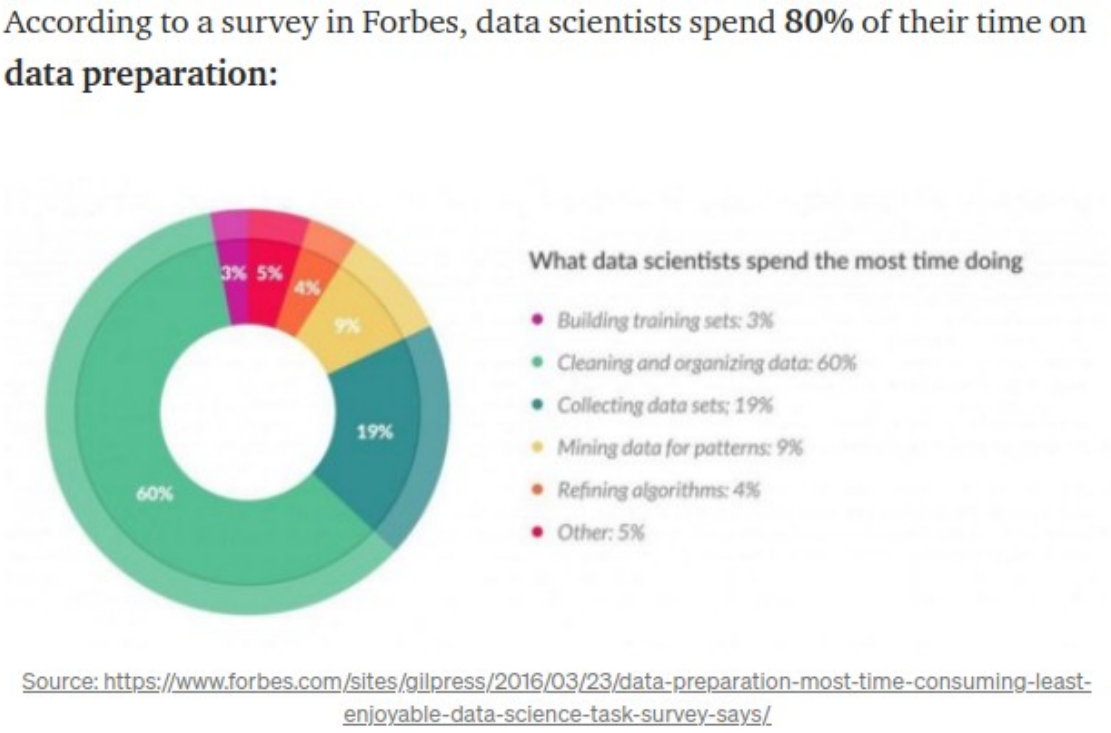
1) **Διαχείριση δεδομένων** (**Data Wrangling**)

2) **Διερευνητική Ανάλυση Δεδομένων, ΔΑΔ** (**Exploratory Data Analysis, EDA**)

3) **Μηχανική χαρακτηριστικών** (**Feature Engineering**)

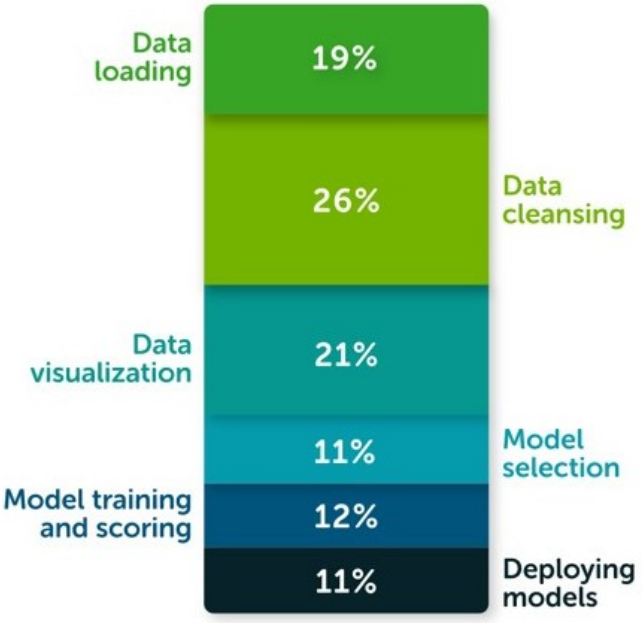
Ένας άλλος παράγοντας που τονίζει τη σημαντικότητα της προεπεξεργασίας των δεδομένων είναι το γεγονός ότι της αναλογεί πάνω από τα 2/3 του συνολικού χρόνου μιας ανάλυσης δεδομένων. Συνήθως, οι **αναλυτές δεδομένων** (**data analysts**) αφιερώνουν για την προεπεξεργασία το 65% - 80% του συνολικού χρόνου. ***[2]***

Από αυτό συμπεραίνουμε πως παρότι η μοντελοποίηση είναι το κύριο τμήμα της συγκεκριμένης επιστήμης, αφού μας προσφέρει τα ζητούμενα συμπεράσματα, απαιτεί ένα σαφώς μικρότερο χρονικό διάστημα για να υλοποιηθεί.



Εικόνα 2.1: Καταμερισμός χρόνου στην ανάλυση δεδομένων

According to The State of Data Science 2020 survey, data management, exploratory data analysis (EDA), feature selection, and feature engineering accounts for more than 66% of a data scientist's time (see the following diagram).



Εικόνα 2.2: Καταμερισμός χρόνου στην ανάλυση δεδομένων

2.3.1 Διαχείριση δεδομένων (Data Wrangling)

Η **διαχείριση δεδομένων** (**data wrangling**), αναφέρεται επίσης και ως **καθαρισμός δεδομένων** (**data cleaning**) και **διαμόρφωση δεδομένων** (**data munging**). Πρόκειται για τη διαδικασία καθαρισμού και μετατροπής των δεδομένων από τον τύπο των **ακατέργαστων δεδομένων** (**raw data**) σε συγκεκριμένη μορφοποίηση, προκειμένου να καταστούν κατάλληλα για τη διαδικασία της ανάλυσης τους. ***[3]***

Ο στόχος της διαχείρισης δεδομένων είναι να διασφαλιστεί η καλή ποιότητα και η ορθή χρησιμότητα τους.

Η διαχείριση δεδομένων, γενικότερα ακολουθεί κάποια συγκεκριμένα στάδια, ώστε τα ακατέργαστα δεδομένα, μετά την κατάλληλη επεξεργασία, να καταχωρηθούν σε βάσεις δεδομένων για αποθήκευση και μελλοντική χρήση.

Το καινούργια δεδομένα μπορεί να παρέχουν ουσιαστικότερες πληροφορίες και ονομάζονται **μεταδεδομένα** (**metadata**). Είναι σημαντικό να διασφαλιστεί ότι τα μεταδεδομένα ανταποκρίνονται στην πληροφορία που περιέχουν τα αρχικά δεδομένα, διαφορετικά μπορεί να αλλοιώσουν τα αποτελέσματα της ανάλυσης δεδομένων. ***[4]***

Εκτός των άλλων, η διαχείριση δεδομένων επιτρέπει στους επιστήμονες να αναλύουν πιο πολύπλοκα δεδομένα πιο γρήγορα και να επιτυγχάνουν πιο ακριβή αποτελέσματα.

Ίσως, το πιο σημαντικό κομμάτι της όλης διαδικασίας είναι να καταλήξουν τα δεδομένα να είναι **“τακτοποιημένα”** (**“tidy”**). Αυτό σημαίνει πως σε ένα σύνολο δεδομένων που θα έχει τη μορφή ενός πίνακα:

i) Κάθε γραμμή, θα αποτελεί μια εγγραφή ή μια παρατήρηση ή ένα δείγμα

ii) Κάθε στήλη, θα είναι ένα **χαρακτηριστικό** (**feature**) ή ένα πεδίο ή μία μεταβλητή

iii) Κάθε πίνακας, θα αποτελεί το σύνολο των δεδομένων.

2.3.1.1 Βασικά στάδια της διαχείρισης δεδομένων

Μια γενικά αποδεκτή λίστα έξι βασικών δραστηριοτήτων της διαχείρισης δεδομένων είναι η παρακάτω: ***[5]***

1) **Ανακάλυψη** (**Discovery**)

Πριν μπούμε βαθειά στην ανάλυση των δεδομένων, πρωταρχικός στόχος είναι να αντιληφθούμε πλήρως τα δεδομένα μας, κάτι που θα μας οδηγήσει στο να καταλάβουμε πως θα τα αναλύσουμε. Για παράδειγμα, σε αυτό το κομμάτι μας παρέχεται η δυνατότητα να ορίσουμε τον τύπο του κάθε χαρακτηριστικού, δηλαδή αν είναι αριθμητικός, κατηγορικός ή κάτι άλλο.

Εδώ, πρέπει να επισημανθεί πως στη συγκεκριμένη δραστηριότητα, πολλοί επιστήμονες περιλαμβάνουν και ορισμένες διεργασίες που άλλοι θεωρούν πως ανήκουν στη **Διερευνητική Ανάλυση Δεδομένων** (**Exploratory Data Analysis, EDA**). Ένα τέτοιο παράδειγμα είναι η ανακάλυψη προτύπων και συσχετίσεων μεταξύ των **χαρακτηριστικών** (**features**) των δεδομένων.

2) **Δομή** (**Structuring**)

Γνωρίζοντας πως τα δεδομένα διατίθενται σε διαφορετικούς τύπους και μορφές, θα πρέπει να μπορούν να οργανωθούν και να τυποποιηθούν, ώστε να έχουμε αργότερα την ευχέρεια να τα επεξεργαστούμε. Έτσι για παράδειγμα, θα πρέπει για κάθε χαρακτηριστικό μετά από αυτό το στάδιο να είναι εφικτό να υπάρχει για όλα τα δεδομένα ενιαίος τρόπος να συγχωνευτούν, να ταξινομηθούν ή να μετασχηματιστούν. Ακόμα, να μπορούμε να διασπάσουμε κάποιο χαρακτηριστικό σε περισσότερα χαρακτηριστικά με τις ίδιες κοινές εντολές για όλες τις **εγγραφές** (**row**) των δεδομένων. Αυτό το κομμάτι εξυπηρετεί τόσο τους υπολογισμούς, όσο και την ανάλυση των δεδομένων.

3) **Καθαρισμός** (**Cleansing**)

Τα ακατέργαστα δεδομένα τις περισσότερες φορές είναι προβληματικά στη χρήση τους και είναι αδύνατη η απευθείας επεξεργασία τους. Συνηθισμένα παραδείγματα δυσκολιών που υπάρχουν στα ακατέργαστα δεδομένα είναι τα παρακάτω:

- **Κενές τιμές** (**null values**) σε κάποια χαρακτηριστικά
- Ύπαρξη **ακραίων τιμών** (**outliers**) σε ορισμένα πεδία
- Διαφορετική **μορφοποίηση** (**formatting**) για τιμές στο ίδιο χαρακτηριστικό (π.χ. 2-5-2021, 02/05/2021, 2 Μαΐου 2021 κ.α.)

Τέτοια και άλλα παραδείγματα πρέπει να διευθετηθούν προκειμένου να μπορέσουν τα δεδομένα να χρησιμοποιηθούν σωστά στα επόμενα στάδια.

4) **Εμπλουτισμός** (**Enriching**)

Ο ρόλος αυτής της διαδικασίας είναι να εξάγουμε νέα δεδομένα, που ενδεχομένως θα μας είναι πιο χρήσιμα. Για παράδειγμα, αν έχουμε κάποιες διευθύνσεις θα μπορούσαμε να προσθέσουμε δυο ακόμα χαρακτηριστικά στα δεδομένα μας, όπως είναι οι συντεταγμένες του γεωγραφικού μήκους και πλάτους των διευθύνσεων.

Αυτή τη διαδικασία, πολλοί επιστήμονες την εντάσσουν στη **μηχανική χαρακτηριστικών** (**feature engineering**), που όπως θα δούμε και σε επόμενα κεφάλαια μοιάζει λογικότερο.

5) **Επικύρωση** (**Validation**)

Με την επικύρωση των δεδομένων μπορούμε να επαληθεύσουμε τη συνέπεια, την ποιότητα και την ασφάλεια των δεδομένων. Για παράδειγμα σε κάποια δεδομένα που ένα χαρακτηριστικό αποτελεί η ημερομηνία γέννησης θα πρέπει να εξασφαλίσουμε πως τα δείγματα των ανθρώπων που διαθέτουμε βρίσκονται σε μια λογική ηλικία και όχι να βρίσκουμε ανθρώπους με μελλοντική ημερομηνία γέννησης.

6) **Δημοσίευση** (**Publishing**)

Οι **επιστήμονες των δεδομένων** (**data scientists**) πρέπει πριν το τελευταίο στάδιο της διαχείρισης δεδομένων να έχουν δημιουργήσει δεδομένα έτοιμα να χρησιμοποιηθούν από αλγόριθμους και προγραμματιστικές εφαρμογές χωρίς προβλήματα, ώστε να γίνει η τελική φάση της ανάλυσης των δεδομένων και να εξαχθούν τα συμπεράσματα. Επίσης, πρέπει να επεξηγούν τα βήματα και τη λογική που ακολουθούν, καθώς και να τεκμηριώνουν τα χαρακτηριστικά που έχουν δημιουργήσει.

Τέλος, καλό είναι να υπάρχει μια **τυποποιημένη προγραμματιστική επεξεργασία** (**pipeline**), ώστε όταν επικαιροποιηθούν τα ακατέργαστα δεδομένα να μπορεί να ξανατρέξει η ίδια ακολουθία και να παραχθούν εκ νέου δεδομένα που στη συνέχεια θα χρησιμοποιηθούν για την καινούργια ανάλυση δεδομένων.

2.3.1.2 Βασικές τεχνικές της διαχείρισης δεδομένων

Στη διαχείριση δεδομένων συναντάμε διάφορες τεχνικές που είναι ιδιαιτέρως σημαντικές και ορισμένες από αυτές δεν πρέπει να τις παραλείψουμε.

Μερικές από τις σημαντικότερες είναι: ***[6]***

1) **Εύρεση του τύπου για κάθε χαρακτηριστικό του συνόλου δεδομένων**

Θα πρέπει να εντοπιστούν οι τύποι των χαρακτηριστικών και να διορθωθούν τυχόν σφάλματα, όπως για παράδειγμα, αν ένα αριθμητικό χαρακτηριστικό έχει αρχικά καταχωρηθεί ως κείμενο.

Οι κυριότεροι τύποι των δεδομένων των χαρακτηριστικών είναι:

i) **Αριθμητικός** τύπος (**Numerical**)

- **Συνεχείς** τιμές (**Continuous**)
- **Διακριτές** τιμές (**Discrete**)

ii) **Κατηγορικός** τύπος (**Categorical**)

- **Ονομαστικές** τιμές (**Nominal**)
- **Διατεταγμένες** τιμές (**Ordinal**)

iii) **Ημερομηνία** (**Date**)

iv) **Χρόνος** – Ώρα (**Time**)

2) **Συμπλήρωση ελλειπουσών τιμών** (**Misssing Data imputation**)

Οι ελλείπουσες τιμές, δηλαδή οι τιμές που λείπουν από κάποιο **χαρακτηριστικό** (**feature**) για κάποιες **εγγραφές** (**rows**), είναι ένα από τα πιο κοινά προβλήματα που μπορείτε να αντιμετωπίσετε, όταν προσπαθείτε να προετοιμάσετε τα δεδομένα σας. Ο λόγος για τις τιμές που λείπουν μπορεί να είναι ανθρώπινα λάθη, διακοπές στη ροή δεδομένων, ζητήματα απορρήτου και πολλοί άλλοι λόγοι. Δυστυχώς, οι τιμές που λείπουν επηρεάζουν την απόδοση των μοντέλων μηχανικής μάθησης, δημιουργώντας πρόβλημα στην αποτελεσματικότητα του μοντέλου. Επίσης, πολλοί αλγόριθμοι της μηχανικής μάθησης δε δέχονται σύνολα δεδομένων με ελλείπουσες τιμές και επιστρέφουν σφάλμα, με αποτέλεσμα να μην μπορούν να χρησιμοποιηθούν.

Αν λοιπόν για κάποιο χαρακτηριστικό έχουμε κενά σε πάρα πολλές εγγραφές μπορούμε να αφαιρέσουμε τελείως το συγκεκριμένο χαρακτηριστικό από τα δεδομένα μας.

Ένας άλλος τρόπος αντιμετώπισης είναι η αυτόματη διαγραφή κάθε εγγραφής που έχει κάποιο χαρακτηριστικό του **κενό** (**null value**). Έτσι, μπορούμε να διαγράψουμε τις συγκεκριμένες εγγραφές. Δεν υπάρχει κάποιο συγκεκριμένο όριο, αλλά ένα ποσοστό αποδεκτό είναι να διατηρήσουμε κατ' ελάχιστο ένα 70% του συνόλου δεδομένων.

Το πρόβλημα με αυτόν τον τρόπο, που είναι σχετικά και ο πιο εύκολος στην εφαρμογή του, είναι πως μειώνονται οι αρχικές εγγραφές του **συνόλου δεδομένων εκπαίδευσης** (**training dataset**), χάνοντας πολύτιμες πληροφορίες που μας παρέχουν οι διαγραμμένες εγγραφές για όλα τα υπόλοιπα χαρακτηριστικά τους που περιέχουν τιμές. Για αυτό προτείνονται οι δυο παρακάτω τρόποι συμπλήρωσης των τιμών.

i) Συμπλήρωση σε αριθμητικό χαρακτηριστικό (Numerical Imputation)

Προφανώς, ο συγκεκριμένος τρόπος αναφέρεται στα χαρακτηριστικά που οι τιμές τους είναι αριθμητικές.

Η συμπλήρωση είναι προτιμότερη επιλογή σε σχέση με τη διαγραφή, επειδή διατηρεί το μέγεθος των δεδομένων. Ωστόσο, η επιλογή του αριθμού που θα συμπληρώσει το κενό της **μη διαθέσιμης** (**Not Available, NA**) τιμής είναι κάτι που πρέπει να εξεταστεί. Αν για παράδειγμα, έχουμε ένα πεδίο που δείχνει το "πλήθος επισκέψεων πελατών εντός του μήνα", στις τιμές που λείπουν μπορούμε να συμπληρώσουμε το 0, αν θεωρούμε πως είναι λογικό.

Εκτός από την περίπτωση που έχουμε μια προεπιλεγμένη τιμή για τις τιμές που λείπουν, υπάρχουν δυο τεχνικές για να συμπληρώσουμε τα κενά. Η πρώτη είναι με τη χρήση του διάμεσου της στήλης για όσες εγγραφές δεν είναι κενές. Με αυτόν τον τρόπο δεν επηρεαζόμαστε και από τις **ακραίες τιμές** (**outliers**), όπως συμβαίνει με τη μέση τιμή.

Η δεύτερη τεχνική είναι βρίσκοντας τη μέση τιμή και τη διακύμανση και διαλέγοντας τυχαίες τιμές από ένα διάστημα που διαφέρει απόλυτα από τη μέση τιμή, για κάποια τιμή που είναι συνάρτηση της τυπικής απόκλισης. Ένα καλό παράδειγμα διαστήματος είναι με κέντρο τη μέση τιμή να είναι 4 φορές η τυπική απόκλιση $(\bar{x}-2*s,\bar{x}+2*s)$

ii) Συμπλήρωση σε κατηγορικό χαρακτηριστικό (Categorical Imputation)

Η συμπλήρωση των τιμών που λείπουν με την κατηγορία που έχει τη μεγαλύτερη συχνότητα στη στήλη είναι μια καλή επιλογή για το χειρισμό κατηγορικών χαρακτηριστικών. Αν όμως, οι τιμές στη στήλη

κατανέμονται ομοιόμορφα και δεν υπάρχει κυρίαρχη τιμή, το να υπολογίζετε μια κατηγορία όπως "Άλλο" μπορεί να είναι πιο λογικό, γιατί σε μια τέτοια περίπτωση, επειδή θα είναι καθαρά τυχαία η επιλογή μπορεί να μας οδηγήσει σε εσφαλμένα συμπεράσματα.

Βεβαίως, υπάρχουν και άλλες πιο προχωρημένες τεχνικές, όπως το KNN imputation και το MICE, που λαμβάνουν υπ' όψιν τους τις τιμές των υπόλοιπων μη κενών χαρακτηριστικών της εγγραφής.

3) **Διαχείριση ακραίων τιμών (Handling outliers)**

Πριν αναφερθούμε στον τρόπο χειρισμού των ακραίων τιμών, πρέπει να τονίσουμε πως ο καλύτερος τρόπος για τον εντοπισμό των ακραίων τιμών είναι η **οπτικοποίηση** (**visualization**) των δεδομένων. Όλες οι άλλες στατιστικές μεθοδολογίες είναι πιθανόν να μας οδηγήσουν σε σφάλμα, ενώ η οπτικοποίηση μας παρουσιάζει ευκρινώς τις ακραίες τιμές. Το πλεονέκτημα που έχουν οι στατιστικές μεθοδολογίες είναι πως δίνουν γρήγορα αποτελέσματα. Εδώ, θα αναφερθούμε σε δύο στατιστικά μεγέθη που βοηθούν στην εύρεση των ακραίων τιμών.

i) Με τυπική απόκλιση (standard deviation)

Εάν μια τιμή έχει απόσταση μεγαλύτερη από τη μέση τιμή κατά ένα πολλαπλάσιο **A** της τυπικής απόκλισης, τότε θεωρείται ακραία τιμή. Δηλαδή οι τιμές που είναι εκτός του διαστήματος $(\bar{x} - A * s, \bar{x} + A * s)$.

Αυτό το πολλαπλάσιο **A** πρακτικά συνήθως είναι μεταξύ των αριθμών 3 και 4, ανάλογα πόσο μεγάλο θεωρούμε το εύρος των φυσιολογικών τιμών. Για παράδειγμα, μπορούμε να θεωρήσουμε ως ακραία τιμή, οποιαδήποτε τιμή δε βρίσκεται εντός του διαστήματος $(\bar{x} - 4 * s, \bar{x} + 4 * s)$

ii) Με εκατοστημόρια (percentiles)

Μια άλλη μαθηματική μέθοδος για την ανίχνευση ακραίων τιμών είναι η χρήση εκατοστημορίων. Μπορείτε να υποθέσετε ένα ορισμένο ποσοστό στην αρχή και στο τέλος των ταξινομημένων τιμών, στο οποίο θα θεωρούμε ότι βρίσκονται οι ακραίες τιμές. Τις εγγραφές που περιλαμβάνουν αυτές τις τιμές θα τις διαγράψουμε και άρα θα παραμείνουν υπόλοιπες. Αν για παράδειγμα ορίσουμε το ποσοστό στο 5% θα αφαιρέσουμε το 5% των μικρότερων και των μεγαλύτερων τιμών και θα κρατήσουμε το υπόλοιπο 90%. Αυτό βέβαια, μπορεί να

μην είναι σωστό, αφού οι αφαιρούμενες τιμές ενδέχεται να μην είναι ακραίες.

Μια άλλη εναλλακτική στην περίπτωση που δε θέλουμε να διαγράψουμε τις παραπάνω τιμές είναι να κρατήσουμε τις συγκεκριμένες εγγραφές και να αντικαταστήσουμε τις τιμές τους με τις τιμές που αντιστοιχούν στα εκατοστημόρια του 5% για τις μικρότερες τιμές από αυτό και στο εκατοστημόριο του 95% για τις μεγαλύτερες τιμές από αυτό. Πρέπει να λάβουμε υπ' όψιν πως με αυτόν τον τρόπο επηρεάζουμε την κατανομή των τιμών, οπότε θα πρέπει να μην υπερβάλλουμε και ανά περίπτωση, να ελέγχουμε αν θα το κάνουμε ή όχι. ***[7]***

2.3.2 Διερευνητική Ανάλυση Δεδομένων, ΔΑΔ (Exploratory Data Analysis, EDA)

Όταν τα δεδομένα συλλεχθούν, καθαριστούν και υποβληθούν σε επεξεργασία, είναι έτοιμα για ανάλυση. Καθώς χειριζόμαστε τα δεδομένα, μπορεί να βρούμε ότι έχουμε τις ακριβείς πληροφορίες που χρειαζόμαστε, αλλά μπορεί και να χρειαστεί να συλλέξουμε περισσότερα δεδομένα. Κατά τη διάρκεια αυτής της φάσης, μπορούμε να χρησιμοποιήσουμε εργαλεία ανάλυσης δεδομένων και λογισμικό που θα μας βοηθήσουν να κατανοήσουμε, να ερμηνεύσουμε και να εξαγάγουμε συμπεράσματα με βάση τα ζητούμενα.

Όπως αναφέραμε προηγουμένως ένα σημαντικό κομμάτι της διαχείρισης δεδομένων αποτελεί η **Διερευνητική Ανάλυση Δεδομένων, ΔΑΔ** (**Exploratory Data Analysis, EDA**). Η ΔΑΔ αναλύει και διερευνά σύνολα δεδομένων και συνοψίζει τα κύρια χαρακτηριστικά τους, χρησιμοποιώντας συχνά μεθόδους **οπτικοποίησης** (**visualization**).

Χρησιμοποιείται κυρίως για να φανεί ποια δεδομένα μπορούν να αποκαλύψουν κάτι πέρα από τα τυπικά στοιχεία που προσφέρει η μοντελοποίηση και να παράσχουν μια καλύτερη κατανόηση των μεταβλητών των δεδομένων και των σχέσεων μεταξύ τους ή ακόμα να μας υποδείξει ποιες μεταβλητές είναι άχρηστες και μπορούν να

αφαιρεθούν από το σύνολο των δεδομένων, αφού δεν προσφέρουν στη μελέτη μας. Μπορεί επίσης να μας βοηθήσει να προσδιορίσουμε εάν οι τεχνικές που εξετάζουμε για την ανάλυση δεδομένων είναι κατάλληλες.

Ο κύριος σκοπός της ΔΑΔ είναι να βοηθήσει στην εξέταση των δεδομένων, πριν κάνουμε οποιεσδήποτε παραδοχές. Μπορεί να βοηθήσει στον εντοπισμό προφανών λαθών, στην εύρεση προτύπων στα υπάρχοντα δεδομένα, στον εντοπισμό ακραίων τιμών, ανωμαλιών που εμφανίζουν τα δεδομένα, στην εύρεση αξιόλογων σχέσεων μεταξύ των μεταβλητών ή να μας οδηγήσει σε άλλες σημαντικές παρατηρήσεις.

Οι επιστήμονες δεδομένων μπορούν να χρησιμοποιήσουν τη ΔΑΔ για να διασφαλίσουν ότι τα αποτελέσματα που παράγουν είναι έγκυρα και εφαρμόσιμα και σε άλλα αντίστοιχα σύνολα δεδομένων. Για παράδειγμα, μπορούμε να βρούμε στατιστικά στοιχεία, όπως είναι οι τυπικές αποκλίσεις των δεδομένων, η ανάλυση των κατηγορικών μεταβλητών ή η εύρεση διαστημάτων εμπιστοσύνης.

2.3.2.1 Τεχνικές Διερευνητικής Ανάλυσης Δεδομένων

Υπάρχουν γενικά δύο κατηγορίες ΔΑΔ. Οι γραφικές και οι μη γραφικές, που ουσιαστικά αναφέρονται στην **οπτικοποίηση δεδομένων** (**data visualization**) ή όχι. Κάθε μια από τις παραπάνω κατηγορίες χωρίζονται σε δυο υποκατηγορίες. Τις **μονομετάβλητες** (**univariate**) και τις **πολυμετάβλητες** (**multivariate**), με βάση την ανεξαρτησία μεταξύ των μεταβλητών που υπάρχει στα δεδομένα μας. ***[8]***

1) **Μονομετάβλητη χωρίς γραφήματα** (**Univariate non-graphical**)

Εδώ, τα δεδομένα διαθέτουν μία μόνο μεταβλητή και το ΔΑΔ γίνεται κατά κύριο λόγο σε μορφή πινάκων. Έτσι, για παράδειγμα μπορούμε να έχουμε την καταγραφή στατιστικών στοιχείων.

2) **Μονομετάβλητη με γραφήματα** (**Univariate graphical**)

Το ΔΑΔ περιλαμβάνει γραφικά εργαλεία όπως **ραβδογράμματα** (**bar charts**) και **ιστογράμματα** (**histograms**) για να φανεί μια εικόνα των ιδιοτήτων αυτής της μεταβλητής, καθώς και άλλα οπτικοποιημένα στατιστικά στοιχεία.

3) **Πολυμετάβλητη χωρίς γραφήματα** (**Multivariate non-graphical**)

Μη γραφικές μέθοδοι όπως οι **πίνακες διασταύρωσης** (**crosstabs**) χρησιμοποιούνται για την απεικόνιση της σχέσης μεταξύ δύο ή περισσότερων μεταβλητών. Επίσης, κάποιες στατιστικές τιμές, όπως ο συντελεστής συσχέτισης, μπορούν να δείξουν εάν υπάρχει πιθανή σχέση ανάμεσα σε διαφορετικές μεταβλητές, καθώς και το μέτρο της συσχέτισης τους.

4) **Πολυμετάβλητη με γραφήματα** (**Univariate graphical**)

Μια γραφική αναπαράσταση μας δίνει μια σημαντικά καλύτερη κατανόηση της σχέσης μεταξύ πολλαπλών μεταβλητών. Τέτοιου είδους γραφήματα αποτελούν τα **διαγράμματα διασποράς** (**scatter plots**), τα **ραβδογράμματα** (**bar charts**) και οι **θερμοχάρτες** (**heatmaps**), όπως για παράδειγμα στον **πίνακα συσχέτισης** (**correlation matrix**).

2.3.2.2 Επιλογή χαρακτηριστικών (Feature selection)

Ένα μεγάλο πρόβλημα που αντιμετωπίζουν οι επιστήμονες των δεδομένων είναι πως σε ένα σύνολο δεδομένων, ενδεχομένως να υπάρχουν δεκάδες ή εκατοντάδες χαρακτηριστικά που προφανώς δεν επηρεάζουν όλα το ζητούμενο **στόχο** (**target**).

Η **επιλογή χαρακτηριστικών** (**feature selection**) είναι μια από τις βασικές έννοιες της μηχανικής μάθησης που επηρεάζει σημαντικά την απόδοση του μοντέλου μας. Τα χαρακτηριστικά που τελικά περιλαμβάνονται στο **εκπαιδευτικό σύνολο δεδομένων** (**training dataset**) που θα χρησιμοποιήσουμε για το μοντέλο της **εξόρυξης**

(**data mining**), ουσιαστικά καθορίζουν την απόδοση που μπορούμε να επιτύχουμε. ***[9]***

Είναι ιδιαιτέρως σημαντικό τα χαρακτηριστικά που είναι καθόλου ή ελάχιστα σημαντικά να αφαιρεθούν από το σύνολο δεδομένων, αφού μπορούν να επηρεάσουν αρνητικά την απόδοση του μοντέλου.

Η ύπαρξη τους μπορεί να μειώσει την ακρίβεια των μοντέλων και να κάνει το μοντέλο να μάθει λανθασμένα βάσει χαρακτηριστικών που στην πραγματικότητα δε σχετίζονται με το ζητούμενο στόχο. Τα πλεονεκτήματα της επιτυχημένης επιλογής χαρακτηριστικών είναι τα εξής:

- Μειώνει την **υπερπροσαρμογή** (**overfitting**), αφού με λιγότερα αχρείαστα δεδομένα είναι πιθανότερο το μοντέλο να αποφύγει να μάθει να επιλέγει με βάση **θορύβους** (**noises**).

- Βελτιώνει την ακρίβεια, αφού τα αχρείαστα δεδομένα μπορεί να παραπλανήσουν το μοντέλο και να χειροτερέψουν την απόδοση.

- Μειώνει το χρόνο εκπαίδευσης, αφού μειώνεται η πολυπλοκότητα του αλγορίθμου και οι αλγόριθμοι εκπαιδεύονται πιο γρήγορα.

Επίσης, ενδεικτικά θα αναφέρουμε κάποιες τεχνικές για την επιλογή χαρακτηριστικών:

1) **Μονομετάβλητη επιλογή (Univariate Selection)**

Τα στατιστικά τεστ μπορούν να χρησιμοποιηθούν για την επιλογή των χαρακτηριστικών που έχουν την ισχυρότερη σχέση με τη μεταβλητή εξόδου. Υπάρχουν βιβλιοθήκες στις προτεινόμενες γλώσσες προγραμματισμού που μπορούν να χρησιμοποιηθούν με μια σειρά διαφορετικών στατιστικών δοκιμών για την επιλογή ενός συγκεκριμένου αριθμού χαρακτηριστικών.

2) **Σημαντικότητα χαρακτηριστικού (Feature Importance)**

Μπορείτε να λάβετε τη σημαντικότητα του χαρακτηριστικού για κάθε χαρακτηριστικό από το σύνολο δεδομένων, χρησιμοποιώντας την αντίστοιχη ιδιότητα. Η σημαντικότητα χαρακτηριστικού μας δίνει μια βαθμολογία για κάθε χαρακτηριστικό των δεδομένων μας. Όσο

μεγαλύτερη είναι η βαθμολογία, τόσο πιο σχετικό είναι το χαρακτηριστικό σε σχέση με τη μεταβλητή εξόδου.

Στις γλώσσες προγραμματισμού, η σημαντικότητα χαρακτηριστικού είναι ενσωματωμένη στους **κατηγοριοποιητές βασισμένους σε δέντρα** (**Tree Based Classifiers**) και μπορούν να μας δείξουν τα Ν σημαντικότερα χαρακτηριστικά που θέλουμε, όπου Ν είναι ο αριθμός που εμείς έχουμε επιλέξει.

3) **Πίνακας συσχέτισης με θερμοχάρτη (Correlation Matrix with Heatmap)**

Η συσχέτιση δηλώνει πώς τα χαρακτηριστικά σχετίζονται μεταξύ τους ή με τη μεταβλητή εξόδου. Η συσχέτιση μπορεί να είναι ανάλογη ή αντιστρόφως ανάλογη της σχέσης των μεταβλητών. Ο θερμοχάρτης, χάρις στα χρώματα του, διευκολύνει τον εντοπισμό των χαρακτηριστικών που σχετίζονται περισσότερο με τη μεταβλητή εξόδου.

2.3.3 Μηχανική χαρακτηριστικών (Feature Engineering)

Όπως έχουμε ήδη δει, οι αλγόριθμοι της μοντελοποίησης που στηρίζονται στη **μηχανική μάθηση** (**machine learning**) αναμένουν τα δεδομένα εισόδου με κάποια καθορισμένη μορφοποίηση για να λειτουργήσουν σωστά. Έτσι, προκύπτει η ανάγκη για τη **μηχανική χαρακτηριστικών** (**feature engineering**).

Οι κυριότεροι στόχοι της μηχανικής χαρακτηριστικών είναι δύο:

1) Η προετοιμασία του όσο το δυνατόν πιο κατάλληλου συνόλου δεδομένων εισόδου, προκειμένου να είναι συμβατό με τις απαιτήσεις του αλγορίθμου μηχανικής μάθησης και τα πεδία του να είναι όσο γίνεται πιο ουσιαστικά.

2) Η βελτίωση της απόδοσης των μοντέλων μηχανικής μάθησης.

Εξάλλου, η απόδοση του αλγόριθμου οφείλεται εξ ολοκλήρου στα χαρακτηριστικά που θα επιλέξουμε και στην επεξεργασία που θα υποστούν.

2.3.3.1 Τεχνικές της μηχανικής χαρακτηριστικών

Στη μηχανική χαρακτηριστικών συναντάμε διάφορες τεχνικές που μπορούμε να χρησιμοποιήσουμε. Κάποιες από αυτές μπορεί να λειτουργήσουν καλύτερα μόνο για συγκεκριμένους αλγόριθμους, ενώ κάποιες μπορεί να είναι αξιόλογες για τις περισσότερες περιπτώσεις ανάλυσης δεδομένων.

Γενικά, το ιδανικό για όσους θέλουν να ασχοληθούν με την ανάλυση δεδομένων είναι να εξασκήσουν όσες περισσότερες τεχνικές μηχανικής δεδομένων μπορούν.

Παρακάτω θα παρουσιάσουμε κάποιες πολύ σημαντικές τεχνικές ανάλυσης:

1) **Λογαριθμικός μετασχηματισμός (Log Transform)**

Ο λογαριθμικός μετασχηματισμός είναι ένας από τους πιο συχνά χρησιμοποιούμενους μαθηματικούς μετασχηματισμούς στη μηχανική χαρακτηριστικών. Τα πλεονεκτήματα του συγκεκριμένου μετασχηματισμού είναι:

i) Βοηθάει στον χειρισμό των **δεδομένων που ανήκουν σε ασύμμετρη κατανομή** (**skewed data**). Μετά τον μετασχηματισμό, η κατανομή προσεγγίζει περισσότερο την κανονική.

ii) Περιορίζει το σχετικό μέγεθος της διαφοράς μεταξύ δύο τιμών. Για παράδειγμα, η διαφορά των 3 ετών ανάμεσα σε δυο άτομα ηλικίας 10 και 13 ετών είναι σημαντικότερη σε σχέση με τη διαφορά δυο ατόμων 70 και 73 ετών. Το νέο αυτό χαρακτηριστικό, που προέρχεται από μια πολλαπλασιαστική διαδικασία και το λογαριθμικό μετασχηματισμό, κανονικοποιεί τις διαφορές μεγέθους δίνοντας τους την αντίστοιχη σχετική αξία.

iii) Μειώνει την επίδραση των ακραίων τιμών, λόγω της κανονικοποίησης στις διαφορές μεγέθους. Αυτό βοηθάει το μοντέλο να γίνει πιο ισχυρό.

Εδώ, πρέπει να σημειώσουμε πως για το λογαριθμικό μετασχηματισμό οι τιμές θα πρέπει να είναι θετικές, αλλιώς θα υπάρξει σφάλμα. Ένας τρόπος που δεν επηρεάζει το αποτέλεσμα, όταν πρόκειται για μη αρνητικές τιμές, είναι να προστίθεται η μονάδα στις αρχικές τιμές, δηλαδή **Log(x+1)**. Έτσι, όλες οι μετασχηματισμένες τιμές θα είναι εκ νέου μη αρνητικές. ***[10]***

2) **Κωδικοποίηση One-Hot (One-Hot Encoding)**

Η κωδικοποίηση One-hot είναι μια από τις πιο κοινές μεθόδους κωδικοποίησης στη μηχανική μάθηση. Η συγκεκριμένη κωδικοποίηση χρησιμοποιείται για να μπορούμε να διαχειριστούμε καλύτερα τα **κατηγορικά χαρακτηριστικά** (**categorical features**). Η γενική λογική αυτής της μεθοδολογίας είναι, αντί το χαρακτηριστικό να καταλαμβάνει μια στήλη μέσα στο σύνολο δεδομένων, να καταλαμβάνει τόσες στήλες όσες και οι διαφορετικές τιμές που έχει το κατηγορικό χαρακτηριστικό. Η κάθε νέα στήλη πλέον έχει το όνομα μίας τιμής του χαρακτηριστικού. Στη συνέχεια, για κάθε εγγραφή συμπληρώνεται 1 στη νέα στήλη που έχει για όνομα την τιμή που υπήρχε σε αυτή την εγγραφή και σε όλες τις υπόλοιπες στήλες μπαίνει μηδέν.

Αυτή η μέθοδος αλλάζει τα κατηγορικά χαρακτηριστικά, τα οποία είναι δύσκολο να αντιληφθούν οι αλγόριθμοι, σε αριθμητικά χαρακτηριστικά, χωρίς να χάσουμε καμία πληροφορία. ***[11]***

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Εικόνα 2.3: Παράδειγμα κωδικοποίησης One-Hot

3) **Πράξεις ομαδοποίησης (Grouping Operations)**

Όπως έχουμε αναφέρει, συνήθως στην ανάλυση δεδομένων κάθε εγγραφή του συνόλου δεδομένων είναι μια **γραμμή** (**row**), που περιέχει ένα συγκεκριμένο χαρακτηριστικό που την κάνει μοναδική, για παράδειγμα τον αριθμό ταυτότητας του ατόμου. Κάθε **στήλη** (**column**) αποτελεί ένα διαφορετικό χαρακτηριστικό της κάθε εγγραφής και συνολικά το σύνολο των δεδομένων αποτελούν έναν **πίνακα** (**table**). Αυτός ο τρόπος αναπαράστασης των δεδομένων θεωρείται **"Τακτοποιημένος"** (**"Tidy"**). ***[A]***

Υπάρχουν όμως σύνολα δεδομένων, όπου για το χαρακτηριστικό που θεωρούμε κύριο, υπάρχουν πολλές διαφορετικές εγγραφές. Παραδείγματα αποτελούν οι τραπεζικές συναλλαγές ή τα ημερομίσθια που δίνει σε κάθε υπάλληλο μια εταιρεία, όταν θέλουμε να τα εξετάσουμε ως προς το κάθε άτομο ξεχωριστά, αφού τότε κάθε άτομο με τον ίδιο αριθμό ταυτότητας έχει πολλαπλές εγγραφές μέσα στον πίνακα.

Σε μια τέτοια περίπτωση, ομαδοποιούμε τα δεδομένα ανά άτομο και στη συνέχεια, κάθε άτομο αντιπροσωπεύεται από μία μόνο σειρά. Το βασικό σημείο για τις πράξεις ομαδοποίησης είναι να αποφασιστεί η συνάρτηση με βάση την οποία θα γίνει η ομαδοποίηση. Όπως θα αναλύσουμε στη συνέχεια, για αριθμητικά χαρακτηριστικά, η πράξη του μέσου όρου και του αθροίσματος είναι οι επικρατέστερες επιλογές, ενώ για κατηγορικά χαρακτηριστικά χρειάζονται πιο περίπλοκες πράξεις.

i) Ομαδοποίηση κατηγορικών χαρακτηριστικών

Για τις κατηγορικές στήλες μπορούν να χρησιμοποιηθούν 3 διαφορετικοί τρόποι:

a) Να επιλεχθεί η τιμή που έχει τη μεγαλύτερη συχνότητα, δηλαδή η επικρατούσα τιμή.

b) Να ακολουθήσουμε τη λογική της κωδικοποίησης one-hot και να δημιουργήσουμε ένα **συγκεντρωτικό πίνακα** (**pivot table**), όπου θα βάλουμε σε κάθε νέα στήλη που θα έχει όνομα από τις τιμές του χαρακτηριστικού, τη συχνότητα που έχει στο σύνολο δεδομένων για το συγκεκριμένο άτομο. Αυτή είναι μια καλή πρακτική, εάν θέλουμε να επεξεργαστούμε περισσότερο το αρχικό χαρακτηριστικό, ώστε να μη χάσουμε πληροφορίες.

User	City	Visit Days
1	Roma	1
2	Madrid	2
1	Madrid	1
3	Istanbul	1
2	Istanbul	4
1	Istanbul	3
1	Roma	3

→

User	Istanbul	Madrid	Roma
1	3	1	4
2	4	2	0
3	1	0	0

Pivot table example: Sum of Visit Days grouped by Users

Εικόνα 2.4: Παράδειγμα συγκεντρωτικού πίνακα

c) Ο τελευταίος τρόπος ομαδοποίησης είναι να χρησιμοποιήσουμε τη συνάρτηση **group by**, μετά την εφαρμογή κωδικοποίησης one-hot. Αυτή η μέθοδος δίνει τα ίδια αποτελέσματα με το συγκεντρωτικό πίνακα, που είδαμε παραπάνω, απλά είναι σε διαφορετική μορφή.

ii) Ομαδοποίηση αριθμητικών χαρακτηριστικών

Οι αριθμητικές στήλες συνήθως ομαδοποιούνται χρησιμοποιώντας τις συναρτήσεις **sum** (άθροισμα) και **mean** (μέση τιμή).

4) **Διαχωρισμός χαρακτηριστικών (Feature Split)**

Ο διαχωρισμός των χαρακτηριστικών γίνεται, ώστε κάποια χαρακτηριστικά που είναι δυσνόητα για τους αλγόριθμους μηχανικής μάθησης να μπορούν να αξιοποιηθούν και να συμβάλλουν στην αποδοτικότητα του μοντέλου. Πάρα πολλές φορές, το σύνολο των δεδομένων περιέχει στήλες με **συμβολοσειρές** (**string**) που παραβιάζουν τις αρχές των **“τακτοποιημένων”** (**“tidy”**) δεδομένων. Στόχος μας είναι να μπορέσουμε να αξιοποιήσουμε τα χρήσιμα τμήματα της συμβολοσειράς, δημιουργώντας ένα ή περισσότερα καινούργια χαρακτηριστικά στηριζόμενοι στις παρακάτω επιδιώξεις:

i) Να μπορούν οι αλγόριθμοι να αντιληφθούν τα δεδομένα.

ii) Να έχουμε τη δυνατότητα να τα ομαδοποιήσουμε.

iii) Να βελτιώσουμε την απόδοση του μοντέλου, αποκαλύπτοντας κρυμμένες πληροφορίες.

Ένα παράδειγμα που σχετίζεται με τη χρήση του διαχωρισμού χαρακτηριστικών είναι αν σε ένα χαρακτηριστικό περιλαμβάνεται το όνομα και το επώνυμο μαζί, τότε προσπαθούμε να τα διαχωρίσουμε

σε δυο καινούργιες διαφορετικές στήλες, όπου τόσο το όνομα, όσο και το επώνυμο θα είναι ένα νέο ξεχωριστό χαρακτηριστικό.

5) **Εξαγωγή χαρακτηριστικών από ημερομηνία / ώρα**

Η συγκεκριμένη τεχνική υπάγεται εν μέρει στην τεχνική του διαχωρισμού χαρακτηριστικών, αλλά έχει και άλλες πρόσθετες δυνατότητες. Γενικά, τα χαρακτηριστικά που ο τύπος τους είναι ημερομηνία παρέχουν πολύτιμες πληροφορίες, αλλά πολλές φορές δεν αξιοποιούνται από τους αλγόριθμους μηχανικής μάθησης. Ίσως, ο λόγος να είναι η μορφοποίηση τους ακόμα κι αν είναι στο πιο σύνηθες για τους ανθρώπους τρόπο γραφής, όπως "31-12-2017".

Για να μπορέσουν οι αλγόριθμοι να επεξεργαστούν πλήρως τα δεδομένα ενδεικτικά θα μπορούσαμε να πούμε τρεις τρόπους:

i) Εξαγωγή των τμημάτων της ημερομηνίας σε τρία νέα χαρακτηριστικά. Ένα για την ημέρα, ένα για το μήνα και ένα για το έτος.

ii) Εξαγωγή του χρονικού διαστήματος από την τρέχουσα ημερομηνία, που ενδεχομένως να είναι σημαντικό, όπως για παράδειγμα αν έχουμε την ημερομηνία γέννησης να εμφανίζεται η ηλικία. Αν θέλουμε, αυτό μπορεί να συμβεί πάλι σε τρία χαρακτηριστικά μετρώντας τα έτη, τους μήνες και τις μέρες από την τρέχουσα ημερομηνία.

iii) Εξαγωγή του ονόματος της ημέρας ή κάποιων άλλων αντίστοιχων χαρακτηριστικών που να στηρίζονται σε αυτό. Για παράδειγμα αν είναι Σαββατοκύριακο ή καθημερινή.

Με αυτό τον τρόπο οι αλγόριθμοι είναι πιο εύκολο να αξιοποιήσουν το σύνολο της ημερομηνίας.

6) **Κλιμάκωση (Scaling)**

Στις περισσότερες περιπτώσεις, όσα χαρακτηριστικά είναι αριθμητικά σε ένα σύνολο δεδομένων έχουν εντελώς διαφορετικά μεταξύ τους στατιστικά μεγέθη. Προφανές είναι πως διαφέρουν στη μέση τιμή, στη διακύμανση και στο εύρος τους. Γι’ αυτό το λόγο, χρησιμοποιώντας στον αλγόριθμο του μοντέλου ανόμοια μεταξύ τους χαρακτηριστικά μεγέθη, ενδέχεται να εξαχθούν λάθος συμπεράσματα.

Για παράδειγμα, μια διαφορά 30 χρόνων στην ηλικία δεν είναι το ίδιο σημαντική με μια διαφορά 30 ευρώ στο ετήσιο εισόδημα.

Για να υπάρξει μια ενιαία αντιμετώπιση στις αριθμητικές τιμές χρησιμοποιούμε την τεχνική της κλιμάκωσης. Με αυτόν τον τρόπο, τα συνεχή αριθμητικά χαρακτηριστικά συμπεριφέρονται όλα με τον ίδιο τρόπο. Η κλιμάκωση δεν είναι υποχρεωτική, αλλά κάποιες φορές είναι ιδιαιτέρως σημαντική να πραγματοποιηθεί. Έτσι, οι αλγόριθμοι που βασίζονται σε υπολογισμούς απόστασης όπως ο **αλγόριθμος k-πλησιέστερων γειτόνων** (**k-nearest neighbors algorithm, k-NN**) ή ο **αλγόριθμος συσταδοποίησης k-μέσων** (**k-means clustering**) πρέπει να έχουν κλιμακωτά, συνεχή αριθμητικά χαρακτηριστικά για να λειτουργήσουν.

Οι πιο συνήθεις τρόποι κλιμάκωσης είναι δυο:

i) Κανονικοποίηση (Normalization)

Η κανονικοποίηση αλλάζει την κλίμακα των τιμών του χαρακτηριστικού και αντιστοιχεί όλες τις τιμές στο κλειστό σύνολο μεταξύ 0 και 1. Αυτός ο μετασχηματισμός δεν αλλάζει την κατανομή του χαρακτηριστικού.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Όμως, λόγω των μειωμένων τυπικών αποκλίσεων, οι επιδράσεις των ακραίων τιμών γίνονται εντονότερες. Επομένως, πριν από την κανονικοποίηση, είναι ιδιαιτέρως σημαντικό να έχουμε αναλύσει και να έχουμε διαχειριστεί τις ακραίες τιμές.

ii) Κανονικοποίηση z-score (standarization or z-score normalization)

Η κανονικοποίηση z-score αντιστοιχεί τις τιμές του χαρακτηριστικού λαμβάνοντας υπόψη την τυπική απόκλιση. Εάν η τυπική απόκλιση των χαρακτηριστικών είναι διαφορετική, τότε και το εύρος τους θα διαφέρει. Αυτό μειώνει την επίδραση των ακραίων τιμών.

$$z = \frac{X - \mu}{\sigma}$$

7) **Διαχείριση μη ισορροπημένων συνόλων δεδομένων (Handling imbalanced datasets)**

Για τη συγκεκριμένη περίπτωση, θα εξετάσουμε στο 3ο κεφάλαιο της εργασίας μας αναλυτικά τις διάφορες τεχνικές, αφού το συγκεκριμένο αντικείμενο αποτελεί και την κύρια έρευνα της παρούσας διπλωματικής.

2.4 Μηχανική μάθηση (Machine learning)

Η μηχανική μάθηση έχει πλέον μπει σε μεγάλο βαθμό στη ζωή μας, αφού παίζει ρόλο σε πάρα πολλούς επιστημονικούς τομείς, όπως για παράδειγμα ανθρωπιστικούς, οικονομικούς και θετικών επιστημών.

Σύμφωνα με τον ορισμό που έδωσε ο το 1959 ο Άρθουρ Σάμουελ, η μηχανική μάθηση είναι το πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν προγραμματιστεί ρητά για το αποτέλεσμα που θα δώσουν.

Ειδικότερα, η μηχανική μάθηση είναι υποπεδίο της επιστήμης των υπολογιστών που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες σε αυτά και στη συνέχεια, να εξάγουν αποφάσεις για το αποτέλεσμα.

Ο τομέας της Μηχανικής Μάθησης χωρίζεται στις παρακάτω τρεις κατηγορίες:

1) **Επιτηρούμενη μάθηση (Supervised Learning)**

Το υπολογιστικό πρόγραμμα δέχεται στην είσοδο του τα **δεδομένα εκπαίδευσης** (**training data**), καθώς και τα αποτελέσματα από έναν οδηγό και ο στόχος είναι να δημιουργήσει ένα γενικότερο μονοπάτι, προκειμένου να αντιστοιχίσει τις εισόδους με τα αποτελέσματα.

2) **Μη επιτηρούμενη μάθηση (Unsupervised Learning)**

Χωρίς να παρέχεται κάποια εμπειρία στον αλγόριθμο μάθησης, πρέπει να βρει την δομή των δεδομένων εισόδου. Η μη επιτηρούμενη μάθηση μπορεί να είναι αυτοσκοπός (ανακαλύπτοντας κρυμμένα μοτίβα σε δεδομένα) ή ενδιάμεσο στάδιο για ένα επέλθει ένα οριστικό τέλος της διαδικασίας.

3) **Ενισχυτική μάθηση (Reinforcement Learning)**

Ένα πρόγραμμα υπολογιστή αλληλεπιδράει με ένα δυναμικό περιβάλλον στο οποίο πρέπει να επιτευχθεί ένας συγκεκριμένος στόχος, όπως είναι η οδήγηση ενός οχήματος, χωρίς κάποιος δάσκαλος να του λέει ρητά αν έχει προσεγγίσει το στόχο του. Ένα άλλο παράδειγμα είναι να μάθει να παίζει ένα παιχνίδι, όπως είναι το σκάκι, εναντίον κάποιου αντιπάλου.

Η κύρια διαφορά μεταξύ των δύο πρώτων κατηγοριών, που είναι και οι κυρίαρχες, είναι πως η επιτηρούμενη μάθηση γίνεται χρησιμοποιώντας μια **ισχύουσα αλήθεια** (**ground truth**), δηλαδή έχουμε εκ των προτέρων τη γνώση των τιμών εξόδου για τα δείγματά μας. Επομένως, ο στόχος της επιτηρούμενης μάθησης είναι να δημιουργήσει έναν αλγόριθμο, που έχοντας ως οδηγό ένα δείγμα εισόδου που δίνει γνωστές εξόδους, να μπορεί να προσεγγίσει καλύτερα τη σχέση μεταξύ εισόδου και εξόδου και να εφαρμοστεί σε καινούργια, παρόμοια δεδομένα τα οποία δεν περιλαμβάνονται στο δείγμα μας και για τα οποία αγνοούμε τις εξόδους τους.

Η μη επιτηρούμενη μάθηση, από την άλλη πλευρά, δεν έχει **ετικέτες** (**labels**) ή αλλιώς **εξόδους** (**outputs**), οπότε στόχος της είναι να συμπεράνει τη φυσική δομή που υπάρχει σε ένα σύνολο δεδομένων.

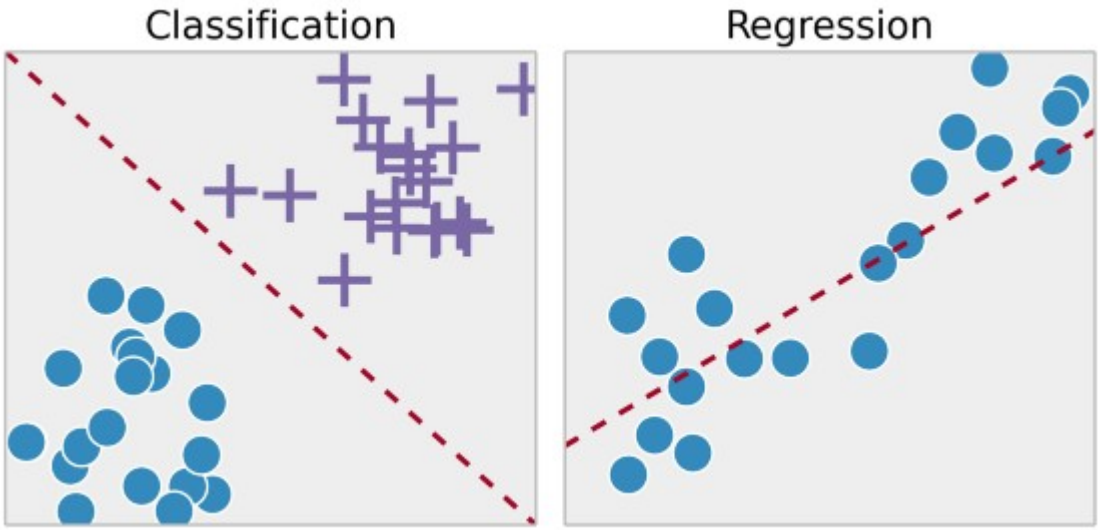
Στη συνέχεια, αναλύονται οι βασικές παράμετροι των ειδών της μάθησης.

2.4.1 Επιτηρούμενη μάθηση (Supervised Learning)

Η επιτηρούμενη μάθηση χρησιμοποιείται συνήθως:

1) στην **κατηγοριοποίηση** (**classification**), δηλαδή όταν θέλουμε να αντιστοιχίσουμε τα δεδομένα εισόδου σε **ετικέτες δεδομένων** (**data labels**), δηλαδή σε ποια κατηγορία εξόδου ανήκουν τα καινούργια δεδομένα μας.

2) στην **παλινδρόμηση** (**regression**), όταν θέλουμε για τα δεδομένα εισόδου να υπολογίσουμε την πιθανή τιμή σε μια συνεχή έξοδο.



Εικόνα 2.5: Παραδείγματα κατηγοριοποίησης και παλινδρόμησης

Οι συνηθέστεροι αλγόριθμοι στην επιτηρούμενη μάθηση περιλαμβάνουν τη **λογιστική παλινδρόμηση** (**Logistic Regression**), τον αλγόριθμο **Naive Bayes**, τις **Μηχανές Διανυσμάτων Υποστήριξης ΜΔΥ** (**Support Vector Machines, SVM**), τα **τεχνητά νευρικά δίκτυα** (**Artificial Neural Networks**) και τα **τυχαία δάση** (**Random Forests**).

Τόσο στην κατηγοριοποίηση, όσο και στην παλινδρόμηση, ο στόχος είναι να βρούμε συγκεκριμένες σχέσεις ή δομές στα δεδομένα εισόδου που να μας επιτρέπουν να παράγουμε επιτυχείς προβλέψεις για τα δεδομένα εξόδου. Να σημειωθεί ότι η εκμάθηση για την εύρεση της επιτυχούς εξόδου καθορίζεται εξ ολοκλήρου από τα δεδομένα εκπαίδευσης. Έτσι, ενώ έχουμε μια ισχύουσα αλήθεια, χρησιμοποιώντας το μοντέλο μας θα εξάγουμε κάποια αποτελέσματα που θα θεωρήσουμε πως είναι αληθή, αλλά αυτό δε σημαίνει ότι οι ετικέτες δεδομένων είναι πάντα σωστές σε καινούργια δεδομένα. Οι **θόρυβοι** (**noises**) ή οι λανθασμένες ετικέτες σε ορισμένα από τα δεδομένα εκπαίδευσης ενδέχεται να προκαλέσουν σαφή μείωση της αποτελεσματικότητας του μοντέλου.

Κατά τη διεξαγωγή της επιτηρούμενης μάθησης, το σημαντικότερο είναι η **πολυπλοκότητα** (**complexity**) του μοντέλου και η **εξισορρόπηση συστηματικού σφάλματος και διακύμανσης**

(**bias-variance tradeoff**). Ουσιαστικά, αυτές οι δυο παράμετροι είναι αλληλένδετες και αντιστρόφως ανάλογες ως προς τη σχέση τους.

Η πολυπλοκότητα του μοντέλου αναφέρεται στην πολυπλοκότητα της συνάρτησης που δημιουργείται για να προβλέπει τις πιθανές εξόδους και είναι αντίστοιχη με τον βαθμό μιας πολυωνυμικής συνάρτησης.

Το κατάλληλο επίπεδο πολυπλοκότητας του μοντέλου καθορίζεται γενικά από τη φύση των δεδομένων εκπαίδευσης. Εάν ο αριθμός των δεδομένων είναι μικρός ή αν τα δεδομένα μας δεν κατανέμονται ομοιόμορφα, θα πρέπει να επιλεγεί ένα μοντέλο χαμηλής πολυπλοκότητας. Αυτό συμβαίνει επειδή αν χρησιμοποιηθεί ένα μοντέλο υψηλής πολυπλοκότητας και ο αριθμός δεδομένων εισόδου είναι μικρός, τότε θα υπάρξει υπερπροσαρμογή. Η **υπερπροσαρμογή** (**overfitting)** είναι η προσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης σε υπερβολικό βαθμό, το οποίο δημιουργεί πρόβλημα, καθώς δεν είναι ικανό να προβλέψει σωστά καινούργια δεδομένα.

Η **εξισορρόπηση συστηματικού σφάλματος και διακύμανσης** (**bias-variance tradeoff**) σχετίζεται επίσης με τη γενίκευση του μοντέλου. Σε οποιοδήποτε μοντέλο, υπάρχει μια ισορροπία μεταξύ του **συστηματικού σφάλματος (bias)**, που είναι το μέγεθος του σταθερού σφάλματος, και της διακύμανσης, που είναι το ποσό κατά το οποίο το σφάλμα μπορεί να κυμαίνεται με βάση διαφορετικά σύνολα **δεδομένων εκπαίδευσης**. Έτσι, το υψηλό συστηματικό σφάλμα και η χαμηλή διακύμανση θα ήταν ένα μοντέλο που θα έβγαινε λάθος σταθερά σε ένα ποσοστό 20% των φορών, ενώ ένα μοντέλο χαμηλού συστηματικού σφάλματος και υψηλής διακύμανσης θα μπορούσε να ήταν λάθος σε ένα ποσοστό που θα κυμαινόταν ανάμεσα στο 5% και στο 50% ανάλογα με το σύνολο των δεδομένων που θα χρησιμοποιούσαμε κάθε φορά για να το εκπαιδεύσουμε.

Η εύρεση του σωστού σημείου εξισορρόπησης βρίσκεται εμπειρικά και συνήθως σχετίζεται με το πρόβλημα που αντιμετωπίζουμε, αλλά και τα ίδια τα δεδομένα μας. Όμως σαφώς, η αύξηση του συστηματικού σφάλματος και η αναμενόμενη μείωση της διακύμανσης, οδηγεί σε μοντέλα με σχετικά σταθερά επίπεδα λάθος εκτίμησης, τα οποία μπορεί να είναι προτιμότερα για ορισμένους τύπους προβλημάτων.

Επιπλέον, προκειμένου να παραχθούν μοντέλα που να μπορούν να γενικευθούν αρκετά καλά, η διακύμανση του μοντέλου πρέπει να κλιμακώνεται ανάλογα με το μέγεθος και την πολυπλοκότητα των δεδομένων εκπαίδευσης. Για παράδειγμα, τα μικρά σε μέγεθος και

απλά σύνολα δεδομένων πρέπει συνήθως να σχετίζονται με μοντέλα χαμηλής διακύμανσης και αντίστοιχα τα μεγάλα σε αριθμό και πολύπλοκα σύνολα δεδομένων, πρέπει να σχετίζονται με μοντέλα υψηλότερης διακύμανσης, ώστε να μπορούν να μάθουν πλήρως τη δομή των δεδομένων. ***[12]***

2.4.2 Μη επιτηρούμενη μάθηση (Unsupervised Learning)

Οι πιο συνηθισμένες εργασίες στο πλαίσιο της μη επιτηρούμενης μάθησης είναι η **συσταδοποίηση** (**clustering**), η **μάθηση αναπαράστασης** (**representation learning**) και η **εκτίμηση πυκνότητας** (**density estimation**). Σε όλες αυτές τις περιπτώσεις, θέλουμε να μάθουμε την υπάρχουσα δομή των δεδομένων, χωρίς να χρησιμοποιούμε συγκεκριμένες ετικέτες στις εξόδους. Μερικοί συνηθισμένοι αλγόριθμοι περιλαμβάνουν την **συσταδοποίηση k-μέσων** (**k-means clustering**), την **ανάλυση κύριων συνιστωσών** (**Principal Component Analysis, PCA**) και τους **αυτόματους κωδικοποιητές** (**Autoencoders**). Δεδομένου ότι δεν παρέχονται ετικέτες, στις περισσότερες μη επιτηρούμενες μεθόδους μάθησης δεν υπάρχει συγκεκριμένος τρόπος σύγκρισης της απόδοσης του μοντέλου. Δύο συνηθισμένες περιπτώσεις στις οποίες χρησιμοποιείται η μη επιτηρούμενη μάθηση είναι η **διερευνητική ανάλυση δεδομένων, ΔΑΔ** (**Exploratory Data Analysis, EDA**) και η **μείωση διαστάσεων**(**Dimensionality Reduction, DR**).

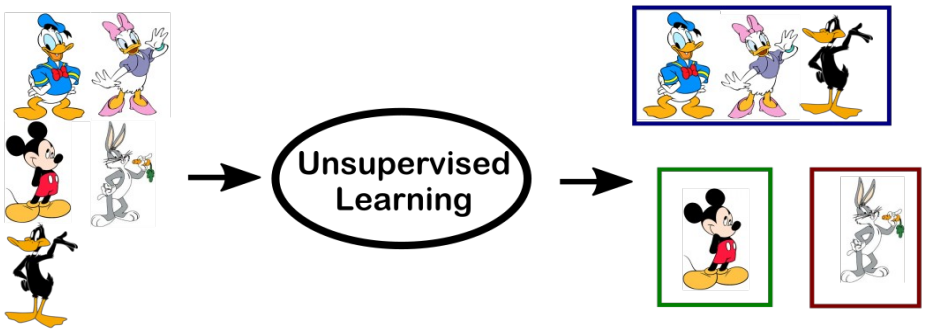
Η μη επιτηρούμενη μάθηση είναι πολύ χρήσιμη στη ΔΑΔ, επειδή μπορεί να αναγνωρίσει αυτόματα τη δομή. Για παράδειγμα, εάν ένας αναλυτής προσπαθούσε να διαχωρίσει σε ομάδες τους καταναλωτές, οι μη επιτηρούμενες μέθοδοι ομαδοποίησης θα ήταν ένα εξαιρετικό σημείο εκκίνησης για την ανάλυσή τους. Σε καταστάσεις όπου είναι αδύνατο ή μη εφαρμόσιμο για έναν άνθρωπο να προτείνει τάσεις στα δεδομένα, η μη επιτηρούμενη μάθηση μπορεί να παρέχει αρχικές εκτιμήσεις που μπορούν στη συνέχεια να χρησιμοποιηθούν για τη δοκιμή πιο συγκεκριμένων υποθέσεων.

Η μείωση διαστάσεων, αναφέρεται στις μεθόδους που χρησιμοποιούνται για την αναπαράσταση δεδομένων χρησιμοποιώντας λιγότερα **χαρακτηριστικά** (**features**) των δεδομένων. Ουσιαστικά, πρόκειται για την επιλογή χαρακτηριστικών που είδαμε σε προηγούμενο κεφάλαιο.

Η μη επιτηρούμενη μάθηση μπορεί να χρησιμοποιηθεί στο πρώτο στάδιο της ανάλυσης δεδομένων που έχει να κάνει με την προεπεξεργασία των δειγμάτων. ***[13]***

2.4.3 Ενισχυτική μάθηση (Reinforcement Learning)

Η ενισχυτική μάθηση αφορά τη δημιουργία κατάλληλων ενεργειών για τη μεγιστοποίηση της ανταμοιβής σε μια συγκεκριμένη κατάσταση. Σε αντίθεση με την επιτηρούμενη μάθηση, η ενισχυτική μάθηση δεν έχει ετικέτες δεδομένων και γι' αυτό υπάρχει κάποιος παράγοντας ενίσχυσης και αυτός αποφασίζει τα βήματα που θα κάνει για να εκτελεστεί μια δεδομένη εργασία. Επειδή σε αυτή την περίπτωση δεν υπάρχουν δεδομένα εκπαίδευσης, το σύστημα μάθησης είναι υποχρεωμένο να μάθει από την εμπειρία του. ***[14]***



Εικόνα 2.6: Ενισχυτική μάθηση

2.5 Μοντέλα εξόρυξης δεδομένων (Data mining models)

Στην εξόρυξη δεδομένων, ένα μοντέλο ή ένας αλγόριθμος είναι ένα σύνολο **ευρετικών** (**heuristics**) και υπολογιστικών μεθόδων που μπορούν να αναλύσουν τα εκπαιδευτικά δεδομένα και να καταλήξουν σε συγκεκριμένους τύπους ή πρότυπα που θα μας βοηθήσουν να προσδιορίσουμε τη μεταβλητή εξόδου σε καινούργια δεδομένα.

Ορισμένοι από τους κυριότερους αλγόριθμους που χρησιμοποιούνται είναι οι παρακάτω:

1) **Απλός Bayes (Naive Bayes)**

Πρόκειται για μια μέθοδο κατηγοριοποίησης που στηρίζεται στο **Μπεϋζιανό θεώρημα** (**Bayes' Theorem**) με μια υπόθεση ανεξαρτησίας μεταξύ των μεταβλητών. Στο αντικείμενο μας, μεταβλητές αποτελούν τα χαρακτηριστικά του συνόλου δεδομένων.

Το μοντέλο Naive Bayes είναι εύκολο στην κατασκευή και ιδιαίτερα χρήσιμο για πολύ μεγάλα σύνολα δεδομένων. Παρά την απλότητα του, το Naive Bayes θεωρείται ένας σχετικά αξιόπιστος κατηγοριοποιητής.

Το Μπεϋζιανό θεώρημα απεικονίζεται στην επόμενη εξίσωση:

$$P(c \vee x)=\frac{P(x \vee c) * P(c)}{P(x)}, \text{ με } P(c \vee x)=P(x_1 \vee c)\times P(x_2 \vee c)\times ...\times P(x_n \vee c)\times P(c)$$

όπου

- P(c|x) είναι η **εκ των υστέρων πιθανότητα** (**posterior probability**) της κλάσης c δοθέντος του χαρακτηριστικού x.

- P(c) είναι η **εκ των προτέρων πιθανότητα** (**prior probability**) της κλάσης c.

- P(x|c) είναι η **πιθανοφάνεια** (**likelihood**) που είναι η πιθανότητα του χαρακτηριστικού x δοθέντος της κλάσης c.

- P(x) είναι η **εκ των προτέρων πιθανότητα** (**prior probability**) του χαρακτηριστικού x.

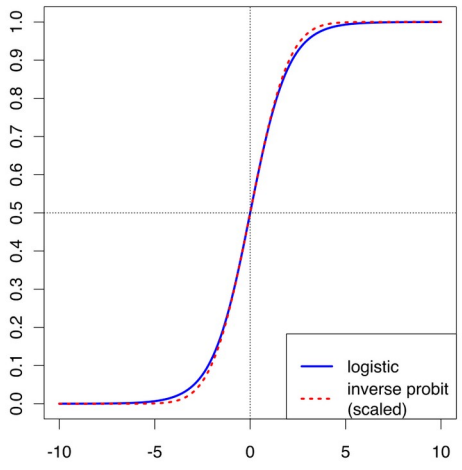
2) **Λογιστική παλινδρόμηση (Logistic Regression)**

Η λογιστική παλινδρόμηση χρησιμοποιεί τις στατιστικές μεθόδους, έχοντας ως στόχο να φτιάξει ένα μοντέλο για ένα εκπαιδευτικό σύνολο δεδομένων, σε περίπτωση που η μεταβλητή εξόδου έχει δυο κλάσεις. Τα χαρακτηριστικά εισόδου θα πρέπει να είναι ανεξάρτητα μεταξύ τους.

Το αποτέλεσμα προσδιορίζεται χάρη στη χρήση μιας λογιστικής συνάρτησης, η οποία εκτιμά μια πιθανότητα και κατηγοριοποιεί την εγγραφή στην αντίστοιχη κλάση. Η λογιστική συνάρτηση, που ονομάζεται και σιγμοειδής, είναι μια καμπύλη που δίνει αριθμητικές τιμές στο ανοιχτό διάστημα (0,1). Εν τέλει, το αποτέλεσμα δίνει για το ζητούμενο χαρακτηριστικό την τιμή 0 ή 1, αποδίδοντας κάθε δείγμα σε μια από τις δυο κλάσεις. Γενικά, το λογιστικό μοντέλο είναι ένα μη γραμμικό μοντέλο, τα σφάλματα, του οποίου δεν υπακούν στην κανονική κατανομή και η μεταβλητή απόκρισης είναι διακριτή.

Ο τύπος της λογιστικής συνάρτησης είναι $f(z)=\frac{e^z}{1+e^z}=\frac{1}{1+e^{-z}}$

όπου z είναι η μεταβλητή εισόδου και f(z) το αποτέλεσμα αυτής.



Εικόνα 2.7: Γραφική απεικόνιση της λογιστικής συνάρτησης

3) **k-πλησιέστερων γειτόνων (K-Nearest Neighbours, k-NN)**

Ο αλγόριθμος k-NN αναφέρεται ως ένας **τεμπέλης μαθητής** (**lazy learner**), αφού στην πραγματικότητα δεν κάνει κάτι άλλο κατά τη διάρκεια της εκπαίδευσης, εκτός από το να αποθηκεύει εγγραφές που θα χρησιμοποιήσει για την εκπαιδευτική διαδικασία. Μόνο όταν εισάγονται νέα δεδομένα, για τα οποία δεν έχουμε αποτέλεσμα στη μεταβλητή εξόδου ενεργεί διαφορετικά. Τότε μόνο, ο αλγόριθμος κάνει κατηγοριοποίηση και δίνει αποτέλεσμα. Ο τρόπος λειτουργίας του στην περίπτωση που έχει δεδομένο χωρίς ετικέτα εξόδου, ακολουθεί δύο βήματα.

α) Κοιτάζει τις k πλησιέστερες εγγραφές που έχουν ετικέτα εξόδου από τα εκπαιδευτικά δεδομένα, δηλαδή όπως αναφέρουμε τους k-πλησιέστερους γείτονες.

β) Χρησιμοποιώντας τα παραπάνω k δεδομένα, υπολογίζει ποια κλάση θα εκτιμήσει για το νέο δεδομένο.

Για να βρει τους πλησιέστερους γείτονες, για συνεχείς μεταβλητές χρησιμοποιεί μια μέτρηση απόστασης, όπως η Ευκλείδεια απόσταση. Η επιλογή της μέτρησης απόστασης εξαρτάται σε μεγάλο βαθμό από τα δεδομένα. Για διακριτά δεδομένα, μετατρέπει τα διακριτά δεδομένα σε συνεχή. Κάτι τέτοιο μπορεί να επιτευχθεί με την απόσταση Hamming.

4) **Μηχανές Διανυσμάτων Υποστήριξης ΜΔΥ (Support Vector Machines, SVM)**

Οι μηχανές διανυσμάτων υποστήριξης χρησιμοποιούν ένα υπερεπίπεδο για να κατηγοριοποιήσουν τα δεδομένα σε 2 κλάσεις. Ένα υπερεπίπεδο είναι μια συνάρτηση, η οποία χωρίζει το ν-διάστατο χώρο σε δύο υποχώρους και που ο καθένας αντιπροσωπεύει μια κλάση της μεταβλητής εξόδου.

Αν για παράδειγμα, έχουμε μόνο δυο χαρακτηριστικά στο σύνολο των δεδομένων, τότε δημιουργείται μια ευθεία **f(x) = mx + b** και οι δυο κλάσεις αντιστοιχούν στα σημεία πάνω ή κάτω από την ευθεία. Αν έχουμε τρία χαρακτηριστικά δημιουργείται ένα επίπεδο που έχει δυο υποχώρους, πάνω και κάτω από το επίπεδο, και ο καθένας αντιστοιχεί σε μια κλάση. Με την ίδια λογική, δημιουργείται υπερεπίπεδο για όσα χαρακτηριστικά κι αν διαθέτει το σύνολο δεδομένων.

5) **Δέντρο αποφάσεων (Decision tree)**

Η εκπαίδευση του δέντρου αποφάσεων είναι μια κλασική μέθοδος εξόρυξης δεδομένων. Ο στόχος είναι να δημιουργηθεί ένα μοντέλο που να προβλέπει την τιμή της μεταβλητής εξόδου με βάση τις δοθείσες μεταβλητές εισόδου. Η εκτιμώμενη τιμή της μεταβλητής εξόδου καθορίζεται με μια σειρά από ερωτήματα και συνθήκες.

Κάθε εσωτερικός κόμβος του δέντρου, που δεν αποτελεί φύλλο, αφορά ένα χαρακτηριστικό. Αν για παράδειγμα πρόκειται για χαρακτηριστικό που παίρνει αριθμητικές τιμές, ο διαχωρισμός προς τους απόγονους, συνήθως γίνεται με βάση κάποια συγκεκριμένη τιμή του. Οι απόγονοι του κόμβου μπορεί να είναι ή κάποιες από τις κλάσεις της μεταβλητής εισόδου ή κάποιος νέος εσωτερικός κόμβος που αναφέρεται σε κάποιο άλλο χαρακτηριστικό του συνόλου δεδομένων.

Το δέντρο αποφάσεων δημιουργείται χωρίζοντας το αρχικό σύνολο, που αποτελεί τη ρίζα του δέντρου, σε υποσύνολα. Ο διαχωρισμός γίνεται από ένα σύνολο αποφάσεων που βασίζεται στον τρόπο κατηγοριοποίησης του συνόλου δεδομένων. Αυτή η διαδικασία επαναλαμβάνεται μέχρι κάποιος κόμβος να αποδίδει μόνο μια τιμή για την κλάση ή ο διαχωρισμός να μη συμβάλλει σημαντικά στο αποτέλεσμα.

Μια εναλλακτική ονομασία τους είναι **Δέντρο Κατηγοριοποίησης και Παλινδρόμησης** (**Classification And Regression Tree, CART**).

Στο σημείο αυτό θα κάνουμε μια σύντομη αναφορά στις έννοιες των **συνδυαστικών μεθόδων** (**Ensemble methods**) και των **μεθόδων δειγματοληψίας Bootstrap** (**Bootstrap sampling methods**) για να γίνουν πιο κατανοητά τα επόμενα μοντέλα εξόρυξης δεδομένων.

<u>Συνδυαστικές μέθοδοι (Ensemble methods)</u>

Οι συνδυαστικές μέθοδοι στη μηχανική μάθηση μπορούν και συνθέτουν ορισμένους από τους προηγούμενους ή και άλλους αλγόριθμους με αποτέλεσμα να δίνουν πιο ορθά αποτελέσματα.

Για παράδειγμα, η χρήση περισσότερων του ενός δέντρων αποφάσεων αυξάνει κατά πολύ την αποτελεσματικότητα του συνδυαστικού

αλγόριθμου. Στη συνέχεια, θα δούμε μερικά πιο συγκεκριμένα παραδείγματα τέτοιων αλγόριθμων. ***[15]***

Μέθοδοι δειγματοληψίας Bootstrap (Bootstrap sampling methods)

Η μέθοδος δειγματοληψίας bootstrap προέρχεται από την αντίστοιχη στατιστική μέθοδο και ουσιαστικά είναι μια μέθοδος δειγματοληψίας που χρησιμοποιεί την τυχαία δειγματοληψία με επανάθεση. Αυτό που κάνει τη μέθοδο Bootstrap εξαιρετικά σημαντική είναι πως:

i) Αποτελεί το κύριο στοιχείο για πολλούς σύγχρονους αλγόριθμους μηχανικής μάθησης, όπως τα **τυχαία δάση** (**random forests**) και οι αλγόριθμοι AdaBoost, gradient boost και XGBoost.

ii) Μπορεί να χρησιμοποιηθεί για την εκτίμηση των παραμέτρων ενός πληθυσμού, όπως το μέσο και το τυπικό σφάλμα. Επίσης, μπορούμε να κάνουμε υποθέσεις για την κατανομή ενός δείγματος που δεν είναι αρκετά μεγάλο. Ουσιαστικά, με την υπόθεση ότι το δείγμα είναι αντιπροσωπευτικό του πληθυσμού, η δειγματοληψία Bootstrap διεξάγεται για την παροχή εκτίμησης της κατανομής του στατιστικού δείγματος.

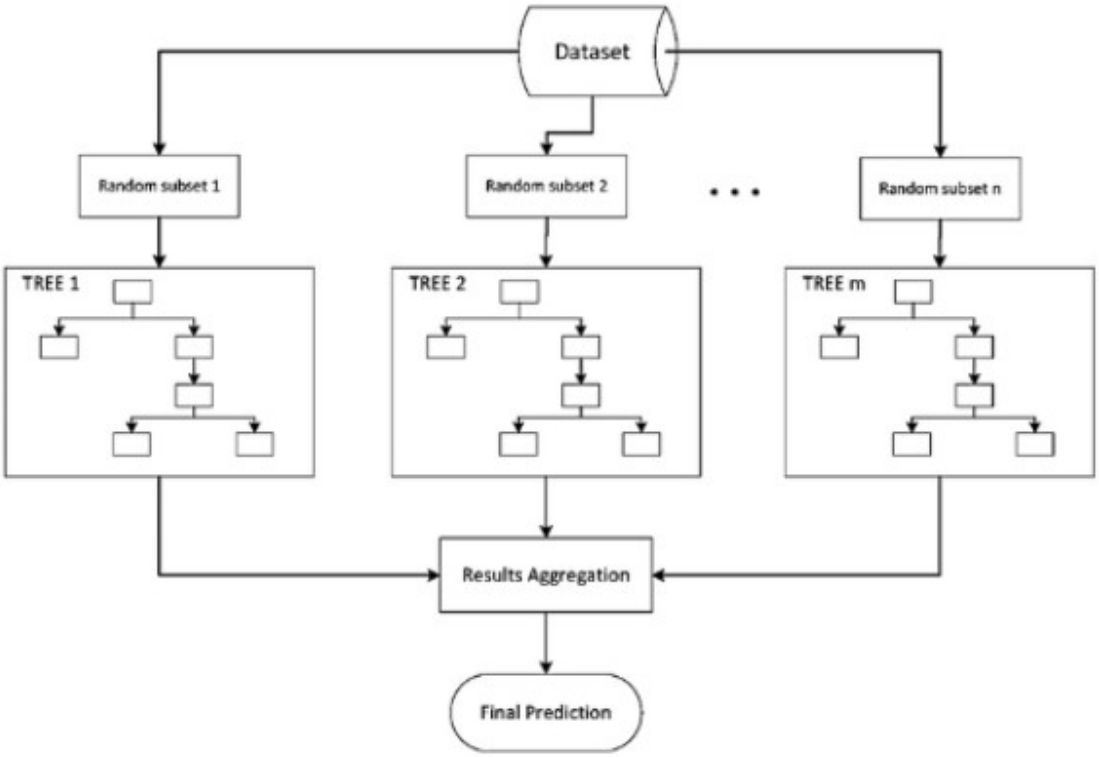
6) **Δέντρο Αποφάσεων Bagging (Bagging Decision Tree)**

Το δέντρο αποφάσεων Bagging συνδυάζει Bootstrapping και **συσσωμάτωση** (**Aggregation**). Το όνομα προέρχεται από το συνδυασμό των λέξεων Bootstrap και Aggregating. Με αυτόν τον τρόπο, δημιουργείται ένα μοντέλο **συνδυαστικής** (**ensemble**) μεθόδου.

Η λειτουργία του ακολουθεί την εξής λογική. Από ένα δείγμα δεδομένων, επιλέγονται πολλά δειγματοληπτικά υποσύνολα με τη μέθοδο Bootstrap, δηλαδή με επανάθεση. Ένα Δέντρο Αποφάσεων σχηματίζεται σε καθένα από τα Bootstrap υποσύνολα δειγμάτων. Αφού δημιουργηθεί κάθε Δέντρο Αποφάσεων, ένας αλγόριθμος χρησιμοποιείται για να συσσωρεύσει τα Δέντρα Αποφάσεων για να σχηματίσει τον πιο αποτελεσματικό παράγοντα πρόβλεψης.

Το δέντρο αποφάσεων Bagging είναι ένα μοντέλο που μας βοηθάει να μειώσουμε τη διακύμανση των αλγορίθμων των απλών δέντρων αποφάσεων που έχουν ιδιαίτερα υψηλή διακύμανση. Τα δέντρα

αποφάσεων είναι ιδιαιτέρως ευαίσθητα στα συγκεκριμένα δεδομένα στα οποία εκπαιδεύονται, οπότε αν φέρουμε νέα δεδομένα, πιθανότατα να αλλάξει το δέντρο απόφασης και κατ' επέκταση και οι προβλέψεις μας.



Εικόνα 2.8: Τρόπος λειτουργίας του δέντρου αποφάσεων Bagging

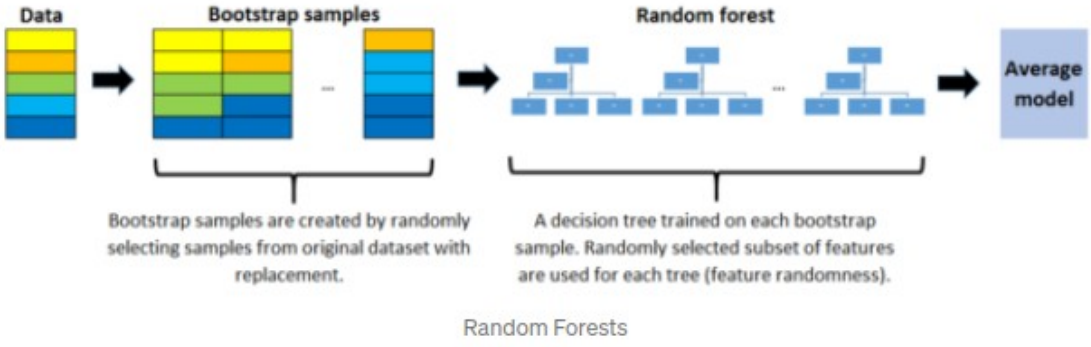
7) **Τυχαίο Δάσος (Random Forest)**

Τα μοντέλα Τυχαίου Δάσους μπορούν να θεωρηθούν ως βελτίωση του αλγόριθμου Bagging.

Τα δέντρα αποφάσεων αποτελούν **άπληστους** (**greedy**) αλγόριθμους. Άπληστοι αλγόριθμοι ονομάζονται αυτοί που σε κάθε βήμα τους επιλέγουν την πιο άμεσα ωφέλιμη λύση, χωρίς να κρίνουν το πιο μακροπρόθεσμο αποτέλεσμα. Με αυτόν τον τρόπο, επιλέγουν και τα χαρακτηριστικά που θα χρησιμοποιήσουν, ανάλογα με την απόδοση τους σε κάθε βήμα, ελαχιστοποιώντας το σφάλμα. Έτσι, τα δέντρα αποφάσεων μπορούν να έχουν πολλές ομοιότητες και να έχουν υψηλή συσχέτιση στις προβλέψεις τους.

Ο συνδυασμός προβλέψεων από πολλαπλά μοντέλα είναι αποτελεσματικότερος, εάν οι προβλέψεις δε συσχετίζονται ή είναι ασθενώς συσχετισμένες. Το Τυχαίο Δάσος αλλάζει το μοντέλο εκμάθησης του κάθε δέντρου, ώστε οι προβλέψεις από το κάθε δέντρο να έχει μικρότερη συσχέτιση.

Στο **δέντρο κατηγοριοποίησης και παλινδρόμησης** (**CART**), σε κάθε κόμβο διάσπασης, το μοντέλο εξετάζει όλες τις μεταβλητές και όλες τις τιμές μεταβλητών, για να επιλέξει το βέλτιστο σημείο διάσπασης. Η αλλαγή στον αλγόριθμο τυχαίου δάσους είναι πως περιορίζεται σε ένα τυχαίο δείγμα χαρακτηριστικών από τα οποία θα πρέπει να επιλέξει.



Εικόνα 2.9: Λογική λειτουργίας του τυχαίου δάσους

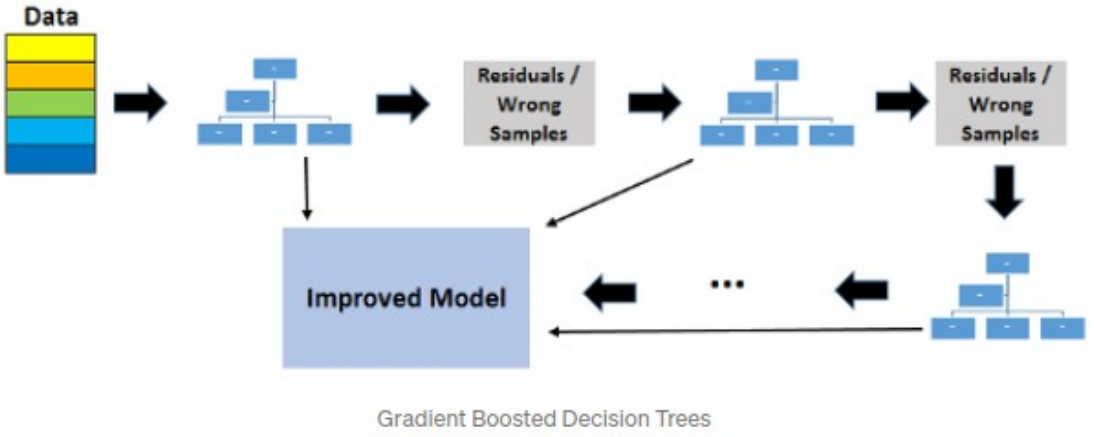
8) **Ενισχυμένα Δέντρα Αποφάσεων (Boosted Decision Tree)**

Με τον όρο **ενίσχυση** (**boosting**) ορίζεται ο διαδοχικός συνδυασμός πολλαπλών ασθενών μοντέλων σε σειρά, ώστε να δημιουργηθεί ένα ισχυρό μοντέλο. Στην περίπτωση των **κλιμακούμενων ενισχυμένων** (**gradient boosted**) δέντρων αποφάσεων, τα αδύναμα μοντέλα είναι επίσης, δέντρα αποφάσεων.

Κάθε δέντρο προσπαθεί να ελαχιστοποιήσει τα σφάλματα του προηγούμενου δέντρου. Το κάθε δέντρο είναι αδύναμο μοντέλο, αλλά η προσθήκη πολλών δέντρων διαδοχικά σε σειρά, όπου το καθένα εστιάζει στα σφάλματα του προηγούμενου, κάνει τα ενισχυμένα δέντρα εξαιρετικά αποδοτικά και ακριβή μοντέλα. Σε αντίθεση με τα δέντρα αποφάσεων bagging, η ενίσχυση δεν περιλαμβάνει δειγματοληψία με επανάθεση, δηλαδή bootstrap. Κάθε φορά που

προστίθεται ένα νέο δέντρο, αναλύει μια τροποποιημένη έκδοση του αρχικού συνόλου δεδομένων.

Σημαντικό στοιχείο είναι πως επειδή τα δέντρα προστίθενται διαδοχικά, οι ενισχυμένοι αλγόριθμοι μαθαίνουν αργά.



Εικόνα 2.10: Λογική λειτουργίας των ενισχυμένων δέντρων αποφάσεων

9) **Κατηγοριοποίηση ψηφοφορίας (Voting Classification)**

Ένας κατηγοριοποιητής ψηφοφορίας είναι μια μέθοδος κατηγοριοποίησης που χρησιμοποιεί πολλαπλούς κατηγοριοποιητές για να κάνει προβλέψεις. Χρησιμοποιείται όταν ένας επιστήμονας δεδομένων δεν μπορεί να διασαφηνίσει πια μέθοδο είναι η ενδεδειγμένη. Επομένως, χρησιμοποιώντας τις προβλέψεις από πολλαπλούς κατηγοριοποιητές, ο κατηγοριοποιητής ψηφοφορίας αποφασίζει για τον πιο χρήσιμο.

10) **Νευρωνικά Δίκτυα (Neural Networks)**

Το νευρωνικό δίκτυο είναι ένα δίκτυο από απλούς υπολογιστικούς κόμβους διασυνδεδεμένους μεταξύ τους. Είναι εμπνευσμένο από το Κεντρικό Νευρικό Σύστημα (ΚΝΣ), το οποίο προσπαθεί να προσομοιώσει. Οι νευρώνες είναι τα δομικά στοιχεία του δικτύου. Κάθε τέτοιος κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων από διαφορετικές πηγές επιτελεί έναν υπολογισμό με βάση αυτές τις

εισόδους και παράγει μία έξοδο. Η εν λόγω έξοδος είτε κατευθύνεται στο περιβάλλον, είτε τροφοδοτείται ως είσοδος σε άλλους νευρώνες του δικτύου.

Το κύριο χαρακτηριστικό των νευρωνικών δικτύων είναι η εγγενής ικανότητα μάθησης. Η μάθηση επιτυγχάνεται μέσω της εκπαίδευσης, μίας επαναληπτικής διαδικασίας σταδιακής προσαρμογής των παραμέτρων του δικτύου σε τιμές κατάλληλες, ώστε να επιλύεται με επαρκή επιτυχία το προς εξέταση πρόβλημα. Αφού ένα δίκτυο εκπαιδευτεί, οι παράμετροί του συνήθως «παγώνουν» στις κατάλληλες τιμές και από εκεί κι έπειτα είναι σε λειτουργική κατάσταση. Το ζητούμενο είναι το λειτουργικό δίκτυο να χαρακτηρίζεται από μία ικανότητα γενίκευσης. Αυτό σημαίνει πως είναι ικανό να δίνει ορθές εξόδους για καινούργια σύνολα δεδομένων.

2.6 Μετρικές αξιολόγησης (Evaluation Metrics) αλγόριθμων εξόρυξης δεδομένων

Όπως ήδη αναφέραμε, οι αλγόριθμοι κατηγοριοποίησης λαμβάνουν τα εκπαιδευτικά σύνολα δεδομένων και χρησιμοποιούν τις διαθέσιμες πληροφορίες για τη δημιουργία μοντέλων πρόβλεψης, προκειμένου να κατηγοριοποιήσουν καινούργια σύνολα δεδομένων. Ο τρόπος για να εκτιμήσουμε την ορθότητα των μοντέλων μας είναι η χρήση των μετρικών αξιολόγησης.

Πριν παραθέσουμε ορισμένους τύπους των μετρικών αξιολόγησης, θα επεξηγήσουμε κάποιους απαραίτητους συμβολισμούς, θεωρώντας γενικά την περίπτωση, όπου η μεταβλητή εξόδου μπορεί να πάρει τιμές από δυο μόνο κλάσεις τη Θετική και την Αρνητική.

P (**Condition Positive**) = ο πραγματικός αριθμός των θετικών εγγραφών στα δεδομένα

N (**Condition Negative**) = ο πραγματικός αριθμός των αρνητικών εγγραφών στα δεδομένα

TP (**True Positive**) = όσες εγγραφές ανήκουν στη Θετική κλάση και κατηγοριοποιήθηκαν στη Θετική κλάση (Κατηγοριοποιήθηκαν σωστά)

FN (**False Negative**) = όσες εγγραφές ανήκουν στη Θετική κλάση και κατηγοριοποιήθηκαν στη Αρνητική κλάση (Κατηγοριοποιήθηκαν λάθος)

FP (**False Positive**) = όσες εγγραφές ανήκουν στην Αρνητική κλάση και κατηγοριοποιήθηκαν στη Θετική κλάση (Κατηγοριοποιήθηκαν λάθος)

TN (**True Negative**) = όσες εγγραφές ανήκουν στην Αρνητική κλάση και κατηγοριοποιήθηκαν στην Αρνητική κλάση (Κατηγοριοποιήθηκαν σωστά)

Γνωρίζοντας πλέον τη σημασία των συμβολισμών, παραθέτουμε τους σημαντικότερους τύπους που χρησιμοποιούνται στις μετρικές αξιολόγησης.

1) **Ορθότητα** (**Accuracy**)

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN}$$

2) **Ακρίβεια (Precision) / PPV (Positive Predictive Value)**

$$PPV = \frac{TP}{TP + FP}$$

3) **Ανάκληση (Recall) / Ευαισθησία (Sensitivity) / TPR** (**True Positive Rate**)

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

4) **Εξειδίκευση** (**Specificity**) **/ TNR** (**True Negative Rate**)

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = 1 - FPR$$

5) **FPR** (**False Positive Rate**)

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - TNR$$

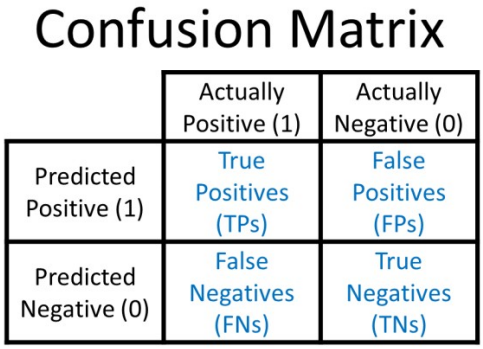
Μετά τις παραπάνω επεξηγήσεις, θα προχωρήσουμε στην αναφορά των κυριότερων μετρικών αξιολόγησης των μοντέλων εξόρυξης δεδομένων. ***[16]***

1) **Πίνακας σύγχυσης (Confusion Matrix)**

Πρόκειται για ένα δισδιάστατο πίνακα που μας δείχνει την απόδοση του μοντέλου μας. Επίσης, αναφέρεται και ως **πίνακας λάθους** (**Error Matrix**). Αν και δεν αποτελεί μια επαρκή αξιολόγηση του μοντέλου, θεωρείται πως είναι ένα σημαντικό πρώτο βήμα για την αξιολόγηση.

Οι προβλέψεις αθροίζονται και κατηγοριοποιούνται ανά κλάση, προτού συγκριθούν με τις πραγματικές τιμές. Το μέγεθος του πίνακα είναι αντίστοιχο με τον αριθμό των κλάσεων της μεταβλητής εξόδου. Αν υπάρχουν δύο μόνο κλάσεις, όπως στο παράδειγμα που αναφέραμε προηγουμένως, ο πίνακας θα είναι 2x2. Αντίστοιχα, αν υπάρχουν 3 κλάσεις, ο πίνακας θα είναι 3x3 και ούτω καθεξής.

Αυτός ο πίνακας μας βοηθάει να προσδιορίσουμε αν το μοντέλο μας βελτιώνεται. Μας δείχνει τα σφάλματα που γίνονται και βοηθάει στον προσδιορισμό του ακριβούς τύπου τους.



Εικόνα 2.11: Πίνακας σύγχυσης

2) **Ορθότητα (Accuracy)**

Η ορθότητα ενός μοντέλου ορίζεται ως το ποσοστό των σωστών προβλέψεων. Πρέπει όμως να λάβουμε υπ' όψιν, πως γίνεται λιγότερο αξιόπιστη όταν η μια κλάση είναι σημαντικά μεγαλύτερη από την άλλη, όπως συμβαίνει στα μη ισορροπημένα σύνολα δεδομένων. Σε αυτή την περίπτωση δε θα πρέπει να χρησιμοποιείται ως μοναδική μετρική.

Ο τύπος της συγκεκριμένης μετρικής αξιολόγησης είναι ο ίδιος που αναφέραμε και προηγουμένως

$$ACC = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN}$$

3) **Λογαριθμική απώλεια (Logarithmic loss)**

Η λογαριθμική απώλεια λειτουργεί θέτοντας κάποια ποινή σε όλες τις εσφαλμένες κατηγοριοποιήσεις. Ο κατηγοριοποιητής εκχωρεί, χάρις στη μετρική, μια συγκεκριμένη πιθανότητα σε κάθε κλάση για όλα τα δείγματα.

Ο υπολογισμός της απώλειας δίνεται από τον τύπο:

$$LogarithmicLoss = \frac{-1}{N} * \sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * \log\left(p_{ij}\right)$$

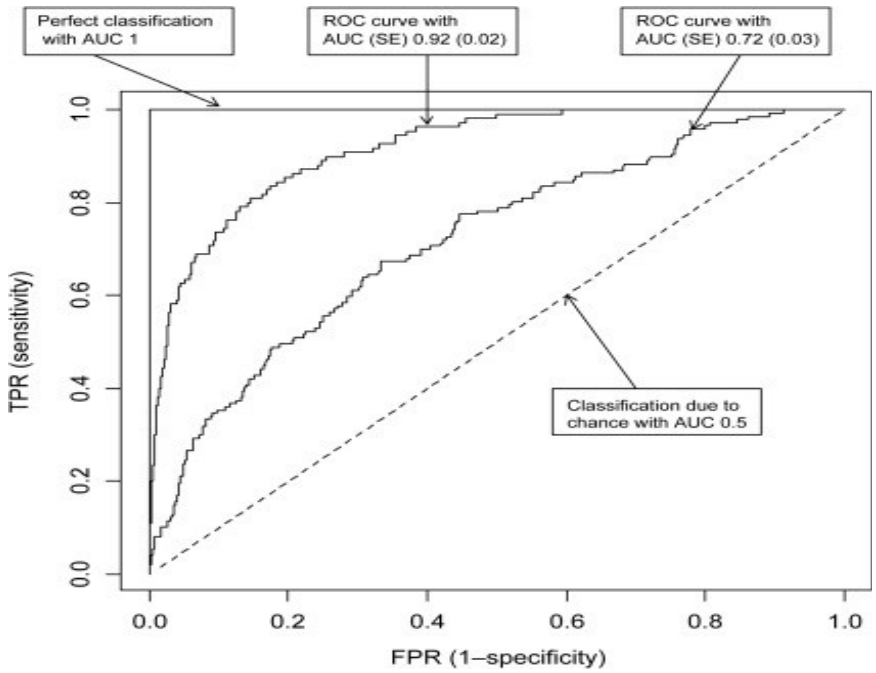
όπου:

- y_{ij} – Δείχνει αν το δείγμα i ανήκει στην κλάση j ή όχι
- p_{ij} - Δείχνει την πιθανότητα του i που ανήκει στην κλάση j

Συνοπτικά, το εύρος της λογαριθμικής απώλειας κυμαίνεται από το 0 έως το άπειρο. Όσο πιο κοντά είναι στο 0, τόσο μεγαλύτερη είναι η ακρίβεια της ορθότητας.

4) **Καμπύλη Λειτουργικού Χαρακτηριστικού Δέκτη ROC (Receiver Operating Characteristic) και η περιοχή κάτω από την καμπύλη AUC (Area Under Curve)**

Η καμπύλη Λειτουργικού Χαρακτηριστικού Δέκτη (ROC) είναι μια μετρική αξιολόγησης για προβλήματα δυαδικής κατηγοριοποίησης. Είναι μια καμπύλη πιθανοτήτων που σχεδιάζει το TPR έναντι του FPR για ποικίλες τιμές του **κατώτατου ορίου** (**threshold**), διαχωρίζοντας ουσιαστικά το θόρυβο από τις παρατηρήσεις. Η περιοχή κάτω από την καμπύλη (AUC) είναι η μετρική που δίνει τη δυνατότητα στον κατηγοριοποιητή να διακρίνει μεταξύ των κλάσεων.



Εικόνα 2.12: Παράδειγμα καμπύλης ROC

Όταν η AUC = 1, τότε ο κατηγοριοποιητής βρίσκει όλα τα σύνολα σε ποια κλάση ανήκουν ακριβώς. Αντίστοιχα, αν η AUC ήταν 0, τότε ο κατηγοριοποιητής θα προέβλεπε όλα τα αρνητικά ως θετικά και όλα τα θετικά ως αρνητικά. Όταν η AUC κυμαίνεται μεταξύ 0,5 και 1, τότε υπάρχει αυξημένη πιθανότητα ο κατηγοριοποιητής να μπορεί να βρίσκει σωστά την κλάση κάθε εγγραφής. Αυτό συμβαίνει επειδή ο κατηγοριοποιητής είναι σε θέση να ανιχνεύσει περισσότερες εγγραφές σωστά από ότι λάθος. Όταν η AUC = 0,5, σημαίνει πως ο κατηγοριοποιητής προβλέπει είτε τυχαία, είτε με σταθερή επιλογή για το σύνολο των δεδομένων.

Γενικά, όσο υψηλότερη είναι η AUC, τόσο καλύτερη είναι η απόδοση του μοντέλου στη διάκριση μεταξύ θετικών και αρνητικών κατηγοριών.

5) **Ευαισθησία (Sensitivity) / Πραγματικό θετικό ποσοστό TPR** (**True Positive Rate**)

Το πραγματικό θετικό ποσοστό, που αναφέρεται ως ευαισθησία, αντιστοιχεί στο ποσοστό των θετικών σημείων δεδομένων που θεωρούνται σωστά ως θετικά, σε σχέση με όλα τα σημεία που όντως ανήκουν στη θετική κλάση, ασχέτως αν κατηγοριοποιήθηκαν σωστά ή όχι.

Ο τύπος της συγκεκριμένης μετρικής αξιολόγησης είναι ο ίδιος που αναφέραμε και προηγουμένως

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

6) **Βαθμολογία F1 (F1 Score)**

Στη στατιστική ανάλυση δυαδικής κατηγοριοποίησης, η βαθμολογία F είναι μετρική της **ορθότητας** (**accuracy**) ενός τεστ. Η βαθμολογία F1 είναι ο αρμονικός μέσος όρος της **ακρίβειας** (**precision**) και της **ανάκλησης** (**recall**).

Η υψηλότερη δυνατή τιμή μιας βαθμολογίας F είναι 1, υποδεικνύοντας τέλεια ακρίβεια και ανάκληση και η χαμηλότερη δυνατή τιμή είναι 0, εάν είτε η ακρίβεια είτε η ανάκληση είναι μηδέν.

$$F_1 = 2\frac{PPV * TPR}{PPV + TPR} = \frac{2 * TP}{2 * TP + FP + FN}$$

3 Μέθοδοι διαχείρισης μη ισορροπημένων δεδομένων κατηγοριοποίησης (Imbalanced classification)

Στη **μηχανική μάθηση**, τα μη ισορροπημένα σύνολα δεδομένων αποτελούν τη συντριπτική πλειοψηφία. Αν πάρουμε ένα παράδειγμα που υπάρχουν δύο πιθανές κλάσεις για τη μεταβλητή εξόδου και το **εκπαιδευτικό σύνολο δεδομένων** (**training dataset**) είναι περίπου στο 50% για κάθε κλάση, τότε αποτελεί ένα ισορροπημένο σύνολο. Αν όμως η μια από τις κλάσεις αντιστοιχεί σε ένα ποσοστό 95% και μόνο το υπόλοιπο 5% ανήκει στην άλλη κλάση, προφανώς έχουμε ένα μη ισορροπημένο σύνολο.

Σύμφωνα με το developers.google.com ***[17]*** τα ποσοστά που μας δείχνουν αν ένα σύνολο είναι μη ισορροπημένο κυμαίνονται ως εξής:

Βαθμός ισορροπίας συνόλου	Ποσοστό μικρότερης κλάσης
Ήπια	20% - 40%
Μέτρια	1% - 20%
Έντονη	<1%

Η πρόκληση της μοντελοποίησης μη ισορροπημένων σύνολων δεδομένων είναι ότι οι περισσότερες τεχνικές μηχανικής μάθησης θα αγνοήσουν τη μικρότερη κλάση και άρα θα μοιάζει πως έχουν καλή απόδοση. Όμως, σημαντικότερο είναι η επιτυχημένη πρόβλεψη των δειγμάτων της μικρότερης τάξης, οπότε το μοντέλο δε θα είναι αποτελεσματικό.

Η εξισορρόπηση των κλάσεων είναι υψίστης σημασίας για την αποτελεσματική μάθηση και την αμερόληπτη λήψη αποφάσεων για την κατηγοριοποίηση. Η ανισορροπία των κλάσεων μπορεί να επιλυθεί με έναν από τους παρακάτω τρεις γενικούς τρόπους:

- Τεχνικές στο επίπεδο των δεδομένων που περιλαμβάνουν την επεξεργασία της δειγματοληψίας πριν τη μηχανική μάθηση.
- Αλγοριθμικές τεχνικές που περιλαμβάνουν την τροποποίηση των αλγορίθμων μηχανικής μάθησης.

- Μάθηση με ευαισθησία κόστους, που αφορά κατά κύριο λόγο τη χρήση ποινών κατά την εκπαίδευση.

Παρακάτω, θα περιγράψουμε πιο αναλυτικά τις βασικότερες μεθόδους που μπορούμε να χρησιμοποιήσουμε και προέρχονται από τους προαναφερόμενους γενικούς τρόπους. ***[18]***

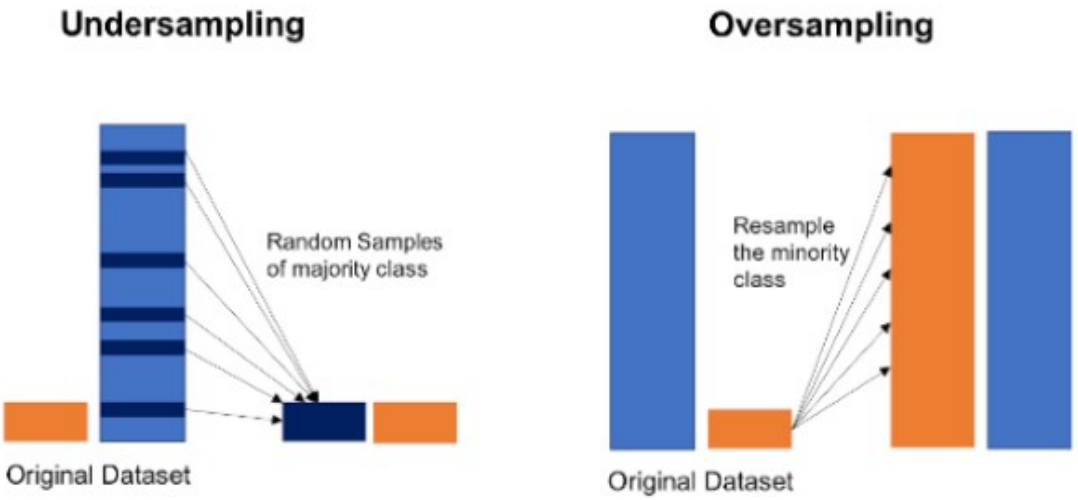
1) **Υπο-δειγματοληψία** (**under-sampling**) **και υπερ-δειγματοληψία** (**over-sampling**)

Οι δυο σημαντικότερες προσεγγίσεις για τη διαχείριση ενός μη ισορροπημένου συνόλου δεδομένων είναι η **υπο-δειγματοληψία** (**under-sampling**) και η **υπερ-δειγματοληψία** (**over-sampling**).

Η υπο-δειγματοληψία εξισορροπεί το σύνολο δεδομένων μειώνοντας τον αριθμό των εγγραφών της επικρατούσας κλάσης. Αυτή η μέθοδος χρησιμοποιείται όταν η συνολική ποσότητα δεδομένων είναι αρκετά μεγάλη. Η κύρια λογική είναι να παραμείνει το σύνολο των εγγραφών της μικρότερης κλάσης και με κάποιο τρόπο, από αυτούς που θα δούμε στη συνέχεια, να κρατήσουμε έναν ίδιο αριθμό εγγραφών και για την επικρατούσα κλάση ή τουλάχιστον να μειωθεί η αναλογία μεταξύ των δυο κλάσεων. Με το καινούργιο εξισορροπημένο σύνολο δεδομένων, μπορούμε να προχωρήσουμε στη μοντελοποίηση.

Αντίθετα, η υπερ-δειγματοληψία χρησιμοποιείται όταν η ποσότητα των δεδομένων δεν είναι επαρκής και άρα, θα υπάρχει πρόβλημα αν αφαιρέσουμε δείγματα από την επικρατούσα κλάση. Η συγκεκριμένη τεχνική για να εξισορροπήσει το σύνολο δεδομένων, αυξάνει το μέγεθος των εγγραφών της μικρότερης κλάσης, ώστε να φτάσει σε ίδιο αριθμό εγγραφών με την επικρατούσα κλάση ή τουλάχιστον να μειώσει την αναλογία μεταξύ τους.

Σημειώστε ότι δεν υπάρχει απόλυτο πλεονέκτημα της μιας μεθόδου δειγματοληψίας έναντι της άλλης, αν και επικρατεί η άποψη πως η υπερ-δειγματοληψία είναι προτιμότερη, αφού η υπο-δειγματοληψία δημιουργεί απώλεια πληροφορίας που πιθανόν να είναι δυσαναπλήρωτη για το σύνολο δεδομένων. Γενικά, η εφαρμογή αυτών των δύο μεθόδων εξαρτάται από το σύνολο των δεδομένων στο οποίο εφαρμόζεται. Επίσης, ένας συνδυασμός υπερ-δειγματοληψίας και υπο-δειγματοληψίας είναι συχνά μια επιτυχημένη, εναλλακτική προσέγγιση.



Εικόνα 3.1: Εξισορρόπηση των δυο κλάσεων με χρήση α) υπο-δειγματοληψίας και β) υπερ-δειγματοληψίας

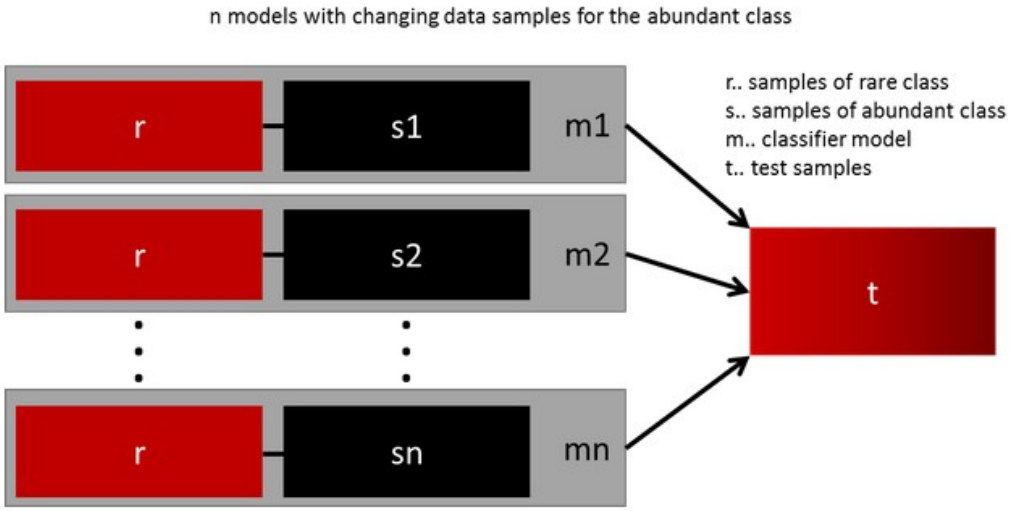
2) **Σωστή χρήση της k-fold διασταύρωση επικύρωσης (K-fold Cross-Validation)**

Αξίζει να σημειωθεί ότι για τα μη ισορροπημένα σύνολα δεδομένων, η **διασταύρωση επικύρωσης** (**Cross-Validation**) πρέπει να εφαρμόζεται πριν χρησιμοποιήσουμε τη μέθοδο της υπερ-δειγματοληψίας. Αφού, η υπερ-δειγματοληψία παίρνει τα λίγα δεδομένα της μικρότερης κλάσης και με βάση αυτά δημιουργεί νέα δεδομένα της ίδιας κλάσης ή απλά τα αντιγράφει, όταν η διασταύρωση επικύρωσης εφαρμοστεί αμέσως μετά την υπερ-δειγματοληψία, βασικά αυτό που κάνουμε είναι να υπερπροσαρμόζουμε το μοντέλο μας σε ένα σύνολο που έχει δημιουργηθεί με κάποια τεχνική από εμάς και πιθανότατα, δε θα μας δώσει αξιόπιστα αποτελέσματα. Πολλές φορές, όμως η συγκεκριμένη σειρά ενεργειών είναι δύσκολο να εφαρμοστεί στην πράξη.

3) **Συγκέντρωση πολλαπλών δημιουργηθέντων δειγμάτων από το συνόλων δεδομένων**

Ο ευκολότερος τρόπος για την επιτυχή γενίκευση ενός μοντέλου είναι με τη χρήση περισσότερων δεδομένων. Το πρόβλημα είναι ότι πολλοί

κλασσικοί αλγόριθμοι κατηγοριοποίησης, όπως ο αλγόριθμος **τυχαίου δάσους** (**random forest**) ή η **λογιστική παλινδρόμηση** (**logistic regression**) τείνουν να εστιάζουν στην πλειοψηφία, με αποτέλεσμα να αγνοείται η μικρότερη κλάση. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι να φτιαχτούν υποομάδες της επικρατούσας κλάσης που η καθεμία να αποτελείται από τόσα δεδομένα όσα έχει και η μικρότερη κλάση. Στη συνέχεια, να χρησιμοποιούμε τον αλγόριθμο εκπαίδευσης για κάθε σύνολο που θα αποτελείται σταθερά από τη μικρότερη κλάση και κάθε φορά από μια διαφορετική υποομάδα της επικρατούσας κλάσης. Έτσι, θα έχουμε πολλά ισορροπημένα σύνολα δεδομένων που το καθένα θα δίνει ένα μοντέλο. Από όλα αυτά, μπορούμε να καταλήξουμε σε ένα γενικότερο μοντέλο για το αρχικό σύνολο δεδομένων μας.

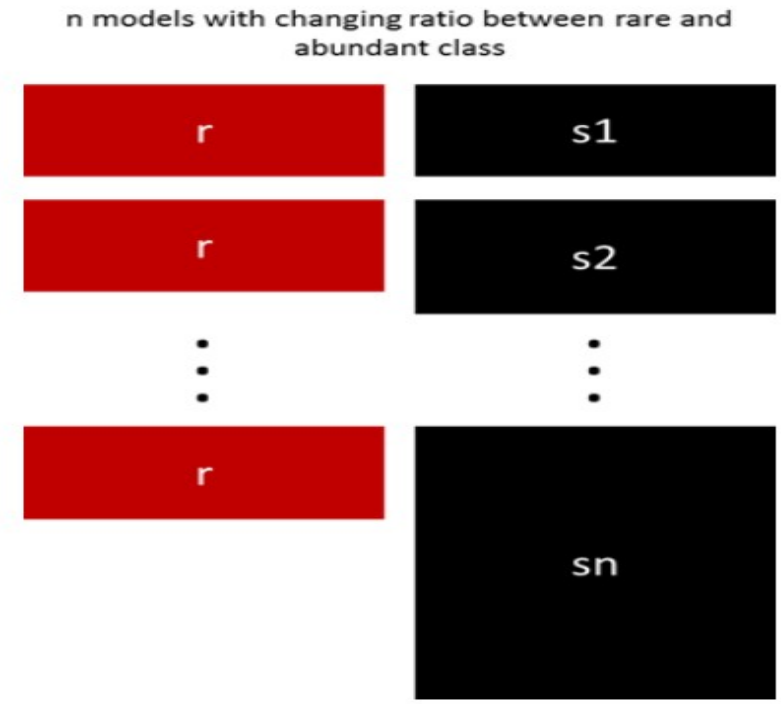


Εικόνα 3.2: Εκπαίδευση της μικρότερης κλάσης με υποομάδες της επικρατούσας κλάσης που έχουν το ίδιο πλήθος δεδομένων

4) **Δημιουργία πολλαπλών δειγμάτων από το συνόλων δεδομένων με διαφορετικές αναλογίες μεταξύ των κλάσεων**

Η προηγούμενη προσέγγιση μπορεί να προσαρμοστεί παίζοντας με την αναλογία μεταξύ της μικρότερης και της επικρατούσας κλάσης. Η καλύτερη αναλογία εξαρτάται σε μεγάλο βαθμό από τα δεδομένα και τα μοντέλα που χρησιμοποιούνται. Προτείνεται επίσης, η εκπαίδευση των μοντέλων να μη γίνεται σε όλες τις υποπεριπτώσεις με την αναλογία 1:1, αλλά να υπάρχουν και κάποιες υποπεριπτώσεις όπου

τα μοντέλα θα εκπαιδεύονται με αναλογίες 1:2 ή 3:1, γενικεύοντας ακόμα περισσότερο την αποδοτικότητα των μοντέλων.



Εικόνα 3.3: Εκπαίδευση της μικρότερης κλάσης με υποομάδες της επικρατούσας κλάσης που έχουν διαφορετικά πλήθη δεδομένων

5) **Χρήση της ενδεδειγμένης μετρικής αξιολόγησης (evaluation metric)**

Η χρήση ακατάλληλων μετρήσεων αξιολόγησης για τη δημιουργία μοντέλου, όταν έχουμε μη ισορροπημένα σύνολα δεδομένων, μπορεί να δημιουργήσει εντελώς εσφαλμένες εκτιμήσεις. Για παράδειγμα, αν έχουμε ένα σύνολο δεδομένων, όπου η μια κλάση υπάρχει σε ποσοστό 98%, τότε αν απλά αξιολογήσουμε το μοντέλο επιλέγοντας για όλο το εκπαιδευτικό σύνολο δεδομένων να βάλουμε ότι ανήκουν στην πρώτη κλάση, τότε θα λάβουμε το εκπληκτικό ποσοστό ακρίβειας στο 98%. Παρ' όλ' αυτά, προφανώς το μοντέλο μας θα είναι εντελώς λάθος.

Σε αυτήν την περίπτωση, μπορούν να εφαρμοστούν άλλες εναλλακτικές μετρήσεις αξιολόγησης που έχουμε αναφέρει και στο

κεφάλαιο 2.6, όπως **Ακρίβεια** (**Precision**), **Ανάκληση (Recall), Βαθμολογία F1** (**F1 Score**) κλπ.

6) **Χρήση μοντέλων με ποινή (Penalized models)**

Μπορούμε να χρησιμοποιήσετε τους ίδιους αλγόριθμους, αλλά να τους κατευθύνουμε με μια διαφορετική αντιμετώπιση του προβλήματος. Η κατηγοριοποίηση με ποινή επιβάλλει ένα επιπλέον κόστος στο μοντέλο, σε περίπτωση που το λάθος αφορά τη μικρότερη κατηγορία, κατά τη διάρκεια της εκπαίδευσης. Αυτές οι ποινές μπορούν να κάνουν το μοντέλο να δώσει μεγαλύτερη προσοχή στη μικρότερη κλάση. Συνήθως, η χρήση των ποινών ή των βαρών εξειδικεύεται στον αλγόριθμο εκμάθησης. Υπάρχουν εκδοχές των αλγορίθμων με ποινή, όπως το penalized-SVM και το penalized-LDA.

Η χρησιμοποίηση ποινών είναι επιθυμητή, εάν είμαστε εγκλωβισμένοι σε έναν συγκεκριμένο αλγόριθμο και δεν μπορούμε να κάνουμε διαφοροποιήσεις στη δειγματοληψία ή έχουμε άσχημα αποτελέσματα. Ουσιαστικά, προσφέρει έναν ακόμη τρόπο να εξισορροπηθούν οι κλάσεις. Το πρόβλημα είναι πως η διαδικασία της σωστών τιμών για τις ποινές μπορεί να είναι αρκετά περίπλοκη. Πιθανότατα, θα πρέπει να δοκιμάσουμε πολλαπλές τιμές για τις ποινές και να δούμε ποιες λειτουργούν καλύτερα για το πρόβλημά μας.

7) **Χρήση διαφορετικής θεώρησης των μη ισορροπημένων μοντέλων**

Υπάρχουν τομείς μελέτης αφιερωμένοι σε μη ισορροπημένα σύνολα δεδομένων. Έχουν τους δικούς τους αλγόριθμους, μέτρα και ορολογία.

Δύο σημαντικοί τομείς είναι η **ανίχνευση ανωμαλιών** (**anomaly detection**) και η **ανίχνευση αλλαγών** (**change detection**).

Η ανίχνευση ανωμαλιών είναι η ανίχνευση σπάνιων γεγονότων. Αυτό θα μπορούσε να είναι μια κακόβουλη δραστηριότητα από ένα πρόγραμμα. Η κακόβουλη δραστηριότητα θα πρέπει να συμβαίνει σε πολύ μικρό ποσοστό σε σύγκριση με την ομαλή ροή του συστήματος. Αυτή η διαφορετικότητα στον τρόπο προσέγγισης θεωρεί τη μικρότερη κλάση ως την κλάση των **ακραίων τιμών** (**outliers**) που μπορεί να μας βοηθήσει να επινοήσουμε νέους τρόπους διαχωρισμού και κατηγοριοποίησης των δειγμάτων.

Η ανίχνευση αλλαγών είναι παρόμοια με την ανίχνευση ανωμαλιών, εκτός από τον τρόπο θεώρησης, όπου αντί να ψάχνει για μια ανωμαλία, ψάχνει για αλλαγή ή διαφορά. Αυτό μπορεί να είναι μια αλλαγή στη συμπεριφορά ενός χρήστη, όπως παρατηρείται από τα πρότυπα χρήσης ή τις τραπεζικές συναλλαγές.

Και οι δύο αυτές αλλαγές ταιριάζουν περισσότερο σε προβλήματα πραγματικού χρόνου της κατηγοριοποίησης που μπορεί να μας δώσει μερικούς νέους τρόπους σκέψης για το πρόβλημά μας.

3.1 Ανάλυση μεθόδων υπο-δειγματοληψίας, υπερ-δειγματοληψίας και μεικτών

Στόχος της συγκεκριμένης διπλωματικής εργασίας είναι η εφαρμογή και η αξιολόγηση των βασικότερων προτεινόμενων τεχνικών υπο-δειγματοληψίας, υπερ-δειγματοληψίας και μεικτών τεχνικών, όπως αναφέρονται στην ερευνητική εργασία των **Susan** και **Kumar** με τίτλο "**The balancing trick: Optimized sampling of imbalanced datasets – A brief survey of the recent State of the Art**". ***[B]*** Ήδη, από το προηγούμενο κεφάλαιο έχουμε κάνει μια γενική εισαγωγή στις παραπάνω τεχνικές για σύνολα δεδομένων που η εξαρτημένη μεταβλητή αποτελείται από δυο κλάσεις. Σε αυτό το σημείο, θα αναφερθούμε σε κάποιες παραμέτρους που θα μας είναι χρήσιμες.

Η **αναδειγματοληψία** (**resampling**) είναι απαιτούμενη για να επιτευχθεί η εξισορρόπηση των δύο κλάσεων. Όμως, θα πρέπει να υπάρχει η προϋπόθεση πως **η διάρθρωση τόσο της κάθε κλάσης ξεχωριστά** (**intra-class diversity**), όσο και **μεταξύ των δυο κλάσεων** (**inter-class diversity**) δεν έχουν αλλοιωθεί στο καινούργιο σύνολο δεδομένων. Οπότε, η αναδειγματοληψία θα πρέπει να γίνει προσεκτικά και με έξυπνους τρόπους. Μια επιλογή ορισμένων από τα υπάρχοντα δείγματα της επικρατούσας κλάσης που να διατηρεί τις παραπάνω προϋποθέσεις, θα ήταν ιδανική.

Οι τεχνικές που θα αναφερθούν αναλυτικά στη συνέχεια είναι κλασσικές μέθοδοι όπως το SMOTE για τη μικρότερη κλάση και η τυχαία υπο-δειγματοληψία της επικρατούσας κλάσης ή και τεχνικές που έχουν να κάνουν με την **οριογραμμή** (**borderline**) που διαχωρίζει τις δυο κλάσεις, τον τρόπο επιλογής της ή τα δείγματα που βρίσκονται κοντά της και δημιουργούν το όριο για την τελική επιλογή κλάσης.

Και σε αυτό το ζήτημα, ο "έξυπνος" τρόπος με τον οποίο λειτουργεί ο κάθε αλγόριθμος δειγματοληψίας διαφέρει. Με τον όρο "έξυπνος" εννοούμε έναν αλγόριθμο που διατηρεί τις διαρθρώσεις τόσο στο εσωτερικό των κλάσεων, όσο και μεταξύ των κλάσεων, όπως αναφέραμε προηγουμένως.

Εδώ, να σημειώσουμε πως το πλήθος των δειγμάτων του τελικού συνόλου δεδομένων που εκπαιδεύεται δεν παίζει κανένα ρόλο στην απόδοση.

3.1.1 Μέθοδοι υπο-δειγματοληψίας

Οι μέθοδοι της υπο-δειγματοληψίας διαφέρουν ως προς τον τρόπο επιλογής των δειγμάτων που θα παραμείνουν από την επικρατούσα κλάση. Η μια λογική είναι να εστιάζουν στα δείγματα που θα διαγραφούν, ενώ η άλλη είναι να εστιάζει στα δείγματα που θα παραμείνουν. Η τρίτη μέθοδος συνδυάζει τα δυο παραπάνω.

Στη συνέχεια, θα αναλύσουμε ορισμένες σημαντικές μεθόδους υπο-δειγματοληψίας. ***[19]***

1) **Τυχαία υπο-δειγματοληψία (Random under-sampling, RUS)**

Στην τυχαία υπο-δειγματοληψία αφαιρούμε με τυχαίο τρόπο από την επικρατούσα κλάση τόσα δείγματα, ώστε να εξισορροπήσουμε το πλήθος της επικρατούσας κλάσης με τον αριθμό του δείγματος της μικρότερης κλάσης. Αυτή η μέθοδος δημιουργεί μεγάλη πιθανότητα απώλειας πληροφορίας από το αρχικό σύνολο δεδομένων, κάτι που προφανώς έχει μεγάλο αντίκτυπο στην απόδοση του μοντέλου μας.

Στην Python η συγκεκριμένη μέθοδος υλοποιείται με την κλάση **RandomUnderSampler**.

2) **Near Miss υπο-δειγματοληψία**

Η τεχνική της Near Miss υπο-δειγματοληψίας βασίζεται στη μέθοδο των **k-πλησιέστερων γειτόνων** (**k-Nearest Neighbors, KNN**). Τη συγκεκριμένη μέθοδο την έχουμε αναφέρει στο κεφάλαιο 2.5 όπου χρησιμοποιείται για την εξόρυξη δεδομένων. Εδώ θα δούμε τη χρήση της για μη ισορροπημένα σύνολα δεδομένων.

Το Near Miss αποτελεί μια συλλογή τριών παραπλήσιων τεχνικών που επιλέγουν τα δείγματα με βάση την απόσταση των δειγμάτων της επικρατούσας κλάσης από τα δείγματα της μικρότερης κλάσης. Αναλυτικότερα, έχουμε τις παρακάτω τεχνικές:

i) Το NearMiss-1 επιλέγει να διατηρήσει τα δείγματα της επικρατούσας κλάσης χρησιμοποιώντας ως λογική τη μικρότερη μέση απόσταση από τα τρία πλησιέστερα δείγματα της μικρότερης κλάσης. Σε αυτή την έκδοση του Near Miss αναμένουμε οι συστάδες των

δειγμάτων της επικρατούσας κλάσης που θα παραμείνουν να είναι αυτές που βρίσκονται γύρω από τα δείγματα της μικρότερης κλάσης, δηλαδή στην περιοχή της επικάλυψης.

ii) Το NearMiss-2 επιλέγει να διατηρήσει τα δείγματα της επικρατούσας κλάσης χρησιμοποιώντας ως λογική τη μικρότερη μέση απόσταση από τα τρία πιο απομακρυσμένα δείγματα της μικρότερης κλάσης. Εδώ, αν και δεν είναι προφανής η λογική, αναμένουμε τα επιλεγμένα δείγματα της επικρατούσας κλάσης να βρίσκονται στο κεντρικότερο σημείο της επικάλυψης των δυο κλάσεων.

ii) Το NearMiss-3 επιλέγει να διατηρήσει ένα συγκεκριμένο αριθμό δειγμάτων της επικρατούσας κλάσης για κάθε δείγμα της μικρότερης κλάσης που είναι πιο κοντά. Σε αυτή την περίπτωση, θα δούμε πως τα δείγματα της μικρότερης κλάσης που βρίσκονται στην περιοχή της επικάλυψης των δυο κλάσεων έχουν σχετικά κοντά τους μέχρι ν γείτονες από την επικρατούσα κλάση.

Ο τρόπος μέτρησης της απόστασης καθορίζεται στο ν-διάστατο χώρο των χαρακτηριστικών, χρησιμοποιώντας είτε την Ευκλείδεια απόσταση, είτε κάποια παρόμοια μετρική. ***[C]***

Στην Python η συγκεκριμένη μέθοδος υλοποιείται με την κλάση **NearMiss**.

3) **Μέθοδος συμπυκνωμένων πλησιέστερων γειτόνων (Condensed Nearest Neighbors, CNN)**

Η μέθοδος CNN είναι μια τεχνική υπο-δειγματοληψίας που αναζητάει ένα υποσύνολο δειγμάτων που να επιτυγχάνει 100% επιτυχία του μοντέλου. Αυτό το υποσύνολο αναφέρεται ως το ελάχιστο συνεπές σύνολο. Η συγκεκριμένη μέθοδος αποτελεί μια εναλλακτική της KNN μεθόδου, αλλά έχει το πλεονέκτημα ότι μειώνει τις απαιτήσεις στη χρήση της μνήμης του υπολογιστή. Η μέθοδος CNN διαγράφει τα δείγματα της επικρατούσας κλάσης που βρίσκονται μακριά από την οριογραμμή και δεν προσφέρουν κάτι στην κατηγοριοποίηση των νέων δειγμάτων.

Ο τρόπος λειτουργίας είναι αρχικά, η απαρίθμηση των δειγμάτων στο σύνολο δεδομένων και μετά, η αποθήκευση τους μόνο στην περίπτωση που δεν μπορούν να κατηγοριοποιηθούν σωστά από τα ήδη αποθηκευμένα δείγματα. Όταν χρησιμοποιείται για μη ισορροπημένα σύνολα δεδομένων, τα αποθηκευμένα δείγματα αποτελούνται από

όλα τα δείγματα της μικρότερης κλάσης και στη συνέχεια, αποθηκεύονται δείγματα της επικρατούσας κλάσης που κατηγοριοποιούνται εσφαλμένα. Στο τελικό στάδιο του αλγόριθμου ενδέχεται η αναλογία των δυο κλάσεων να μην είναι ένα προς ένα, αλλά σίγουρα θα είναι μικρότερη από την αρχική αναλογία του συνόλου δεδομένων.

Γενικά, το CNN αποτελεί μια σχετικά αργή διαδικασία, οπότε δεν επιλέγεται για πολύ μεγάλα σύνολα δεδομένων, ενώ συνήθως προτιμάται να μη ζητείται μεγάλος αριθμός γειτόνων, αλλά κάποιος μικρός αριθμός. Ένα επίσης σημαντικό μειονέκτημα της συγκεκριμένης μεθόδου είναι πως επιλέγει δείγματα με τυχαίο τρόπο, αφού έχει σημασία η σειρά με την οποία ελέγχονται τα δείγματα, και ειδικά το αρχικό υποσύνολο δειγμάτων που επιλέγεται. Αυτό έχει ως αποτέλεσμα να αποθηκεύονται μη σημαντικά δείγματα, καθώς και δείγματα που είναι στο εσωτερικό της επικρατούσας κλάσης και όχι στην **οριογραμμή** (**borderline**). ***[D]***

Στην Python η συγκεκριμένη μέθοδος υλοποιείται με την κλάση **CondensedNearestNeighbour**.

4) **Σύνδεσμοι Tomek** (**Tomek links**)

Στη μέθοδο των συνδέσμων Tomek, αναζητούμε τα δείγματα που είναι πολύ κοντινά μεταξύ τους, αλλά προέρχονται από διαφορετικές κλάσεις. Η κοντινή απόσταση μπορεί να αξιολογηθεί με κάποια μετρική, όπως είναι η Ευκλείδεια απόσταση.

Για να θεωρηθεί ένα ζευγάρι δειγμάτων ως σύνδεσμο Tomek θα πρέπει να πληρούνται οι εξής προϋποθέσεις:

i) Το δείγμα Α να έχει πιο κοντινό δείγμα το Β

ii) Το δείγμα Β να έχει πιο κοντινό δείγμα το Α

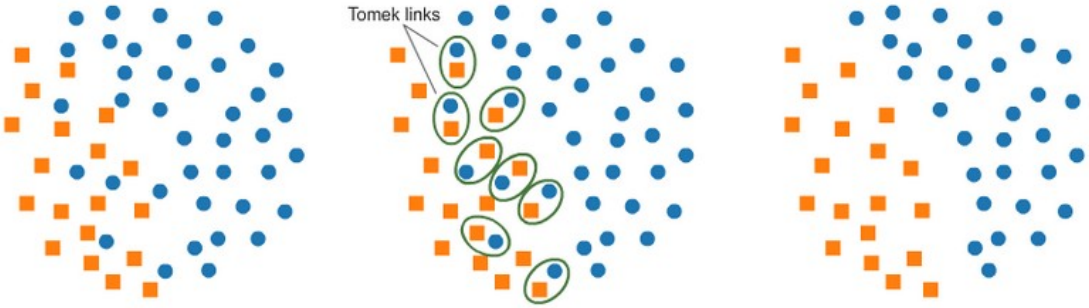
iii) Τα δείγματα Α και Β να ανήκουν σε διαφορετικές κλάσεις

Σε θεωρητικό επίπεδο, τα ζευγάρια Tomek ορίζουν την οριογραμμή μεταξύ των δυο κλάσεων. Υπάρχει όμως και η άλλη περίπτωση, όπου κάποιο από τα δυο δείγματα αποτελεί θόρυβο. Αυτό είναι λογικό, καθώς μόνο οριακά δείγματα και δείγματα που αποτελούν θορύβους θα έχουν τους πλησιέστερους γείτονες τους στην άλλη κλάση.

Ο τρόπος χειρισμού στη μέθοδο των συνδέσμων Tomek είναι:

i) είτε να σβηστούν και τα δυο δείγματα, προκειμένου να αυξηθεί η απόσταση μεταξύ των δυο κλάσεων, διευκολύνοντας τη διαδικασία της κατηγοριοποίησης.

ii) είτε να σβηστεί μόνο το δείγμα που ανήκει στην επικρατούσα κλάση, θεωρώντας πως αποτελεί θόρυβο.



Εικόνα 3.4: Παράδειγμα υπο-δειγματοληψίας με τη μέθοδο συνδέσμων Tomek.

Προφανώς, στη συγκεκριμένη μέθοδο δεν αναμένουμε να υπάρχει πλήρης εξισορρόπηση μεταξύ των δυο κλάσεων, αφού συνήθως η αφαίρεση των δειγμάτων της επικρατούσας κλάσης είναι πολύ μικρή, οπότε συνδυάζουμε αυτή τη μέθοδο με κάποια άλλη, όπως είναι η KNN ή κάποιες άλλες που θα δούμε στην υπερ-δειγματοληψία. ***[E]***

Στην Python η συγκεκριμένη μέθοδος υλοποιείται με την κλάση **TomekLinks**.

5) **Μέθοδος επεξεργασμένων πλησιέστερων γειτόνων (Edited Nearest Neighbors, ENN)**

Μια ακόμα μέθοδος που χρησιμοποιείται για την εύρεση δειγμάτων πάνω στην οριογραμμή που χωρίζει τις δυο κλάσεις, καθώς και των δειγμάτων που αποτελούν θορύβους είναι η μέθοδος των επεξεργασμένων πλησιέστερων γειτόνων, ENN.

Αυτή η μέθοδος αρχικά εφαρμόζει μια διαδικασία για τους 3 πλησιέστερους γείτονες των δειγμάτων που η κατηγοριοποίηση τους μέσω του μοντέλου δεν είναι σωστή και λειτουργεί ως μια πρώτη μείωση του πλήθους των δεδομένων, αφού αφαιρούνται δείγματα από

την επικρατούσα κλάση. Στη συνέχεια, εφαρμόζεται ένας κατηγοριοποιητής για μόνο ένα πλησιέστερο δείγμα που αποδίδει τις τελικές εκτιμήσεις.

Η πρώτη φάση με τους 3 πλησιέστερους γείτονες λειτουργεί ως εξής. Υπολογίζονται οι 3 πλησιέστεροι γείτονες για κάθε δείγμα. Αν το δείγμα είναι της επικρατούσας κλάσης και δεν κατηγοριοποιείται σωστά, τότε διαγράφεται. Αν το δείγμα ανήκει στη μικρότερη κλάση και δεν κατηγοριοποιείται σωστά, τότε διαγράφονται οι γείτονες που ανήκουν στην επικρατούσα κλάση.

Η δεύτερη φάση είναι μια κατηγοριοποίηση με έναν μόνο πλησιέστερο γείτονα, από όπου προκύπτει η τελική εκτίμηση.

Και σε αυτή τη μέθοδο, όπως και στους συνδέσμους Tomek, υπάρχει μείωση της αναλογίας μεταξύ των κλάσεων, αλλά όπως είναι προφανές η αναλογία δε φτάνει το ένα προς ένα. Οπότε, καλύτερο είναι να χρησιμοποιείται σε συνδυασμό με κάποια μέθοδο υπο-δειγματοληψίας. ***[F]***

Στην Python η συγκεκριμένη μέθοδος υλοποιείται με την κλάση **EditedNearestNeighbours**.

6) **Μέθοδος μονόπλευρης επιλογής (One-Sided Selection, OSS)**

Η μέθοδος της μονόπλευρης επιλογής, OSS αποτελεί ένα συνδυασμό των συνδέσμων Tomek και της μεθόδου CNN. Οι σύνδεσμοι Tomek φροντίζουν να απομακρύνουν τα δείγματα της επικρατούσας κλάσης που βρίσκονται στην οριογραμμή μεταξύ των δυο κλάσεων, καθώς και τα δείγματα που αποτελούν θόρυβο. Αντίστοιχα, όπως αναφέραμε και προηγουμένως, η μέθοδος CNN διαγράφει όσα δείγματα της επικρατούσας κλάσης δεν προσφέρουν στην κατηγοριοποίηση των νέων δειγμάτων.

Η χρήση της μεθόδου CNN πραγματοποιείται σε ένα βήμα και περιλαμβάνει αρχικά την αποθήκευση όλων των δειγμάτων της μικρότερης κλάσης και ενός μικρού υποσυνόλου δειγμάτων της επικρατούσας κλάσης. Στη συνέχεια, κατηγοριοποιούνται όλα τα υπόλοιπα δείγματα της επικρατούσας κλάσης με τη μέθοδο KNN για ένα μοναδικό γείτονα και προστίθενται στα αποθηκευμένα δείγματα μόνο εκείνα που κατατάχθηκαν σε λάθος κλάση, δηλαδή στη μικρότερη κλάση. ***[G]***

Στην Python η συγκεκριμένη μέθοδος υλοποιείται με την κλάση **OneSidedSelection**.

7) **Μέθοδος εκκαθάρισης γειτόνων (Neighborhood Cleaning Rule, NCR)**

Η μέθοδος εκκαθάρισης γειτόνων, NCR συνδυάζει τη μέθοδο των συμπυκνωμένων πλησιέστερων γειτόνων, CNN και τη μέθοδο των επεξεργασμένων πλησιέστερων γειτόνων, ENN. Όπως αναφέραμε και προηγουμένως, η μέθοδος CNN διαγράφει τα δείγματα της επικρατούσας κλάσης που είναι μακριά από την οριογραμμή των δυο κλάσεων, ενώ η μέθοδος ENN διαγράφει τα δείγματα της επικρατούσας κλάσης που βρίσκονται στην οριογραμμή ή αποτελούν θορύβους.

Η διαφορά με τη μέθοδο μονόπλευρης επιλογής, OSS είναι πως διαγράφονται λιγότερα από τα περιττά δείγματα που βρίσκονται μακριά από την οριογραμμή. Η βασική εστίαση της μεθόδου NCR είναι στον “καθαρισμό” των δειγμάτων που διατηρούνται, ώστε το σύνολο που θα παραμείνει να είναι πιο ποιοτικό και κατ’ επέκταση πιο αποδοτικό. Στο τέλος, έχουμε ένα λιγότερο εξισορροπημένο σύνολο δεδομένων μεταξύ των δυο κλάσεων, αλλά τα δείγματα που παραμένουν είναι πιο συγκεκριμένα ως προς τα όρια της κάθε κλάσης, ενώ έχουν διαγραφεί και οι θόρυβοι και άρα η κατηγοριοποίηση είναι αποτελεσματικότερη.

Και σε αυτή τη μέθοδο αποθηκεύονται όλα τα δείγματα της μικρότερης κλάσης και μετά, χρησιμοποιώντας το ENN, όσα δείγματα της επικρατούσας κλάσης βρίσκονται στην οριογραμμή διαγράφονται. Τέλος, η χρήση της μεθόδου CNN πραγματοποιείται σε ένα βήμα, όπου τα υπόλοιπα δείγματα της επικρατούσας κλάσης που έχουν κατηγοριοποιηθεί λάθος, σε σχέση με το αποθηκευμένο σύνολο, διαγράφονται, αλλά μόνο σε περίπτωση που ο αριθμός των δειγμάτων της επικρατούσας κλάσης είναι μεγαλύτερος από το μισό του πλήθους των δειγμάτων της μικρότερης κλάσης. ***[H]***

Στην Python η συγκεκριμένη μέθοδος υλοποιείται με την κλάση **NeighbourhoodCleaningRule**.

8) **Μέθοδος με χρήση συστάδων (Clustering)**

Μια άλλη μέθοδος, που περιέχει μεγαλύτερο βαθμό τυχαιότητας, είναι να **συσταδοποιήσουμε** (**clustering**), δηλαδή να ομαδοποιήσουμε, τα δείγματα της επικρατούσας κλάσης. Στη συνέχεια, αφαιρούμε κάποια δείγματα από κάθε **συστάδα** (**cluster**), μειώνοντας τον αριθμό, αλλά παράλληλα επιδιώκοντας το δείγμα που θα παραμείνει να είναι αντιπροσωπευτικό.

Επιλέγουμε ο αριθμός των συστάδων να είναι ίσος με το πλήθος των δειγμάτων της μικρότερης κλάσης. Ο τρόπος επιλογής των τελικών δειγμάτων της επικρατούσας κλάσης είναι:

i) με τη μέθοδο των **κεντρικών δειγμάτων των συστάδων (Cluster Centroids**), όπου διατηρούμε μόνο το κεντρικό σημείο κάθε συστάδας, δημιουργώντας το καινούργιο σύνολο δεδομένων της επικρατούσας κλάσης

ii) με την αντικατάσταση του κεντρικού σημείου κάθε συστάδας με το πλησιέστερο γειτονικό δείγμα που ήδη υπάρχει, διατηρώντας μόνο αυτά τα δείγματα για να έχουμε ένα υποσύνολο του αρχικού συνόλου δεδομένων.

9) **Μέθοδος βελτιστοποίησης του υποσυνόλου δειγμάτων (Sample Subset Optimization, SSO)**

Για τη μέθοδο SSO η βασική ιδέα είναι η επιλογή ενός συγκεκριμένου υποσυνόλου από όλα τα διαθέσιμα δείγματα, χρησιμοποιώντας ως κριτήριο την ελαχιστοποίηση του αναμενόμενου σφάλματος. Αυτό προκύπτει με μια διαδικασία **διασταύρωσης επικύρωσης** (**cross-validation**) στα δεδομένα εκπαίδευσης.

Αν η μέθοδος SSO εφαρμοστεί σε μια προσέγγιση βελτιστοποίησης, όπως η μέθοδος PSO μπορούν να ληφθούν παράλληλα πολλά βελτιστοποιημένα υποσύνολα δειγμάτων με μια μόνο εκτέλεση της μεθόδου PSO.

Ανάλυση της μεθόδου βελτιστοποίησης ατόμου και σμήνους (Particle Swarm Optimization, PSO)

Η συγκεκριμένη μέθοδος αποτελεί έναν ευρετικό αλγόριθμο, ο οποίος είναι εμπνευσμένος από τη συμπεριφορά του σμήνους των πτηνών. Η

λογική της μεθόδου δίνει έμφαση στο συνδυασμό της κίνησης τόσο του σμήνους ως ενιαίο σύνολο, όσο και του κάθε μεμονωμένου πτηνού ξεχωριστά. Για παράδειγμα, όταν ψάχνουν για τροφή και βρίσκουν κάτι να φάνε, όλα ακολουθούν το σμήνος αλλά παράλληλα κοιτάζει και το κάθε πτηνό ξεχωριστά μήπως βρει τροφή, προκειμένου να αυξήσει την ικανοποίηση του κορεσμού της πείνας, τόσο του ίδιου, όσο και συνολικά του σμήνους.

Αντίστοιχα λοιπόν, η λύση για ένα δεδομένο πρόβλημα αναπτύσσεται από τις αλληλεπιδράσεις μεταξύ ατόμων, ουσιαστικά παραπέμποντας στα μεμονωμένα δείγματα, που κινούνται σε έναν ν-διάστατο χώρο για την αναζήτηση λύσεων για ολόκληρο το σμήνος, όπου αναφέρεται στο σύνολο δεδομένων. Μια λύση σε αυτή την περίπτωση θεωρείται η θέση των ατόμων στο ν-διάστατο χώρο χαρακτηριστικών. Στους αλγόριθμους PSO, κάθε άτομο στο σμήνος κινείται λαμβάνοντας υπόψη τη δική του εμπειρία, καθώς και την εμπειρία του πιο επιτυχημένου γείτονα. Η δική του εμπειρία αφορά τη μνήμη του ατόμου για την καλύτερη θέση που είχε στο παρελθόν. ***[I]***

3.1.2 Μέθοδοι υπερ-δειγματοληψίας

Ο κύριος λόγος που ένα μη ισορροπημένο σύνολο δεδομένων προκαλεί πρόβλημα στους αλγόριθμους εξόρυξης είναι επειδή η μικρότερη κλάση δεν έχει επαρκή αντιπροσωπευτικά δείγματα για να μπορέσει να οριστεί η οριογραμμή μεταξύ των κλάσεων. Ένας τρόπος αντιμετώπισης είναι η υπερ-δειγματοληψία που αυξάνει τον αριθμό των δειγμάτων της μικρότερης κλάσης.

Στη συνέχεια, θα αναλύσουμε ορισμένες σημαντικές μεθόδους υπερ-δειγματοληψίας. ***[20]***

1) **Τυχαία υπερ-δειγματοληψία** (**Random over-sampling, ROS**)

Στην τυχαία υπερ-δειγματοληψία δημιουργούμε πρόσθετα αντίγραφα των δειγμάτων της μικρότερης κλάσης, με τυχαίο τρόπο, ώστε ο αριθμός τους στο εκπαιδευτικό σύνολο δεδομένων να φτάσει εκείνον της επικρατούσας κλάσης. Αυτή η μέθοδος δεν προσφέρει τίποτε από πλευράς καινούργιας πληροφορίας για το διαχωρισμό των κλάσεων

και δημιουργεί πιθανό πρόβλημα **υπερπροσαρμογής** (**overfitting**), οπότε το μοντέλο που θα δημιουργήσουμε θα έχει μικρή απόδοση σε νέα δεδομένα. Η συγκεκριμένη μέθοδος θα μπορούσε να χρησιμοποιηθεί ως έσχατη λύση, μόνο στην περίπτωση που θα είχαμε υπερβολικό αριθμό δεδομένων και λόγω του πλήθους τους, οποιαδήποτε περαιτέρω επεξεργασία των δεδομένων θα ήταν πολύ δύσκολο να τη διαχειριστεί η υπολογιστική ισχύς που θα είχαμε στη διάθεση μας.

Στην Python η συγκεκριμένη μέθοδος υλοποιείται με την κλάση **RandomOverSampler**.

2) **Τεχνική υπερ-δειγματοληψία συνθετικής μικρότερης κλάσης (Synthetic Minority Oversampling Technique, SMOTE)**

Μια βελτιωμένη μέθοδος σε σχέση με την απλή δημιουργία αντιγράφων των δειγμάτων της μικρότερης κλάσης, είναι η σύνθεση νέων δειγμάτων από τη μικρότερη κλάση. Αυτό είναι ένα είδος πιο ουσιαστικής αύξησης των δειγμάτων σε σύνολα δεδομένων και είναι η βάση για πολλές εναλλακτικές μεθόδους.

Η τεχνική υπερ-δειγματοληψία συνθετικής μικρότερης κλάσης, η οποία καλείται SMOTE, είναι η πιο διαδεδομένη τεχνική αύξησης του συνόλου δεδομένων. Ο τρόπος που η SMOTE παράγει καινούργια συνθετικά δείγματα στο ν-διάστατο χώρο των χαρακτηριστικών είναι ο εξής. Με ορισμένα κριτήρια, επιλέγει δυο ήδη υπάρχοντα δείγματα και πάνω στη νοητή ευθεία του ν-διάστατου χώρου τοποθετεί καινούργια συνθετικά δείγματα, αυξάνοντας έτσι το πλήθος της μικρότερης κλάσης.

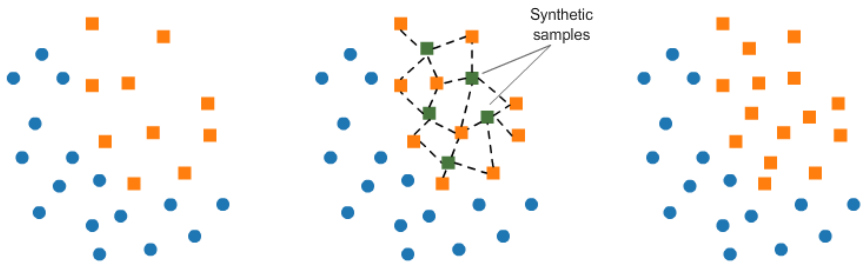
Συγκεκριμένα, η SMOTE επιλέγει πρώτα τυχαία ένα δείγμα από τη μικρότερη κλάση και βρίσκει τους k πλησιέστερους γείτονες της που επίσης ανήκουν στη μικρότερη κλάση. Από εκεί επιλέγει πάλι τυχαία έναν από τους πλησιέστερους γείτονες και στην ευθεία που δημιουργούν τα δυο δείγματα στο χώρο των χαρακτηριστικών επιλέγεται ένα σημείο που είναι το καινούργιο συνθετικό δείγμα που δημιουργείται. Δηλαδή, αποτελεί έναν κυρτό συνδυασμό των δύο επιλεγμένων δειγμάτων. Με αυτή τη μέθοδο δημιουργούμε όσα συνθετικά δείγματα της μικρότερης κλάσης χρειαζόμαστε.

Η αποτελεσματικότητα της συγκεκριμένης μεθόδου οφείλεται στο γεγονός πως τα καινούργια συνθετικά δείγματα βρίσκονται μεταξύ των ήδη υπαρχόντων δειγμάτων της μικρότερης κλάσης. Έτσι, το νέο

μετασχηματισμένο σύνολο δεδομένων δείχνει να διαθέτει περισσότερα δείγματα της μικρότερης κλάσης τα οποία όμως διαθέτουν τιμές που ανταποκρίνονται στη δομή της συγκεκριμένης κλάσης.

Ένα μειονέκτημα που έχει αυτή η μέθοδος είναι πως δε λαμβάνει καθόλου υπ' όψιν τα δείγματα της επικρατούσας κλάσης. Έτσι, σε περίπτωση που οι κλάσεις επικαλύπτονται, τα συνθετικά δείγματα μπορεί να συμπίπτουν ή να είναι υπερβολικά κοντά με κάποια από τα δείγματα της επικρατούσας κλάσης. ***[J]***

Στην Python η συγκεκριμένη μέθοδος υλοποιείται με την κλάση **SMOTE**.



Εικόνα 3.5: Παράδειγμα υπερ-δειγματοληψίας με τη μέθοδο SMOTE.

3) **SMOTE Οριογραμμής (Borderline SMOTE)**

Μια βελτιωμένη εκδοχή της μεθόδου SMOTE είναι η SMOTE Οριογραμμής που περιλαμβάνει την επιλογή εκείνων των δειγμάτων της μικρότερης κλάσης που έχουν κατηγοριοποιηθεί λανθασμένα, για παράδειγμα με ένα μοντέλο όπως του k-πλησιέστερου γείτονα. Στη συνέχεια, μπορούμε να αυξήσουμε με υπερ-δειγματοληψία μόνο αυτά τα δείγματα που έδωσαν λάθος εκτίμηση.

Γενικά, τα δείγματα που εκτιμήθηκαν λάθος είναι πιο πιθανό να ανήκουν στα δείγματα της οριογραμμής ή κοντά σε αυτήν, αφού αυτές είναι περιοχές όπου προσεγγίζονται ή επικαλύπτονται οι δυο κλάσεις. Έτσι, τα δείγματα που έχουν καταχωρηθεί λάθος είναι πιο σημαντικά και σε αυτά εστιάζει η συγκεκριμένη μέθοδος για να δημιουργηθεί ένα πιο αποτελεσματικό μοντέλο κατηγοριοποίησης.

Επίσης, έχουμε την εκδοχή της SMOTE1 Οριογραμμής, όπου υπάρχει δημιουργία συνθετικών δειγμάτων με υπερ-δειγματοληψία στα δείγματα της επικρατούσας κλάσης που οδηγούν τα κοντινά τους της μικρότερης κλάσης να κατηγοριοποιηθούν λανθασμένα.

Ακόμα, έχουμε την εκδοχή της SMOTE2 Οριογραμμής, όπου δημιουργεί συνθετικά παραδείγματα αποκλειστικά και μόνο για τα δείγματα της μικρότερης κλάσης που βρίσκονται πάνω στην οριογραμμή. Το SMOTE2 Οριογραμμής, εκτός από τα παραπάνω δείγματα, δημιουργεί συνθετικά δείγματα και από τα δείγματα του πλησιέστερου γείτονα που ανήκει στην επικρατούσα κλάση και έχει κατηγοριοποιηθεί σωστά. ***[K]***

Στην Python η συγκεκριμένη μέθοδος υλοποιείται με την κλάση **BorderlineSMOTE**.

4) **Προσαρμοστική συνθετική δειγματοληψία (Adaptive Synthetic Sampling, ADASYN)**

Μια άλλη εκδοχή της μεθόδου SMOTE αποτελεί η προσαρμοστική συνθετική δειγματοληψία, ADASYN. Η μέθοδος αυτή δημιουργεί συνθετικά δείγματα έχοντας, σε κάθε σημείο του χώρου, ως κριτήριο το πλήθος των συνθετικών δειγμάτων να είναι αντιστρόφως ανάλογο με την πυκνότητα των δειγμάτων της μικρότερης κλάσης. Δηλαδή, δημιουργεί περισσότερα συνθετικά δείγματα σε περιοχές του χώρου χαρακτηριστικών όπου η πυκνότητα των δειγμάτων της μικρότερης κλάσης είναι χαμηλή και λιγότερα ή καθόλου εκεί όπου η παρουσία τους είναι υψηλή. Ουσιαστικά, τα δείγματα της μικρότερης κλάσης σταθμίζονται ανάλογα με την πυκνότητα τους και όπου υπάρχει χαμηλή πυκνότητα η δημιουργία συνθετικών δειγμάτων είναι αυξημένη.

Το ADASYN έχει ως λογική να προσαρμόζει τη δημιουργία συνθετικών δειγμάτων ανάλογα με την κατανομή τους, αφού το κριτήριο είναι η κατανομή πυκνότητας. Έτσι, παράγει περισσότερα συνθετικά δείγματα της μικρότερης κλάσης στα σημεία που είναι πιο δύσκολο να εκπαιδευτούν σε σύγκριση με εκείνα που έχουν αρκετά δείγματα για να εκπαιδευτούν εύκολα.

Στη συγκεκριμένη μέθοδο παρατηρούμε πως στα δείγματα, όπου οι κλάσεις επικαλύπτονται, έχουμε τη μεγαλύτερη δημιουργία συνθετικών δειγμάτων. Ένα πρόβλημα που μπορεί να υπάρξει είναι με τις ακραίες τιμές, αφού για αυτές η μέθοδος ADASYN θεωρεί ότι

είναι σημεία με χαμηλή πυκνότητα και λανθασμένα δημιουργεί πολλά νέα δείγματα σε αυτά τα σημεία. Αυτό επηρεάζει αρνητικά την αποτελεσματικότητα του αλγόριθμου, οπότε χρειάζεται ιδιαίτερη προσοχή να αφαιρούνται οι ακραίες τιμές πριν τη χρήση της συγκεκριμένης μεθόδου υπερ-δειγματοληψίας. ***[L]***

Στην Python η συγκεκριμένη μέθοδος υλοποιείται με την κλάση **ADASYN**.

5) **Ενισχυμένο SMOTE (SMOTEboost)**

Η μέθοδος SMOTEboost περιλαμβάνει την έννοια της **ενίσχυσης** (**boosting**) στη μηχανική μάθηση. Όπως αναφέραμε και στο κεφάλαιο 2.5, με την έννοια του boosting ορίζουμε το συνδυασμό πολλαπλών ασθενών μοντέλων που είναι διαδοχικά σε σειρά, ώστε να δημιουργηθεί ένα ισχυρό μοντέλο.

Η μέθοδος SMOTE, όπως προαναφέραμε, αφορά τη βελτίωση της πρόβλεψης των δειγμάτων της μικρότερης κλάσης. Η χρήση του boosting γίνεται για να μη κοστίσει αυτή η παρέμβαση σε **ακρίβεια** (**accuracy**) σε ολόκληρο το σύνολο δεδομένων. Γενικά, βελτιώνει τη συνολική ακρίβεια του συνόλου δεδομένων εστιάζοντας στις αμφισβητούμενες περιπτώσεις της μικρότερης κλάσης.

Ο στόχος είναι να μειωθεί η μεροληψία προς την επικρατούσα κλάση που οφείλεται στην ανισορροπία του συνόλου δεδομένων, οπότε ενώ η διαδικασία της ενίσχυσης δίνει ίσα βάρη σε όλα τα λανθασμένα δείγματα, στη μέθοδο SMOTEboost αυξάνουμε τα βάρη στα δείγματα της μικρότερης κλάσης. Η εισαγωγή της μεθόδου SMOTE σε κάθε επανάληψη της διαδικασίας της ενίσχυσης επιτρέπει στο μοντέλο εκπαίδευσης να δοκιμάσει περισσότερες περιπτώσεις δειγμάτων της μικρότερης κλάσης.

6) **Ενίσχυση της βαθμολογημένης υπερ-δειγματοληψίας της μικρότερης κλάσης (Ranked Minority Oversampling in Boosting, RAMOBoost)**

Η μέθοδος RAMOBoost έχει ως στόχο να μειώσει τη μεροληψία που προκαλείται λόγω της ανισορροπίας των κλάσεων και να προσαρμόσει την εκπαίδευση του μοντέλου με βάση την κατανομή. Αυτό επιτυγχάνεται σε δυο σκέλη. Το πρώτο σκέλος είναι μια διαδικασία ρύθμισης ενός προσαρμοστικού βάρους που υπάρχει στη

μέθοδο RAMOBoost και μετατοπίζει την οριογραμμή ανάμεσα στις δυο κλάσεις προς δείγματα που είναι δύσκολα στην εκμάθηση και ανήκουν όχι μόνο στη μικρότερη, αλλά και στην επικρατούσα κλάση.

Το δεύτερο σκέλος αποτελεί μια κατανομή πιθανότητας βαθμολογημένης δειγματοληψίας που χρησιμοποιείται για να δημιουργεί συνθετικά δείγματα της μικρότερης κλάσης, προκειμένου να εξισορροπήσει μια ασύμμετρη κατανομή.

Ωστόσο, σε αντίθεση με τη μέθοδο SMOTE, η οποία εστιάζει στα δείγματα της μικρότερης κλάσης χωρίς κανένα περαιτέρω κριτήριο, το RAMOBoost αξιολογεί τη πιθανή συνεισφορά στην εκπαίδευση κάθε δείγματος της μικρότερης κλάσης και καθορίζει ανάλογα σε κάθε επανάληψη το βάρος του. Αυτό επιτυγχάνεται με τον υπολογισμό της απόστασης κάθε δείγματος της μικρότερης κλάσης από το σύνολο των πλησιέστερων γειτόνων του για να προσδιοριστεί τελικά, πόσο σημαντικό είναι το συγκεκριμένο δείγμα για τη διαδικασία της εκπαίδευσης. ***[M]***

7) **Υπερ-δειγματοληψία βασισμένη στην απόσταση Mahalanobis (Mahalanobis Distance-based Over-sampling, MDO)**

Η συγκεκριμένη μέθοδος υπερ-δειγματοληψίας, MDO, είναι βασισμένη στην απόσταση Mahalanobis. Στα μαθηματικά, η απόσταση Mahalanobis αφορά την απόσταση μεταξύ ενός σημείου και μίας κατανομής. Είναι μια επέκταση της Ευκλείδειας απόστασης. Το μειονέκτημα της Ευκλείδειας απόστασης είναι ότι μετράει αποκλειστικά την απόσταση μεταξύ δυο σημείων, χωρίς να λαμβάνει υπ’ όψιν την περίπτωση της κατανομής, όπως συμβαίνει για σημεία που βρίσκονται σε ένα σύνολο δεδομένων. Αυτό διορθώνεται με τη χρήση της απόστασης Mahalanobis ως μετρική.

Η μέθοδος MDO δημιουργεί συνθετικά δείγματα που έχουν την ίδια απόσταση Mahalanobis από τη θεωρούμενη ως μέση τιμή της μικρότερης κλάσης, λαμβάνοντας υπ’ όψιν όλα τα δείγματα της. Διατηρώντας τη δομή της συνδιακύμανσης της μικρότερης κλάσης και δημιουργώντας, με έξυπνες μεθόδους, συνθετικά δείγματα κατά μήκος των ισοϋψών καμπυλών των πιθανοτήτων, τα καινούργια δείγματα της μικρότερης κλάσης εξυπηρετούν καλύτερα τα μοντέλα εκμάθησης. Επιπλέον, το MDO μειώνει τον κίνδυνο επικάλυψης μεταξύ διαφορετικών κλάσεων, που είναι ένα σημαντικό ζήτημα όταν υπάρχουν πολλαπλές κλάσεις. ***[N]***

8) **Μέθοδος υπερ-δειγματοληψίας με βάρη στη μικρότερη κλάση με βάση την επικρατούσα κλάση (Majority Weighted Minority Oversampling Technique, MWMOTE)**

Η μέθοδος MWMOTE ελέγχει και σταθμίζει με βάρη τα σημαντικά δείγματα της μικρότερης κλάσης με βάση τα πλησιέστερα τους δείγματα της επικρατούσας κλάσης και μετά, τα χρησιμοποιεί για να εφαρμόσει υπερ-δειγματοληψία δημιουργώντας συνθετικά δείγματα. Για να επιτευχθεί αυτή η διαδικασία, το MWMOTE χρησιμοποιεί πληροφορίες και από τις δυο κλάσεις του συνόλου δεδομένων.

Αρχικά, ταυτοποιεί τα δείγματα της μικρότερης κλάσης που δυσκολεύουν την εκμάθηση και τους εκχωρεί βάρη αναλόγως της σημαντικότητας τους με βάση πληροφορίες για την απόσταση τους από το πλησιέστερο δείγμα της επικρατούσας κλάσης.

Έπειτα, το MWMOTE προσδιορίζει τις συστάδες στη μικρότερη κλάση και χρησιμοποιεί σταθμισμένα δείγματα της μικρότερης κλάσης για τη δημιουργία συνθετικών δειγμάτων εντός των συστάδων. Αυτό γίνεται για να διασφαλιστεί ότι τα παραγόμενα δείγματα βρίσκονται πάντα μέσα σε κάποια συστάδα της μικρότερης κλάσης και δεν επικαλύπτονται με τις περιοχές της επικρατούσας κλάσης. ***[O]***

3.1.3 Μεικτές μέθοδοι υπο-δειγματοληψίας και υπερ-δειγματοληψίας

Γενικότερα, τόσο η υπο-δειγματοληψία, όσο και η υπερ-δειγματοληψία παρουσιάζουν μειονεκτήματα όταν εφαρμόζονται αυτόνομα. Για παράδειγμα, με την υπο-δειγματοληψία έχουμε απώλεια σημαντικής πληροφορίας που παρέχουν τα δείγματα της επικρατούσας κλάσης που διαγράφονται, ενώ στην υπερ-δειγματοληψία μπορεί να συναντήσουμε το φαινόμενο της **υπερπροσαρμογής** (**overfitting**), εξαιτίας των δειγμάτων της μικρότερης κλάσης που δημιουργούνται και στηρίζονται στα λίγα πραγματικά δείγματα. Για αυτό το λόγο, ο συνδυασμός των δυο αυτών γενικότερων μεθόδων είναι πολλές φορές προτιμότερος.

Στη συνέχεια, θα αναλύσουμε ορισμένες σημαντικές μεικτές μεθόδους. ***[21]***

1) **Συνδυασμός τυχαίας υπο-δειγματοληψίας (RUS) και τυχαίας υπερ-δειγματοληψίας (ROS)**

Η μεικτή αυτή μέθοδος είναι η πιο απλή και στηρίζεται στην τυχαία επιλογή τόσο για την υπο-δειγματοληψία, όσο και για την υπερ-δειγματοληψία. Για παράδειγμα, μπορούμε αρχικά να αυξήσουμε τη μικρότερη κλάση με υπερ-δειγματοληψία στο 20% της επικρατούσας κλάσης και στη συνέχεια, να μειώσουμε την επικρατούσα κλάση να γίνει μόνο διπλάσια του πλήθους της καινούργιας μικρότερης κλάσης. Έτσι, σχεδόν εξισορροπούμε τις δυο κλάσεις για να μπορέσουν τα μοντέλα εξόρυξης να λειτουργήσουν αποτελεσματικότερα.

2) **Συνδυασμός SMOTE και τυχαία υπο-δειγματοληψία**

Επειδή η μέθοδος SMOTE είναι η πιο γενική από όσες ανήκουν στην υπερ-δειγματοληψία, θα μπορούσε να χρησιμοποιηθεί αρχικά για να αυξηθεί με συνθετικά δείγματα η μικρότερη κλάση και μετά, με τυχαία υπο-δειγματοληψία να μειώσουμε κατά ένα ποσοστό και την επικρατούσα κλάση. Με αυτό τον τρόπο δεν επιλέγουμε δυο τυχαίες μεθόδους, αλλά χρησιμοποιούμε και μια μέθοδο που διαθέτει μια πιο έξυπνη τεχνική.

3) **Συνδυασμός SMOTE και συνδέσμων Tomek**

Αρχικά, εφαρμόζουμε τη μέθοδο SMOTE για να πάρουμε την υπερ-δειγματοληψία στη μικρότερη κλάση και έπειτα, με τη μέθοδο συνδέσμων Tomek τα δείγματα από την επικρατούσα κλάση που βρίσκονται πλησιέστερα σε δείγματα της μικρότερης κλάσης αφαιρούνται και έχουμε μια υπο-δειγματοληψία που μας βοηθάει και στον καθορισμό της οριογραμμής μεταξύ των δυο κλάσεων. Να υπενθυμίσουμε πως συνήθως, οι σύνδεσμοι Tomek δεν αφαιρούν μεγάλο ποσοστό δειγμάτων της επικρατούσας κλάσης, με αποτέλεσμα να μην υπάρχει εξισορρόπηση των κλάσεων. Όμως, χάρις στην αρχική χρήση της μεθόδου SMOTE επιτυγχάνεται από την αρχή μια σημαντική προσέγγιση του πλήθους των δειγμάτων των δυο κλάσεων.

Στην Python η συγκεκριμένη μέθοδος υλοποιείται με την κλάση **SMOTETomek**.

4) **Συνδυασμός SMOTE και μεθόδου επεξεργασμένων πλησιέστερων γειτόνων (ENN)**

Πρώτα, εφαρμόζουμε τη μέθοδο SMOTE για να πάρουμε την υπερδειγματοληψία στη μικρότερη κλάση και μετά, με τη μέθοδο ENN ελέγχουμε τους πλησιέστερους γείτονες για να δούμε αν θα αφαιρέσουμε το δείγμα. Όπως και στην προηγούμενη μεικτή μέθοδο, εδώ υπεύθυνη για την επί μέρους αριθμητική εξισορρόπηση του μοντέλου είναι η μέθοδος SMOTE.

Στην Python η συγκεκριμένη μέθοδος υλοποιείται με την κλάση **SMOTEENN**.

4 Εφαρμογή ανάλυσης δεδομένων σε μη ισορροπημένα σύνολα δεδομένων

Στο παρόν κεφάλαιο, θα υλοποιήσουμε μια εφαρμογή πάνω σε ένα πραγματικό σύνολο δεδομένων. Το συγκεκριμένο σύνολο δεδομένων είναι αναρτημένο ως διαγωνισμός στην ιστοσελίδα kaggle.com με τίτλο "**HR Analytics: Job Change of Data Scientists**". ***[22]***

4.1 Χαρακτηριστικά συνόλου δεδομένων και επεξήγηση τους

Τα δεδομένα προέρχονται από μια εταιρεία που έχει ως αντικείμενο την Επιστήμη των Δεδομένων και αφορα άτομα υποψήφια να προσληφθούν από την εταιρεία ως υπάλληλοι. Οι υποψήφιοι σε πρώτη φάση έχουν περάσει με επιτυχία κάποια τεστ που διεξάγει η εταιρεία και στη συνέχεια, παρακολουθούν μια αναλυτική εκπαίδευση.

Ο σκοπός αυτής της ανάλυσης δεδομένων είναι η αξιολόγηση των υποψηφίων, όσον αφορά την πιθανότητα, μετά το τέλος της εκπαίδευσης, να δεχτούν να προσληφθούν από την εταιρεία, και όχι να χρησιμοποιήσουν απλά τη δωρεάν εκπαίδευση που τους παρέχεται, προκειμένου να βρουν κάποια άλλη δουλειά. Τα χαρακτηριστικά που περιλαμβάνονται στο σύνολο δεδομένων αφορούν προσωπικά στοιχεία των υποψήφιων. Μια πετυχημένη ανάλυση δεδομένων θα έχει ένα προφανές όφελος για την ίδια την εταιρεία, αφού θα μειωθούν τα έξοδα και ο χρόνος που θα σπαταληθούν σε υποψήφιους που τελικά δε θα παραμείνουν στην εταιρεία.

Ο διαγωνισμός του kaggle, μας παρέχει δύο σύνολα δεδομένων, train και test, αλλά στην εργασία μας θα εξάγουμε συμπεράσματα αποκλειστικά από το εκπαιδευτικό σύνολο δεδομένων (train set), αφού μόνο για αυτό έχουμε τιμές για το ζητούμενο στόχο της ανάλυσης.

Γενικότερα, τα χαρακτηριστικά του συνόλου δεδομένων διαθέτουν τα εξής στοιχεία:

1) Το σύνολο δεδομένων είναι μη εξισορροπημένο σε ένα ποσοστό 75% με 25%, ανάμεσα στις δυο κλάσεις, δηλαδή μια ήπια ανισορροπία μεταξύ τους. Πάνω σε αυτά, θα γίνει η έρευνα για την επιτυχία των μεθόδων εξισορρόπησης των συνόλων των δεδομένων.

2) Τα περισσότερα στοιχεία είναι κατηγορικά (ονομαστικά, διατεταγμένα ή δυαδικά) και ορισμένα εξ αυτών έχουν μεγάλη ποικιλία τιμών.

Τα χαρακτηριστικά των υποψηφίων που μας παρέχονται είναι τα παρακάτω:

1) **enrollee_id** : Μοναδικό id για κάθε εγγραφή.

2) **city** : Κωδικός πόλης στην οποία κατοικεί ο υποψήφιος.

3) **city_ development _index** : Δείκτης ανάπτυξης πόλης. Πρόκειται για ένα μαθηματικό τύπο που περιλαμβάνει ως μεταβλητές τις υποδομές, τα απόβλητα, την υγεία, την εκπαίδευση και την παραγωγικότητα της κάθε πόλης. Η τιμή του είναι δεκαδικός αριθμός και κυμαίνεται από 0 έως 1.

4) **gender** : Φύλο.

5) **relevent_experience** : Εμπειρία σχετική με το αντικείμενο. Οι τιμές είναι 1 αν υπάρχει εμπειρία, αλλιώς είναι 0.

6) **enrolled_university** : Κατηγοριοποίηση εκπαίδευσης, όσον αφορά την εβδομαδιαία χρονική διάρκεια των σπουδών. Δηλαδή πλήρους ωραρίου, μειωμένου ωραρίου ή χωρίς εκπαίδευση.

7) **education_level** : Επίπεδο ακαδημαϊκών σπουδών.

8) **major_discipline** : Τομέας σπουδών.

9) **experience** : Έτη εργασιακής εμπειρίας.

10) **company_size** : Αριθμός υπαλλήλων παρούσας εργασίας.

11) **company_type** : Είδος εταιρείας στην οποία εργάζεται.

12) **last*new*job** : Απόσταση ετών μεταξύ παρούσας και προηγούμενης εργασίας.

13) **training_hours** : Συμπληρωμένες ώρες εκπαίδευσης.

14) **target** : 0 αν δεν επέλεξε να εργαστεί αλλού, αλλιώς 1.

Ως προς το ρόλο του κάθε χαρακτηριστικού, το σύνολο των δεδομένων αποτελείται από 12 ανεξάρτητες μεταβλητές, μια μεταβλητή που δεν έχει καμία επίδραση, αφού χρησιμοποιείται ως τιμή μοναδικής εγγραφής (ID) και πρόκειται για την **enrollee_id** και την εξαρτημένη μεταβλητή στόχο, που είναι η **target**.

4.2 Διερευνητική Ανάλυση Δεδομένων, ΔΑΔ (Exploratory Data Analysis, EDA)

Όπως γνωρίζουμε, η Διερευνητική Ανάλυση Δεδομένων (ΔΑΔ) μας δίνει πολύτιμες πληροφορίες για τα χαρακτηριστικά. Κυρίως αφορά μεθόδους και τεχνικές που οπτικοποιούν τα δεδομένα μέσω γραφημάτων. Στην εφαρμογή που μελετάμε, θα αναφέρουμε τα σημαντικότερα αποτελέσματα στα οποία καταλήξαμε.

Αρχικά, πρέπει να τονίσουμε την ύπαρξη ελλειπουσών τιμών. Παρατηρούμε ότι από τις 12 ανεξάρτητες μεταβλητές οι 8 έχουν κενές τιμές και μάλιστα οι 7 από αυτές κυμαίνονται σε ένα σημαντικό ποσοστό από 20% μέχρι 32%.

```
Rate of Missing Values per column:

 company_type              0.3205
company_size               0.3099
gender                     0.2353
major_discipline           0.1468
education_level            0.0240
last_new_job               0.0221
enrolled_university        0.0201
experience                 0.0034
enrollee_id                0.0000
city                       0.0000
city_development_index     0.0000
relevent_experience        0.0000
training_hours             0.0000
target                     0.0000
dtype: float64
```

Εικόνα 4.1: Ποσοστά ελλειπουσών τιμών ανά χαρακτηριστικό

Ένα άλλο στοιχείο είναι ότι για κάθε συγκεκριμένη τιμή της μεταβλητής **city** αντιστοιχεί πάντα μια συγκεκριμένη δεκαδική τιμή της μεταβλητής **city_development_index.** Αυτό πρέπει να ληφθεί υπ’ όψιν, αφού επηρεάζει τη μεροληψία του συνόλου δεδομένων και να αφαιρέσουμε τη μεταβλητή **city** σε επόμενο στάδιο.

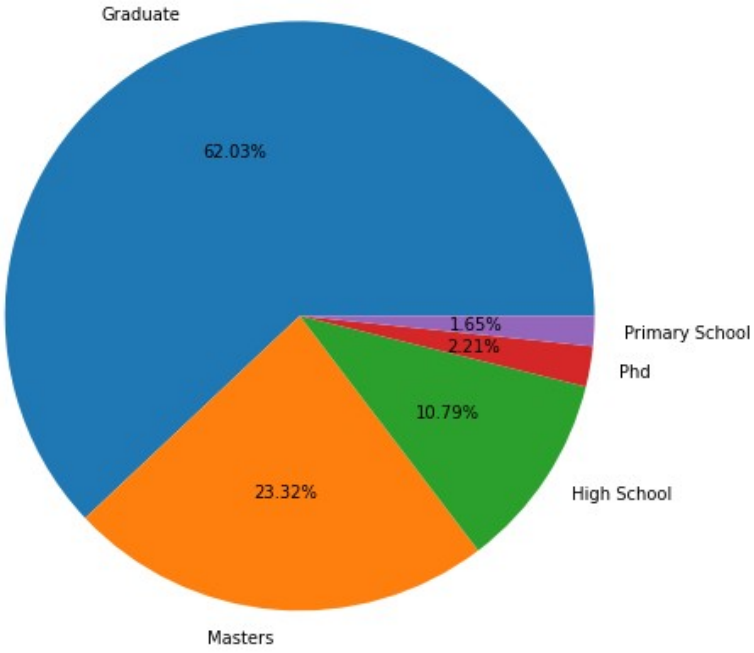
Το **city_development_index** είναι ένας δείκτης που όλες οι τιμές του κυμαίνονται πάνω από το 0,448 και η διάμεσος του είναι μεγαλύτερη από 0,9.

```
count    19158.000000
mean         0.828848
std          0.123362
min          0.448000
25%          0.740000
50%          0.903000
75%          0.920000
max          0.949000
Name: city_development_index, dtype: float64
```

Εικόνα 4.2: Βασικά στατιστικά μεγέθη του city_development_index

Για το χαρακτηριστικό **gender** παρατηρούμε πως η συντριπτική πλειοψηφία των υποψήφιων είναι άνδρες σε ποσοστό άνω του 90%, ενώ από το χαρακτηριστικό **relevent_experience** προκύπτει ότι σε ποσοστό άνω του 70%, υπάρχει προηγούμενη σχετική εμπειρία στο αντικείμενο.

Η μεταβλητή **enrolled_university** μας δείχνει πως σχεδόν τα τρία τέταρτα των υποψήφιων υπαλλήλων δεν έχει κάνει ακαδημαϊκά μαθήματα σχετικά με την Επιστήμη των Δεδομένων, ενώ το ένα πέμπτο έχει κάνει full time μαθήματα.

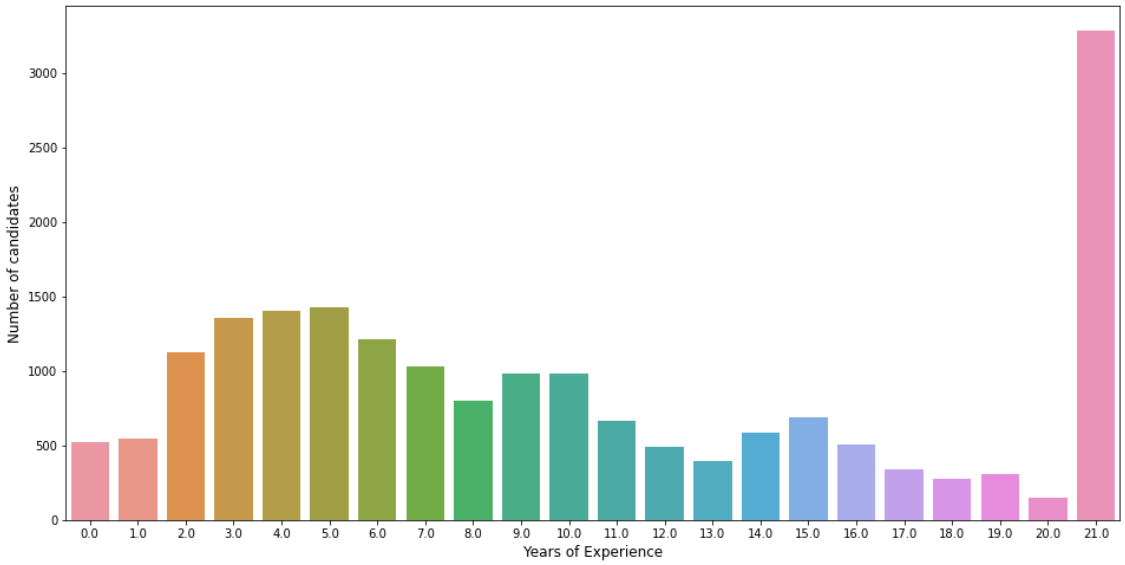


Εικόνα 4.3: Ποσοστά σε πίτα μεταβλητής education_level

Το χαρακτηριστικό **education_level** μας δείχνει πως τα άτομα είναι κυρίως απόφοιτοι πανεπιστημίου σε ποσοστό άνω του 60%, ενώ κάτι λιγότερο από το ένα τέταρτο διαθέτει μεταπτυχιακό τίτλο σπουδών.

Όπως προκύπτει από τη μεταβλητή **major_discipline**, ο τομέας σπουδών σε ένα ποσοστό σχεδόν 90% είναι οι θετικές επιστήμες ή όπως αναγράφεται στα αγγλικά STEM, δηλαδή “Επιστήμη, Τεχνολογία, Μηχανική, Μαθηματικά”.

Η διάμεσος της μεταβλητής **experience** κυμαίνεται στα 9 χρόνια. Πάντως, δεν μπορούμε να εκτιμήσουμε τη μέση τιμή και κάποια ακόμα στατιστικά στοιχεία με ακρίβεια, γιατί τα χρόνια εργασιακής εμπειρίας που υπερβαίνουν τα 20 έτη δεν αναφέρονται αναλυτικά, ενώ αποτελούν το 17,2% του δείγματος που είναι αρκετά υψηλό ποσοστό.

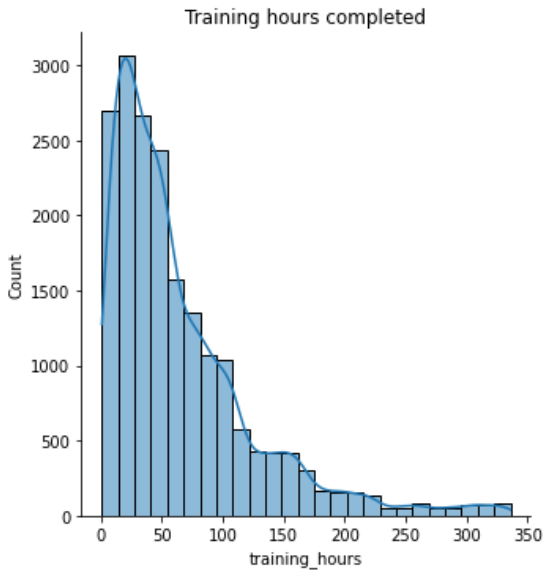


Εικόνα 4.4: Ραβδόγραμμα συχνοτήτων του χαρακτηριστικού experience

Η μεταβλητή **company_size** μας δείχνει πως το μέγεθος της εταιρείας της παρούσας εργασίας των υποψήφιων υπαλλήλων είναι κυρίως στα γκρουπ 50-100 ή 100-500 άτομα που αθροιστικά καλύπτουν το 43% των συνολικών δηλώσεων τους, ενώ ο τύπος της παρούσας εταιρείας για τα τρία τέταρτα των ατόμων είναι Εταιρεία Περιορισμένης Ευθύνης, ΕΠΕ (Pvt Ltd), όπως φαίνεται και από το χαρακτηριστικό **company_type**.

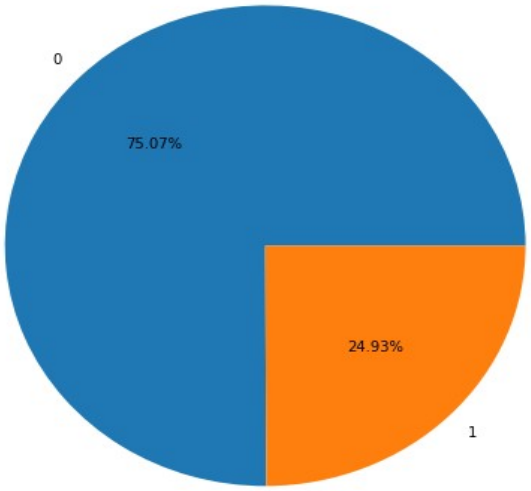
Από τη μεταβλητή **last*new*job** προκύπτει πως σε ποσοστό 43% οι υποψήφιοι χρειάστηκαν μόλις ένα έτος από την τελευταία φορά που άλλαξαν εργασία για να βρουν νέα δουλειά.

Τελευταία από τις ανεξάρτητες μεταβλητές, η **training_hours** δείχνει πως υπάρχει μια έντονη ασυμμετρία αριστερά και έχει ως διάμεσο τις 47 ώρες, ενώ η μέση τιμή των ωρών εκπαίδευσης είναι περίπου 65 ώρες.



Εικόνα 4.5: Ασύμμετρη κατανομή της μεταβλητής training_hours

Η εξαρτημένη μεταβλητή **target** που αποτελεί το αντικείμενο της έρευνας, δίνει 75% ποσοστό των ατόμων που μετά το τέλος της εκπαίδευσης ξεκινάνε να εργάζονται στην εν λόγω εταιρεία.



Εικόνα 4.6: Ποσοστά σε πίτα εξαρτημένης μεταβλητής target

4.3 Μηχανική χαρακτηριστικών (Feature Engineering)

Η μηχανική χαρακτηριστικών όπως έχουμε αναφέρει θα μας βοηθήσει στην προετοιμασία των δεδομένων, προκειμένου να χρησιμοποιηθούν από τους αλγόριθμους μηχανικής μάθησης, καθώς επίσης και για να μετασχηματιστούν τα δεδομένα, ούτως ώστε να βελτιωθεί η απόδοση των μοντέλων.

Ήδη, κατά τη ΔΑΔ έχουμε προβεί σε αλλαγές των πεδίων των χαρακτηριστικών. Έτσι, έχουμε επιλέξει για τις ποσοτικές μεταβλητές που είχαν τιμές που ήταν διαστήματα περιοχών, να αντιστοιχήσουμε αυτά τα διαστήματα σε κάποιους ενδεικτικούς αριθμούς ή διάστημα αριθμών που θα παραπέμπουν σε ταξινομημένες τιμές. Αυτός ο μετασχηματισμός έγινε στις μεταβλητές **experience**, **company_size** και **last*new*job**.

Το επόμενο βήμα είναι η επιλογή των πέντε κατηγορικών μεταβλητών **city**, **gender**, **relevent_experience**, **major_discipline** και **company_type** και η μετατροπή των τιμών τους σε αριθμητικές, κάνοντας χρήση της συνάρτησης της Python LabelEncoder() και σε πρώτη φάση, χωρίς να αλλάξουμε τις Null τιμές τους.

	city	gender	relevent_experience	major_discipline	company_type
0	city_103	Male	Has relevent experience	STEM	NaN
1	city_40	Male	No relevent experience	STEM	Pvt Ltd
2	city_21	NaN	No relevent experience	STEM	NaN
3	city_115	NaN	No relevent experience	Business Degree	Pvt Ltd
4	city_162	Male	Has relevent experience	STEM	Funded Startup

	city	gender	relevent_experience	major_discipline	company_type
0	5.0	1.0	0.0	5.0	NaN
1	77.0	1.0	1.0	5.0	5.0
2	64.0	NaN	1.0	5.0	NaN
3	14.0	NaN	1.0	1.0	5.0
4	50.0	1.0	0.0	5.0	1.0

Εικόνα 4.7: Απεικόνιση των δεδομένων των κατηγορικών μεταβλητών πριν και μετά τη χρήση του LabelEncoder()

Στη συνέχεια, συμπληρώνουμε στις παραπάνω κατηγορικές μεταβλητές τις ελλείπουσες τιμές με τη μέθοδο των k-πλησιέστερων γειτόνων (**KNN imputation**). Έτσι, αντιμετωπίζουμε το πρόβλημα που έχουν πολλοί αλγόριθμοι που δε δέχονται κενές τιμές.

4.4 Μέθοδοι εξισορρόπησης συνόλου δεδομένων

Όπως έχουμε αναφέρει, το αντικείμενο με το οποίο καταπιανόμαστε είναι οι τεχνικές διαχείρισης των μη ισορροπημένων συνόλων. Επομένως, θα εστιάσουμε κυρίως στη διαδικασία υλοποίησης των συγκεκριμένων μεθόδων. Για να τις ονομάσουμε θα χρησιμοποιήσουμε το αντίστοιχο όνομα κλάσης της Python.

Οι μέθοδοι που επιλέχθηκαν είναι 5 υπο-δειγματοληψίας, 4 υπερ-δειγματοληψίας και 2 μεικτές. Δηλαδή, η συνολική σύγκριση θα αφορά 11 διαφορετικές μεθόδους που με διάφορους τρόπους που έχουν αναφερθεί σε προηγούμενα κεφάλαια, αλλάζουν το αρχικό σύνολο δεδομένων. Άρα, συνολικά θα ελέγξουμε τα αποτελέσματα για 12 **μεθόδους**, αφού θα εξάγουμε συμπεράσματα και για το αρχικό σύνολο δεδομένων.

Στον παρακάτω πίνακα θα δούμε τις μεταβολές που προκαλεί η κάθε μέθοδος στο πλήθος των εγγραφών του συνόλου δεδομένων ανά κλάσεις και αθροιστικά.

Μέθοδος	Αριθμός εγγραφών		Σύνολο εγγραφών	Ποσοστό εγγραφών		Μεταβολή εγγραφών
	Κλάση 0	Κλάση 1		Κλάση 0	Κλάση 1	
Αρχικό σύνολο	14381	4777	19158	75,07%	24,93%	
Μέθοδοι Υπο-δειγματοληψίας						
RandomUnderSampler	4777	4777	9554	50,00%	50,00%	-50,13%
NearMiss	4777	4777	9554	50,00%	50,00%	-50,13%
CondensedNearestNeighbour	5433	4777	10210	53,21%	46,79%	-46,71%
OneSidedSelection	12999	4777	17776	73,13%	26,87%	-7,21%
NeighbourhoodCleaningRule	7878	4777	12655	62,25%	37,75%	-33,94%
Μέθοδοι Υπερ-δειγματοληψίας						
RandomOverSampler	14381	14381	28762	50,00%	50,00%	50,13%
SMOTE	14381	14381	28762	50,00%	50,00%	50,13%
BorderlineSMOTE	14381	14381	28762	50,00%	50,00%	50,13%
ADASYN	14890	14381	29271	50,87%	49,13%	52,79%
Μεικτές Μέθοδοι						
SMOTETomek	14002	14002	28004	50,00%	50,00%	46,17%
SMOTEENN	10577	6155	16732	63,21%	36,79%	-12,66%

Πίνακας 4.1: Αριθμός εγγραφών και ποσοστά μετά τις μεθόδους υπο-δειγματοληψίας, υπερ-δειγματοληψίας και τις μεικτές μεθόδους

4.5 Αλγόριθμοι μηχανικής μάθησης

Μετά τη χρησιμοποίηση των τεχνικών εξισορρόπησης των δεδομένων, προχωράμε σε **διασταύρωση επικύρωσης** (**cross-validation**) για κάθε ένα από τα 12 σύνολα που έχουν δημιουργηθεί, σε ποσοστό 80% για το μέγεθος του συνόλου εκπαίδευσης και 20% για το σύνολο του τεστ.

Όσον αφορά τους αλγόριθμους μηχανικής μάθησης που χρησιμοποιήσαμε είναι δυο ευρέως χρησιμοποιούμενοι αλγόριθμοι. Ο πρώτος προέρχεται από το σύνολο των μεθόδων δειγματοληψίας bootstrap και είναι τα **τυχαία δάση** (**Random Forest**), ενώ ο δεύτερος είναι τα **κλιμακούμενα ενισχυμένα δέντρα αποφάσεων** (**Gradient Boosting**). Για κάθε ένα από τα παραπάνω, δοκιμάσαμε διαφορετικές παραμέτρους, ώστε να πετύχουμε τη μεγαλύτερη δυνατή απόδοση.

1) **Αλγόριθμος Random Forest**

Χρησιμοποιήσαμε διαφορετικές τιμές για 5 παραμέτρους του συγκεκριμένου αλγόριθμου, οι οποίες είναι:

i) **n_estimators**: Ο αριθμός των δέντρων στο δάσος.

ii) **max_features** : Ο αριθμός των χαρακτηριστικών που λαμβάνονται υπ’ όψιν κατά τη διαδικασία των διαχωρισμών σε κάθε κόμβο του δέντρου.

iii) **max_depth** : Ο μέγιστος αριθμός κόμβων σε βάθος για κάθε δέντρο.

iv) **min_samples_split** : Το μικρότερο πλήθος εγγραφών που απαιτούνται, ώστε ο αλγόριθμος να διαχωρίσει έναν εσωτερικό κόμβο.

v) **min_samples_leaf** : Ο μικρότερος αριθμός εγγραφών που απαιτούνται να υπάρχουν σε ένα φύλλο. Ένας κόμβος διαχωρίζεται μόνο αν και στα δυο κλαδιά που δημιουργούνται περιέχεται ο συγκεκριμένος αριθμός εγγραφών.

Οι τελικές τιμές των παραμέτρων που δοκιμάζουμε στην εφαρμογή μας είναι:

i) **n_estimators**: [50, 70, 90, 110, 130, 150, 170]

ii) **max_features** : ['sqrt', 'log2']

iii) **max_depth** : [null, 5, 10, 15 ,20, 25, 30]

iv) **min_samples_split** : [2, 5, 10]

v) **min_samples_leaf** : [1, 2, 4]

Ο συνολικός αριθμός δοκιμών που έχουμε από τις παραπάνω παραμέτρους είναι 7*2*7*3*3 = **882 διαφορετικοί συνδυασμοί**. Για να τρέξουμε όλους αυτούς τους συνδυασμούς, παράλληλα με τις 11 **διαφορετικές μεθόδους εξισορρόπησης του συνόλου δεδομένων και τα αρχικά δεδομένα**, θα ήταν ένα ιδιαιτέρως χρονοβόρο εγχείρημα. Για αυτό το λόγο χρησιμοποιήσαμε την κλάση της Python **RandomizedSearchCV** που με δειγματοληπτικό τρόπο επιλέγει κάποιους τυχαίους συνδυασμούς των τιμών των παραμέτρων και εκτελεί μόνο αυτές τις περιπτώσεις.

Στην εφαρμογή μας επιλέγουμε να εκτελέσουμε **40 συνδυασμούς** από τους 882 για κάθε μια από τις **12 μεθόδους εξισορρόπησης του συνόλου δεδομένων**. Στο τέλος, έχουμε επιλέξει να ενημερωθούμε για τον καλύτερο από τους 40 συνδυασμούς αντίστοιχα για την κάθε μέθοδο, ώστε να το χρησιμοποιήσουμε στην τελική αξιολόγηση της αποδοτικότητας της μεθόδου.

2) **Αλγόριθμος Gradient Boosting**

Για να μπορέσουμε να έχουμε μια πιο αντικειμενική σύγκριση μεταξύ των δυο αλγορίθμων, συντονίσαμε ακριβώς τις ίδιες παραμέτρους με τον αλγόριθμο **Random Forest** , δηλαδή τις **n_estimators**, **max_features**, **max_depth**, **min_samples_split**, και **min_samples_leaf**.

Από πλευράς τιμών που ελέγξαμε και σε αυτή την περίπτωση είχαμε ακριβώς τις ίδιες επιλογές τιμών στην κάθε παράμετρο, όπως και στον προηγούμενο αλγόριθμο.

4.6 Μετρικές αξιολόγησης

Οι μετρικές αξιολόγησης που χρησιμοποιήσαμε για να εκτιμήσουμε την απόδοση των συγκεκριμένων αλγόριθμων είναι η **Βαθμολογία F1** (**F1 Score**) και η **Καμπύλη Λειτουργικού Χαρακτηριστικού Δέκτη με την περιοχή κάτω από την καμπύλη** (**ROC-AUC,Receiver Operating Characteristic - Area Under Curve**).

Γενικότερα, για τα μη ισορροπημένα σύνολα δεδομένων η προτεινόμενη μετρική αξιολόγησης είναι η **ROC-AUC**, η οποία επηρεάζεται λιγότερο από την ανισοκατανομή των δεδομένων. Παρ' όλ' αυτά, επιλέξαμε να χρησιμοποιήσουμε και μια δεύτερη μετρική, ώστε να έχουμε ένα πιο αντικειμενικό αποτέλεσμα, οπότε χρησιμοποιήσαμε και την **F1 Score**.

Βέβαια, πρέπει να τονίσουμε πως μετά τη χρήση των μεθόδων εξισορρόπησης, το πλήθος των κλάσεων των νέων συνόλων δεδομένων είναι ομοιόμορφο ως προς την εξαρτημένη μεταβλητή, άρα οι μετρικές μεταξύ τους δεν έχουν τη μεγάλη διαφορά που θα αναμέναμε. Πάντως, όπως θα δούμε από τα αποτελέσματα, στη συντριπτική πλειοψηφία των εκτιμήσεων η μετρική **ROC-AUC** είχε καλύτερη απόδοση.

Εδώ να τονίσουμε πως για κάθε μια από τις παραπάνω αποδόσεις, ανά μοντέλο και ανά μέθοδο εξισορρόπησης, έχει επιλεγεί ο αποδοτικότερος συνδυασμός παραμέτρων από τους 40 που δοκιμάστηκαν σε κάθε περίπτωση από την εφαρμογή, χρησιμοποιώντας τη συνάρτηση **RandomizedSearchCV()** της **Python**.

Παρακάτω, παραθέτουμε τις καλύτερες αποδόσεις για κάθε συνδυασμό των παραμέτρων που εξετάζουμε.

Μέθοδος	**Αποδόσεις Αλγόριθμων**			
	Random Forest		**Gradient Boosting**	
	F1-score	**ROC-AUC**	**F1-score**	**ROC-AUC**
Αρχικό σύνολο δεδομένων	49,00%	66,03%	47,76%	65,30%
Μέθοδοι Υπο-δειγματοληψίας				
RandomUnderSampler	71,90%	73,11%	72,77%	73,12%
NearMiss	72,44%	70,46%	71,70%	70,33%
CondensedNearestNeighbour	66,81%	69,92%	68,51%	70,81%
OneSidedSelection	54,34%	68,56%	56,50%	69,88%
NeighbourhoodCleaningRule	70,88%	76,69%	71,39%	76,89%
Μέθοδοι Υπερ-δειγματοληψίας				
RandomOverSampler	90,43%	89,96%	91,04%	90,75%
SMOTE	81,59%	81,61%	83,08%	83,10%
BorderlineSMOTE	82,18%	81,96%	82,97%	82,86%
ADASYN	81,58%	81,15%	82,44%	82,09%
Μεικτές Μέθοδοι				
SMOTETomek	81,77%	81,80%	83,19%	83,24%
SMOTEENN	93,91%	91,68%	94,59%	92,61%

Πίνακας 4.2: Εκτίμηση αποδοτικότητας μοντέλων ανά μέθοδο εξισορρόπησης και ανά αλγόριθμο με χρήση δυο μετρικών

5 Συμπεράσματα

Το κεφάλαιο αυτό ολοκληρώνει τη μεταπτυχιακή εργασία, συνοψίζοντας τα συμπεράσματα που αποκομίσαμε, κυρίως από την εφαρμογή. Υπενθυμίζουμε πως στην εφαρμογή η έρευνα αποτελούνταν από τρία διαφορετικά επίπεδα. Βεβαίως, η εργασία μας ήταν επικεντρωμένη στο πρώτο επίπεδο από τα παρακάτω:

α) 11 **μέθοδοι εξισορρόπησης συνόλου δεδομένων μη ισορροπημένων μοντέλων**

β) 2 αλγόριθμοι μηχανικής μάθησης. Για κάθε αλγόριθμο εξετάζουμε **40** διαφορετικούς συνδυασμούς από τους 882 δυνατούς, οι οποίοι αφορούν επιλεγμένες τιμές σε 5 παραμέτρους τους.

γ) 2 μετρικές αξιολόγησης μοντέλων μηχανικής μάθησης

5.1 Συμπεράσματα για τις μεθόδους εξισορρόπησης των συνόλων δεδομένων

Κάνοντας μια ανακεφαλαίωση, θα πρέπει να επισημάνουμε ορισμένα ζητήματα, πριν προχωρήσουμε στα συμπεράσματα.

Η συγκεκριμένη εφαρμογή βασίζεται σε ένα ηπίως μη ισορροπημένο σύνολο δεδομένων, αφού η αναλογία των δύο κλάσεων είναι 75% προς 25%. Όπως είναι αναμενόμενο, στις μεθόδους υπο-δειγματοληψίας από το αρχικό πλήθος δεδομένων της επικρατούσας κλάσης παραμένει το ένα τρίτο των εγγραφών, ώστε να εξισορροπηθεί το πλήθος των δύο κλάσεων. Αυτό έχει ως γενικό αποτέλεσμα μια μείωση του συνόλου των εγγραφών κατά 50%. Στις μεθόδους υπερ-δειγματοληψίας έχουμε έναν τριπλασιασμό των εγγραφών της μικρότερης κλάσης που οδηγεί σε μια αύξηση των συνολικών εγγραφών κατά 50%.

Για να κάνουμε τη σύγκριση της αυξομείωσης του πλήθους των δεδομένων για ένα υποθετικό σύνολο δεδομένων με έντονο βαθμό ανισορροπίας της τάξης του 1% για τη μικρότερη κλάση, στις αντίστοιχες μεθόδους το σύνολο δεδομένων θα έχει αύξηση ή μείωση

που προσεγγίζει το 100%. Οπότε, στην εφαρμογή μας έχουμε πολύ μεγαλύτερο ποσοστό πραγματικών δεδομένων προς επεξεργασία.

Στη συνέχεια, θα χωρίσουμε τις μεθόδους σε τέσσερις υποκατηγορίες, ανάλογα με το είδος της εξισορρόπησης του συνόλου δεδομένων στο οποίο ανήκουν. Για τις αποδόσεις των μεθόδων θα χρησιμοποιήσουμε τον **πίνακα 4.2** που παρουσιάσαμε στο προηγούμενο κεφάλαιο.

5.1.1 Αποδόσεις για το αρχικό σύνολο δεδομένων

Αρχικά, θα επικεντρωθούμε μόνο στα αποτελέσματα που προέκυψαν από το αρχικό σύνολο δεδομένων, πριν αλλοιωθεί. Αν υποθέταμε πως, χωρίς να χρησιμοποιήσουμε κανέναν αλγόριθμο, για όλα τα άτομα πως η εξαρτημένη μεταβλητή έχει τιμή 0, δηλαδή κανείς δεν έδειχνε ενδιαφέρον να κοιτάξει για άλλη εταιρεία, τότε θα επιτυγχάναμε μια απόδοση 75%.

Έτσι, τα αποτελέσματα που παίρνουμε από τους δυο αλγόριθμους είναι πολύ άσχημα (66% και 48%) και ειδικότερα, στην περίπτωση που χρησιμοποιούμε ως μετρική το **F1-score** (48%). Στα αρχικά δεδομένα είναι ορατός και ο λόγος που για τα μη ισορροπημένα μοντέλα προτείνεται η μετρική **ROC-AUC**, αφού μια διαφορά στην απόδοση της τάξης του 18% είναι τεράστια.

5.1.2 Αποδόσεις για τις μεθόδους υπο-δειγματοληψίας

Σε μια γενική εικόνα, οι συγκεκριμένες μέθοδοι υπολείπονται του ποσοστού του 75% που θα παίρναμε, αν θεωρούσαμε πως όλες οι εγγραφές ανήκουν στην επικρατούσα κλάση. Οι αποδόσεις που μας δίνουν είναι από 66% έως 77% ανάλογα με τον αλγόριθμο και τη

μετρική που επιλέγουμε. Η μόνη απόδοση που κυμαίνεται πολύ χαμηλότερα είναι της μεθόδου **OneSidedSelection**, όταν χρησιμοποιούμε τη μετρική **F1-score** που μας δίνει περίπου **55%**. Αντίθετα, καλύτερη από τις μεθόδους μας είναι η **NeighbourhoodCleaningRule**, όταν χρησιμοποιούμε ως μετρική την **ROC-AUC** που δίνει μια επιτυχία της τάξης του 77% περίπου.

Αν θέλουμε να κατατάξουμε τις μεθόδους θα λέγαμε πως μετά την **NeighbourhoodCleaningRule**, έπεται οι **RandomUnderSampler**, με ποσοστό περίπου 73%, που ενδεχομένως εξηγείται από την τυχαία επιλογή των εγγραφών που βοηθάει να προσεγγιστεί το ποσοστό της επικρατούσας κλάσης.

Ακολουθεί η **NearMiss** με ποσοστό 72% και είναι η μόνη που δίνει καλύτερα αποτελέσματα αν πάρουμε ως μετρική την **F1-score**. Η **CondensedNearestNeighbour** είναι η επόμενη με επιτυχία γύρω στο 70%, ενώ όπως προαναφέραμε η χειρότερη μέθοδος είναι η **OneSidedSelection**.

5.1.3 Αποδόσεις για τις μεθόδους υπερ-δειγματοληψίας

Στην υπερ-δειγματοληψία, οι μέθοδοι δείχνουν μια ιδιαιτέρως σημαντική βελτίωση στην επιτυχή εκτίμηση των τιμών της εξαρτημένης μεταβλητής. Ίσως, σημαντικό ρόλο σε αυτό να παίζει το σαφώς μεγαλύτερο δείγμα που εξετάζουμε, αφού σε σχέση με την υπο-δειγματοληψία το δείγμα είναι τριπλάσιο. Σαν επισήμανση να αναφέρουμε πως πάντα το μεγαλύτερο δείγμα είναι πιθανότερο να δημιουργήσει **υπερπροσαρμογή** (**overfitting**), χωρίς όμως στις τρεις από τις τέσσερις μεθόδους να υποστηρίζεται μια τέτοια θεωρία.

Η μόνη μέθοδος στην οποία υπάρχει έντονη πιθανότητα υπερπροσαρμογής των δεδομένων που θα μας δώσει λανθασμένα ένα μεγάλο ποσοστό είναι η μέθοδος **RandomOverSampler**, η οποία μας δίνει και το μεγαλύτερο ποσοστό επιτυχίας, που προσεγγίζει ακόμα και το 91%. Όμως, αυτό μπορεί να είναι δείγμα υπερπροσαρμογής στο

εκπαιδευτικό σύνολο δεδομένων με αποτέλεσμα να μην υπάρχει η αντίστοιχη πολύ μεγάλη απόδοση σε κάποιο νέο σύνολο δεδομένων.

Οι άλλες τρεις μέθοδοι κινούνται σε παραπλήσιες αποδόσεις που είναι καλύτερες από το 75%. Οι αποδόσεις των μεθόδων **SMOTE**, **BorderlineSMOTE** και **ADASYN** κυμαίνονται μεταξύ 81% και 83% ανεξαρτήτως των άλλων δυο παραμέτρων, δηλαδή των αλγόριθμων και των μετρικών που θα χρησιμοποιήσουμε.

5.1.4 Αποδόσεις για τις μεικτές μεθόδους

Για τις δυο μεικτές μεθόδους έχουμε τεράστια διαφορά μεταξύ των αποδόσεων τους, όμως οι δυο αποτελούν επιτυχημένες μεθόδους. Η μέθοδος **SMOTEENN** έχει σχεδόν άριστη απόδοση και βεβαίως αποτελεί με διαφορά την καλύτερη μέθοδο εξισορρόπησης δεδομένων από όλες όσες έχουμε χρησιμοποιήσει. Το ποσοστό της απόδοσης της αγγίζει το 95% που είναι υπερβολικά μεγάλο ποσοστό επιτυχίας.

Η άλλη μέθοδος, η **SMOTETomek** επίσης έχει βελτιώσει το ποσοστό της επικρατούσας κλάσης, επιτυγχάνοντας απόδοση 83% που είναι αρκετά καλή, αν και υπολείπεται κατά πολύ της μεθόδου **SMOTEENN**.

5.2 Εξάρτηση των αποδόσεων από τους αλγόριθμους και τις μετρικές

Παρότι, το αντικείμενο της εργασίας δεν αφορά αυτή τη μελέτη, θα πρέπει να εκτιμήσουμε κατά πόσο οι διαφορετικοί αλγόριθμοι ή οι μετρικές αξιολόγησης μπορούν να δημιουργήσουν μεροληψία και να αλλοιώσουν τα συμπεράσματα που έχουμε εξάγει για τις μεθόδους εξισορρόπησης. Γι' αυτό επιλέξαμε να ελέγξουμε δυο διαφορετικούς αλγόριθμους μηχανικής μάθησης και δυο διαφορετικές μετρικές για να έχουμε μια πιο αντικειμενική προσέγγιση των αποτελεσμάτων που μας δίνουν οι μέθοδοι εξισορρόπησης.

Παρακάτω, δείχνουμε τη σύγκριση των αποδόσεων μεταξύ τόσο των αλγορίθμων, όσο και των μετρικών που χρησιμοποιήσαμε. Τα αποτελέσματα εξάγονται από τον **πίνακα 4.2** που παραθέσαμε στο προηγούμενο κεφάλαιο.

Μέθοδος	**Σύγκριση αποδόσεων**							
	Random Forest – Gradient Boosting				**F1-score – ROC-AUC**			
	F1-score		**ROC-AUC**		**Random Forest**		**Gradient Boosting**	
Αρχικό σύνολο δεδομένων	RF	1,24%	RF	0,73%	RA	-17,03%	RA	-17,54%
Μέθοδοι Υπο-δειγματοληψίας								
RandomUnderSampler	GB	-0,87%	GB	-0,01%	RA	-1,21%	RA	-0,35%
NearMiss	RF	0,74%	RF	0,13%	F1	1,98%	F1	1,37%
CondensedNearestNeighbour	GB	-1,70%	GB	-0,89%	RA	-3,11%	RA	-2,30%
OneSidedSelection	GB	-2,16%	GB	-1,32%	RA	-14,22%	RA	-13,38%
NeighbourhoodCleaningRule	GB	-0,51%	GB	-0,20%	RA	-5,81%	RA	-5,50%

Μέθοδος	Σύγκριση αποδόσεων					
	Random Forest – Gradient Boosting			F1-score – ROC-AUC		
	F1-score		ROC-AUC	Random Forest		Gradient Boosting
Μέθοδοι Υπερ-δειγματοληψίας						
RandomOverSampler	GB	-0,61%	GB -0,79%	F1	0,47%	F1 0,29%
SMOTE	GB	-1,49%	GB -1,49%	RA	-0,02%	RA -0,02%
BorderlineSMOTE	GB	-0,79%	GB -0,90%	F1	0,22%	F1 0,11%
ADASYN	GB	-0,86%	GB -0,94%	F1	0,43%	F1 0,35%
Μεικτές Μέθοδοι						
SMOTETomek	GB	-1,42%	GB -1,44%	RA	-0,03%	RA -0,05%
SMOTEENN	GB	-0,68%	GB -0,93%	F1	2,23%	F1 1,98%

Πίνακας 5.1: Σύγκριση μεταξύ των αποδόσεων στους αλγόριθμους και στις μετρικές αξιολόγησης

Αυτό που συμπεραίνουμε είναι πως για το αρχικό σύνολο δεδομένων η διαφορά στην απόδοση που οφείλεται στους αλγόριθμους είναι περίπου 1% με καλύτερο τον RandomForest, ενώ η διαφορά μεταξύ των μετρικών είναι τεράστια με ποσοστό 17% επιβεβαιώνοντας πως για ένα μη ισορροπημένο σύνολο δεδομένων η καλύτερη μετρική είναι η ROC-AUC σε αντίθεση με την F1score που έχει πάρα πολύ χαμηλή απόδοση.

Σε ορισμένες μεθόδους υπο-δειγματοληψίας παρατηρούμε μεγάλες έως τεράστιες διαφορές στις αποδόσεις, όπου υπερισχύει η ROC-AUC. Ειδικά, για την OneSidedSelection έχουμε 14% διαφορά, για την CondensedNearestNeighbour έχουμε 3% και για τη NeighbourhoodCleaningRule έχουμε 5,50%. Οι διαφορές αυτές είναι ιδιαιτέρως αξιοσημείωτες και είναι προφανές πως η μετρική ROC-AUC μας δίνει γενικά πολύ καλύτερα αποτελέσματα.

Για τους αλγόριθμους, οι διαφορές είναι μικρότερες του 2% ενώ για τις περισσότερες μεθόδους εξισορρόπησης μας δίνουν παραπλήσια αποτελέσματα.

Αντιθέτως, στις μεθόδους υπερ-δειγματοληψίας τόσο οι μετρικές, όσο και οι αλγόριθμοι δεν επηρεάζουν σημαντικά τις αποδόσεις ανάμεσα στις διαφορετικές μεθόδους εξισορρόπησης του συνόλου δεδομένων.

Οι αποδόσεις που δίνουν οι μέθοδοι εξισορρόπησης δεν εξαρτώνται από την επιλογή της μετρικής, ενώ με εξαίρεση τη μέθοδο SMOTE, όπου ο αλγόριθμος GradientBoosting μας δίνει καλύτερες αποδόσεις κατά 1,50%, στις υπόλοιπες μεθόδους δε φαίνεται ο αλγόριθμος να επηρεάζει την απόδοση.

Τέλος για τις μεικτές μεθόδους, η SMOTETomek με τον αλγόριθμο GradientBoosting δίνει καλύτερα αποτελέσματα κατά 1,50%, σε αντίθεση με τη μέθοδο SMOTEENN που είναι κάτω από 1% η διαφορά της χρήσης διαφορετικών αλγόριθμων.

Στη σύγκριση των μετρικών αξιολόγησης, η SMOTETomek δίνει παρόμοιες αποδόσεις και στις δυο περιπτώσεις, ενώ για τη SMOTEENN η οποία υπενθυμίζουμε ότι μας δίνει την καλύτερη πρόβλεψη με σχεδόν 95%, την πρόβλεψη αυτή την παίρνουμε με τη μετρική F1-score με διαφορά που κυμαίνεται γύρω στο 2%. Μάλιστα, αξίζει να σημειώσουμε πως σε ελάχιστες άλλες περιπτώσεις η συγκεκριμένη μετρική μας δίνει τις καλύτερες αποδόσεις.

Με ελάχιστες εξαιρέσεις, μπορούμε να συμπεράνουμε πως ούτε οι μετρικές αξιολόγησης, αλλά ούτε και οι αλγόριθμοι επηρεάζουν έντονα τις μεθόδους εξισορρόπησης. Έτσι, γίνεται κατανοητό πως οι αποδόσεις που λαμβάνουμε από τις μεθόδους εξισορρόπησης του συνόλου δεδομένων είναι αποτέλεσμα της δικής τους ικανότητας να εκτιμήσουν σωστά τις τιμές της εξαρτημένης μεταβλητής.

Ερευνητικές εργασίες (Research Papers)

[A] **Hadley Wickham,"Tidy data",** ***The Journal of Statistical Software*****, vol. 59, 2014**

[B] **S. Susan and A. Kumar, "The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent State of the Art", 2020**

[C] **J.P. Zhang and I. Mani, "KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction", 2003**

[D] **C. Hilborn and D. Lainiotis, "The Condensed Nearest Neighbor Rule", 1968**

[E] **I. Tomek, "Two Modifications of CNN", 1976**

[F] **D. Wilson, "Asymptotic Properties of Nearest Neighbor Rules Using Edited Data", 1972**

[G] **M. Kubat and S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection", 1997**

[H] **J. Laurikkala, "Improving Identification of Difficult Small Classes by Balancing Class Distribution", 2001**

[I] J. Kennedy and R. Eberhart**, "Particle swarm optimization", 1995**

[J] **N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer , "SMOTE: Synthetic Minority Over-sampling Technique", June 2002**

[K] **H, Han, W. Wang and B. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning", August 2005**

[L] **H. He, Y. Bai, E. Garcia and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning", 2008**

[M] **S. Chen, H. He and E. A. Garcia, "RAMOBoost: Ranked Minority Oversampling in Boosting", Oct. 2010**

[N] **L. Abdi and S. Hashemi, "To Combat Multi-Class Imbalanced Problems by Means of Over-Sampling Techniques", Jan. 2016**

[O] **S. Barua, "MWMOTE--Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning", Feb. 2014**

Βιβλιογραφία (Βιβλία & Ιστοσελίδες)

[1] N. Leech, K. Barrett and G. A. Morgan, "SPSS for Intermediate Statistics", σελ. 42-50, Routledge, 2004

[2] https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=5e2435f46f63

[3] **T. Rattenbury, J. M. Hellerstein, J. Heer, S. Kandel and C. Carreras, "Principles of Data Wrangling: Practical Techniques for Data Preparation", O'Reilly Media, Inc., σελ. ix-x, July 2017**

[4] **T. Rattenbury, J. M. Hellerstein, J. Heer, S. Kandel and C. Carreras, "Principles of Data Wrangling: Practical Techniques for Data Preparation", O'Reilly Media, σελ. 12-15, July 2017**

[5] **T. Rattenbury, J. M. Hellerstein, J. Heer, S. Kandel and C. Carreras, "Principles of Data Wrangling: Practical Techniques for Data Preparation", O'Reilly Media, σελ. 8-10 και 31-35, July 2017**

[6] Wes McKinney**, "Python for Data Analysis", σελ. 191-195, O'Reilly Media, October 2017 - 2nd edition**

[7] Wes McKinney**, "Python for Data Analysis", σελ. 205-206, O'Reilly Media, October 2017 - 2nd edition**

[8] **H. J. Seltman, "Experimental Design and Analysis", σελ. 61-62, Carnegie Mellon University, 2012**

[9] **https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e**

[10] **A. Zheng and A. Casari, "Feature Engineering for Machine Learning", σελ. 29-30, O'Reilly Media, April 2018**

[11] **A. Zheng and A. Casari, "Feature Engineering for Machine Learning", σελ. 104-105, O'Reilly Media, April 2018**

[12] N. Nilsson**, "Introduction to Machine Learning. An early draft of a proposed textbook", σελ. 5-6, Stanford University, 1996**

[13] N. Nilsson**, "Introduction to Machine Learning. An early draft of a proposed textbook", σελ. 119, Stanford University, 1996**

[14] N. Nilsson**, "Introduction to Machine Learning. An early draft of a proposed textbook", σελ. 143, Stanford University, 1996**

[15] **Ian H. Witten, Eibe Frank and Mark A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques", σελ. 61-62, Elsevier, 2017 - 3rd edition**

[16] **https://www.explorium.ai/blog/top-10-evaluation-metrics-for-classification-models/**

[17] **https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data**

[18] **https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html**

[19] H. He and Y. Ma**, "Imbalanced Learning: Foundations, Algorithms, and Applications", σελ. 45-46, Wiley, 2013**

[20] H. He and Y. Ma**, "Imbalanced Learning: Foundations, Algorithms, and Applications", σελ. 46-47, Wiley, 2013**

[21] H. He and Y. Ma**, "Imbalanced Learning: Foundations, Algorithms, and Applications", σελ. 47-48, Wiley, 2013**

[22] **https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists**

Εικόνες

Πίνακες