ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΥΠΟΛΟΓΙΣΤΩΝ

Πρόβλεψη εξάρσεων του ιού του Δυτικού Νείλου βάσει περιβαλλοντικών παραμέτρων με την χρήση μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Εμμανουήλ Α. Τσόλια

Επιβλέπων: Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2022

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΥΠΟΛΟΓΙΣΤΩΝ

Πρόβλεψη εξάρσεων του ιού του Δυτικού Νείλου βάση περιβαλλοντικών παραμέτρων με την χρήση μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Εμμανουήλ Α. Τσόλια

Επιβλέπων : Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 14η Νοεμβρίου 2022

.....	.....	.....
Στέφανος Κόλλιας	Αθανάσιος Βουλόδημος	Γιώργος Στάμου
Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2022

.....
Εμμανουήλ Α. Τσόλιας

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

Περίληψη

Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας, οι μεταδιδόμενες από φορείς ασθένειες προκαλούν περισσότερους από 700.000 θανάτους τον χρόνο, καθώς και αρκετά εκατομμύρια νοσήσεις παγκοσμίως. Ως 'φορείς' ορίζονται ζωντανοί οργανισμοί που δύνανται να μεταφέρουν μολυσματικά παθογόνα μεταξύ ανθρώπων ή από κάποιο ζώο προς τον άνθρωπο. Αυτά τα παθογόνα μπορούν να είναι παράσιτα, ιοί ή βακτήρια που προκαλούν ασθένειες όπως η ελονοσία και ο κίτρινος πυρετός. Συνήθως υπάρχουν σε ένα ποσοστό του πληθυσμού σε μόνιμη κατάσταση αλλά συχνά εμφανίζονται μέσω εξάρσεων που κατακλύζουν το σύστημα προκαλώντας θανάτους, χρόνια προβλήματα υγείας και ενίοτε κοινωνικό στίγμα.

Καθώς η κατανομή αυτών των ασθενειών συσχετίζεται με παραμέτρους όπως το κλίμα, οι οικονομικές συνθήκες και τα δημογραφικά χαρακτηριστικά, πολλές ασθένειες περιορίζονται εκ των πραγμάτων σε συγκεκριμένες γεωγραφικές περιοχές. Ωστόσο η κλιματική αποσταθεροποίηση έχει δημιουργήσει πρόσφορες συνθήκες για την εγκατάσταση και εξάπλωση αρκετών ασθενειών σε νέες περιοχές όπου οι ιστορικά οι περιβαλλοντικές παράμετροι δεν ευνοούσαν -ή ακόμα και δεν επέτρεπαν- την παρουσία τους. Επιπροσθέτως, η ραγδαία αύξηση της μετακίνησης τόσο ανθρώπων όσο και εμπορευμάτων έχει προκαλέσει την εισαγωγή διάφορων παθογόνων σε χώρες που ιστορικά δεν υπήρχε παρουσία.

Ένα παράδειγμα τέτοιας ασθένειας που έχει απασχολήσει την ιατρική κοινότητα είναι αυτό του ιού του Δυτικού Νείλου. Πρόκειται για έναν ιό που στην γενική περίπτωση προκαλεί παρόμοια συμπτωματολογία με αυτή της κοινής γρίπης, αλλά σε ένα μικρό ποσοστό του πληθυσμού μπορεί να προκαλέσει εγκεφαλίτιδα ή μηνιγγίτιδα, και τελικά θάνατο. Ο συγκεκριμένος ιός που για πρώτη φορά ανιχνεύθηκε το 1936 στην Ουγκάντα, έχει επεκταθεί λόγω όσων αναφέραμε σε όλες σχεδόν τις ηπείρους, με χαρακτηριστική την εισαγωγή στις ΗΠΑ το 1999 όπου πλέον αποτελεί την κύρια ασθένεια με φορέα τα κουνούπια. Ακόμα έχει ισχυρή παρουσία σε χώρες της ανατολικής, νότιας και δυτικής Ευρώπης όπου εμφανίζεται κατά κύματα τους θερινούς μήνες. Η Ελλάδα ανήκει στις χώρες με τα εντονότερα καταγεγραμμένα ξεσπάσματα με χιλιάδες κρούσματα σε ανθρώπους και ιπποειδή. Προς το παρόν, δεν υπάρχουν αποτελεσματικές θεραπείες αλλά ούτε και εμβόλια για τον συγκεκριμένο ιό.

Καθώς τα μέσα για την αντιμετώπιση της νόσου είναι ανεπαρκή και το κόστος στον άνθρωπο μεγάλο, προκύπτει η ανάγκη για την ανάπτυξη εργαλείων έγκαιρης προειδοποίησης. Η υψηλή συσχέτιση των εξάρσεων του ιού με τις εκάστοτε κλιματικές παραμέτρους υποδεικνύει μια πιθανή βάση των προβλεπτικών μοντέλων. Στο παρόν πόνημα επιχειρείται η ανάπτυξη ενός τέτοιου εργαλείου που αξιοποιεί τα κλιματικά δεδομένα της Ελλάδας, ανά δήμο, μέσω ενός μοντέλου μηχανικής μάθησης, του αλγορίθμου XGBoost, με σκοπό την πρόβλεψη επερχόμενων εξάρσεων του ιού του Δυτικού Νείλου.

Λέξεις κλειδιά: Μηχανική Μάθηση, XGBoost, Πρόβλεψη, Ιός του Δυτικού Νείλου, Έξαρση, Κλιματικά δεδομένα.

Abstract

According to the World Health Organization, vector-borne diseases cause more than 700,000 deaths worldwide and sicken millions of people. The term ‘vector’ refers to any organism that can transmit infectious pathogens between humans, or from animals to humans. These pathogens can be viruses, bacteria, and parasites that cause diseases like malaria and yellow fever. Usually, a fraction of the population is infected at any time but its often the case that outbreaks occur, which then overwhelm the healthcare system causing deaths, lifelong conditions and even stigmatization.

As the distribution of these diseases depends on climatic, economic, and demographic factors many diseases were constrained to certain geographic regions. However, climate destabilization has altered the weather characteristics of many regions in a way that infestation is no longer hindered. Additionally, the interconnected global economy results in people, as well as products, being transported at an unprecedented scale which in turn facilitates the spread of diseases.

One such disease that has raised concern among the medical community is the West Nile Virus. While this virus will typically cause only mild symptoms, a small fraction of patients presents with severe neuroinvasive disease such as encephalitis or meningitis which occasionally result in death. It was first isolated in Uganda in 1937 but is now present every continent except Antarctica. One characteristic example of its expansive potential is the USA spillover where it is now the leading cause of mosquito-borne diseases. Similarly, many countries of east, south and west Europe now report seasonal outbreaks during the summer months. Greece has recorded some of the most severe outbreaks with high number of human as well as animal cases. As of now, no effective treatment or vaccine is available.

Since the therapeutic means are limited and the toll in human life and well-being is so high, it follows that prevention is of essence. This can be achieved through the development of early warning systems. The high correlation between outbreaks and climatic features points to a promising basis for the predictive models. In this exercise we attempt to develop a predictive tool that utilizes the climatic data of Greece at a municipal level, through the XGBoost machine learning algorithm, in order to predict coming outbreaks of the West Nile Virus.

Key words: Machine learning, XGBoost, Prediction, West Nile Virus, Outbreak, Climatic data.

Ευχαριστίες

Με αυτήν την εργασία ολοκληρώνεται η φοιτητική πορεία μου στο Εθνικό Μετσόβιο Πολυτεχνείο, πράγμα το οποίο δεν θα ήταν εφικτό χωρίς τον κόπο των ανθρώπων που εργάζονται καθημερινά στις αίθουσες, στα εργαστήρια και στην γραμματεία και για αυτό τους είμαι ευγνώμων.

Ιδιαίτερα θέλω να ευχαριστήσω τον καθηγητή κ. Στέφανο Κόλλια για την εμπιστοσύνη στην ανάθεση αυτής της εργασίας αλλά και την καθοδήγηση στην εκπόνηση της. Ακόμα θέλω να ευχαριστήσω τον καθηγητή κ. Χρίστο Χατζηχριστοδούλου για την ευκαιρία που μου έδωσε να συμμετάσχω το πρόγραμμα του West Nile που αποτέλεσε την βάση για την συγκεκριμένη εργασία, και τα μέλη της ομάδας του ερευνητικού προγράμματος ΕΜΠΡΟΣ Γεώργιο Γεωργακίλα, Μιχάλη Κουρέα και Γεώργιο Χαρβαλή για την καθοδήγηση στην διάρκεια του ερευνητικού προγράμματος.

Τέλος, ευχαριστώ τους γονείς μου στους οποίους χρωστάω τα πάντα.

Αφιερώνεται στην γυναίκα μου, Ιωάννα

Contents

Κεφάλαιο 1. Εισαγωγή

1.1 Ιός του Δυτικού Νείλου στην Ελλάδα

Ο ιός του Δυτικού Νείλου αποτελεί μια πλέον ενδημική νόσο σε διάφορες χώρες παγκοσμίως και ιδιαίτερα στην Ευρώπη, με κρούσματα να εμφανίζονται σε ετήσια βάση, κατά κανόνα τους θερινούς μήνες [1]. Πρόκειται για έναν RNA ιό που προκαλεί τον πυρετό του Δυτικού Νείλου [2]. Η ασθένεια αυτή που μεταδίδεται μέσω των κουνουπιών [3], εκτυλίσσεται με καθόλου ή λίγα συμπτώματα στο 80% των ασθενών , ωστόσο σε λιγότερο από το 1% θα προσβάλει το Κεντρικό Νευρικό Σύστημα (ΚΝΣ) το οποίο συνεπάγεται βαριά κλινική εικόνα και ορισμένες φορές θάνατο [4]. Στην Ελλάδα, αν και τα έτη 2010-2014 καταγραφόντουσαν κρούσματα, αυτά μηδενίστηκαν αναπάντεχα τις επόμενες δύο χρονιές. Ωστόσο το 2017 είχαμε ξανά καταγραφή κρουσμάτων ενώ το 2018 σημειώθηκε μία έντονη αναζωπύρωση και τις επόμενες χρονιές καταγράφηκε σταθερή παρουσία, ενώ ακόμα καταγράφηκε παρουσία του ιού σε όλες τις περιφέρειες [5]. Από τα παραπάνω προκύπτει πως ο ιός έχει εγκατασταθεί στην χώρα. Ακόμα, σύμφωνα με τα στοιχεία του Εθνικού Οργανισμού Δημόσιας Υγείας (ΕΟΔΥ) μόνο τις χρονιές 2018-2021 σημειώθηκαν περίπου 120 θάνατοι και περισσότερες από 500 περιπτώσεις με προσβολή του ΚΝΣ [5]. Επομένως, πρόκειται για ένα πρόβλημα που χρήζει ουσιαστικής αντιμετώπισης. Ωστόσο, μέχρι και την στιγμή της συγγραφής αυτής της εργασίας, δεν υπάρχουν εμβόλια ή φάρμακα για την πρόληψη ή την θεραπεία της νόσου [6]–[8]. Ακόμα, η αλυσίδα μετάδοσης αυτού προς τον άνθρωπο είναι ιδιαίτερα περίπλοκη καθώς περιλαμβάνει πέρα από τον ίδιο τον ιό, τα πτηνά από τα οποία προέρχεται και τα κουνούπια μέσω των οποίων μεταδίδεται [3], [9], [10]. Κάθε ένα από αυτά τα στάδια εισάγει πολυπλοκότητα στον συνολικό μηχανισμό ενώ ταυτόχρονα όλα επηρεάζονται από τις κλιματικές συνθήκες της ευρύτερης περιοχής [11]. Για αυτόν τον αυτό δεν είναι δυνατή μέχρι τώρα η πρόβλεψη της κυκλοφορίας του ιού στην Ελλάδα.

1.2 Σκοπός της εργασίας

Όπως αναφέρθηκε, ο πυρετός του Δυτικού Νείλου κοστίζει την ζωή σε δεκάδες ανθρώπους κάθε χρόνο ενώ προκαλεί βαριά νόσο σε εκατοντάδες [5]. Προς το παρόν δεν υπάρχουν μέσα για την θεραπεία αυτού ενώ η συμπεριφορά του σε επίπεδο έτους, αν και εξαρτάται από τις κλιματικές συνθήκες, φαίνεται να είναι δυσπρόβλεπτη. Σε αυτήν την εργασία

επικεντρωνόμαστε στην πρόληψη φαινομένου μέσω την ανάπτυξης ενός εργαλείου έγκαιρης προειδοποίησης. Αξιοποιώντας τα κλιματικά δεδομένα των τελευταίων 12 ετών, σε συνδυασμό με τα στοιχεία των κρουσμάτων αυτού του διαστήματος, επιχειρούμε να εκπαιδεύσουμε ένα μοντέλο μηχανικής μάθησης προκειμένου να μπορεί με βάση νέα καιρικά δεδομένα να προβλέπει την πιθανότητα εξάρσεως της νόσου, σε επίπεδο δήμου. Ένα τέτοιο εργαλείο μπορεί να παρέχει χρήσιμο χρόνο στις αρχές προκειμένου αυτές να προβούν σε ενέργειες ελέγχου του πληθυσμού των κουνουπιών, σε στελέχωση και προετοιμασία των υποδομών υγείας, ενημέρωση του πληθυσμού κ.α.

1.3 Σχετικές έρευνες

Αν και η μοντελοποίηση της κυκλοφορίας του ιού δεν έχει ακόμα επιτευχθεί για την περιοχή της Ελλάδας, η παγκόσμια έκταση του φαινομένου έχει οδηγήσει διάφορες
ερευνητικές ομάδες στην μελέτη του ζητήματος. Προκειμένου να αξιοποιήσουμε αυτές τις προσπάθειες αναζητήσαμε την σχετική βιβλιογραφία και μελετήσαμε 1) τα δεδομένα που χρησιμοποιήθηκαν 2) τα μοντέλα που δοκιμάστηκαν και 3) τα αποτελέσματα που επιτεύχθηκαν. Τα δύο πρώτα αποτέλεσαν επιβεβαίωση των αρχικών υποθέσεων που κάναμε, ενώ το τρίτο μας έδωσε μία γενική εικόνα των προσδοκιών που μπορούσαμε να έχουμε για τις επιδόσεις του μοντέλου μας.

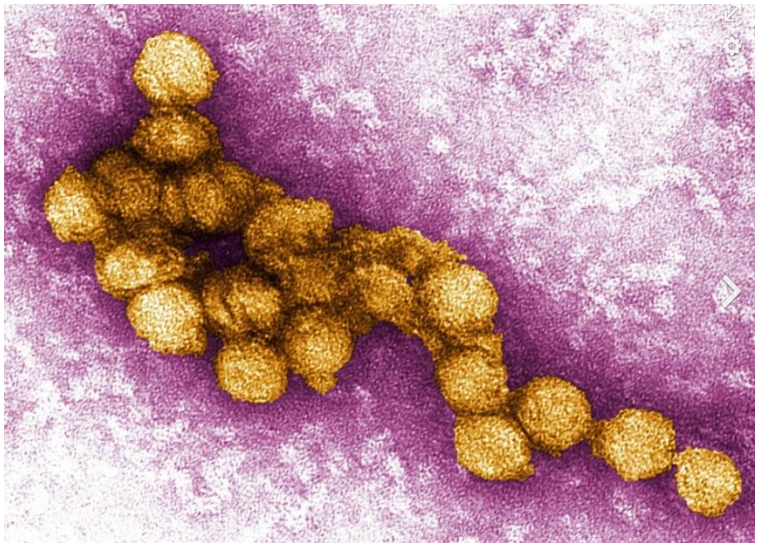
1.4 Οργάνωση αναφοράς

Στο κεφάλαιο 2 παρατίθεται η υπάρχουσα γνώση σχετικά με τον ιό, τις νόσους που προκαλεί και τα επιδημιολογικά χαρακτηριστικά του. Στο κεφάλαιο 3 εξηγείται ο τρόπος λειτουργίας του αλγορίθμου XGBoost, οι επιπλέον τεχνικές που βελτιώνουν την κλιμακωσιμότητά και την ευελιξία του, και μαζί με τον ίδιο συνθέτουν την βιβλιοθήκη XGBoost, καθώς και οι υπερπαράμετροι που τον καθορίζουν. Στο κεφάλαιο 4 μελετάται η σχετική βιβλιογραφία και στο κεφάλαιο 5 αναλύεται η προσέγγισή μας. Αυτό περιλαμβάνει την εύρεση και προεπεξεργασία των δεδομένων, την επιλογή και επεξήγηση των μετρικών και της μεθόδου βελτιστοποίησης των υπερπαραμέτρων, και τα αποτελέσματα που προκύπτουν. Στο κεφάλαιο 6 κατατίθενται τα συμπεράσματα και οι σκέψεις για επόμενες έρευνες.

Κεφάλαιο 2. Ο ιός και η ασθένεια του Δυτικού Νείλου

2.1 Ο ιός

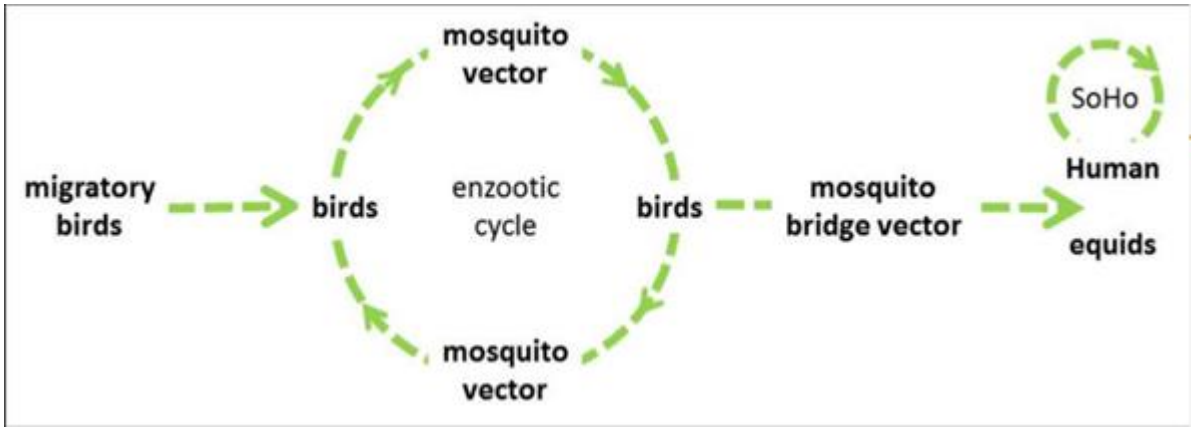
Ο ιός του Δυτικού Νείλου (εικόνα 1), που προκαλεί την ομώνυμη ασθένεια, είναι ένας μονόκλωνος RNA ιός της οικογένειας Flaviviridae. Ανήκει στο γένος Flavivirus που περιέχει και άλλους γνωστούς ιούς όπως τον Ζίκα, τον Δάγκειο και τον ιό του κίτρινου πυρετού [12].



Εικόνα 1: Φωτομικρογραφία του WNV (εμφανίζεται με κίτρινο). [13]

Ο πρωτεύων ξενιστής του ιού είναι τα πτηνά. Συγκεκριμένα, συνήθως ανιχνεύεται στα κοράκια και στις κίσσες, ωστόσο έχει βρεθεί σε πάνω από 300 διαφορετικά είδη πτηνών [14], [15]. Ακόμα ανιχνεύεται συχνά σε διάφορα ήδη κουνουπιών, ωστόσο τα είδη που διαδραματίζουν σημαντικό ρόλο στην εξάπλωση του ιού είναι τα είδη του γένους Culex που τρέφονται τόσο από πτηνά, όσο και από ανθρώπους ή άλογα (C. Pipiens, C. Tarsalis κλπ.) καθώς αυτά αποτελούν τον λεγόμενο φορέα-γέφυρα (bridge-vector) [9], [10], [16]. Τα πτηνά μαζί με τα κουνούπια αποτελούν τους λεγόμενους φυσικούς ξενιστές του ιού [17], και έτσι, στον συνήθη αγροτικό-ενζωοτικό κύκλο μετάδοσης, ο ιός μεταπηδά από πτηνό σε φορέα και από φορέα σε πτηνό κ.ο.κ. ενώ πιο σπάνια μπορεί να συμβεί μετάδοση από πτηνό σε πτηνό μέσω μολυσμένης τροφής ή νερού [18]. Ωστόσο, σε περιβάλλοντα που συνυπάρχουν οι βασικοί ξενιστές με άλλα ήδη, μπορεί να συμβεί διάχυση, δηλαδή όταν ένα μολυσμένο κουνούπι τραφεί από ένα τρίτο ζώο, ο ιός να μεταφερθεί σε αυτό (εικόνα 2) [3], [10], [18]. Τα θηλαστικά που μολύνονται με αυτόν τον τρόπο ονομάζονται συμπτωματικοί -με την έννοια της σύμπτωσης- ‘αδιέξοδοι ξενιστές’, καθώς ο ιός δεν μπορεί να αναπαραχθεί σε αυτά σε βαθμό ώστε επόμενο τσίμπημα να οδηγήσει σε μόλυνση του κουνουπιού, πράγμα που

σημαίνει πως η αλυσίδα μετάδοσης διακόπτεται [16]. Η χαμηλή αυτή ιαιμία (επίπεδα υιικού φορτίου στο αίμα) που παρατηρείται στα θηλαστικά δεν σημαίνει ότι αυτά δεν νοσούν, καθώς έχει βρεθεί πως οι άνθρωποι και τα άλογα μπορούν να εμφανίσουν βαριά συμπτωματολογία. Από την άλλη, οι σκύλοι και οι γάτες σπάνια εμφανίζουν συμπτώματα [19].



Εικόνα 2: Ο κύκλος μετάδοσης της νόσου. SoHo: μετάδοση από άνθρωπο σε άνθρωπο (σπάνιο) [20]

2.2 Η ασθένεια

Όταν ένα μολυσμένο κουνούπι τραφεί από άνθρωπο είναι πιθανό ο ιός να μεταδοθεί σε αυτόν. Όμως, το 80% των μολύνσεων εκτυλίσσονται ασυμπτωματικά και δεν καταγράφονται [21][4]. Το υπόλοιπο 20% θα εμφανίσει συμπτώματα τα οποία ποικίλουν σε σοβαρότητα και διάρκεια, και συνήθως ξεκινάνε 48 ώρες με 2 εβδομάδες μετά το τσίμπημα [3]. Τα πιο ήπια από τα συμπτώματα περιλαμβάνουν πυρετό, εμετό και δερματικά εξανθήματα και οι ασθενείς που βιώνουν μόνο αυτά συνήθως αναρρώνουν πλήρως, αν και ο οργανισμός μπορεί να χρειαστεί από εβδομάδες έως και μήνες για να ανακάμψει πλήρως.

Στις σοβαρότερες περιπτώσεις, που αποτελούν λιγότερο από το 1% των μολύνσεων, μπορεί να επέλθει παράλυση και κόμμα. Αυτές οι περιπτώσεις κρατάνε αρκετές εβδομάδες και ενίοτε προκαλούν μόνιμες εγκεφαλικές βλάβες ενώ το 10% καταλήγει σε θάνατο. Υποκείμενα νοσήματα και προχωρημένη ηλικία αυξάνουν τον κίνδυνο βαριάς νόσου [4], [22]. Οι περιπτώσεις βαριάς νόσησης κατηγοριοποιούνται ως εξής:

- Πυρετός του Δυτικού Νείλου (WNF): Επηρεάζει το 20% των κρουσμάτων. Συμπτώματα παρόμοια με αυτά της γρίπης [23].
- Νευροεπεμβατική ασθένεια του Δυτικού Νείλου (WNND): Εμφανίζεται σε ποσοστό μικρότερο του 1% και προσβάλει το ΚΝΣ προκαλώντας μηνιγγίτιδα, εγκεφαλίτιδα

μηνιγγοεγκεφαλίτιδα ή ένα σύνδρομο παρόμοιο με την πολιομυελίτιδα [24]. Η εγκεφαλίτιδα (WNE) αποτελεί και την πιο συχνή εκδοχή εκδήλωση της νευροεπεμβατικής νόσου και παρουσιάζει τα συνήθη συμπτώματα εγκεφαλίτιδας και κινητικές δυσλειτουργίες [25].

- Πολιομυελίτιδα του Δυτικού Νείλου (WNP): Αποτελεί ένα σύνδρομο που προκύπτει από μόλυνση με WNV και προκαλεί οξεία χαλαρή παράλυση (flaccid paralysis) με πιθανό αποτέλεσμα την αναπνευστική ανεπάρκεια [26].

Στις μη-νευρολογικές επιπλοκές της νόσου αναφέρονται η καλπάζουσα ηπατίτιδα, η Παγκρεατίτιδα [27], η μυοκαρδίτιδα[28], η ραβδομυόλυση[28], η νεφρίτιδα κ.α., ωστόσο αυτές εμφανίζονται σπάνια . Παράλληλα, αρκετά συχνά μπορούν να εμφανιστούν και δερματικές εκδηλώσεις [29].

Όπως αναφέραμε παραπάνω, η έκβαση στην πλειοψηφία των περιπτώσεων δεν είναι αρνητική. Ωστόσο πρόσφατες έχουν καταλήξει πως οι μακροπρόθεσμες επιπτώσεις μπορεί να είναι περισσότερο σοβαρές και διαρκείς από όσο είχε αρχικά εκτιμηθεί [30]. Πολλοί ασθενείς που βιώσαν μέτρια συμπτώματα, αναφέρουν κινητικές και νοητικές δυσκολίες για περισσότερους από 12 μήνες μετά νόσηση, ενώ το διάστημα της ανάρρωσης στιγματίζεται από έντονη σωματική κούραση [31].

Σε αυτές τις περιπτώσεις χρησιμοποιούνται συχνά αναλγητικά φάρμακα, μηχανήματα υποστήριξης της αναπνοής και ενδοφλέβιοι οροί για την αντιμετώπιση των συμπτωμάτων [4], [32]. Τα φαρμακευτικά μέσα περιορίζονται σε αυτά που αναφέρθηκαν καθώς δεν υπάρχει θεραπεία για τον ιό [33]. Επιπλέον, αν και έχουν γίνει διάφορες προσπάθειες για την δημιουργία εμβολίου, καμία από αυτές δεν έχει περάσει την φάση 2 των κλινικών δοκιμών [7]. Αντίθετα για τα άλογα διατίθενται αρκετά διαφορετικά εμβόλια [34]. Πριν την ανακάλυψη αυτών, το 40% των αλόγων που μολυνόταν στην βόρεια Αμερική πέθαινε [16]. Παρόμοια είναι η έκβαση και ορισμένα από τα πτηνά που αποτελούν τους φυσικούς ξενιστές του ιού [14][16]. Λόγω της μη ύπαρξης εμβολίων και θεραπευτικών μέσων για τους ανθρώπους, οι περισσότερες προσπάθειες εστιάζουν στην πρόληψη της νόσου [35]. Αυτό αναλύεται σε δύο άξονες: αφενός την προσωπική προστασία, που περιλαμβάνει την χρήση εντομοαπωθητικών, την αποφυγή χώρων όπου υπάρχει δραστηριότητα των εντόμων και τη χρήση κατάλληλου εξοπλισμού ή ρουχισμού, και αφετέρου τον έλεγχο των κουνουπιών μέσω εντομοκτόνων, παγίδων, και περιορισμού των στάσιμων υδάτων.

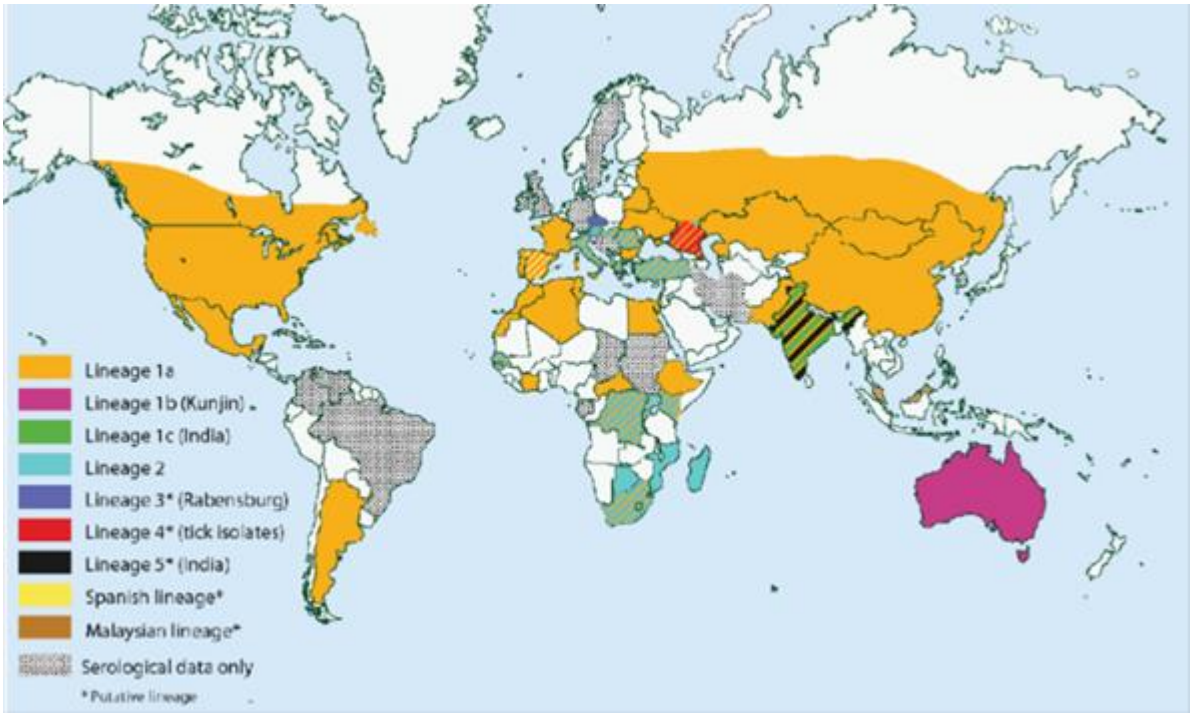
2.3 Η επιδημιολογία του ιού

Ο ιός ανιχνεύθηκε για πρώτη φορά στην Ουγκάντα το 1937 [36]. Επακόλουθες οροεπιδημιολογικές μελέτες ανίχνευσαν αντισώματα σε πληθυσμούς πολλών κρατών της Αφρικής [37], [38]. Στην συνέχεια καταγράφηκαν περιπτώσεις κρουσμάτων σε άλογα στην Αίγυπτο αλλά και στην Γαλλία στο πρώτο μισό της δεκαετίας του 1960, ενώ κρούσματα επιβεβαιώθηκαν και σε άλλες χώρες της νότιας Ευρώπης, της νοτιοδυτικής Ασίας και της Αυστραλίας [39], [40]. Τέλος, το 1999 ο ιός μεταφέρθηκε και στην Νέα Υόρκη των Η.Π.Α., από όπου ξεκίνησε η εξάπλωσή του στις δύο αμερικανικές ηπείρους [41]. Έκτοτε έχουν καταγραφεί ξεσπάσματα σε ένα πλήθος από χώρες [42] όπως δείχνει η εικόνα 3.

Ο ιός του Δυτικού Νείλου εμφανίζεται κατά εποχιακά κύματα στις χώρες με εύκρατο κλίμα [3] τα οποία ξεκινάνε το Ιούλιο και τερματίζονται τον Οκτώβριο, ενώ σε χώρες με ιδιαίτερα υγρό ή ζεστό κλίμα αυτά μπορεί να διαρκούν περισσότερο [1]. Αυτό οφείλεται στο ότι οι υψηλότερες θερμοκρασίες επιτρέπουν τον ταχύτερο πολλαπλασιασμό του ιού στους φυσικούς ξενιστές [11]. Από την άλλη, περίοδοι ξηρασίας έχουν επίσης συσχετιστεί με αυξημένα κρούσματα την επόμενη χρονιά [43]. Αυτό αποδίδεται στην μείωση των πληθυσμών σε είδη που τρώνε τα κουνούπια, ή στην μετατόπιση των ενδιαιτημάτων διαφόρων πτηνών τα οποία για αυτόν τον λόγο έρχονται σε επαφή με πληθυσμούς μολυσμένων κουνουπιών. Στους ανθρώπους, τα άτομα κάθε ηλικίας έχουν ίδιες πιθανότητες να μολυνθούν από την νόσο αλλά τα άτομα μεγαλύτερων ηλικιών έχουν περισσότερες πιθανότητες να εμφανίσουν βαριά συμπτώματα [1]. Πέρα από την μόλυνση από φορέα, έχουν καταγραφεί περιπτώσεις μόλυνσης από μετάγγιση αίματος, μεταμόσχευση οργάνων και θηλασμό [35]. Ωστόσο αυτές οι περιπτώσεις είναι σπάνιες, και στην έρευνά μας δεν θα τις διακρίνουμε από τις υπόλοιπες.

Υπάρχει η εκτίμηση οτι η κλιματική αλλαγή αυξάνει την χωρική έκταση των τροπικών ασθενειών, και αυτό εκτιμάται ότι συμβαίνει και με τον WNV [11]. Οι προβλέψεις για τα καιρικά φαινόμενα υποδεικνύουν για το μέλλον συχνότερες πλημμύρες, ξηρασίες και αυξητική τάση της θερμοκρασίας [44], που με την σειρά τους μεταφράζονται σε αύξηση των πληθυσμών των κουνουπιών, και αλλαγή των χαρακτηριστικών που σχετίζονται με την τον βαθμό επικινδυνότητας του ιού όπως ο χρόνος επώασης, ο ρυθμός μετάλλαξης του ιού, η

αποδοτικότητα της μετάδοσης του ιού κ.α. [11]. Επομένως υπάρχουν φόβοι ότι η συγκεκριμένη ασθένεια θα αποτελεί μεγαλύτερη απειλή από ότι σήμερα στο μέλλον.



Εικόνα 3: Παγκόσμια κατανομή του WNV και των γενεαλογιών του. [42]

Κεφάλαιο 3. Η βιβλιοθήκη XGBoost

3.1 Γενικά στοιχεία

Ο αλγόριθμος eXtreme Gradient Boosting (XGBoost) παρουσιάστηκε για πρώτη φορά το 2014 σαν μία εκδοχή των Μηχανών Ενίσχυσης Κλίσης (Gradient Boosting Machines – GBM) που παρείχε βελτιστοποιήσεις ως προς την αποδοτικότητα των υπολογισμών αλλά και την επίδοση των αποτελεσμάτων. Έκτοτε έχει διακριθεί για τις πρωτιές που έχει επιτύχει σε μία πληθώρα διαγωνισμών του Kaggle [45], μίας πλατφόρμας για επαγγελματίες και μαθητές του τομέα της επιστήμης δεδομένων στην οποία συχνά φιλοξενούνται διαγωνισμοί μηχανικής μάθησης. Ακόμα αναγνωρίστηκε από το CERN ως η βέλτιστη προσέγγιση για την κατηγοριοποίηση σημάτων από τον Μεγάλο Επιταχυντή Αδρονίων χάρη στην εξαιρετική κλιμακωσιμότητά του [46]. Ανήκει στην κατηγορία των μεθόδων συνόλου (ensemble) και χρησιμοποιεί την μέθοδο καθόδου κλίσης για την βελτιστοποίηση της συνάρτησης απώλειας. Στην επόμενη ενότητα θα αναλύσουμε αυτές τις δύο τεχνικές και στην συνέχεια θα παραθέσουμε και τα ιδιαίτερα χαρακτηριστικά που τον διακρίνουν.

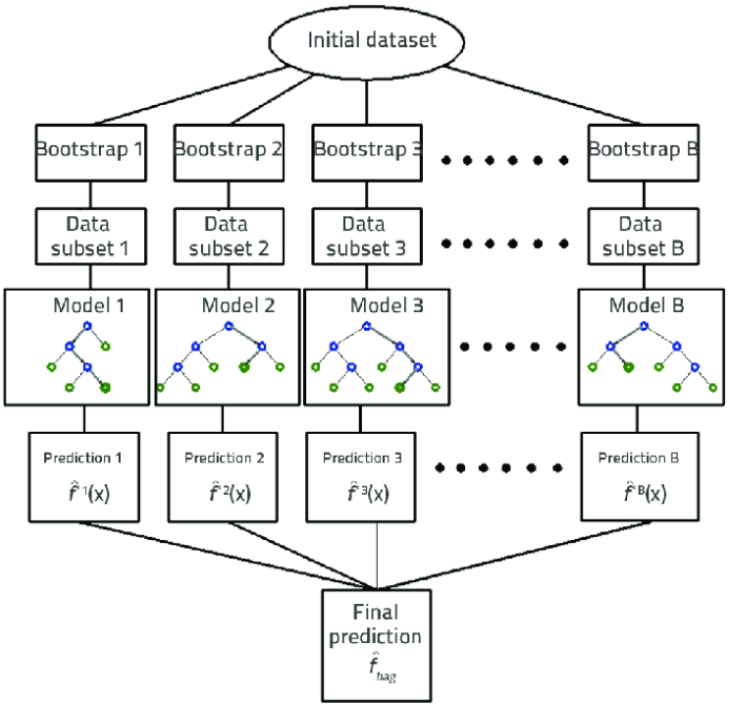
3.2 Gradient Boosting στον αλγόριθμο XGBoost

3.2.1 Μάθηση συνόλου

Ο αλγόριθμος XGBoost ανήκει στην κατηγορία των αλγορίθμων μάθησης συνόλου (ensemble learning). Η συγκεκριμένη προσέγγιση ξεκινάει από την παραδοχή ότι είναι πιθανό, ένα μοντέλο μηχανικής μάθησης να μην είναι σε θέση να προσφέρει αξιόπιστα αποτελέσματα από μόνο του. Είναι δυνατό τότε να συνδυάσουμε περισσότερους μαθητές (learners) για την δημιουργία ενός άλλου μοντέλου, η έξοδος του οποίου προκύπτει βάση των επιμέρους εξόδων των μοντέλων που αποτελούν το σύνολο, των λεγόμενων βασικών μαθητών (base learners). Οι βασικοί μαθητές συνήθως είναι διαφορετικές υποστάσεις του ίδιου μοντέλου, αν και μπορούν να είναι και διαφορετικά μοντέλα, και μαζί αποτελούν τον μαθητή συνόλου (ensemble learner). Έχουν αναπτυχθεί δύο διαφορετικές τεχνικές για μάθηση συνόλου, η ενίσχυση και η ενσάκιση.

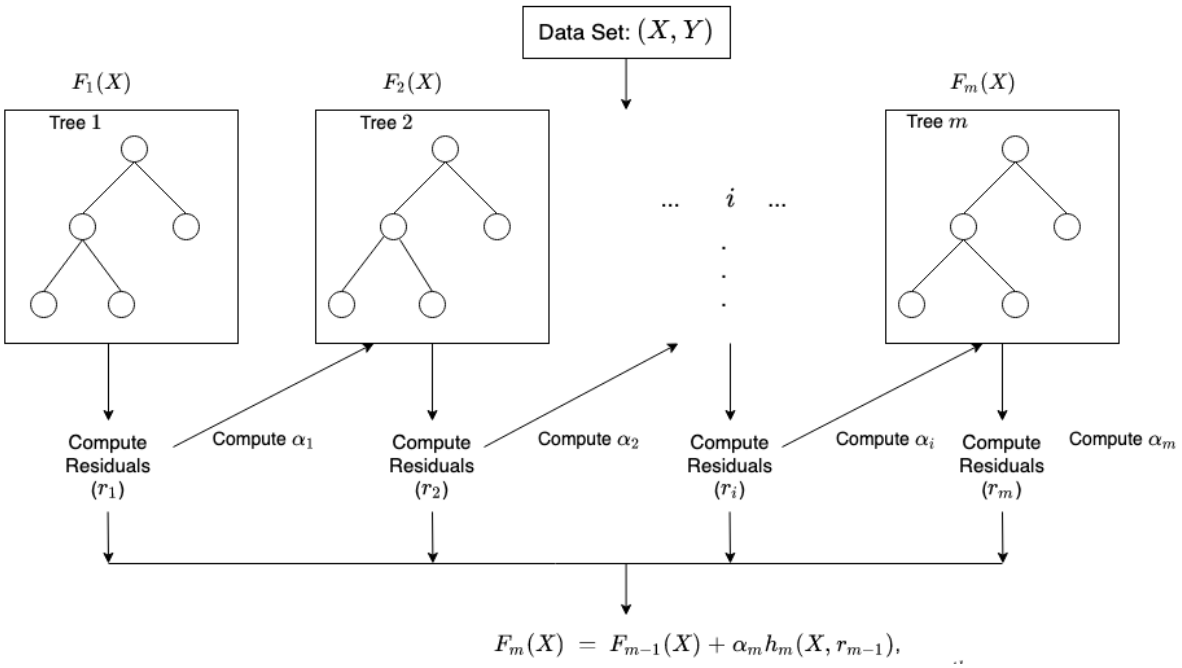
3.2.2 Ενίσχυση και Ενσάκιση

Η ενσάκιση συνήθως βασίζεται σε σύνολα από δένδρα αποφάσεων (decision trees). Ο λόγος είναι ότι τα δένδρα αποφάσεων, αν και συνήθως έχουν καλές επιδόσεις, στις περιπτώσεις που αστοχούν αυτό συμβαίνει σε σημαντικό βαθμό. Η επίδοση, που θα μπορούσε να αναφερθεί σαν 'βαθμός κατανόησης' του μοντέλου, ονομάζεται πόλωση (bias) ενώ ο βαθμός αστοχίας διακύμανση (variance). Έτσι μπορούμε να πούμε ότι τα δένδρα αποφάσεων χαρακτηρίζονται από χαμηλή πόλωση αλλά υψηλή διακύμανση. Στην πράξη παρατηρείται μία αντιστρόφως ανάλογη σχέση στα δύο. Ένας τρόπος να εκμεταλλευτούμε τις συνήθως καλές επιδόσεις αυτών των δέντρων και να μειώσουμε την σημασία των αστοχιών είναι να εκπαιδεύουμε παράλληλα ένα πλήθος από δέντρα, το καθένα σε ένα υποσύνολο των δεδομένων με επανατοποθέτηση, και να θεωρούμε σαν έξοδο του συνολικού μοντέλου τον μέσο όρο των επιμέρους εξόδων. Η εικόνα 4 παρουσιάζει αυτήν την τεχνική.



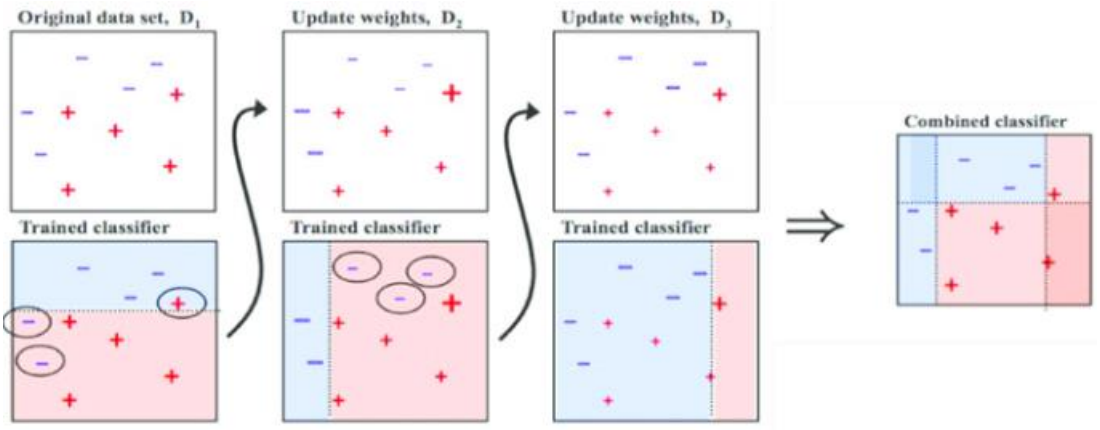
Εικόνα 4: Συνάθροιση αποτελεσμάτων σε σύνολο δέντρων αποφάσεων [58]

Αναφέραμε πως εμπειρικά παρατηρείται μία αντίστροφη σχέση ανάμεσα στην πόλωση και την διακύμανση. Η τεχνική της ενίσχυσης αποτελεί το αντίστροφό της ενσάκισης ως προς τις δύο αυτές ιδιότητες, δηλαδή αξιοποιεί μαθητές με υψηλή πόλωση αλλά χαμηλή διακύμανση με τέτοιο τρόπο ώστε το συνολικό μοντέλο να έχει χαμηλή πόλωση διατηρώντας την χαμηλή διακύμανση των βασικών μαθητών. Συγκεκριμένα, ξεκινάει με έναν βασικό μαθητή, ο οποίος είναι πάλι ένα δένδρο αποφάσεων αλλά με μικρό βάθος. Το μικρό αυτό μέγεθος σημαίνει ότι οι επιδόσεις του είναι οριακά καλύτερες από τα αποτελέσματα που θα είχαμε αν μαντεύαμε στην τύχη, ταυτόχρονα όμως, η γενικού μόνο επιπέδου πληροφορία που μαθαίνει εξασφαλίζει ότι η διακύμανση θα είναι μικρή. Στην συνέχεια προστίθενται επιπλέον μαθητές,



Εικόνα 5: Στην ενίσχυση τα μοντέλα προστίθενται διαδοχικά και το αποτέλεσμα είναι το άθροισμα των επιμέρους εξόδων [47].

σειριακά, και με τρόπο ώστε καθένας από αυτούς να βελτιώνει κάπως την συνολική επίδοση. Η έξοδος του μοντέλου είναι το άθροισμα όλων των βασικών μαθητών. Αν και η χρήση μικρών δέντρων προσφέρει υψηλή ερμηνευσιμότητα και εξασφαλίζει ότι το δέντρο δεν θα υπερπροσαρμοστεί, η δημιουργία μεγάλου πλήθους τέτοιων δέντρων μπορεί επίσης να οδηγήσει σε υπερπροσαρμογή, οπότε απαιτείται προσεκτική επιλογή κριτηρίων τερματισμού του αλγορίθμου. Στην εικόνα 5 εμφανίζεται η διαδικασία προσθήκης νέων μοντέλων και στην εικόνα 6 παρουσιάζονται η επίδραση του κάθε νέου βασικού μαθητή στα αποτελέσματα του συνολικού μοντέλου. Πρέπει να σημειώσουμε πως στην περίπτωση του XGBoost, τα επόμενα δέντρα μετά το πρώτο δεν εκπαιδεύονται στην πρόβλεψη της εξόδου, αλλά στην πρόβλεψη του υπολείμματος του μέχρι τότε συνόλου.



Εικόνα 6: Κάθε βασικός μαθητής που προστίθεται μειώνει λίγο από το συνολικό σφάλμα [48].

3.2.3 Αντικειμενική Συνάρτηση

Μετά την περιγραφή του γενικού τρόπου μάθησης των μοντέλων συνόλου μπορούμε να αναλύσουμε πως αυτός υλοποιείται στην πράξη [49]. Η τακτική που χρησιμοποιείται στα προβλήματα επιβλεπόμενης μάθησης συνίσταται στον ορισμό μίας αντικειμενικής συνάρτησης και στην βελτιστοποίησή της. Στην γενική περίπτωση η αντικειμενική συνάρτηση περιέχει δύο όρους, την απώλεια της εκπαίδευσης (training loss) και τον όρο κανονικοποίησης (regularization term), και λαμβάνει την μορφή:

$$obj(\theta) = L(\theta) + \Omega(\theta)$$

όπου L είναι η συνάρτηση απώλειας (loss function) της εκπαίδευσης και Ω όρος κανονικοποίησης. Οι συναρτήσεις απώλειας μετράνε το πόσο κοντά είναι η πρόβλεψη στην πραγματική τιμή. Έχουν προταθεί και χρησιμοποιούνται διάφορες συναρτήσεις, όπως το Άθροισμα των Τετραγώνων του Σφάλματος (Sum Squared Error - SSE), το Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error – MSE), η Ρίζα του Μέσου Τετραγωνικού Σφάλματος (Root Mean Squared Error – RMSE) κ.α.. Το μέσο τετραγωνικό σφάλμα (ΜΤΣ) χρησιμοποιείται αρκετά συχνά και δίνεται από την εξίσωση (1):

$$L(\theta) = \sum_{j} (y_i - \hat{y_i})^2$$

ενώ στην περίπτωση που πραγματοποιούμε λογιστική παλινδρόμηση (logistic regression) υπολογίζουμε την λογαριθμική απώλεια (log loss) μέσω της εξίσωσης (2):

$$L(\theta) = \sum_{i} \left[y_i \, ln\left(i + e^{-\hat{y_i}}\right) + (1 - y_i) ln\left(i + e^{\hat{y_i}}\right) \right]$$

όπου με y_i συμβολίζεται η πραγματική τιμή του i-οστού δείγματος και με $\hat{y_i}$ η πρόβλεψη του μοντέλου.

Ο όρος κανονικοποίησης λαμβάνει μία τιμή ανάλογη με την πολυπλοκότητα του μοντέλου. Αυτό σημαίνει ότι τα πιο πολύπλοκα μοντέλα αυξάνουν την τιμή της αντικειμενικής συνάρτησης ακριβώς όπως συμβαίνει με τα μοντέλα που έχουν χαμηλές προβλεπτικές επιδόσεις. Το αποτέλεσμα είναι κατά την προσθήκη νέων μοντέλων να τείνουν να προτιμώνται αυτά που παρέχουν έναν καλό συνδυασμό επιδόσεων και απλότητας, το οποίο είναι και το ζητούμενο, αφού οι βασικοί μαθητές θέλουμε να είναι απλά μοντέλα. Ο τρόπος με τον οποίον ορίζεται η πολυπλοκότητα ενός μοντέλου θα εξηγηθεί παρακάτω.

Στην παράγραφο των μοντέλων συνόλου αναφέρθηκε ότι η έξοδός τους προκύπτει ως άθροισμα των εξόδων των βασικών μαθητών. Έτσι, η πρόβλεψη $\hat{y}_i$ δίνεται από τον τύπο (3):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)\,, f_k \in \mathcal{F}$$

όπου K είναι το πλήθος των δέντρων, f_k είναι μία συνάρτηση του χώρου συναρτήσεων $\mathcal{F}$, και $\mathcal{F}$ είναι το σύνολο όλων των δυνατών δέντρων ταξινόμησης/κατηγοριοποίησης (Classification And Regression Trees – CART). Θεωρώντας $\omega(f_k)$ την πολυπλοκότητα του δέντρου f_k έχουμε την κατωτέρω (εξίσωση (4)) αντικειμενική συνάρτηση προς βελτιστοποίηση:

$$obj(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \omega(f_k)$$

3.2.4 Εκπαίδευση μέσω προσθηκών

Καθώς η εκπαίδευση συνίσταται στην προσθήκη του κατάλληλου δέντρου, το αντικείμενο της μάθησης δεν είναι παρά οι συναρτήσεις f_i που περιέχουν την δομή του δέντρου και τις τιμές των φύλλων. Αυτό είναι αρκετά δύσκολο να πραγματοποιηθεί γιατί στις συναρτήσεις αυτές δεν μπορεί να εφαρμοστεί η μέθοδος της καθόδου κλίσης (gradient descent) καθώς δεν είναι διαφορίσιμες. Επειδή δεν μπορούμε να μάθουμε όλα τα δέντρα σε μία επανάληψη, δρούμε μέσω διαδοχικών προσθηκών όπου η κάθε μία διορθώνει, σε κάποιον βαθμό, το σφάλμα της μέχρι τώρα μάθησης. Έτσι προκύπτουν οι συναρτήσεις:

$$\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\dots \\
\hat{y}_i^{(t)} &= \textstyle\sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned}$$

Αν συνδυάσουμε την αντικειμενική συνάρτηση στο βήμα t με τον τύπο του μέσου τετραγωνικού σφάλματος, δηλαδή τις εξισώσεις (1) και (4), προκύπτει η εξίσωση (5):

$$obj^{(t)} = \sum_{i=1}^{n} \left(y_i - \left(\hat{y}_i^{(t-1)} + f_t(x_i)\right)\right)^2 + \sum_{i=1}^{t} \omega(f_i)$$

$$= \sum_{i=1}^{n} \left[2\left(\hat{y}_i^{(t-1)} - y_i \right) f_t(x_i) + f_t(x_i)^2 \right] + \omega(f_t) + c$$

Σε αυτήν την μορφή το ΜΤΣ αναλύεται διακριτά σε έναν πρωτοβάθμιο όρο, ή αλλιώς το υπόλειμμα, και σε έναν τετραγωνικό όρο. Αντίθετα αν χρησιμοποιηθεί άλλη συνάρτηση απώλειας ο τύπος που προκύπτει δεν έχει τόσο καθαρή μορφή οπότε, παίρνοντας το ανάπτυγμα Taylor της συνάρτησης απώλειας μέχρι τον δευτεροβάθμιο όρο προκύπτει ο τύπος (6):

$$obj^{(t)} = \sum_{(i=1)}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)} \right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \omega(f_t) + c$$

Όπου τα g_i και h_i ορίζονται ως:

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)} \right)$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)} \right)$$

Αγνοώντας τις σταθερές η αντικειμενική συνάρτηση γίνεται στο βήμα t:

$$obj = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \omega(f_t) \qquad (7)$$

Το οποίο αποτελεί την νέα και αρκετά απλή αντικειμενική συνάρτηση προς βελτιστοποίηση. Το μεγάλο πλεονέκτημα αυτής της μορφής είναι ότι εξαρτάται μόνο από τους όρους g_i και h_i. Αυτό επιτρέπει την βελτιστοποίηση οποιασδήποτε συνάρτησης απώλειας προβαίνοντας στην εύρεση των προαναφερθέντων όρων και λύνοντας στην συνέχεια την εξίσωση (7).

3.2.5 Πολυπλοκότητα Δέντρου

Για να ολοκληρωθεί ο ορισμός της αντικειμενικής συνάρτησης πρέπει να ορίσουμε και την πολυπλοκότητα του δέντρου $\omega(f)$. Το ίδιο το δέντρο $f(x)$ ορίζεται ως

$$f_t(x) = w_{q(x)}, w \in R^T, q: R^d \to \{1,2,\dots,T\} \qquad (8)$$

Όπου w είναι το διάνυσμα με τις τιμές των φύλλων, q είναι η συνάρτηση που αναθέτει κάθε δείγμα στο φύλλο στο οποίο αντιστοιχεί, και T είναι το πλήθος των φύλλων. Έτσι στο XGBoost η πολυπλοκότητα δίνεται από τον τύπο (9):

$$\omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

Σημειώνεται ότι ο παραπάνω ορισμός είναι αυθαίρετος και μπορούν να οριστούν διαφορετικές συναρτήσεις για τον ίδιο σκοπό. Η παράμετρος γ ρυθμίζει την κανονικοποίηση και επιλέγεται από τον χρήστη με βάση τη εμπειρία από διαφορετικές εκτελέσεις του αλγορίθμου. Ο Chen την αναφέρει ως το «κόστος της πολυπλοκότητας λόγω της εισαγωγής επιπλέον φύλλου» [50]. Μεγαλύτερες τιμές επιβάλουν μεγαλύτερη ποινή ανάλογα με το T με αποτέλεσμα η επιλογή μεγαλύτερων δέντρων να εμποδίζεται περισσότερο.

Συνδυάζοντας τις εξισώσεις (7) (8) και (9) παίρνουμε την αντικειμενική συνάρτηση στον βήμα t ως:

$$obj^{(t)} \approx \sum_{i=1}^{n} \left[g_i w_{q(x_i)} + \frac{1}{2} h_i w_{q(x_i)}^2 \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

$$= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T$$

Όπου $I_j = \{i | q(x_i) = j\}$ είναι το σύνολο των δεικτών των δειγμάτων που αντιστοιχούν στο j-οστό φύλλο. Προκειμένου να συμπτύξουμε ακόμα την παραπάνω έκφραση ορίζουμε $G_i = \sum_{i \in I_j} g_i$ και $H_j = \sum_{i \in I_j} h_i$ οπότε προκύπτει:

$$obj^{(t)} = \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2} \left(H_j + \lambda \right) w_j^2 \right] + \gamma T \quad (10)$$

Με τα w_j να είναι ανεξάρτητα μεταξύ τους και το τμήμα μέσα στις αγκύλες να αποτελεί δευτεροβάθμια εξίσωση. Τα βέλτιστα w_j για δομή δέντρου $q(x)$ προκύπτουν με επίλυση της δευτεροβάθμιας:

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

και η ελάχιστη τιμή της αντικειμενικής:

$$obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \qquad (11)$$

όπου μικρότερες τιμές υποδεικνύουν καλύτερη δομή. Αυτό παρέχει έναν ολοκληρωμένο τρόπο αξιολόγησης ενός δέντρου. Επομένως σε κάθε βήμα μπορούμε να υπολογίσουμε τα σκορ όλων των πιθανών δέντρων και να επιλέξουμε το καλύτερο. Αυτό πρακτικά δεν είναι εφικτό οπότε επιχειρούμε να βελτιστοποιήσουμε ένα επίπεδο του δέντρου κάθε φορά. Αυτό μπορεί να γίνει υπολογίζοντας το κέρδος της κατάτμησης ενός φύλλου σε δύο νέα φύλλα:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \qquad (12)$$

το οποίο αποτελείται από το άθροισμα των σκορ στα δύο νέα φύλλα (δεξί και αριστερό) πλην το αρχικό, και αφαιρώντας ακόμα την σταθερά κανονικοποίησης. Το νόημα της παραπάνω εξίσωσης είναι αρκετά εύληπτο: η παράσταση στο εσωτερικό των αγκύλων εκτιμά αν η κατάτμηση του αρχικού φύλλου αυξάνει τις επιδόσεις του δέντρου και η αφαίρεση της σταθεράς κανονικοποίησης υποδεικνύει αν το κέρδος από αυτήν την κατάτμηση υπερβαίνει ένα κατώφλι. Αν η τελευταία συνθήκη δεν ισχύει η κατάτμηση δεν πραγματοποιείται. Το κέρδος αυτό υπολογίζεται για κάθε πιθανή κατάτμηση των ταξινομημένων δειγμάτων και επιλέγεται η βέλτιστη.

3.3 Βελτιστοποιήσεις και ιδιαίτερα χαρακτηριστικά

Τα προαναφερθέντα εξηγούν την ιδέα πίσω από τον τρόπο λειτουργίας του μοντέλου XGBoost. Αυτό, προσφέρεται στους χρήστες μέσω της βιβλιοθήκης XGBoost στην οποία είναι επιπλέον υλοποιημένες διάφορες τεχνικές που επιχειρούν να βελτιώσουν την απόδοση του μοντέλου ή να του προσθέσουν ευελιξία. Το σύνολο αυτών των δυνατοτήτων, που δηλώνονται με το επίθετο eXtreme, έχει βρεθεί στην πράξη να το διακρίνει σε σχέση με άλλα μοντέλα ως προς τις απαιτήσεις του σε χρόνο και την κλιμακωσιμότητά του [50]. Οι τεχνικές αυτές είναι [51]:

- **Προσεγγιστικός αλγόριθμος για την εύρεση σημείων κατάτμησης:** Ο διαχωρισμός ενός φύλλου σε δύο νέα πραγματοποιείται βάση μίας τιμής που αποτελεί το κατώφλι για τον διαχωρισμό των παρατηρήσεων. Η εύρεση αυτής της τιμής για συνεχόμενα χαρακτηριστικά προϋποθέτει την ταξινόμησή και μετακίνησή τους στην μνήμη, το οποίο για μεγάλα σετ δεδομένων μπορεί να μην είναι εφικτό. Στην βιβλιοθήκη περιλαμβάνεται προσεγγιστικός αλγόριθμος ο οποίος προτείνει τιμές με βάση τα εκατοστημόρια της κατανομής του χαρακτηριστικού. Τα συνεχή χαρακτηριστικά διακριτοποιούνται μέσω αντιστοίχισης στα διαστήματα που προκύπτουν από τις προτεινόμενες τιμές και η βέλτιστη τιμή επιλέγεται με βάση τα συγκεντρωτικά στατιστικά των διαστημάτων που προκύπτουν κάθε φορά.
- **Δομή τμημάτων για παράλληλη μάθηση:** για ταχύτερη επεξεργασία, το XGBoost αξιοποιεί πολλαπλούς πυρήνες του επεξεργαστή. Αυτό είναι δυνατό χάρη στο ότι έχει σχεδιαστεί να αξιοποιεί τις δομές τμημάτων (τμήματα – blocks στην μνήμη ονομάζονται υποσύνολα αυτής που αποτελούνται από διαδοχικές θέσεις μνήμης και επιτρέπουν την αποθήκευση πληροφορίας ‘συνεχόμενα’. Όταν αυτό πραγματοποιείται, οι λειτουργίες της μνήμης εκτελούνται ταχύτερα). Τα δεδομένα ταξινομούνται και αποθηκεύονται στα τμήματα, πράγμα που επιτρέπει την επαναχρησιμοποίησή τους σε μετέπειτα επαναλήψεις, αντί αυτά να επαναϋπολογίζονται. Αυτό βελτιώνει και τον παραπάνω προσεγγιστικό αλγόριθμο αλλά και την υποδειγματοληψία των στηλών.
- **Βεβαρυμμένη σκιαγράφηση ποσοστών για προσεγγιστική μάθηση:** Οι περισσότεροι υπάρχοντες αλγόριθμοι εύρεσης σημείων διαχωρισμού δουλεύουν για δεδομένα χωρίς βάρη (ή αλλιώς, για μοναδιαία βάρη) μέσω ενός αλγορίθμου σκιαγράφησης τεταρτημόριων. Ο XGBoost χρησιμοποιεί μία εναλλακτική αυτού που λειτουργεί και για δεδομένα με βάρη, και μπορεί να εκτελεστεί κατανεμημένα.
- **Κανονικοποίηση:** Παρέχεται η δυνατότητα της υποπροτεραιοποίησης των πιο περίπλοκων μοντέλων μέσω L1 ή και L2 κανονικοποίησης, για την πρόληψη της υπερπροσαρμογής.
- **Διαχείριση αραιών δεδομένων:** Απουσιάζουσες τιμές και τεχνικές προεπεξεργασίας όπως η κωδικοποίηση ανά μία τιμή (one-hot encoding) δημιουργούν κενά ανάμεσα στα χρήσιμα δεδομένα. Ο αλγόριθμος για την εύρεση σημείων διαχωρισμού που χρησιμοποιεί ο XGBoost διαχειρίζεται μοτίβα τέτοιων κενών με τρόπο που αυξάνει της αποδοτικότητά του.

- **Επίγνωση κρυφής μνήμης:** Κατά την εκτέλεση του XGBoost απαιτούνται προσβάσεις στις θέσεις της μνήμης όπου βρίσκονται τα στατιστικά των κλίσεων οι οποίες δεν είναι συνεχόμενες. Μέσω της παραχώρησης εσωτερικών απομονωτών στο κάθε νήμα και αποθήκευσης των στατιστικών σε αυτούς επιτυγχάνεται καλύτερη αξιοποίηση του υλικού.
- **Επεξεργασία εκτός πυρήνα:** Για δεδομένα που δεν χωράνε στην κύρια μνήμη, πραγματοποιείται καταμερισμός αυτών σε πολλαπλά τμήματα και αποθήκευση αυτών στον δίσκο. Όταν αυτά ζητηθούν από τον αλγόριθμο ένα ξεχωριστό νήμα αναλαμβάνει την αποσυμπίεσή και εγγραφή τους στην μνήμη

Εκτός από τις παραπάνω λειτουργίες, υπάρχουν ορισμένα πιο απλά χαρακτηριστικά που προσφέρουν στον αλγόριθμο κάποιο πλεονέκτημα. Αυτά είναι [52]:

- **Ευελιξία στην επιλογή των κριτηρίων αξιολόγησης:** Όπως αναφέραμε έχουν οριστεί διάφορες συναρτήσεις απώλειας. Η υλοποίηση του αλγορίθμου επιτρέπει την επιλογή οποιασδήποτε από αυτές, και ακόμα επιτρέπει τον ορισμό νέων κριτηρίων αξιολόγησης.
- **Κλάδεμα δέντρων:** Σε άλλους αλγορίθμους ενίσχυσης κλίσης η ανάπτυξη του δέντρου σταματάει όταν ένας διαχωρισμός επιφέρει αρνητικό κέρδος. Στο XGBoost δημιουργούνται πρώτα όλα τα φύλλα που επιτρέπονται με βάση το ορισμένο μέγεθος του δέντρου και στην συνέχεια αφαιρούνται, από κάτω προς τα πάνω, όσα έχουν αρνητικό κέρδος. Το αποτέλεσμα είναι ότι με αυτόν τον τρόπο ανακαλύπτονται διαχωρισμοί που αρχικά επιφέρουν αρνητικό κέρδος αλλά θέτουν την βάση για διαχωρισμό με μεγαλύτερο κέρδος παρακάτω.
- **Ενσωματωμένος μηχανισμός διασταυρούμενης επικύρωσης:** Υπάρχει η δυνατότητα να εκτελείται διασταυρούμενη επικύρωση σε κάθε επανάληψη, το οποίο διευκολύνει την εύρεση του βέλτιστου αριθμού επαναλήψεων σε κάθε τρέξιμο.

3.4 Υπερπαράμετροι

Οι υπερπαράμετροι του μοντέλου κατευθύνουν τον τρόπο λειτουργίας του και είναι κρίσιμες για την τελική επίδοσή του. μπορούν να χωριστούν σε τρεις κατηγορίες

- Γενικές υπερπαράμετροι που επηρεάζουν της συνολική λειτουργία

- Υπερπαράμετροι ενίσχυσης που κατευθύνουν την ανάπτυξη των βασικών μαθητών
- Υπερπαράμετροι μάθησης που ορίζουν τον τρόπο που πραγματοποιείται η εκπαίδευση και η αξιολόγηση

Γενικές υπερπαράμετροι:

- **booster:** πρόκειται για το είδος των βασικών μαθητών. Αυτοί μπορούν να είναι, είτε δεντρικά μοντέλα, είτε γραμμικά, ωστόσο στην πράξη σχεδόν πάντα χρησιμοποιούνται τα πρώτα καθώς υπερέχουν ως προς τις αποδόσεις.
- **silent:** αφορά στην εκτύπωση μηνυμάτων εκτέλεσης
- **nthread:** πλήθος νημάτων που θέλουμε να χρησιμοποιήσει ο αλγόριθμος για παράλληλη επεξεργασία

Υπερπαράμετροι ενίσχυσης:

- **eta:** αντίστοιχο του ρυθμού μάθησης που γνωρίζουμε από άλλα μοντέλα
- **min_child_weight:** το ελάχιστο συνολικό άθροισμα των βαρών των παρατηρήσεων που απαιτείται σε ένα παιδί. Μεγαλύτερες τιμές προλαμβάνουν την υπερπροσαρμογή διότι εμποδίζουν την εκμάθηση σχέσεων που ενδεχομένως είναι ειδικές σε συγκεκριμένα δείγματα
- **max_depth:** είναι το μέγιστο βάθος που μπορεί να έχει ένα δέντρο. Στην πράξη όλα τα δέντρα αναπτύσσονται μέχρι αυτό και στην συνέχεια κλαδεύονται, όπως αναφέραμε.
- **max_leaf_nodes:** το μέγιστο πλήθος φύλλων είναι ένας άλλος τρόπος να οριστεί το μέγιστο βάθος καθώς τα δέντρα είναι δυαδικά
- **max_delta_step:** πρόκειται για το μέγιστο επιτρεπτό βάρος στο κάθε δέντρο. Σε περιπτώσεις κατηγοριοποίησης που η μία κλάση είναι υπερβολικά μεγαλύτερη από τις άλλες, το βάρος ενός δέντρου μπορεί να γίνει αρκετά μεγάλο ώστε παρά τον πολλαπλασιασμό του με τον ρυθμό μάθησης, το γινόμενο τους να κυριαρχεί στην συνολική συνεισφορά. Με αυτήν την παράμετρο θέτουμε ένα άνω όριο σε αυτήν.
- **subsample:** εισάγει στοχαστικότητα στην εκπαίδευση υποδειγματοληπτόντας τις παρατηρήσεις
- **colsample_bytree:** αντίστοιχα με το subsample αλλά για χαρακτηριστικά
- **colsample_bylevel:** αντίστοιχα με το colsample_bytree αλλα για το κάθε επίπεδο του δέντρου

- **lambda:** πρόκειται για όρο L2 κανονικοποίησης στα βάρη (βλ. τύπο 9 παρ. 3.2.3). Επηρεάζει την κανονικοποίηση
- **alpha:** όρος L1 κανονικοποίησης στα βάρη. Χρησιμοποιείται σε περιπτώσεις υψηλής διαστατικότητας.
- **scale_pos_weight:** επιταχύνει την σύγκλιση σε περιπτώσεις ανισορροπίας κλάσεων.

Υπερπαράμετροι μάθησης:

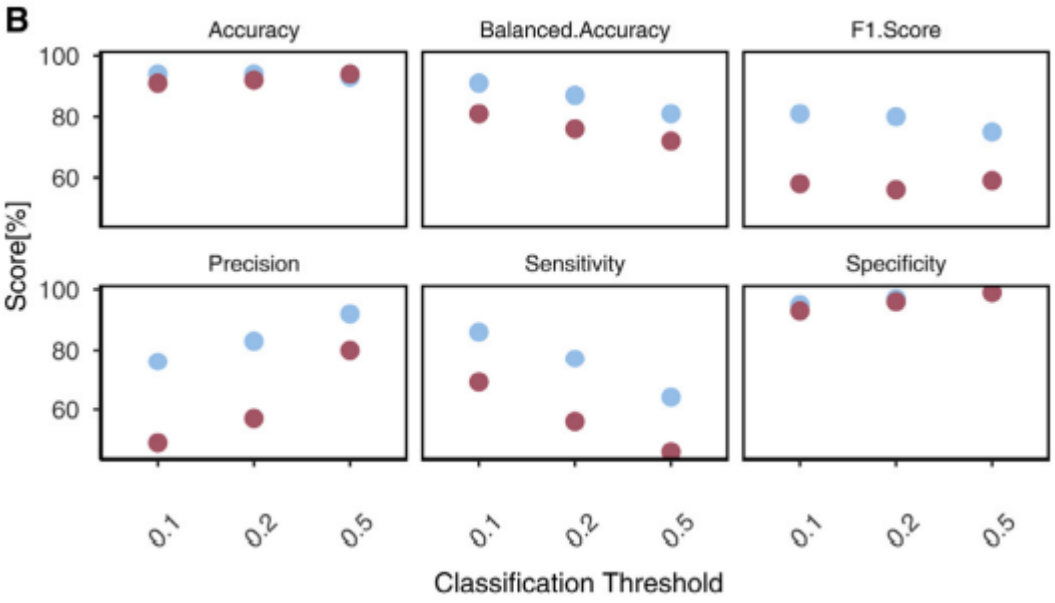
- **objective:** ορίζει την συνάρτηση απώλειας προς ελαχιστοποίηση. Συνήθεις τιμές είναι
  - **binary:logistic** - λογιστική παλινδρόμηση για δυαδική κατηγοριοποίηση. Επιστρέφει την προβλεπόμενη πιθανότητα
  - **multi:softmax** - κατηγοριοποίηση πολλαπλών κλάσεων
  - **multi:softprob** - αντίστοιχο με το παραπάνω αλλά επιστρέφει την προβλεπόμενη πιθανότητα κάθε δείγμα να ανήκει σε κάθε μία από τις κλάσεις.
- **eval_metric:** Πρόκειται για την μετρική που θα χρησιμοποιηθεί στα δεδομένα επικύρωσης. Οι προεπιλεγμένες τιμές είναι RMSE για παλινδρόμηση και Error για κατηγοριοποίηση. Δυνατές τιμές είναι:
  - **rmse** – ρίζα μέσου τετραγωνικού σφάλματος
  - **mae –** μέσο απόλυτο σφάλμα
  - **logloss –** αρνητική log-πιθανότητα
  - **error –** ρυθμός σφάλματος δυαδικής κατηγοριοποίησης
  - **merror –** error για πολλαπλές κατηγορίες
  - **mlogloss –** logloss για πολλαπλές κατηγορίες
  - **auc –** εμβαδό υπό την καμπύλη
- **seed:** το φύτρο για την παραγωγή τυχαίων αριθμών.

Κεφάλαιο 4. Βιβλιογραφική μελέτη

Η αφορμή για την παρούσα εργασία είναι η ανάγκη για ένα προβλεπτικό μοντέλο για τα ξεσπάσματα του ιού του Δυτικού Νείλου. Για την πραγματοποίηση αυτού απαιτείται αφενός ένα σύνολο δεδομένων που θεωρούμε ότι θα περιέχουν πληροφορία σχετική με αυτά τα ξεσπάσματα, τα οποία στην συνέχεια μπορούμε να τα αναφέρουμε και σαν στόχο, και αφετέρου ένα μοντέλο μηχανικής μάθησης που θα αξιοποιήσει αυτά τα δεδομένα. Όπως αναφέραμε στο κεφάλαιο 2 υπάρχει θεωρητική και εμπειρική γνώση που υποδεικνύει εξάρτηση του στόχου μας από τις κλιματικές συνθήκες, ωστόσο δεν είναι προφανές ποιες ακριβώς παράμετροι είναι χρήσιμες σε ένα τέτοιο πρόβλημα, ενώ ακόμα είναι πιθανό να υπάρχουν και άλλα χαρακτηριστικά που επηρεάζουν το συνολικό αποτέλεσμα όπως κοινωνικοί και οικονομικοί παράγοντες, πολεοδομία κλπ. Αντίστοιχα στο κομμάτι του μοντέλου, έχει αναπτυχθεί ένα πλήθος αλγορίθμων για διάφορες υποπεριοχές του τομέα της τεχνητής νοημοσύνης, χωρίς όμως αυτό να αποκλείει έναν αλγόριθμο από το να έχει καλές επιδόσεις σε περισσότερες από μία υποπεριοχές, όπως τα συνελικτικά νευρωνικά δίκτυα (Convolutional neural networks – CNNs) που αποδίδουν καλά και σε προβλήματα όρασης αλλά και σε προβλήματα χρονοσειρών.

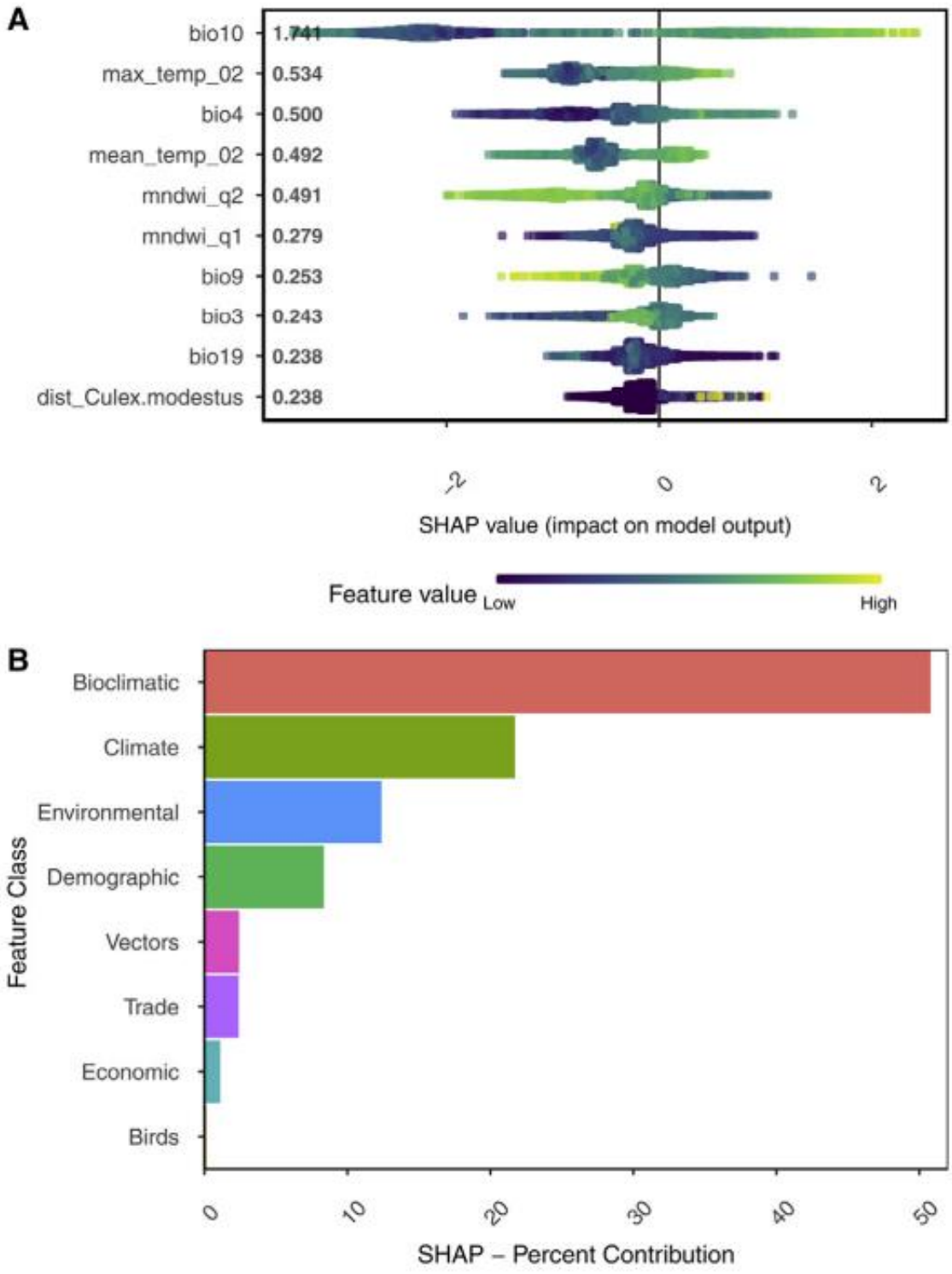
Σαν πρώτο βήμα λοιπόν πραγματοποιήθηκε μία μελέτη της υπάρχουσας βιβλιογραφίας με σκοπό την εξακρίβωση των δεδομένων που χρειάζεται να συλλεχθούν, των μοντέλων τα οποία αποδίδουν καλά σε παρόμοια προβλήματα, αλλά και των επιδόσεων που είναι ρεαλιστικά επιτεύξιμες. Συγκεκριμένα αναζητήσαμε στις ιστοσελίδες IEEE explore, Science Direct και PubMed, έρευνες σχετικές με τις λέξεις-κλειδιά “West Nile virus” και “Prediction”, περιορίζοντας τα αποτελέσματά μας από το έτος 2005 και έπειτα, και επιλέξαμε τις 10 πρώτες από κάθε ιστοσελίδα. Αρκετά αποτελέσματα αφορούσαν στην πρόβλεψη της εξέλιξης άλλων διαδικασιών όπως της εξέλιξης της νόσου σε ασθενείς, στην αλληλεπίδρασης των πρωτεϊνών ανθρώπου-ιού κ.α. οπότε απορρίφθηκαν. Ακόμα, από τα 30 συνολικά αποτελέσματα, πολλά αποτελούσαν επανεμφάνιση της ίδιας έρευνας σε άλλη ιστοσελίδα. Τελικά προέκυψαν 5 έρευνες με αντικείμενο αντίστοιχο του δικού μας, όπως την πρόβλεψη του WNV σε χώρες την νότιας Ευρώπης ή σε πολιτείες των ΗΠΑ.

Η παρούσα εργασία βασίζεται κυρίως στην έρευνα των Farooq et al. [53] που δημοσιεύθηκε τον Ιούνιο το 2022. Σκοπός αυτής την εργασίας ήταν η επιβεβαίωση του ρόλου των κλιματικών παραμέτρων στις εξάρσεις της ασθένειας αλλά και η πρόβλεψη επερχόμενων



Εικόνα 7: Μετρικές επίδοσης για διαφορετικές τιμές κατωφλίου [53].

εξάρσεων. Κατά τα συμπεράσματα των συγγραφέων, αν και υπάρχουν εργασίες πάνω στην συσχέτιση του ιού με το κλίμα για ορισμένες χώρες, το πλήθος των εργασιών που αξιοποιούν κλιματικά δεδομένα υψηλότερης χρονικής ανάλυσης και αλγορίθμους τεχνητής νοημοσύνης είναι περιορισμένο, πράγμα που συμφωνεί με τα δικά μας ευρήματα και καταδεικνύει την σημασία αυτής της μελέτης. Τα χαρακτηριστικά που χρησιμοποίησαν ήταν η μέγιστη και η μέση τιμή της θερμοκρασίας και του υετού, δείκτες της χλωρίδας και της διαθεσιμότητας του νερού, η κατανομή των ηλικιών του ανδρικού και του γυναικείου πληθυσμού, το μέσο ετήσιο εισόδημα και οι μετακινήσεις εμπορικών αγαθών, οι πληθυσμοί των Culex Pipiens, Culex modestus και των πτηνών, καθώς και τα κρούσματα του ιού στους ανθρώπους. Σαν μοντέλο επιλέχθηκε το XGBoost λόγω τις επίδοσής του σε προβλήματα κατηγοριοποίησης και επιπλέον χρησιμοποιήθηκε το εργαλείο SHAP (SHAPley Additive exPlanations) που υπολογίζει την συνεισφορά κάθε χαρακτηριστικού στην έξοδο του μοντέλου. Τα πειράματα έδειξαν υψηλή διακριτική ικανότητα του μοντέλου ανάμεσα στις δύο κλάσεις όπως φαίνεται στην εικόνα 7. Επιπλέον το εργαλείο SHAP ξεχώρισε τις θερμοκρασιακές ανωμαλίες της άνοιξης και δείκτες που σχετίζονται με την διαθεσιμότητα του νερού στα πιο χρήσιμα χαρακτηριστικά. Οι πληθυσμοί των κουνουπιών επίσης φάνηκαν καθοριστικοί για το ιστορικό ξέσπασμα του 2018, ενώ τα δημογραφικά χαρακτηριστικά και οι οικονομικοί δείκτες αξιολογήθηκαν χαμηλά (εικόνα 8). Η χωρική ανάλυση ήταν σε NUTS3 (Nomenclature of territorial units for statistics) που αντιστοιχούν σε επίπεδο περιφέρειας.



Εικόνα 8: Αξιολόγηση σημαντικότητας χαρακτηριστικού σύμφωνα με το SHAP [53]

Σε μία εργασία με αντίστοιχο αντικείμενο, οι Ajith et al. [54] συγκρίνουν τον ταξινομητή τυχαίου δάσους (Random Forest Classifier – RFC), τον αφελή Μπεϊζιανό ταξινομητή (naive Bayes Classifier) και το μοντέλο προσαρμοστικής ενίσχυσης (Adaptive Boost – AdaBoost).

Προκύπτει πως ο RFC επιτυγχάνει τα καλύτερα αποτελέσματα με τον AdaBoost να έρχεται οριακά δεύτερος. Καθώς ο XGBoost στην πλειοψηφία των περιπτώσεων υπερτερεί του RFC [55], οδηγούμαστε και πάλι στην επιλογή του πρώτου.

Οι Keyel et al. σε παρόμοια έρευνα για την περιοχή της Νέας Υόρκης και του Κονέκτικατ καταλήγουν βρίσκουν επίσης παραμέτρους σχετικές με την θερμοκρασία, την υγρασία και τον πληθυσμό των κουνουπιών να έχουν τον μεγαλύτερο αντίκτυπο στην ένταση των ξεσπασμάτων, ωστόσο παρατηρούν διάφορα σφάλματα στις προβλέψεις του μοντέλου το οποίο αποδίδουν στην πολυπλοκότητα των εμπλεκόμενων στο πρόβλημα μηχανισμών.

Κεφάλαιο 5. Η προσέγγισή μας

5.1. Εργαλεία

Η συγγραφή του κώδικα πραγματοποιήθηκε εξ ολοκλήρου σε γλώσσα Python έκδοση 3.10.2 στο περιβάλλον του Microsoft Visual Studio Code και συγκεκριμένα σε Jupyter notebook. Χρησιμοποιήθηκαν οι βιβλιοθήκες:

- random για την δημιουργία του τυχαίου ταξινομητή,
- pandas για επεξεργασία και διαχείριση των δεδομένων,
- sns και matplotlib για την δημιουργία γραφημάτων
- xgboost για την χρήση του XGBoost Classifier
- RocCurveDisplay του πακέτου sklearn.metrics για την δημιουργία και εμφάνιση των Receiver Operating Characteristics (ROC) γραφημάτων

Ακόμα χρησιμοποιήθηκαν οι συναρτήσεις του πακέτου sklearn.metrics:

- roc_auc_score για τον υπολογισμό και την εμφάνιση των αντίστοιχων σκορ
- classification_report για τον υπολογισμό και την εμφάνιση των μετρικών της ακρίβειας και ανάκλησης
- confusion matrix για τον υπολογισμό και την εμφάνιση του πίνακα σύγχυσης.

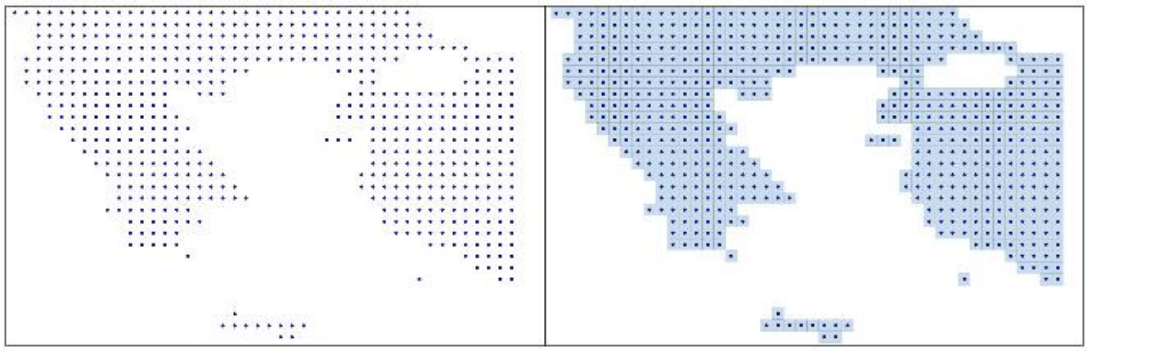
5.2. Συλλογή δεδομένων

Η μελέτη μας επικεντρώνεται στην Ελλάδα, οπότε μετά το πέρας της βιβλιογραφικής μελέτης αναζητήσαμε πηγές οπού να διατίθενται ιστορικές καταγραφές κλιματικών μετρήσεων. Συγκεκριμένα αξιοποιήθηκε η πέμπτη γενιά ατμοσφαιρικών αναλύσεων ERA-5 του Ευρωπαϊκού Κέντρου Πρόγνωσης (European Centre for Medium-Range Weather Forecasts – ECMWF) για το παγκόσμιο κλίμα. Εκεί διατίθενται ημερήσιοι μέσοι όροι 10 μεταβλητών για την περιοχή της Ελλάδας για το διάστημα 01/01/2010 με 31/12/2020 οι οποίες είναι:

- Θερμοκρασία επιφάνειας σε ύψος 2m (°C)
- Σημείο δρόσου σε ύψος 2m (°C)
- Βροχόπτωση (mm)
- Σχετική υγρασία (%)
- Ειδική υγρασία (g/kg)

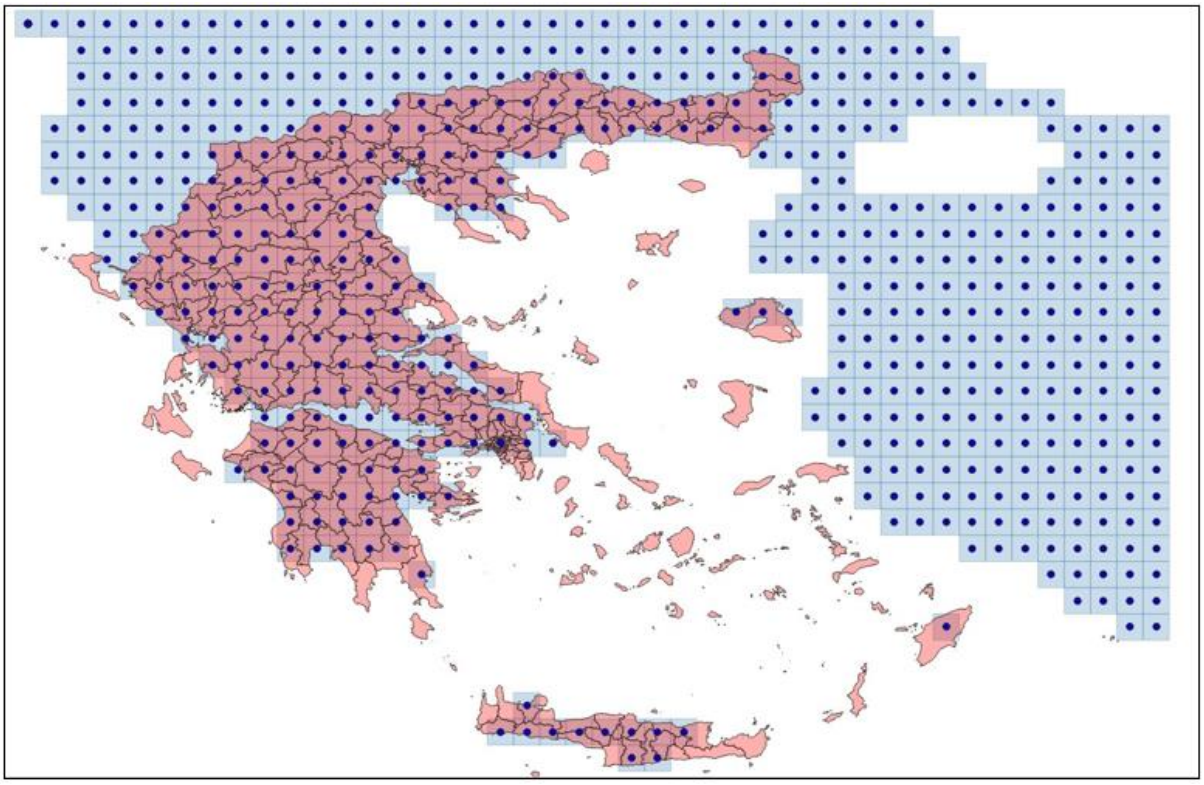
- Πίεση στην μέση στάθμη της θάλασσας (hPa)
- Ταχύτητα ανέμου σε ύψος 10m (m/s)
- Κατεύθυνση ανέμου σε ύψος 10m (deg)
- Θερμοκρασία εδάφους σε βάθος από 0-7cm (°C)
- Ογκομετρικό νερό εδάφους σε βάθος από 0-7cm (°C)

Τα παραπάνω δεδομένα βρίσκονται σε ανάλυση 0,25 x 0,25 δεκαδικές μοίρες (DD) το οποίο αντιστοιχεί σε χωρική γεωγραφική κατανομή 584 σημείων



Εικόνα 9: Σημειακά δεδομένα για την περιοχή της Ελλάδας από το ECMWF

Προκειμένου να εξαχθούν από αυτά οι τιμές των χαρακτηριστικών σε επίπεδο δήμου έγινε χρήση διανυσματικού τύπου .shp Καλλικρατικών ΟΤΑ από τα Ψηφιακά Χαρτογραφικά Υπόβαθρα της ΕΛΣΤΑΤ και στην συνέχεια εκτελέστηκαν υπολογισμοί εμβαδού για κάθε πολύγωνο ΟΤΑ (σύνολο 326)



Εικόνα 10: Υπέρθεση δημοτικών ορίων και πλακιδίων για την περιοχή της Ελλάδας

Ακολούθησε:

- ενοποίηση (Union) των επίπεδων πλακιδίων (tiles) και ΟΤΑ (δήμος)
- υπολογισμός του συντελεστή επικάλυψης ΟΤΑ ανά πλακίδιο
- πολλαπλασιασμός συντελεστή επικάλυψης ΟΤΑ με τις τιμές των μεταβλητών των πλακιδίων που τέμνει

και τελικά προέκυψαν 260 εγγραφές με σταθμισμένες περιβαλλοντικές μεταβλητές

Στην συνέχεια προστέθηκαν οι δείκτες κανονικοποιημένης διαφοράς βλάστησης (Normalized Difference Vegetation Index – NDVI) και νερού (Normalized Difference Water Index) που περιγράφουν ποσοτικοποιημένα τις διαχρονικές τοπικές αλλαγές στην παρουσία της βλάστησης και του νερού αντίστοιχα. Αυτοί υπάρχουν διαθέσιμοι στο Earth Engine Data Catalog της Google και για τον υπολογισμό των δεικτών αυτών ανά ΟΤΑ χρησιμοποιήθηκε η συνάρτηση .zonal_statistics του πακέτου geemap.

Τέλος, τα δεδομένα που αφορούν τα κρούσματα του κάθε δήμου για την κάθε ημέρα αντλήθηκαν από τα αρχεία του Εθνικού Οργανισμού Δημόσιας Υγείας (ΕΟΔΥ)

Σημειώνουμε ότι οι ενέργειες της παραγράφου 5.1. πραγματοποιήθηκαν από τα μέλη της ομάδας του προγράμματος ΕΜΠΡΟΣ Μιχάλη Κουρέα και Γεώργιο Χαρβαλή.

5.3. Προεπεξεργασία δεδομένων

Λανθασμένες – απουσιάζουσες τιμές: Ανάμεσα στην απόκτηση και στην χρησιμοποίηση των δεδομένων μεσολαβεί το στάδιο της προεπεξεργασίας αυτών. Συχνά συμβαίνει κάποιο σφάλμα στην μέτρηση, στην καταγραφή ή στην μετατροπή μίας τιμής. Αυτά δεν είναι πάντα αντιληπτά αλλά όταν υπερβαίνουν το φυσιολογικό εύρος ή ακολουθούν μη φυσιολογική κατανομή μπορούν να ανιχνευθούν μέσω της σύνοψης 5 αριθμών (five numbers summary) η οποία επιστρέφει την τιμή του ελάχιστου, του ορίου του πρώτου τεταρτημόριου, του διάμεσου, του ορίου του τρίτου τεταρτημόριου και του μεγίστου. Έτσι μπορούν να εντοπιστούν και να απορριφθούν άμεσα έκτοπες τιμές. Ακόμα είναι πιθανό κάποιες τιμές να λείπουν εξ ολοκλήρου το οποίο μπορεί να δημιουργήσει πρόβλημα στον αλγόριθμο. Στην δική μας περίπτωση τα δεδομένα παρέχονταν σε μία ήδη καθαρισμένη μορφή οπότε δεν παρατηρήθηκαν τέτοια προβλήματα. Εξαίρεση αποτελούν τα δεδομένα της μεταβλητής NDVI, (που προήλθαν από το Google Earth Engine Data Catalog) στα οποία υπήρχαν απουσιάζουσες τιμές. Ωστόσο αυτές αντιστοιχούσαν σε μήνες που δεν ήταν σημαντικοί για την εκπαίδευση του μοντέλου. Ακόμα, ο αλγόριθμος XGB έχει ενσωματωμένους μηχανισμούς για την βέλτιστη αντιμετώπιση τέτοιων περιπτώσεων, οι οποίοι μαθαίνουν τις τιμές που επιφέρουν τις καλύτερες επιδόσεις και τοποθετούν αυτές στην θέση τους [50].

Παράγωγα χαρακτηριστικά: Εκτός από την απόλυτη τιμή των μεταβλητών συχνά έχουν σημασία και άλλα μεγέθη όπως η μέση τιμή, η τυπική απόκλιση ή η κανονικοποιημένη ανωμαλία διότι περιέχουν πληροφορία για το πόσο εκτός φυσιολογικού είναι μία συγκεκριμένη τιμή. Ασυνήθιστα εκδηλώσεις φαινομένων συχνά προκύπτουν σαν αποτέλεσμα ασυνήθιστων τιμών των μεταβλητών που τα καθορίζουν, όπως είναι λογικό, οπότε είναι χρήσιμο να γνωρίζουμε κατά πόσο μία τιμή είναι τυπική ή όχι. Για τον λόγο αυτό υπολογίσαμε και προσθέσαμε στο φύλλο δεδομένων μας την μέση τιμή και την κανονικοποιημένη ανωμαλία του κάθε χαρακτηριστικού.

Χρονική ανάλυση: Τα δεδομένα παρέχονται από τις πηγές που αξιοποιήσαμε σε επίπεδο ημέρας. Ωστόσο παρατηρήθηκε στις πρώτες δοκιμές πως στην πράξη είναι δύσκολο να επιτευχθεί τέτοια ακρίβεια, ενώ η ανισορροπία των κλάσεων είναι τεράστια, καθώς οι μέρες με καταγεγραμμένα κρούσματα είναι ελάχιστες. Για τον λόγο αυτό υπολογίστηκαν οι

εβδομαδιαίοι μέσοι όροι κάθε ανεξάρτητης μεταβλητής, καθώς και το εβδομαδιαίο άθροισμα της εξαρτημένης και η έρευνα πραγματοποιήθηκε σε αυτήν την χρονική ανάλυση

Τα παραπάνω πραγματοποιηθήκαν μία φορά στα ανεπεξέργαστα δεδομένα ώστε να προκύψουν τα δεδομένα που θα αποτελέσουν την βάση στην συνέχεια. Αυτά, πριν χρησιμοποιηθούν από το μοντέλο υφίστανται μερικά ακόμα στάδια παραμετροποιημένης επεξεργασίας. Κάθε στάδιο επιχειρεί να βελτιώσει τις επιδώσεις οπότε δοκιμάζεται επανειλημμένα για διάφορες τιμές παραμέτρων.

Επιλογή υποσυνόλου δήμων: Από το άθροισμα των κρουσμάτων σε επίπεδο περιφέρειας που παρουσιάζεται στην εικόνα 11 εξακριβώνεται ότι αυτές αποτελούν διαφορετικές περιπτώσεις. Από τις δεκατρείς στο σύνολο περιφέρειες βλέπουμε πως αυτή της Κεντρικής Μακεδονίας έχει τον υψηλότερο αριθμό συνολικών κρουσμάτων φτάνοντας τα ~630, οι περιφέρειες Αττικής και Ανατολικής Μακεδονίας – Θράκης έχουν μικρότερο, κοντά στα 270, ενώ η περιφέρεια Θεσσαλίας μόλις ξεπερνάει τα 100. Ο ακριβής τρόπος με τον οποίο το πλήθος των κρουσμάτων χαρακτηρίζεται ως αρκετά ή όχι έχει να κάνει και με την ισορροπία των κλάσεων αλλά και με την ανάγκη για ύπαρξη ικανού αριθμού δειγμάτων στην κάθε μία, και θα εξηγηθεί περισσότερο στην συνέχεια. Για την ώρα αναφέρουμε ότι οι περιφέρειες Πελοποννήσου και όσες βρίσκονται έπειτα από αυτήν (σύμφωνα με την κατάταξη του πίνακα της εικόνας 9) δεν μπορούν να αξιοποιηθούν κατά την εκπαίδευση ή αξιολόγηση του μοντέλου, πράγμα το οποίο θα αποδειχθεί παρακάτω. Έτσι στα πειράματα που ακολουθούν συνήθως επιλέγονται κάποιοι συνδυασμοί δήμων των πρώτων τεσσάρων περιφερειών, ανάλογα με το πείραμα.

Περιφέρεια	Κρούσματα
ΚΕΝΤΡΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ	632
ΑΤΤΙΚΗΣ	286
ΑΝΑΤΟΛΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ - ΘΡΑΚΗΣ	251
ΘΕΣΣΑΛΙΑΣ	109
ΠΕΛΟΠΟΝΝΗΣΟΥ	45
ΣΤΕΡΕΑΣ ΕΛΛΑΔΑΣ	19
ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ	16
ΙΟΝΙΩΝ ΝΗΣΩΝ	5
ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ	3
ΚΡΗΤΗΣ	3
ΗΠΕΙΡΟΥ	1
ΒΟΡΕΙΟΥ ΑΙΓΑΙΟΥ	0
ΝΟΤΙΟΥ ΑΙΓΑΙΟΥ	0

Εικόνα 11: Αθροιστικά κρούσματα ανά περιφέρεια στο διάστημα 2010-2021

Επιλογή υποσυνόλου εβδομάδων: Καθώς οι περίοδοι εξάρσεων είναι οι μήνες από Ιούνιο έως Οκτώβριο, οι υπόλοιπες εβδομάδες δεν χρειάζεται να εξεταστούν. Επομένως, αφού ενσωματώσουμε τα δεδομένα των χειμερινών εβδομάδες στις (βλ. **Δημιουργία ιστορικών**

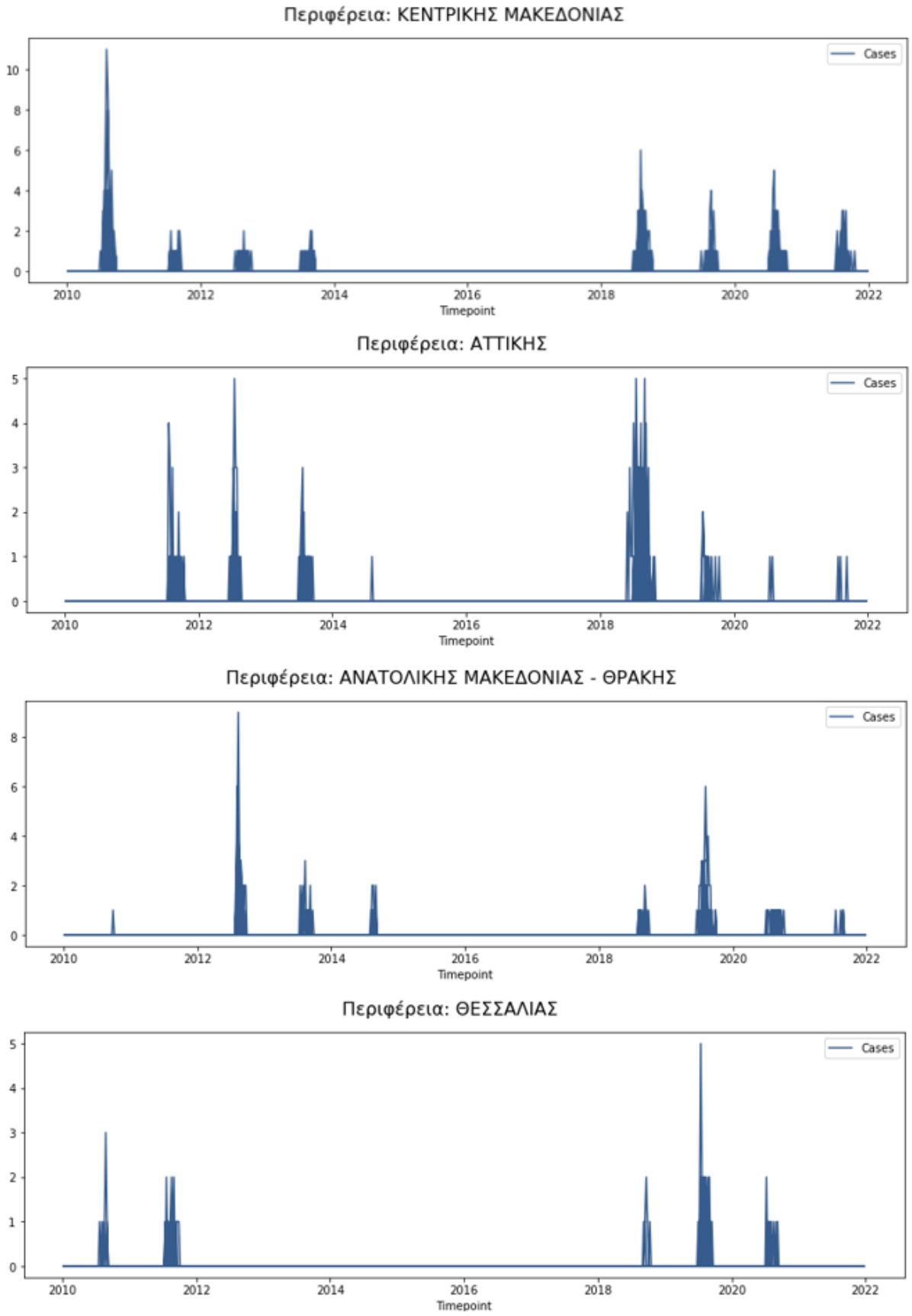
χαρακτηριστικών στην επόμενη σελ.) τις απορρίπτουμε από την βάση ώστε να μην πραγματοποιηθεί εκπαίδευση σε αυτές. Τελικά στην βάση μένουν οι εβδομάδες 20-42.

Δυαδικοποίηση: Η πρόβλεψη του πλήθους των κρουσμάτων αποτελεί πρόβλημα παλινδρόμησης και θεωρείται αρκετά πιο δύσκολο να πραγματοποιηθεί σε σχέση με τα προβλήματα κατηγοριοποίησης. Σε αυτήν την εργασία μετατρέπουμε το πρόβλημα παλινδρόμησης σε πρόβλημα κατηγοριοποίησης ορίζοντας δύο κλάσεων εβδομάδες: αυτές στις οποίες δεν καταγράφηκαν κρούσματα και αυτές στις οποίες καταγράφηκαν. Κάθε εβδομάδα με τουλάχιστον ένα κρούσμα ανήκει στην δεύτερη κλάση.

Χρονικός διαμοιρασμός κρουσμάτων: Σε περιπτώσεις όπου τα δείγματα κάποιας κλάσης είναι μικρά σε αριθμό χρησιμοποιούνται μέθοδοι που δημιουργούν επιπλέον δείγματα. Αυτό μπορεί να γίνει μέσω της υπερδειγματοληψίας (oversampling), δηλαδή της δημιουργίας αντιγράφων, ή μέσω της παραγωγής συνθετικών δεδομένων, δηλαδή δειγμάτων που αποτελούνται από τους συνδυασμούς των χαρακτηριστικών άλλων δειγμάτων. Αυτές οι μέθοδοι όμως δεν είναι τόσο εύκολο να εφαρμοστούν σε δεδομένα χρονοσειρών λόγω των χρονικών εξαρτήσεων που εμπλέκονται. Επιπλέον γνωρίζουμε ότι οι χρόνοι καταγραφής κάθε κρούσματος δεν αντανακλούν την πραγματική στιγμή εμφάνισής τους αλλά περιέχουν το σφάλμα που σχετίζεται με την διαφορά στον χρόνο εμφάνισης συμπτωμάτων, τον χρόνο στον οποίον θα καταγραφούν από τους εργαζομένους του συστήματος υγείας κλπ. Επομένως έχει νόημα σε περιπτώσεις όπου πολλά κρούσματα κατανέμονται σε μία εβδομάδα, αυτά να ανακατανεμηθούν στις γειτονικές. Συγκεκριμένα, εάν τα κρούσματα μίας εβδομάδας υπερβαίνουν ένα (παραμετροποιημένο) κατώφλι ενώ τα κρούσματα της προηγούμενης εβδομάδας ήταν μηδέν, θεωρούμε ότι και η προηγούμενη είχε κρούσματα. Αυτή η μέθοδος, αν και δεν είναι πολύ δραστική, έχει το πλεονέκτημα ότι αντιστοιχεί αρκετά στην πραγματικότητα, και στην πράξη βρέθηκε ότι ενδεχομένως να βελτιώνει τα αποτελέσματα.

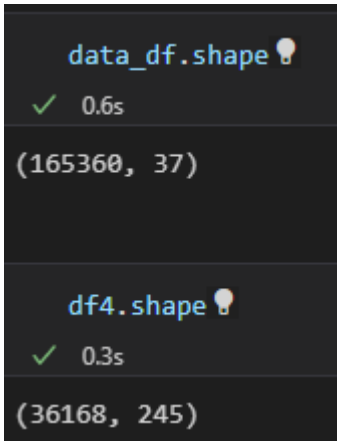
Δημιουργία ιστορικών χαρακτηριστικών: Όπως έχει αναφερθεί τα ξεσπάσματα του ιού είναι το αποτέλεσμα της δράσης αρκετών φαινομένων που το ένα επηρεάζει το άλλο αλυσιδωτά. Επομένως για το κάθε χαρακτηριστικό είναι κρίσιμη όχι μόνο η τιμή του μία δεδομένη χρονική στιγμή αλλά η εξέλιξη αυτής της τιμής στον χρόνο. Μέσω συνάρτησης που εκτελείται κάθε φορά πριν την εκπαίδευση του μοντέλου, προστίθενται οι ιστορικές τιμές ως διαφορετικά χαρακτηριστικά, με το πλήθος των εβδομάδων αυτών να αποτελεί παράμετρο η τιμή της οποίας χρήζει αναζήτησης.

Η μορφή των κρουσμάτων είναι όπως παρουσιάζεται στην εικόνα 12. Παρατηρούμε ότι αυτά καταγράφονται στην μέση κάθε έτους που αντιστοιχεί στους θερινούς μήνες. Ακόμα παρατηρούμε την απουσία του ιού τα έτη 2015-2017 και την επανεμφάνιση το καλοκαίρι του



Εικόνα 12:Η χρονοσειρά των κρουσμάτων στις τέσσερεις περιφέρειες που επικεντρώνεται η μελέτη

2018. Είναι φανερό ότι η χρονοσειρά είναι αρκετά αραιή στην περιφέρεια Θεσσαλίας με ξεσπάσματα σε μόνο 5 από τα 12 έτη, πράγμα που δυσκολεύει την εκπαίδευση. Το φύλλο δεδομένων έχει τις διαστάσεις που παρουσιάζονται στην εικόνα 13. Οι 135,360 στήλες προκύπτουν ως το γινόμενο 53 εβδομάδων επί 12 έτη, επί 260 δήμους, και οι 37 στήλες αντιστοιχούν στα βασικά χαρακτηριστικά συν κάποια πληροφορία όπως ημερομηνία και Καλλικρατικός κωδικός δήμου. Στο φύλλο δεδομένων που χρησιμοποιείται από το μοντέλο έχουμε 36,168 γραμμές που προκύπτουν ως το γινόμενο 22 εβδομάδων επί 12 έτη, επί 137 δήμους και το σύνολο των στηλών αντιστοιχεί στα αρχικά 35 χαρακτηριστικά πλην 2 που αποτελούσαν μεταπληροφορία, συν 30 επί το πλήθος των εβδομάδων από ιστορικά δεδομένα που έχει το συγκεκριμένο σύνολο δεδομένων.



Εικόνα 13: Διαστάσεις δεδομένων πριν την προεπεξεργασία και μετά

Χωρισμός δεδομένων εκπαίδευσης, επικύρωσης και ελέγχου: Στην γενική περίπτωση για τον διαχωρισμό δεδομένων εκπαίδευσης επικύρωσης και ελέγχου χρησιμοποιούνται έτοιμες συναρτήσεις που λαμβάνουν σαν είσοδο ένα σετ δεδομένων και ένα ποσοστό και μοιράζουν τα δεδομένα στην τύχη σε δυο νέα σετ έτσι ώστε η πληθικότητες αυτών να αντιστοιχούν στο δοσμένο ποσοστό. Αντίθετα στις περιπτώσεις των χρονοσειρών που η διαδοχή των δειγμάτων έχει καθοριστική σημασία, συνήθως επιλέγεται ένα σημείο στο χρόνο και τα δείγματα διαχωρίζονται με βάση αυτό. Μία εναλλακτική είναι ο χωρικός διαχωρισμός κατά τον οποίον το σύνολο των δήμων διαχωρίζεται σε τρία (ξένα μεταξύ τους) υποσύνολα και καθένα χρησιμοποιείται έναν από τους παραπάνω σκοπούς. Πρέπει να αναγνωρίσουμε ότι η πρώτη προσέγγιση είναι πλησιέστερη με τον στόχο αυτής της εργασίας καθώς σε ένα σύστημα έγκαιρης προειδοποίησης ο στόχος είναι η παραγωγή προβλέψεων για τους επερχόμενους μήνες ενώ τα παρελθοντικά δεδομένα είναι γνωστά. Παράλληλα όμως, η δεύτερη προσέγγιση μπορεί να έχει νόημα στις περιπτώσεις που υπάρχουν δεδομένα για αρκετές περιοχές αλλά για λίγα σχετικά έτη. Εκεί το ‘θυσιάσουμε’ επιπλέον έτη για επικύρωση και έλεγχο ελλοχεύει τον κίνδυνο να πραγματοποιηθεί εκπαίδευση σε μικρό αριθμό ετών που σημαίνει ότι ο αλγόριθμος θα μάθει τις ιδιαιτερότητες αυτών και δεν θα γενικεύει αποτελεσματικά. Ακόμα, σε ορισμένες περιπτώσεις ενδέχεται να μην απαιτείται η παραγωγή προβλέψεων αλλά να αναζητούνται τα χαρακτηριστικά που προσφέρουν την μεγαλύτερη προβλεπτική ικανότητα, η να αξιολογούνται συγκριτικά οι επιδόσεις

διαφορετικών μοντέλων. Για τους παραπάνω λόγους σε αυτήν την εργασία πραγματοποιήθηκαν πειράματα και των δύο προσεγγίσεων

5.4. Εκπαίδευση και αποτελέσματα

5.4.1 Μετρικές επίδοσης και αξιολόγηση:

Για την εκπαίδευση του μοντέλου, επιλέχθηκε η μετρική αξιολόγησης logloss, μέσω της υπερπαραμέτρου eval_metric, που όπως αναφέραμε (κεφ 3.4) ενδείκνυται για προβλήματα δυαδικής κατηγοριοποίησης

Για την μέτρηση των αποτελεσμάτων χρησιμοποιήθηκε η βαθμολογία $F1$. Πρόκειται για ένα παράγωγο μέγεθος από τις μετρικές της ακρίβειας (precision) και της ανάκλησης (recall) οι οποίες δίνονται από τους τύπους 13 και 14.

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \qquad (13)$$

$$recall = \frac{true\ positives}{true\ positives + false\ negatives} \qquad (14)$$

Η βαθμολογία $F1$ προκύπτει ως:

$$2 * \frac{precision * recall}{precision + recall}$$

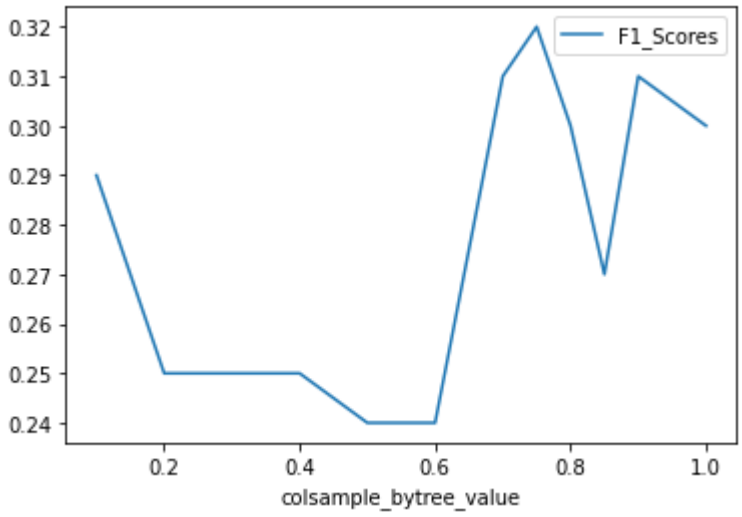
Καθώς υπάρχει ανισορροπία στις κλάσεις, είναι εύκολο να επιτευχθεί υψηλό σκορ χαρακτηρίζοντας όλα τα δείγματα σαν δείγματα της πλειοψηφούσας κλάσης. Για τον λόγο αυτό επικεντρωνόμαστε στις βαθμολογίες που αφορούν την μειοψηφούσα κλάση, δηλαδή τα θετικά δείγματα. Η ανισορροπία των κλάσεων είναι άλλος ένας λόγος για την επιλογή της $F1$ η οποία ενδείκνυται σε τέτοιες περιπτώσεις [56].

Ακόμα, χρησιμοποιείται και ο πίνακας σύγχυσης που περιέχει την πληροφορία για το πλήθος τω δειγμάτων που ανήκουν στον κάθε συνδυασμό πραγματικής και προβλεπόμενης κλάσης.

Τέλος, για την αξιολόγηση του μοντέλου χρησιμοποιούμε τον τυχαίο στρωματοποιημένο ταξινομητή (stratified random classifier) ο οποίος ταξινομεί τα δείγματα στην τύχη αλλά φροντίζοντας να διατηρεί την αναλογία των κλάσεων των δεδομένων εκπαίδευσης

5.4.2 Βελτιστοποίηση υπερπαραμέτρων:

Για την εύρεση του βέλτιστου συνδυασμού τιμών των υπερπαραμέτρων συνήθως πραγματοποιείται αναζήτηση πλέγματος (grid search). Αυτή η τεχνική μπορεί να θεωρηθεί εξαντλητική καθώς πραγματοποιεί εκπαίδευση, αξιολόγηση και σύγκριση αποτελεσμάτων για όλους τους πιθανούς συνδυασμούς, πράγμα το οποίο συνεπάγεται ότι θα βρεθεί ο βέλτιστος. Όμως λόγω της εξαντλητικής αναζήτησης, αποτελεί μια από τις πιο ακριβές μεθόδους, και για την δική μας περίπτωση, θεωρώντας ότι η μέση εκπαίδευση απαιτεί 5 δευτερόλεπτα, η αναζήτηση ανάμεσα σε 10 διαφορετικές τιμές, για τις 11 υπερπαραμέτρους που αναφέραμε στο κεφάλαιο 3.4 αντιστοιχεί σε 100 δισεκατομμύρια εκπαιδεύσεις και περίπου 139 εκατομμύρια ώρες. Επομένως πάντα πραγματοποιείται αναζήτηση σε ένα υποσύνολο τιμών και παραμέτρων. Σε αυτήν την εργασία πραγματοποιούμε αναζήτηση ανά παράμετρο το οποίο είναι από τις πιο οικονομικές μεθόδους. Συγκεκριμένα για κάθε υπερπαράμετρο, ορίζουμε 12 τιμές γύρω από τις προκαθορισμένες και πραγματοποιούμε την αναζήτηση. Με αυτόν τον τρόπο, για τις 11 υπερπαραμέτρους και για 5 δευτερόλεπτα ανά εκπαίδευση προκύπτουν 132 πειράματα για τα οποία χρειάζονται 11 λεπτά. Αυτό επαναλήφθηκε για δεύτερη φορά, χρησιμοποιώντας τις βέλτιστες τιμές της πρώτης επανάληψης για της παραμέτρους που δεν δοκιμαζόντουσαν κάθε φορά, αν και παρατηρήσαμε μηδενικές διαφορές. Σημειώνεται ότι υπάρχουν ενδιάμεσες μέθοδοι που πραγματοποιούν πολλαπλές αναζητήσεις πλέγματος σε υποσύνολα των υπερπαραμέτρων και των υποψήφιων τιμών αυτών, ωστόσο και αυτές μπορούν να διαρκέσουν στην περίπτωσή μας μερικές δεκάδες ώρες, οπότε δεν



Εικόνα 14: Αναζήτηση βέλτιστης τιμής για την υπερπαράμετρο colsample_bytree βάσει της μετρικής F1

χρησιμοποιήθηκαν. Στην εικόνα 14 βλέπουμε την αναζήτηση για την τιμή της υπερπαραμέτρου colsample_bytree, χρησιμοποιώντας την μετρική $F1$ για αξιολόγηση. Αν και η καμπύλη εμφανίζει μεγάλες διακυμάνσεις ξεχωρίζει σαν βέλτιστη τιμή η 0.75. Σημειώνουμε ότι ο αλγόριθμος XGBoost στην γενική περίπτωση είναι στοχαστικός ως προς την εκπαίδευσή, λόγω του ότι η υποδειγματοληψία που πραγματοποιεί ως προς τα δείγματα και τα χαρακτηριστικά πραγματοποιείται τυχαία. Αυτό σημαίνει ότι τα αποτελέσματα δύο εκπαιδεύσεων μπορεί να είναι διαφορετικά, αν και στην πράξη η διαφορά είναι αμελητέα.

Οι υπερπαράμετροι που επιλέχθηκαν είναι:

- subsample = 0.8
- colsample_bytree = 0.75
- learning_rate = 0.05
- max_depth = 2
- n_estimators = 5000
- early_stopping_rounds = 100

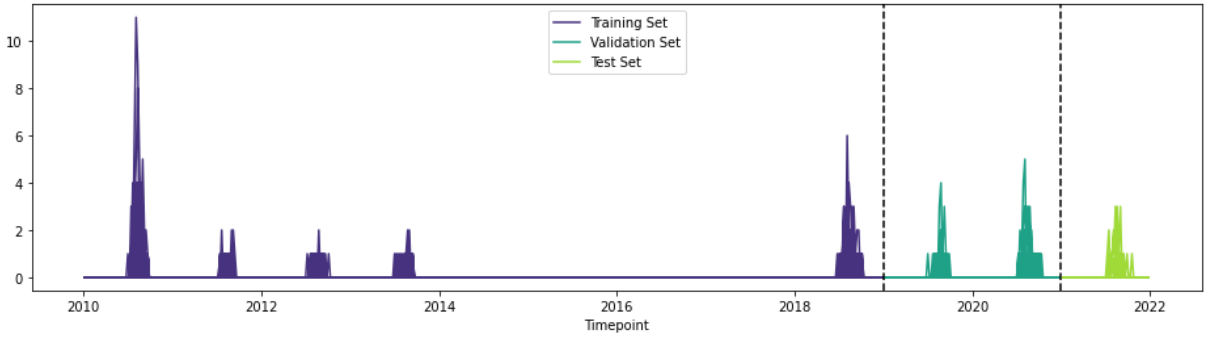
Για τις υπόλοιπες βρέθηκε ότι αλλαγή από την προεπιλεγμένη τιμή δεν βελτίωνε τα αποτελέσματα, όταν οι 6 υπερπαράμετροι που παραθέσαμε είχαν τις βέλτιστες τιμές.

5.4.3 Προσέγγιση Ι (κατά έτη) - Ανά περιφέρεια:

Εξετάζουμε τις επιδόσεις του μοντέλου χρησιμοποιώντας τα δεδομένα των ετών 2010-2017 για εκπαίδευση, 2018-2019 για επικύρωση και 2020-2021 για έλεγχο.

Περιφέρεια Κεντρικής Μακεδονίας:

Στην εικόνα 15 βλέπουμε τον παραπάνω διαχωρισμό για την περιφέρεια της Κεντρικής Μακεδονίας, στην οποία σημειώθηκαν και τα περισσότερα κρούσματα. Η αναλογία των



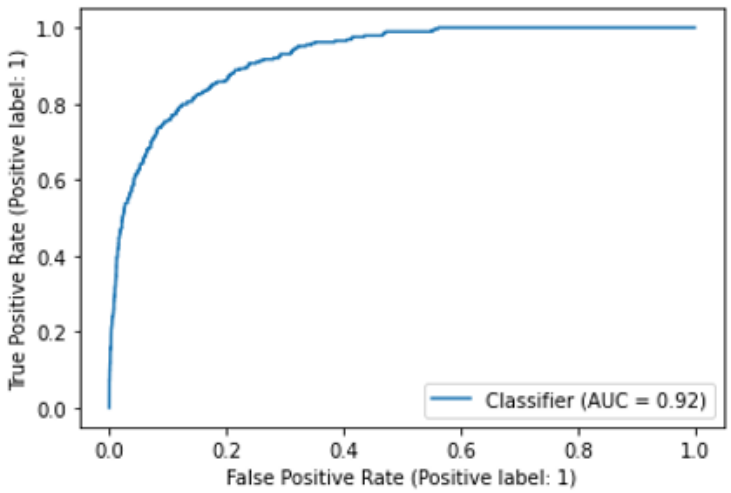
Εικόνα 15: Διαχωρισμός των δεδομένων με βάση την χρονολογία, σε σύνολο εκπαίδευσης, επικύρωσης και ελέγχου

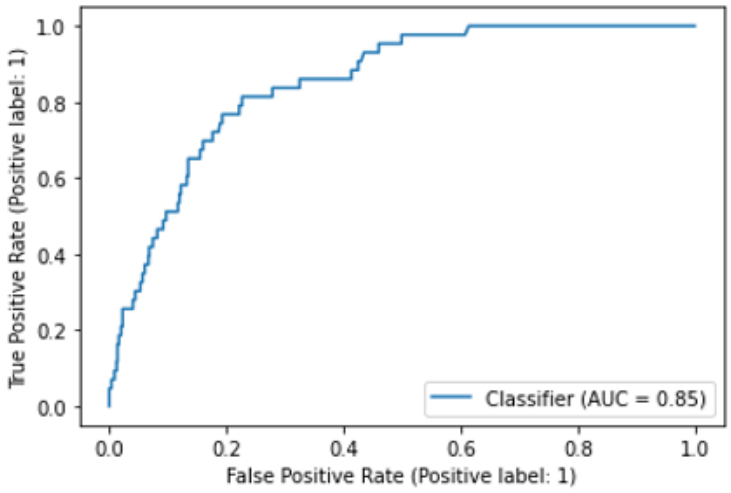
κλάσεων, δηλαδή η αναλογία των εβδομάδων χωρίς κρούσματα προς τις εβδομάδες με κρούσματα, για το διάστημα από την 20η έως την 42η εβδομάδα είναι: 25.1 για τα δεδομένα της εκπαίδευσης, 18.4 για τα δεδομένα της επικύρωσης και 11.5 για τα δεδομένα του ελέγχου. Τα αποτελέσματα φαίνονται στην εικόνα 16. Συγκεκριμένα, στο πάνω μισό εμφανίζονται τα αποτελέσματα του XGBoost και στο κάτω μισό τα αποτελέσματα του τυχαίου στρωματοποιημένου ταξινομητή για σύγκριση. Στο αριστερό πάνω τμήμα παρατίθεται ο πίνακας σύγχυσης. Παρατηρώντας την βαθμολογία F1 και τον πίνακα σύγχυσης βλέπουμε πως το μοντέλο μας είναι σημαντικά πιο ικανό στο να διακρίνει τις δύο κλάσεις σε σχέση με το μοντέλο αναφοράς. Για τις μετρικές της ακρίβειας και της ανάκλησης πρέπει να αναφέρουμε ότι έχουν μία αντιστρόφως ανάλογη σχέση. Αυτό συμβαίνει διότι πριν κάθε δείγμα κατηγοριοποιηθεί, δίνεται σε αυτό μία πιθανότητα από το μοντέλο, η οποία είναι η πιθανότητα να ανήκει στην μία κλάση (η πιθανότητα να ανήκει στην άλλη κλάση είναι η συμπληρωματική της). Στην συνέχεια η πιθανότητα του κάθε δείγματος συγκρίνεται με ένα κατώφλι προκειμένου να γίνει η κατηγοριοποίηση. Με βάση την επιλογή του κατωφλίου μπορεί να επηρεαστεί η αυστηρότητα με την οποία το τα δείγματα κατηγοριοποιούνται ως θετικά. Μεγαλύτερη αυστηρότητα, δηλαδή υψηλότερο κατώφλι, σημαίνει ότι μόνο τα δείγματα για τα οποία υπάρχει μεγάλη βεβαιότητα κατηγοριοποιούνται στην θετική κλάση οπότε πράγμα το οποίο αυξάνει την ακρίβεια. Από την άλλη, όλα τα υπόλοιπα δείγματα κατηγοριοποιούνται σαν αρνητικά, που σημαίνει ότι το μοντέλο χάνει πολλά από τα θετικά, πράγμα που μειώνει την

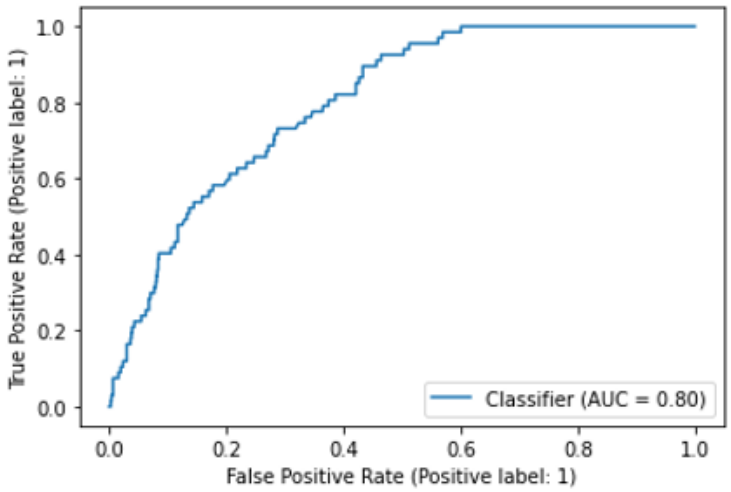
```
 [[633 136]
 [ 29  38]]
              precision    recall  f1-score   support

           0       0.96      0.82      0.88       769
           1       0.22      0.57      0.32        67

    accuracy                           0.80       836
   macro avg       0.59      0.70      0.60       836
weighted avg       0.90      0.80      0.84       836
```

```
[[727  42]
 [ 65   2]]
              precision    recall  f1-score   support

           0       0.92      0.95      0.93       769
           1       0.05      0.03      0.04        67

    accuracy                           0.87       836
   macro avg       0.48      0.49      0.48       836
weighted avg       0.85      0.87      0.86       836
```

Εικόνα 16: Πίνακας σύγχυσης και μετρικές επίδοσης μοντέλου XGBoost και τυχαίου ταξινομητή







Εικόνα 17: Καμπύλες ROC και εμβαδό υπό αυτές για τα δεδομένα εκπαίδευσης, επικύρωσης και ελέγχου, (από πάνω προς τα κάτω).

ανάκληση. Ιδανικά θέλουμε το μοντέλο να δίνει χαμηλή πιθανότητα σε όλα τα αρνητικά και υψηλή σε όλα τα θετικά. Έτσι, με ένα ενδιάμεσο κατώφλι η κατηγοριοποίηση θα γινόταν για όλα τα δείγματα σωστά. Στην πράξη υπάρχουν δείγματα και από τις δύο κλάσεις για τα οποία το μοντέλο δίνει παρόμοια πιθανότητα, οπότε, με βάση το κατώφλι, επιλέγουμε αν θα κατηγοριοποιηθούν προς την μία κλάση ή την άλλη. Το πλήθος των δειγμάτων για τα οποία συμβαίνει αυτό εξαρτάται από την απόδοση του μοντέλου και περιγράφεται από το εμβαδό υπό την καμπύλη ROC. Παραθέτουμε τις καμπύλες ROC για το μοντέλο μας στην εικόνα 17. Στον κατακόρυφο άξονα είναι η ευαισθησία (sensitivity του μοντέλου):

$$sensitivity = \frac{TP}{TP + FN}$$

και στον οριζόντιο η αστοχία (fallout):

$$fallout = \frac{FP}{FP + TN}$$

Ο ιδανικός ταξινομητής επιτυγχάνει ευαισθησία 1 με αστοχία μηδέν το οποίο συνεπάγεται εμβαδό υπό την καμπύλη ίσο με 1 μονάδες εμβαδού (μ.ε.) Αντίθετα ο τυχαίος ταξινομητής

έχει παρόμοια ευαισθησία και αστοχία οπότε η καμπύλη ROC είναι μία ευθεία με κλίση 45° και εμβαδό 0.5 μ.ε. Είναι φανερό ότι όσο καλύτερος είναι ο ταξινομητής τόσο πιο κοντά είναι το εμβαδό υπό την καμπύλη στις 1μ.ε. Τα σημεία της καμπύλης προκύπτουν για τα διαφορετικά κατώφλια, όπου σαν διαφορετικά κατώφλια θεωρούμε αυτά τα οποία προκαλούν διαφορετική κατηγοριοποίηση σε τουλάχιστον ένα δείγμα. Έτσι, η τρίτη καμπύλη μας δείχνει πως αλλάζει η ευαισθησία και η αστοχία στα δεδομένα ελέγχου καθώς μεγαλώνει η τιμή του κατωφλίου. Αυτό μπορούμε να το δούμε και στους πίνακες σύγχυσης της εικόνας 18. Πρέπει να σημειωθεί ότι η επιλογή κατωφλίου πρέπει υποχρεωτικά να γίνεται με την καμπύλη των δεδομένων επικύρωσης και όχι με την καμπύλη των δεδομένων ελέγχου καθώς με το δεύτερο συνεπάγεται διαρροή πληροφορίας. Βλέπουμε ότι με εξαίρεση τις οριακές τιμές, όσο το κατώφλι μειώνεται, μειώνεται και η ακρίβεια ενώ αυξάνεται η ανάκληση. Η βαθμολογία F1 λαμβάνει την μέγιστη τιμή όταν το κατώφλι είναι τέτοιο, ώστε η ακρίβεια και η ανάκληση να έχουν κοντινές τιμές. Ωστόσο,

Threshold = 1.0

```
[[769   0]
[ 67   0]]
              precision    recall  f1-score

           0       0.92      1.00      0.96
           1       0.00      0.00      0.00
```

Threshold = 0.15

```
[[754  15]
[ 61   6]]
              precision    recall  f1-score

           0       0.93      0.98      0.95
           1       0.29      0.09      0.14
```

Threshold = 0.09

```
[[704  65]
[ 41  26]]
              precision    recall  f1-score

           0       0.94      0.92      0.93
           1       0.29      0.39      0.33
```

Threshold = 0.07

```
[[618 151]
[ 28  39]]
              precision    recall  f1-score

           0       0.96      0.80      0.87
           1       0.21      0.58      0.30
```

Threshold = 0.05

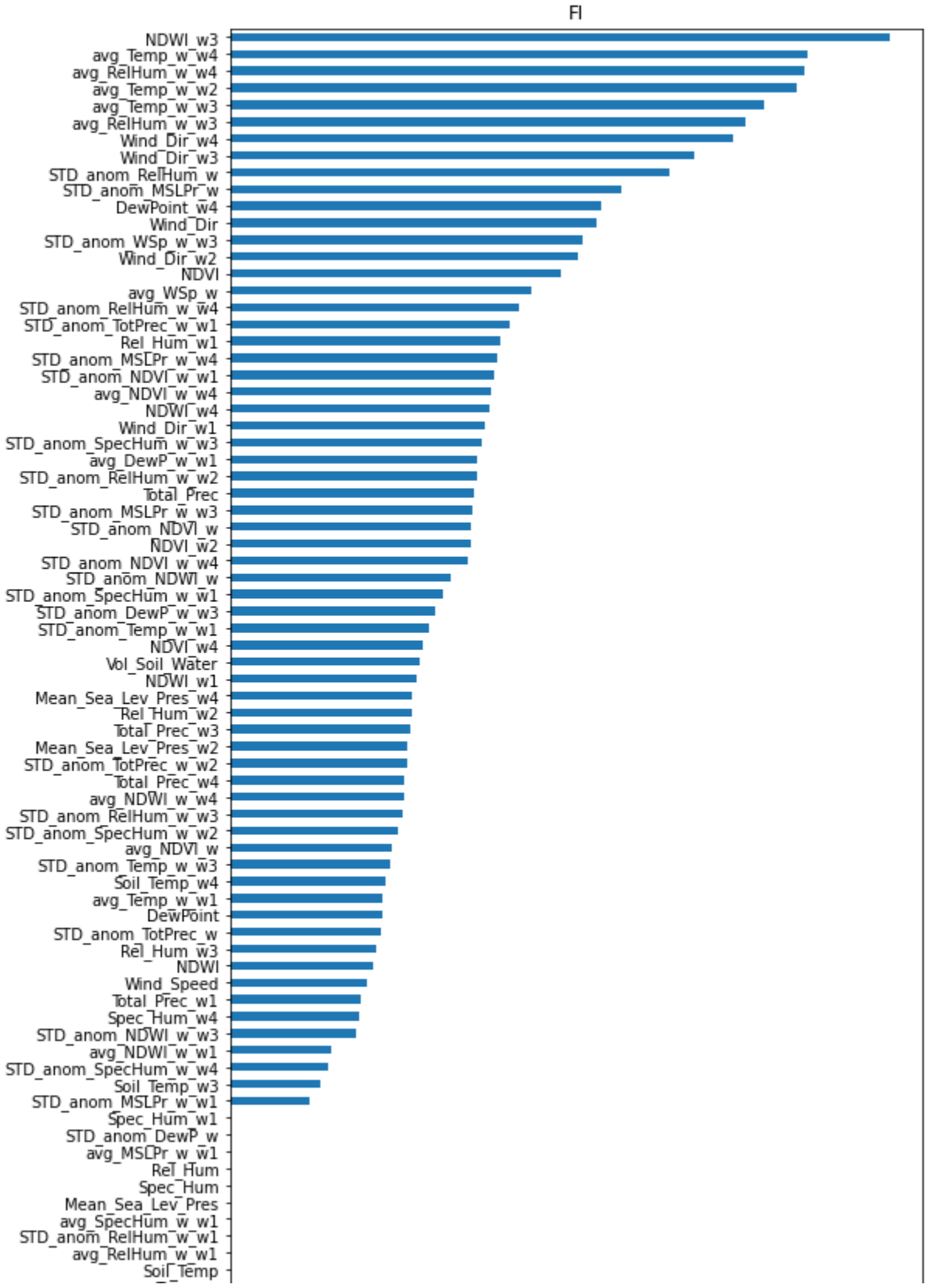
```
[[481 288]
[ 14  53]]
              precision    recall  f1-score

           0       0.97      0.63      0.76
           1       0.16      0.79      0.26
```

Εικόνα 18: Πίνακας σύγχυσης και μετρικές επίδοσης για διαφορετικά κατώφλια στο ίδιο μοντέλο και δεδομένα

ανάλογα με το πρόβλημα αυτό μπορεί να μην είναι επιθυμητό. Στην δική μας περίπτωση, η μη πρόβλεψη μίας εξάρσεως συνεπάγεται μεγάλο κόστος στις ζωές και στην υγεία του πληθυσμού, ενώ η λανθασμένη πρόβλεψη εξάρσεως συνεπάγεται αχρείαστες ενέργειες πρόληψης και ενημέρωσης, το οποίο δεν θεωρείται το ίδιο αρνητικό. Επομένως, χωρίς να υπάρχουν καθορισμένες τιμές, προτιμάται η ανάκληση να είναι υψηλότερη από την ακρίβεια. Στην συνέχεια της εργασία θα επιλέγουμε τα κατώφλια που δίνουν ανάκληση κοντά στο 55-60%, με όποια τιμή ακρίβειας αντιστοιχεί σε αυτό.

Σημαντικότητα χαρακτηριστικών (feature importance):

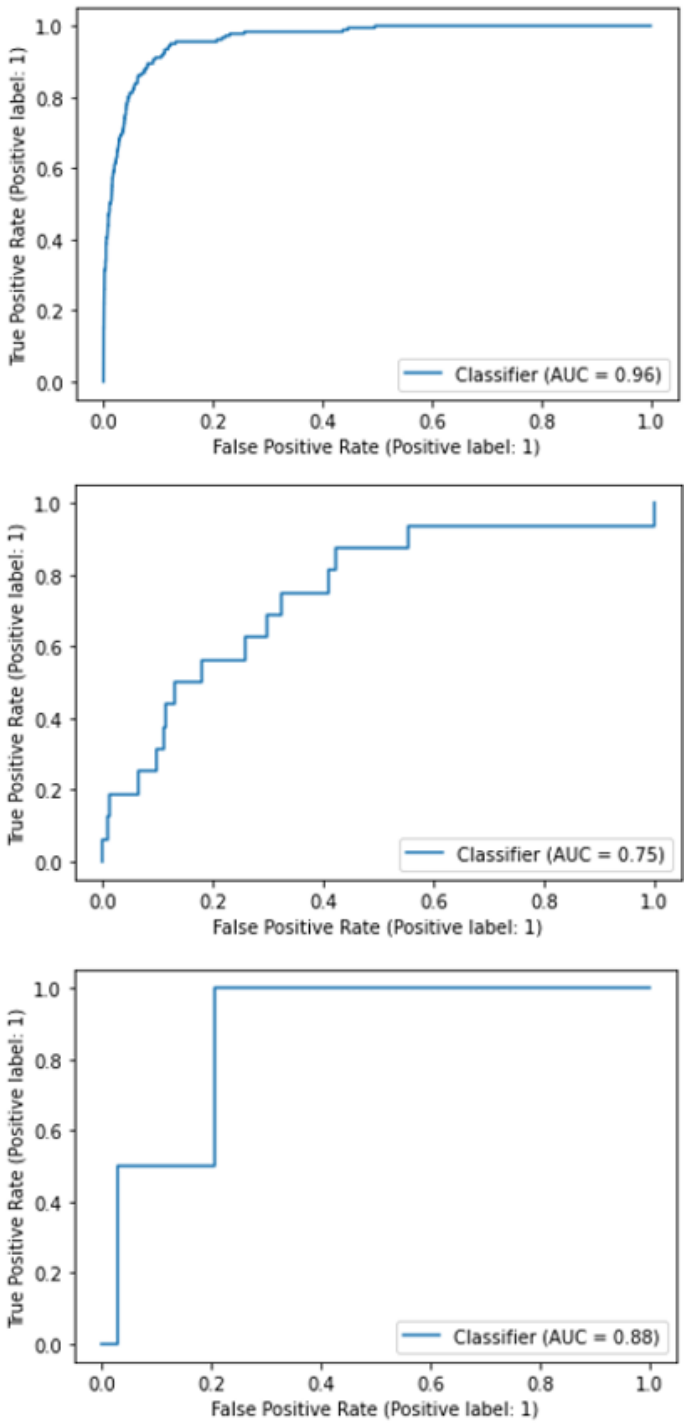
Μετά την εκπαίδευση του μοντέλου θα θέλαμε να γνωρίζουμε πως αυτό δουλεύει, ή πως χρησιμοποιούνται τα δεδομένα που του δώσαμε. Σε αυτό είναι χρήσιμες οι τιμές της σημαντικότητας του κάθε χαρακτηριστικού (feature importance), οι οποίες υπολογίζονται από τον ίδιο τον αλγόριθμο κατά την διάρκεια της εκπαίδευσης. Πρόκειται για τον μέσο όρο σε όλα τα δέντρα, της ικανότητας διάκρισης των δειγμάτων στην σωστή κατηγορία, για το κάθε χαρακτηριστικό. Αυτή η ικανότητα μετράται με βάση την καθαρότητα (purity) και συνήθως υπολογίζεται με βάση τον δείκτη Gini (Gini index) [57]. Για το μοντέλο μας η σημαντικότητα του κάθε χαρακτηριστικού φαίνεται στην εικόνα 19. Παρατηρούμε ότι στις πρώτες θέσεις βρίσκονται χαρακτηριστικά που σχετίζονται με την υγρασία και την θερμοκρασία, που έρχεται σε συμφωνία με την υπάρχουσα γνώση. Από την άλλη, αντίστοιχα χαρακτηριστικά βαθμολογούνται με μηδενική σημαντικότητα. Αυτό συμβαίνει διότι για το σύνολο δεδομένων μας, αυτά παρουσιάζουν συσχέτιση οπότε εάν αξιοποιηθεί το ένα, το άλλο δεν προσφέρει επιπλέον πληροφορία. Καθώς ο αλγόριθμος είναι στοχαστικός, διαδοχικές επαναλήψεις δίνουν διαφορετικές σημαντικότητες. Σε κάθε επανάληψη όμως κάποια χαρακτηριστικά θερμοκρασίας και υγρασίας βρίσκονται στις πρώτες θέσεις, ενώ χαρακτηριστικά όπως η ατμοσφαιρική πίεση και η κατεύθυνση του αέρα βρίσκονται χαμηλά. Αναφέρουμε πως καθώς στο στάδιο της εκπαίδευσης δοκιμάζονται όλα τα χαρακτηριστικά και επιλέγονται τα καλύτερα, η επιλογή χαρακτηριστικών (feature selection) που σε άλλους αλγορίθμους είναι κρίσιμη, δεν είναι απαραίτητη εδώ, και το μόνο που μπορεί να προσφέρει είναι βελτίωση υπολογιστικής φύσης.



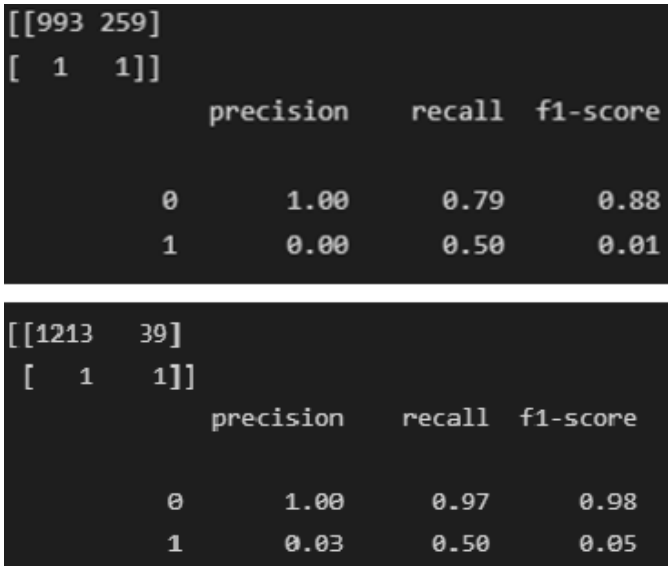
Εικόνα 19: Σημαντικότητα χαρακτηριστικών του μοντέλου

Περιφέρεια Αττικής:

Η περιφέρεια Αττικής παρουσιάζει την πρόκληση ότι κατά τα τελευταία έτη παρουσιάζει ελάχιστο αριθμό κρουσμάτων όπως φαίνεται από την εικόνα 12. Ταυτόχρονα είναι η περιφέρεια με το μεγαλύτερο πλήθος δήμων (57) ενώ οι υπόλοιπες περιφέρειες δεν ξεπερνούν τους 40. Αυτό σημαίνει η κατανομή των κρουσμάτων είναι αρκετά αραιή. Συγκεκριμένα οι αναλογίες των κλάσεων είναι 61.7 για την εκπαίδευση, 77.4 για την επικύρωση και 626 για τον έλεγχο, στον οποίον περιέχονται μόλις 2 κρούσματα. Πραγματοποιώντας εκπαίδευση με τον ίδιο τρόπο που κάναμε για την περιφέρεια της Κεντρικής Μακεδονίας παίρνουμε της καμπύλες ROC της εικόνας 20 που αρχικά φαίνονται ικανοποιητικές. Στην πραγματικότητα τα αποτελέσματα δεν είναι καλά όπως θα φαίνεται στον πίνακα σύγχυσης και στις μετρικές της εικόνας 21. Η συγκεκριμένη περίπτωση παρατίθεται διότι αναδεικνύει την ανάγκη για ενδελεχή συνολικό έλεγχο στα δεδομένα και στις μετρικές καθώς η μελέτη μερικών από αυτές μπορεί να είναι παραπλανητική. Από την εικόνα 21 βλέπουμε πως οι επιδόσεις του μοντέλου (πάνω) είναι στην πραγματικότητα χειρότερες από αυτές του τυχαίου ταξινομητή (κάτω). Συγκριμένα το μοντέλο έχει χαμηλότερη ακρίβεια, ενώ η ανάκληση σαν τιμή δεν έχει σημασία για τόσο μικρό πλήθος θετικών δειγμάτων, πράγμα το οποίο προκύπτει πιο καθαρά από τον πίνακα σύγχυσης.



Εικόνα 20: Καμπύλες ROC της περιφέρειας Αττικής στα δεδομένα εκπαίδευσης, επικύρωσης και ελέγχου (το τελευταίο περιέχει μόνο 2 κρούσματα).

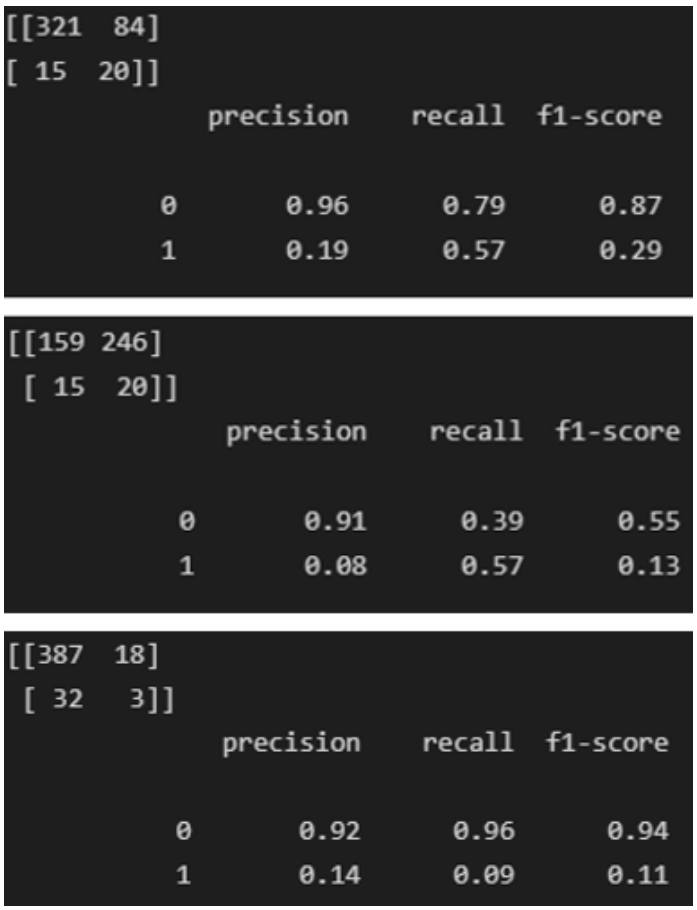


Εικόνα 21: Πίνακας σύγχυσης και μετρικές επίδοσης του μοντέλου μας (πάνω) και του τυχαίου στρωματοποιημένου ταξινομητή (κάτω).

Στην πραγματικότητα το μοντέλο μας έχει μάθει μόνο θόρυβο και δεν διαφέρει από τον τυχαίο ταξινομητή. Ο λόγος για την χαμηλότερη ακρίβεια είναι η επιλογή κατωφλίου, που τυχαίνει να είναι χαμηλότερη. Αν το αυξήσουμε μπορούμε να πάρουμε τα αποτελέσματα του τυχαίου μοντέλου, πράγμα που θα εξηγήσουμε και στην περίπτωση της περιφέρειας Θεσσαλίας όπου συμβαίνει πρακτικά το ίδιο.

Περιφέρεια Ανατολικής Μακεδονίας - Θράκης:

Αν και από την εικόνα 11 βλέπουμε πως έχει μικρότερο αριθμό κρουσμάτων από την περιφέρεια Αττικής κατά λίγο, έχοντας μόνο 20 δήμους προκύπτει ότι έχει μεγαλύτερη καλύτερη ισορροπία στις δύο κλάσεις. Αυτή είναι 49.1, 7.8 και 11.6 για εκπαίδευση, επικύρωση και έλεγχο αντίστοιχα. Βλέπουμε ότι και εδώ η αναλογία διαφέρει αλλά καθώς χωρίζουμε με βάση τα έτη δεν μπορούμε να το ελέγξουμε. Ωστόσο σημειώνουμε ότι ένα ενδιαφέρον πείραμα θα ήταν ο έλεγχος τον επιδόσεων εάν τα δεδομένα μοιράζονταν με τρόπο ώστε η αναλογία να διατηρούταν. Σχετικά με



Εικόνα 22: Πίνακας σύγχυσης και μετρικές επίδοσης για το μοντέλο μας (πάνω), για τυχαίο ταξινομητή ίδιας ανάκλησης με τον δικό μας, και για τυχαίο στρωματοποιημένο ταξινομητή

τα αποτελέσματα έχουμε από την εικόνα 22 ότι το μοντέλο μας πετυχαίνει καλύτερα αποτελέσματα από τον τυχαίο ταξινομητή. Σε αυτήν την περίπτωση περιέχουμε για σύγκριση και έναν τυχαίο με αλλαγμένη την στρωματοποίηση ώστε να έχει την ίδια ανάκληση με αυτό που εκπαιδεύσαμε. Έτσι η διακριτική ικανότητα του μοντέλου παρουσιάζεται ως μείωση του πλήθους των false positive. Οι καμπύλες ROC είναι αντίστοιχες με την περίπτωση της Κεντρικής Μακεδονίας και δεν παρατίθενται.

Περιφέρεια Θεσσαλίας:

Η περιφέρεια της Θεσσαλίας έχει σύμφωνα με την εικόνα 11 τα λιγότερα κρούσματα, μόλις 109 σε 22 δήμους, πράγμα το οποίο μας προδιαθέτει για μη καλές επιδόσεις. Στην εικόνα 23 βλέπουμε ότι αυτές είναι αντίστοιχες του τυχαίου ταξινομητή. Το ότι εμφανίζεται χαμηλότερη ακρίβεια οφείλεται στο μικρό πλήθος θετικών παρατηρήσεων, αφού όταν μειώσαμε το κατώφλι στον τυχαίο ταξινομητή τα false positives αυξήθηκαν μέχρι τα 47 χωρίς να αυξηθούν τα true positives. Το εμβαδό υπό την καμπύλη για το σύνολο ελέγχου ήταν 0.51 το οποίο επίσης υποδεικνύει επιδόσεις αντίστοιχες του τυχαίου. Οι καμπύλες ROC είναι αντίστοιχες με την περίπτωση της Αττικής και δεν παρατίθενται.

```
[[428  47]
[  8   1]]
              precision    recall  f1-score

           0       0.98      0.90      0.94
           1       0.02      0.11      0.04
```

```
[[465  10]
 [  8   1]]
              precision    recall  f1-score

           0       0.98      0.98      0.98
           1       0.09      0.11      0.10
```

Εικόνα 23. Πίνακας σύγχυσης και μετρικές επίδοσης για περιφέρεια Θεσσαλίας. Παρόμοια αποτελέσματα στο μοντέλο XGBoost (πάνω) και στον τυχαίο ταξινομητή (κάτω)

5.4.4 Προσέγγιση Ι (κατά έτη) – Συνδυασμοί περιφερειών:

Εξετάζουμε εάν βελτιώνονται τα αποτελέσματα με την εκπαίδευση σε περισσότερες από μία περιφέρειες ταυτόχρονα. Συγκεκριμένα συμπεριλαμβάνουμε κάθε φορά την περιφέρεια της Κεντρικής Μακεδονίας που είχε τα καλύτερα αποτελέσματα, και μία από τις υπόλοιπες τρεις. Έτσι έχουμε:

Περιφέρειες Κεντρικής Μακεδονίας και Αττικής:

```
Weekly:
 [[1879  142]
 [  32   37]]
              precision    recall  f1-score

           0       0.98      0.93      0.96
           1       0.21      0.54      0.30
```

Εικόνα 24: Πίνακας σύγχυσης και μετρικές επίδοσης για μελέτη Κεντρικής Μακεδονίας & Αττικής

Παρατηρούμε ότι οι επιδόσεις είναι λίγο χειρότερες σε σχέση με όταν μελετήσαμε την Κεντρική Μακεδονία μόνη της (εικόνα 16). Αυτό είναι λογικό καθώς οι δύο περιφέρεις έχουν αρκετά διαφορετικά χαρακτηριστικά και τα δεδομένα της περιφέρειας Αττικής ήταν αρκετά δύσχρηστα.

Περιφέρειες Κεντρικής Μακεδονίας και Ανατολικής Μακεδονίας – Θράκης:

```
[[872 302]
[ 47  55]]
              precision    recall  f1-score

           0       0.95      0.74      0.83
           1       0.15      0.54      0.24
```

Εικόνα 25:Πίνακας σύγχυσης και μετρικές επίδοσης για μελέτη Κεντρικής Μακεδονίας & Ανατολικής Μακεδονίας - Θράκης

Παρατηρούμε ότι και πάλι τα αποτελέσματα είναι χειρότερα. Αυτό είναι αντίθετο από τις προσδοκίες μας καθώς οι δύο περιφέρεις έχουν αρκετά κοινά χαρακτηριστικά, και όταν μελετήθηκαν ξεχωριστά παρατηρήθηκαν καλές επιδόσεις σε αμφότερα.

Περιφέρειες Κεντρικής Μακεδονίας και Θεσσαλίας:

```
[[1009  235]
[  35   41]]
              precision    recall  f1-score

           0       0.97      0.81      0.88
           1       0.15      0.54      0.23
```

Εικόνα 26: Πίνακας σύγχυσης και μετρικές επίδοσης για μελέτη Κεντρικής Μακεδονίας & Θεσσαλίας

Η περίπτωση είναι αντίστοιχη με αυτήν της Αττικής.

Όλες οι περιφέρειες:

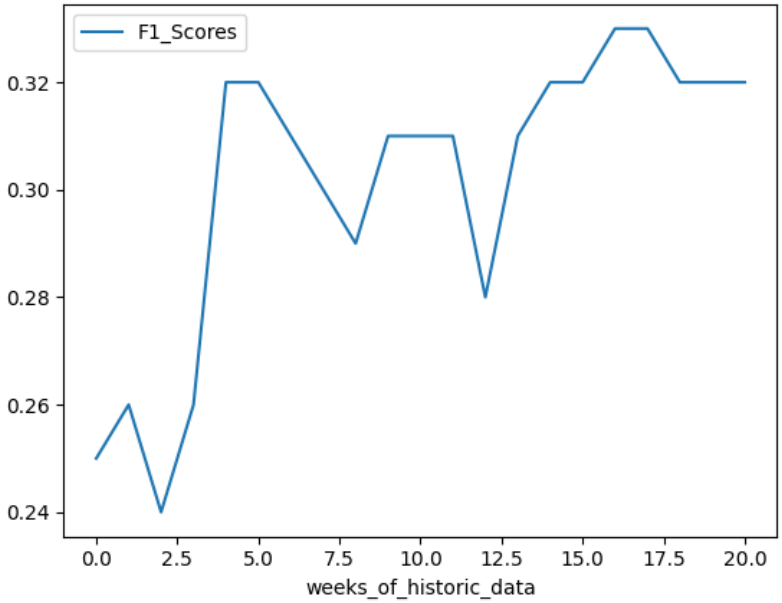
```
[[2477  424]
[  52   61]]
              precision    recall  f1-score

           0       0.98      0.85      0.91
           1       0.13      0.54      0.20
```

Εικόνα 27: Πίνακας σύγχυσης και μετρικές επίδοσης για μελέτη του συνόλου των περιφερειών

Παρατηρούμε ότι τα αποτελέσματα είναι χειρότερα από όταν έγινε μεμονωμένη μελέτη.

5.4.5 Επίδραση των ιστορικών δεδομένων:

Αναφέραμε ότι έχει ενδιαφέρον να μελετηθεί η επίδραση της συμπερίληψης των ιστορικών δεδομένω στην επίδοση του μοντέλου. Εκπαιδεύοντας το στα δεδομενα της Κεντρικής Μακεδονίας για διαφόρες τιμές του πλήθους ιστορικών εβδομάδων παίρνουμε την καμπύλη της εικονας 28 σύμφωνα με την οποία χρειάζονται περισσότερες από 4 εβδομάδες δεδομένων για καλύτερα αποτελέσματα, αν και παρουσιάζεται κάποια διακύμανση. Κάθε μέτρηση εκτελέστηκε 3 φορές για εξάλειψη του θορύβου, αν και τα αποτελέσματα ήταν σταθερά σε κάθε τριάδα. Όλα τα πειράματα που προαναφέραμε έγιναν για πλήθος εβδομάδων ίσο με 16 το οποίο είχε προκύψει από την αναζήτηση για τις βέλτιστες τιμές των υπερπαραμέτρων ότι ήταν το καλύτερο.



Εικόνα 28: Βαθμός F1 για διαφορετικό πλήθος ιστορικών εβδομάδων στα δεδομένα

5.4.6 Προσέγγιση ΙΙ (κατά δήμο) - Ανά περιφέρεια:

Θα εξετάσουμε τώρα έναν διαφορετικό τρόπο χωρισμού των δεδομένων στα τρία σύνολα. Συγκεκριμένα από το σύνολο των δήμων της περιφέρειας θα επιλέγεται κάθε φορά ένας για έλεγχο και άλλοι τέσσερεις για επικύρωση. Οι υπόλοιποι θα χρησιμοποιούνται για την εκπαίδευση. Αυτό θα γίνεται για το σύνολο των δήμων της περιφέρειας και τα αποτελέσματα θα συγκεντρώνονται για εξαγωγή των μετρικών στο τέλος. Με αυτόν τον τρόπο έχουμε:

Περιφέρεια Κεντρικής Μακεδονίας:

```
[[8377  951]
[ 201  239]]
              precision    recall  f1-score

           0       0.98      0.90      0.94
           1       0.20      0.54      0.29
```

Εικόνα 29: Πίνακας σύγχυσης και μετρικές επίδοσης για εκπαίδευση στην περιφέρεια Κεντρικής Μακεδονίας με την δεύτερη προσέγγιση.

Βλέπουμε πως οι επιδόσεις είναι λίγο χειρότερες από την αντίστοιχη περίπτωση της πρώτης προσέγγισης. Σημειώνουμε ότι για την δεύτερη προσέγγιση, καθώς η εκπαίδευση επαναλαμβάνεται για κάθε δήμο, ο συνολικός χρόνος εκτέλεσης του πειράματος υπερβαίνει τα 2.5 λεπτά, πράγμα που καθιστά την αναζήτηση των βέλτιστων τιμών των υπερπαραμέτρων σημαντικά πιο δύσκολο. Για αυτόν τον λόγο χρησιμοποιήσαμε τις υπερπαραμέτρους που βρήκαμε στην πρώτη προσέγγιση, και εξετάσαμε τις γειτονικές τιμές αυτών, για τις οποίες ωστόσο δεν βρέθηκε κάποια διαφορά.

Περιφέρεια Αττικής:

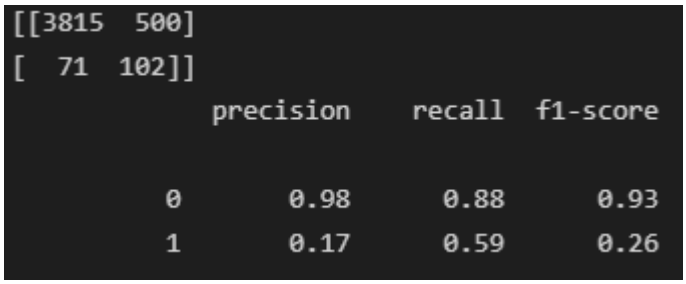
```
[[7392  423]
[  62   43]]
              precision    recall  f1-score

           0       0.99      0.95      0.97
           1       0.09      0.41      0.15
```

Εικόνα 30: Πίνακας σύγχυσης και μετρικές επίδοσης για εκπαίδευση στην περιφέρεια Αττικής με την δεύτερη προσέγγιση

Παρατηρούμε πως σε αντίθεση με την προσέγγιση Ι εδώ έχουμε πολύ καλύτερα αποτελέσματα. Είναι φανερό ότι το μοντέλο έχει καλύτερη προβλεπτική ικανότητα από τον τυχαίο ταξινομητή. Ωστόσο, η συγκεκριμένη προσέγγιση, όπως είχαμε αναφέρει, δεν μπορεί να προσφέρει προβλέψεις για επερχόμενα ξεσπάσματα αλλά μπορεί να φανεί χρήσιμη αν, μέσω της σημαντικότητας των χαρακτηριστικών προσφέρει επιπλέον πληροφορία για το πώς προβλέπει τα ξεσπάσματα, σε σχέση με τα προηγούμενα μοντέλα. Στην συγκεκριμένη

περίπτωση όμως, μελετώντας το διάγραμμα της σημαντικότητας χαρακτηριστικών δεν προκύπτει κάτι τέτοιο.

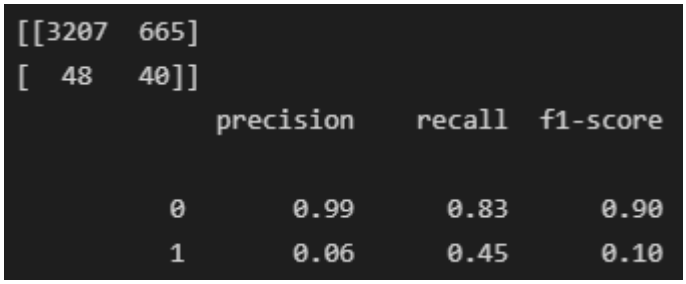
Περιφέρεια Ανατολικής Μακεδονίας – Θράκης:



Εικόνα 31: Πίνακας σύγχυσης και μετρικές επίδοσης για εκπαίδευση στην περιφέρεια Ανατολικής Μακεδονίας – Θράκης με την δεύτερη προσέγγιση

Αντίστοιχα με περιφέρεια Κεντρικής Μακεδονίας.

Περιφέρεια Θεσσαλίας:



Εικόνα 32: Πίνακας σύγχυσης και μετρικές επίδοσης για εκπαίδευση στην περιφέρεια Θεσσαλίας με την δεύτερη προσέγγιση

Αντίστοιχα με περιφέρεια Αττικής.

Από τα παραπάνω προκύπτει ότι η δεύτερη προσέγγιση δεν προσφέρει κάποιο πλεονέκτημα, και ενώ οι επιδόσεις της είναι υποδεέστερες της πρώτης. Για τον λόγο αυτόν δεν θα εφαρμοστεί σε συνδυασμούς περιφερειών.

Κεφάλαιο 6. Συμπεράσματα.

Στα πλαίσια αυτής της εργασίας επιχειρήσαμε να αναπτύξουμε ένα προγνωστικό εργαλείο για τα ξεσπάσματα του ιού του Δυτικού Νείλου για τις περιφέρειες της Ελλάδας στις οποίες υπάρχει τέτοια ανάγκη. Πέρα από την κλασσική μέθοδο για την μελέτη των χρονοσειρών δοκιμάστηκε η τεχνική του χωρισμού των δεδομένων μας ως προς τον χώρο, προκειμένου να αντιμετωπιστεί το πρόβλημα του μικρού χρονικού εύρους το οποίο αυτά καλύπτουν. Αυτή τη τεχνική δεν αποδείχθηκε βοηθητική. Αντίθετα, τα αποτελέσματα της πρώτης είναι σημαντικά καλύτερα από αυτά του τυχαίου ταξινομητή και, αν και δεν ήταν αντίστοιχα άλλων προσπαθειών που μελετήσαμε, παρουσιάζουν πραγματική προγνωστική ικανότητα, στις δύο περιφέρειες για τις οποίες είχαμε αρκετά, και αξιοποιήσιμα δεδομένα. Ακόμα, η παρούσα εργασία πραγματοποιήθηκε χωρίς δεδομένα σημαντικών παραμέτρων όπως αυτά των πληθυσμών των κουνουπιών, τα οποία στις περισσότερες έρευνες που μελετήθηκαν φάνηκαν να βελτιώνουν τις επιδόσεις. Αυτό οφείλεται στο ότι δεν βρέθηκαν διαθέσιμες πηγές τέτοιων δεδομένων για την περιοχή της Ελλάδας. Τα ενθαρρυντικά αποτελέσματα υποδεικνύουν ότι με την διαρκή συλλογή νέων δεδομένων καθώς και με την παρακολούθηση των χαρακτηριστικών που αποδεικνύονται σημαντική από την υπάρχουσα βιβλιογραφία, μοντέλα παρόμοια με αυτό που αναπτύχθηκε μπορούν να αποτελέσουν ένα χρήσιμο εργαλείο του συστήματος υγείας.

Βιβλιογραφία

[1] E. B. Hayes, N. Komar, R. S. Nasci, S. P. Montgomery, D. R. O'Leary, and G. L. Campbell, "Epidemiology and transmission dynamics of West Nile virus disease.," *Emerg Infect Dis*, vol. 11, no. 8, pp. 1167–73, Aug. 2005, doi: 10.3201/eid1108.050289a.

[2] K. C. Smithburn, T. P. Hughes, A. W. Burke, and J. H. Paul, "A Neurotropic Virus Isolated from the Blood of a Native of Uganda 1," *Am J Trop Med Hyg*, vol. s1-20, no. 4, pp. 471–492, Jul. 1940, doi: 10.4269/ajtmh.1940.s1-20.471.

[3] Centers for Disease Control and Prevention, "General Questions About West Nile Virus," Oct. 19, 2017. https://web.archive.org/web/20171026111330/https://www.cdc.gov/westnile/faq/genQuestions.html (accessed Nov. 01, 2022).

[4] Centers for Disease Control and Prevention, "Symptoms, Diagnosis, & Treatment," Jan. 15, 2019. https://web.archive.org/web/20170916140243/https://www.cdc.gov/westnile/symptoms/index.html (accessed Nov. 01, 2022).

[5] Εθνικός Οργανισμός Δημόσιας Υγείας, "Ιός του Δυτικού Νείλου." https://eody.gov.gr/disease/ios-toy-dytikoy-neiloy/ (accessed Nov. 01, 2022).

[6] P. Sampathkumar, "West Nile virus: epidemiology, clinical presentation, diagnosis, and prevention.," *Mayo Clin Proc*, vol. 78, no. 9, pp. 1137–43; quiz 1144, Sep. 2003, doi: 10.4065/78.9.1137.

[7] J. A. Kaiser and A. D. T. Barrett, "Twenty Years of Progress Toward West Nile Virus Vaccine Development.," *Viruses*, vol. 11, no. 9, 2019, doi: 10.3390/v11090823.

[8] L. R. Petersen, A. C. Brault, and R. S. Nasci, "West Nile virus: review of the literature.," *JAMA*, vol. 310, no. 3, pp. 308–15, Jul. 2013, doi: 10.1001/jama.2013.8042.

[9] V. Gamino and U. Höfle, "Pathology and tissue tropism of natural West Nile virus infection in birds: a review.," *Vet Res*, vol. 44, p. 39, Jun. 2013, doi: 10.1186/1297-9716-44-39.

[10] S. C. Weaver, C. Charlier, N. Vasilakis, and M. Lecuit, "Zika, Chikungunya, and Other Emerging Vector-Borne Viral Diseases.," *Annu Rev Med*, vol. 69, pp. 395–408, 2018, doi: 10.1146/annurev-med-050715-105122.

[11] S. Paz, "Climate change impacts on West Nile virus transmission in a global context.," *Philos Trans R Soc Lond B Biol Sci*, vol. 370, no. 1665, Apr. 2015, doi: 10.1098/rstb.2013.0561.

[12] J. S. Mackenzie, D. J. Gubler, and L. R. Petersen, "Emerging flaviviruses: the spread and resurgence of Japanese encephalitis, West Nile and dengue viruses.," *Nat Med*, vol. 10, no. 12 Suppl, pp. S98-109, Dec. 2004, doi: 10.1038/nm1144.

[13] P. E. R. U. Cynthia Goldsmith, "A micrograph of the West Nile Virus." 2014.

[14] Centers for Disease Control and Prevention, "Species of dead birds in which West Nile virus has been detected, United States, 1999-2016." 2016. Accessed: Nov. 01, 2022. [Online]. Available: https://www.cdc.gov/westnile/resources/pdfs/BirdSpecies1999-2016.pdf

[15] K. K. VanDalen, J. S. Hall, L. Clark, R. G. McLean, and C. Smeraski, "West Nile virus infection in American Robins: new insights on dose response.," *PLoS One*, vol. 8, no. 7, p. e68537, 2013, doi: 10.1371/journal.pone.0068537.

[16] A. Marm Kilpatrick, Shannon L. LaDeau, and Peter P. Marra, “Ecology of West Nile Virus Transmission and its Impact on Birds in the Western Hemisphere,” *Auk*, vol. 124, no. 4, pp. 1121–1136, Oct. 2007.
[17] World Health Organization, “West Nile Virus,” Mar. 28, 2019. https://www.who.int/news-room/fact-sheets/detail/west-nile-virus (accessed Nov. 01, 2022).
[18] V. Gamino and U. Höfle, “Pathology and tissue tropism of natural West Nile virus infection in birds: a review.,” *Vet Res*, vol. 44, p. 39, Jun. 2013, doi: 10.1186/1297-9716-44-39.
[19] Centers for Disease Control and Prevention, “Vertebrate Ecology,” Mar. 01, 2013. https://www.cdc.gov/westnile/index.html (accessed Nov. 01, 2022).
[20] Centers for Disease Control and Prevention, “Factsheet about West Nile virus infection.” https://www.ecdc.europa.eu/en/west-nile-fever/facts#:~:text=WNV%20was%20first%20isolated%20in,%2C%20Europe%2C%20Asia%20and%20Oceania. (accessed Nov. 01, 2022).
[21] Sandra Gonzalez Gompf, “West Nile Virus (Encephalitis),” May 08, 2021. https://www.medicinenet.com/west_nile_encephalitis/article.htm (accessed Nov. 01, 2022).
[22] Mayo Clinic, “West Nile Virus”, Accessed: Nov. 01, 2022. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/west-nile-virus/diagnosis-treatment/drc-20350325
[23] E. Olejnik, “Infectious adenitis transmitted by Culex molestus,” *Bull Res Counc Isr*, vol. 2, pp. 210–211, 1952.
[24] L. E. Davis *et al.*, “West Nile virus neuroinvasive disease.,” *Ann Neurol*, vol. 60, no. 3, pp. 286–300, Sep. 2006, doi: 10.1002/ana.20959.
[25] E. M. Flores Anticona, H. Zainah, D. R. Ouellette, and L. E. Johnson, “Two case reports of neuroinvasive west nile virus infection in the critical care unit.,” *Case Rep Infect Dis*, vol. 2012, p. 839458, 2012, doi: 10.1155/2012/839458.
[26] D. K. Mojumder, M. Agosto, H. Wilms, and J. Kim, “Is initial preservation of deep tendon reflexes in West Nile Virus paralysis a good prognostic sign?,” *Neurol Asia*, vol. 19, no. 1, pp. 93–97, Mar. 2014.
[27] D. S. Asnis, R. Conetta, A. A. Teixeira, G. Waldman, and B. A. Sampson, “The West Nile Virus outbreak of 1999 in New York: the Flushing Hospital experience.,” *Clin Infect Dis*, vol. 30, no. 3, pp. 413–8, Mar. 2000, doi: 10.1086/313737.
[28] S. P. Montgomery, C. C. Chow, S. W. Smith, A. A. Marfin, D. R. O’Leary, and G. L. Campbell, “Rhabdomyolysis in patients with west nile encephalitis and meningitis.,” *Vector Borne Zoonotic Dis*, vol. 5, no. 3, pp. 252–7, 2005, doi: 10.1089/vbz.2005.5.252.
[29] R. C. Anderson, K. B. Horn, M. P. Hoang, E. Gottlieb, and B. Bennin, “Punctate exanthem of West Nile Virus infection: report of 3 cases.,” *J Am Acad Dermatol*, vol. 51, no. 5, pp. 820–3, Nov. 2004, doi: 10.1016/j.jaad.2004.05.031.
[30] P. J. Carson *et al.*, “Long-term clinical and neuropsychological outcomes of West Nile virus infection.,” *Clin Infect Dis*, vol. 43, no. 6, pp. 723–30, Sep. 2006, doi: 10.1086/506939.
[31] A. L. Klee *et al.*, “Long-term prognosis for clinical West Nile virus infection.,” *Emerg Infect Dis*, vol. 10, no. 8, pp. 1405–11, Aug. 2004, doi: 10.3201/eid1008.030879.
[32] V. Sambri *et al.*, “West Nile virus in Europe: emergence, epidemiology, diagnosis, treatment, and prevention.,” *Clin Microbiol Infect*, vol. 19, no. 8, pp. 699–704, Aug. 2013, doi: 10.1111/1469-0691.12211.

[33] L. R. Petersen, A. C. Brault, and R. S. Nasci, “West Nile virus: review of the literature.,” *JAMA*, vol. 310, no. 3, pp. 308–15, Jul. 2013, doi: 10.1001/jama.2013.8042.
[34] K. K. Seino *et al.*, “Comparative efficacies of three commercially available vaccines against West Nile Virus (WNV) in a short-duration challenge trial involving an equine WNV encephalitis model.,” *Clin Vaccine Immunol*, vol. 14, no. 11, pp. 1465–71, Nov. 2007, doi: 10.1128/CVI.00249-07.
[35] P. Sampathkumar, “West Nile virus: epidemiology, clinical presentation, diagnosis, and prevention.,” *Mayo Clin Proc*, vol. 78, no. 9, pp. 1137–43; quiz 1144, Sep. 2003, doi: 10.4065/78.9.1137.
[36] K. C. Smithburn, T. P. Hughes, A. W. Burke, and J. H. Paul, “A Neurotropic Virus Isolated from the Blood of a Native of Uganda 1,” *Am J Trop Med Hyg*, vol. s1-20, no. 4, pp. 471–492, Jul. 1940, doi: 10.4269/ajtmh.1940.s1-20.471.
[37] H. BERNKOPF, S. LEVINE, and R. NERSON, “Isolation of West Nile virus in Israel.,” *J Infect Dis*, vol. 93, no. 3, pp. 207–18, doi: 10.1093/infdis/93.3.207.
[38] T. H. WORK, H. S. HURLBUT, and R. M. TAYLOR, “Isolation of West Nile virus from hooded crow and rock pigeon in the Nile delta.,” *Proc Soc Exp Biol Med*, vol. 84, no. 3, pp. 719–22, Dec. 1953, doi: 10.3181/00379727-84-20764.
[39] J. J. Sejvar, “West nile virus: an historical overview.,” *Ochsner J*, vol. 5, no. 3, pp. 6–10, 2003.
[40] M. J. Watts, V. Sarto I Monteys, P. G. Mortyn, and P. Kotsila, “The rise of West Nile Virus in Southern and Southeastern Europe: A spatial-temporal analysis investigating the combined effects of climate, land use and economic changes.,” *One Health*, vol. 13, p. 100315, Dec. 2021, doi: 10.1016/j.onehlt.2021.100315.
[41] D. Nash *et al.*, “The outbreak of West Nile virus infection in the New York City area in 1999.,” *N Engl J Med*, vol. 344, no. 24, pp. 1807–14, Jun. 2001, doi: 10.1056/NEJM200106143442401.
[42] Alexander T. Ciota and Laura D. Kramer, “Vector-Virus Interactions and Transmission Dynamics of West Nile Virus,” *Multidisciplinary Digital Publishing Institute*, vol. 5, no. 12, pp. 3021–3047, Dec. 2013.
[43] L. Brown, J. Medlock, and V. Murray, “Impact of drought on vector-borne diseases--how does one manage the risk?,” *Public Health*, vol. 128, no. 1, pp. 29–37, Jan. 2014, doi: 10.1016/j.puhe.2013.09.006.
[44] “Appendix 6: Topics for Consideration in Future Assessments. Climate Change Impacts in the United States: The Third National Climate Assessment,” Washington, DC, 2014. doi: 10.7930/J06H4FBF.
[45] R. Bekkerman, “The present and the future of the kdd cup competition: an outsider’s perspective.”
[46] Ramya Bhaskar Sundaram, “An End-to-End Guide to Understand the Math behind XGBoost,” Sep. 06, 2018. https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/ (accessed Nov. 06, 2022).
[47] Amazon, “How XGBoost Works.” https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html (accessed Nov. 06, 2022).
[48] Ashish Patel / Medium, “Boosting - Ensemble meta Algorithm for Reducing bias.” https://medium.com/ml-research-lab/boosting-ensemble-meta-algorithm-for-reducing-bias-5b8bfdce281 (accessed Nov. 06, 2022).
[49] xgboost developers, “Introduction to Boosted Trees.” https://xgboost.readthedocs.io/en/latest/tutorials/model.html (accessed Nov. 04, 2022).

[50] Tianqi Chen and Carlos Guestrin, “XGBoost: A Scalable Tree Boosting System,” 2016, Accessed: Nov. 04, 2022. [Online]. Available: https://arxiv.org/pdf/1603.02754.pdf
[51] Reena Shaw, “XGBoost concise technical overview,” Oct. 27, 2017. https://www.kdnuggets.com/2017/10/xgboost-concise-technical-overview.html (accessed Nov. 08, 2022).
[52] Aarshay Jain, “Complete Guide to Parameter Tuning in XGBoost with codes in Python,” *Analytics Vidhya*, Mar. 01, 2016. https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/ (accessed Nov. 08, 2022).
[53] Z. Farooq *et al.*, “Artificial intelligence to predict West Nile virus outbreaks with eco-climatic drivers,” *The Lancet Regional Health - Europe*, vol. 17, p. 100370, Jun. 2022, doi: 10.1016/j.lanepe.2022.100370.
[54] A. Ajith, K. Manoj, H. Kiran, P. J. Pillai, and J. J. Nair, “A Study on Prediction and Spreading of Epidemic Diseases,” in *2020 International Conference on Communication and Signal Processing (ICCSP)*, Jul. 2020, pp. 1265–1268. doi: 10.1109/ICCSP48568.2020.9182147.
[55] Aman Gupta, “XGBoost versus Random Forest,” Apr. 26, 2021.
[56] Motunrayo Olugbenga, “Balanced Accuracy: When Should You Use It,” *Neptune AI*, Jul. 22, 2022.
[57] Notebook Community, “Feature Importance and Feature Selection With XGBoost.” https://notebook.community/minesh1291/MachineLearning/xgboost/feature_importance_v1 (accessed Nov. 13, 2022).
[58] Miljan Kovačević, Nenad Ivanišević, Predrag Petronijević, and Vladimir Despotović, “Construction cost estimation of reinforced and prestressed concrete bridges using machine learning,” *Journal of the Croatian Association of Civil Engineers*, vol. 73, no. 01, pp. 1–13, Feb. 2021, doi: 10.14256/JCE.2738.2019.