



Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης Πολυτεχνική Σχολή

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Γραμμικών Υπολογιστών
Τομέας Τηλεπικοινωνιών

Διπλωματική Εργασία

Πρόβλεψη ρίσκου θνητότητας για νοσούντες της COVID-19 με τη χρήση XGBoost μοντέλων

Εκπόνηση:

Κωνιωτάκης Εμμανουήλ
ΑΕΜ: 8616

Επίβλεψη:

Λεόντιος Χατζηλεοντιάδης
Δημήτρης Ιακωβάκης

Θεσσαλονίκη, Οκτώβρης 2021

ΕΥΧΑΡΙΣΤΙΕΣ

Αρχικά, θα ήθελα να ευχαριστήσω τον Δημήτρη Ιακωβάκη, ο οποίος ήταν εξαιρετικός καθόλη τη διάρκεια της εκπόνησης της διπλωματικής μου εργασίας, καθώς ήταν πάντα διαθέσιμος και πρόσχαρος όποτε χρειαζόμουν κάποια βοήθεια ή συμβουλή με αστραπιαίο χρόνο ανταπόκρισης, δίνοντας μου συνοπτικές και κατατοπιστικές απαντήσεις. Επιπρόσθετα, θα ήθελα να ευχαριστήσω τον κ. Λεόντιο Χατζηλεοντιάδη, που μου έδωσε τη δυνατότητα να ασχοληθώ με τη συγκεκριμένη διπλωματική, η οποία μου επέτρεψε να κάνω μια πρώτη εμβάθυνση στις μεθόδους μηχανικής μάθησης, αλλά και στην επιστήμη των δεδομένων, 2 τομείς που συνδέονται και με τους οποίους θέλω να ασχοληθώ σχεδόν αποκλειστικά στο προσεχές μέλλον. Επίσης, θα ήθελα να τονίσω τη συλλογιστική συνεισφορά του κ. Χατζηλεοντιάδη, ο οποίος ακόμα κι από τις λίγες φορές που ήρθαμε σε επικοινωνία, κάθε φορά μου πρότεινε νέους τρόπους να αντιληφθώ το εκάστοτε πρόβλημα, προσπαθώντας να βρω καλύτερες λύσεις και να μάθω να σκέφτομαι περισσότερο σαν μηχανικός βλέποντας τα πράγματα πιο σφαιρικά και πολύπλευρα.

Πέραν του άμεσου κύκλου συνεργατών, θα ήθελα να ευχαριστήσω τους φίλους μου, οι οποίοι με βοήθησαν σε πολλούς διαφορετικούς τομείς τόσο κατά την εκπόνηση της διπλωματικής μου εργασίας, όσο και κατά τη διάρκεια των φοιτητικών μου χρόνων. Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου, οι οποίοι στέκονταν δίπλα μου σε κάθε δυσκολία πανεπιστημιακής φύσεως ή μη, όντας πάντα διαθέσιμοι και πρόθυμοι να με ακούσουν και να συζητήσουν.

Περίληψη

Η ασθένεια κορονοϊού 2019 (Coronavirus disease 2019, COVID-19 είναι μια μεταδοτική ασθένεια που προκαλείται από τον ιό SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) που όπως δηλώνει το όνομα του ιού, προσβάλλει σοβαρά το αναπνευστικό σύστημα του ξενιστή. Η νόσος αυτή εμφανίστηκε για πρώτη φορά το Δεκέμβρη του 2019 στην πόλη Γουχάν της Κίνας κι έκτοτε διασπάρθηκε σε παγκόσμιο επίπεδο δημιουργώντας μια πανδημία που ξεκίνησε από τις αρχές του 2020 και συνεχίζεται μέχρι και σήμερα.

Από τα αρχικά στάδια της πανδημίας άρχισε να φαίνεται η επικινδυνότητα του συγκεκριμένου ιού οδηγώντας στην προσβολή μεγάλου μέρους του ανθρώπινου πληθυσμού, ένα τμήμα του οποίου χρειαζόταν αναγκαστική νοσηλεία. Δεδομένης της έλλειψης προετοιμασίας σχεδόν πανανθρώπινα για την αντιμετώπιση μιας πανδημίας και της απότομης μετάδοσης του συγκεκριμένου ιού, σε πολλές χώρες υπήρξε αδυναμία στην εξυπηρέτηση όλων των σοβαρά νοσούντων ανά πάσα χρονική στιγμή.

Οι σοβαρά νοσούντες της COVID-19 συνήθως χρειάζονται αναπνευστική υποστήριξη και πολλοί από αυτούς καταλήγουν σε μονάδες εντατικής θεραπείας. Δεδομένου ότι οι εξοπλισμοί αναπνευστικής υποστήριξης κι οι μονάδες εντατικής θεραπείας δεν υπάρχουν σε αφθονία, πολλές φορές το σύστημα υγείας μιας χώρας δεν μπορεί να προσφέρει αυτές τις παροχές σε όλους τους ασθενείς ταυτόχρονα. Σε τέτοιες περιπτώσεις υψηλού νοσοκομειακού φόρτου κι έλλειψης παροχών, πρέπει οι γιατροί να εκτιμήσουν το ρίσκο θνητότητας μεταξύ ώστε να παρέχουν σε εκείνουν με το λιγότερο ρίσκο αναπνευστική υποστήριξη ή να τον εισάγουν σε μονάδα εντατικής θεραπείας.

Δεδομένης της δυσκολίας να παρθούν αποφάσεις τέτοιας φύσεως τόσο σε πρωτικό όσο και σε θεωρητικό επίπεδο, έγιναν προσπάθειες από διάφορες επιστημονικές ομάδες ανά τον κόσμο να κατασκευάσουν μοντέλα μηχανικής μάθησης τα οποία θα μπορούν να κάνουν τέτοιες εκτιμήσεις με αυτοματοποιημένο τρόπο. Λόγω του φαινομένου της πανδημίας σύσσωμη η επιστημονική κοινότητα έσπευσε να κατανοήσει γρήγορα το συγκεκριμένο ιό και να συλλέξει όλων των ειδών δεδομένα, τα οποία μπορούν να χρησιμοποιηθούν από επιστήμονες πολλών διαφορετικών κλάδων, ένας εκ των οποίων είναι κι η επιστήμη των δεδομένων.

Έτσι λοιπόν, η πιθανότητα να πεθάνει ένας ασθενής από την νόσο COVID-19 μπορεί να αναχθεί στην επίλυση ενός προβλήματος μηχανικής μάθησης έχοντας τα κατάλληλα δεδομένα για τον συγκεκριμένο ασθενή. Μέσω της παρούσας διπλωματικής εργασίας, έγιναν διάφορες απόπειρες να φτιαχτούν τέτοια μοντέλα με σκοπό την διερεύνηση πιθανών λύσεων για τέτοιου είδους προβλήματα κι αν οι λύσεις ήταν ικανοποιητικές, πιθανή χρήση των συγκεκριμένων μοντέλων και στην πράξη για την λήψη καλύτερων αποφάσεων. Σε γενικότερο επίπεδο, αναδεικνύεται η χρησιμότητα της συλλογής κι επεξεργασίας δεδομένων για την επίλυση δύσκολων προβλημάτων όπως η εκτίμηση θνητότητας ενός ασθενή, καθώς κι οι δυνατότητες που παρέχει η μηχανική μάθηση.

Title

Mortality risk prediction for COVID-19 patients using XGBoost models

Abstract

Emmanouil Koniotakis
Electrical & Computer Engineering Department,
Aristotle University of Thessaloniki, Greece
October 2021

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by the SARS-CoV-2 virus (Severe Acute Respiratory Syndrome Coronavirus 2) which, as the name implies, seriously affects the host's respiratory system. This disease first appeared in December 2019 in the city of Wuhan, China and has since spread worldwide leading to a pandemic that began in early 2020 and continues to this day.

From the early stages of the pandemic, the danger of this particular virus was evident, leading to the infection of a large part of the human population, a part of which needed compulsory hospitalization. Given the lack of almost universal preparedness to deal with a pandemic and the express transmission of this virus, many countries have been unable to help the entirety of the severely ill patients at any given time.

Severely ill patients from COVID-19 usually need respiratory support and many end up in intensive care units. Given that respiratory support equipment and intensive care units are not in abundance, many times the country's health care system cannot offer these benefits to all patients at the same time. In such cases of elevated hospital load and lack of such services, the doctors need to make assessments among patients about their mortality risk so that the one with the lower risk can be administered with respiratory support or be admitted to an intensive care unit.

Given the difficulty of making decisions of this nature both on a personal and theoretical level, efforts have been made by various scientific groups around the world to construct machine learning models that can make such assessments in an automated manner. Due to the pandemic phenomenon, the scientific community rushed quickly to understand the virus and collected all kinds of data, which can be used by scientists from many different fields, one of which is data science.

Thus, the probability of a patient dying from the COVID-19 disease can be reduced to solving a machine learning problem by having the appropriate data for that particular patient. Through this diploma thesis, various attempts have been made to build such models in order to explore possible solutions to such problems and should the solutions be satisfactory, there could be usage of these models in practice also, in order to make better decisions. In general, the usefulness of data collection and processing for solving difficult problems such as estimating a patient's mortality, as well as the possibilities provided by machine learning, is highlighted.

Περιεχόμενα

Ευχαριστίες	i
Περίληψη	iii
Abstract	v
Λέξεις Κλειδιά	xiii
1 Εισαγωγή	1
1.1 Περιγραφή του Προβλήματος	1
1.2 Σκοπός - Συνεισφορά της Διπλωματικής Εργασίας	2
1.3 Διάρθρωση της Αναφοράς	3
2 Επισκόπηση της Ερευνητικής Περιοχής	4
3 Θεωρητικό Υπόβαθρο	6
3.1 Μηχανική Μάθηση	6
3.1.1 Supervised Learning	7
3.1.2 Random Forest	9
3.1.3 Gradient Boosting	11
3.1.4 Binary Logistic Regression	12
3.2 XGBoost	14
3.2.1 XGBoost Parameters	15
3.2.2 Μετρικές Αξιολόγησης	17
3.3 SMOTE	18
3.4 Cross Validation	20
3.4.1 k-fold Cross Validation	20
4 Κατασκευή δεδομένων	21
4.1 Κατασκευή Δεδομένων	21
4.1.1 Αρχική μορφή δεδομένων	21
4.1.2 Κατασκευή δεδομένων συννοσηροτήτων	23
4.1.3 Κατασκευή δεδομένων σοβαρά νοσούντων	24
4.1.4 Ενοποίηση δεδομένων	24
4.2 Καθαρισμός Δεδομένων	28
4.2.1 Διαδικασία Καθαρισμού Δεδομένων	29
4.2.2 Οπτικοποίηση των τελικών δεδομένων	30
5 Κατασκευή Μοντέλων κι Αποτελέσματα	35
5.1 Εκπαίδευση Μοντέλων	35
5.2 Βελτιστοποίηση υπερπαραμέτρων	37

ΠΕΡΙΕΧΟΜΕΝΑ

5.3	Μέθοδοι Κατασκευής καλύτερων μοντέλων	43
5.4	Αποτελέσματα Μεθόδου 1	48
5.4.1	merged_data_no_td χωρίς το χαρακτηριστικό Severity	48
5.4.2	merged_data_no_td με το χαρακτηριστικό Severity	52
5.5	Αποτελέσματα Μεθόδου 2	56
5.6	Αποτελέσματα Μεθόδου 3	60
5.6.1	merged_data_no_td χωρίς το χαρακτηριστικό Severity για Diabetes = 1	61
5.6.2	merged_data_no_td χωρίς το χαρακτηριστικό Severity για Diabetes = 0	65
5.6.3	merged_data_no_td με το χαρακτηριστικό Severity για Diabetes = 1	69
5.6.4	merged_data_no_td με το χαρακτηριστικό Severity για Diabetes = 0	73
5.7	Σύγκριση Μοντέλων	77
6	Web Application	80
7	Συμπεράσματα και Προεκτάσεις	86
7.1	Συμπεράσματα	86
7.2	Προεκτάσεις	87
	Βιβλιογραφία	90

Κατάλογος Σχημάτων

3.1	Η μηχανική μάθηση ως υποσύνολο του AI	7
3.2	Η μηχανική μάθηση και το AI είναι ξεχωριστοί τομείς με επικάλυψη	7
3.3	Εποπτευόμενη Μάθηση	8
3.4	Τυχαία Δάση με ταξινόμηση	10
3.5	Τυχαία Δάση με παλινδρόμηση	10
3.6	Διαδικασία κατασκευής των gradient boosted trees	11
3.7	Επίδοση των gradient boosted trees ανά επανάληψη	12
3.8	Η λογιστική συνάρτηση με βάση ε κι η αντίστροφή της που είναι η πιθανότητα p	13
4.1	Εμπειρική κατανομή πιθανότητας της χρονικής μετατόπισης πρώτων μετρήσεων των ασθενών	25
4.2	Ιστογράμματα χαρακτηριστικών 0-3 ανά κλάση για merged_data_no_td	30
4.3	Ιστόγραμμα χαρακτηριστικών 4-7 ανά κλάση για merged_data_no_td	31
4.4	Ιστόγραμμα χαρακτηριστικών 8-11 ανά κλάση για merged_data_no_td	31
4.5	Ιστόγραμμα χαρακτηριστικών 12-15 ανά κλάση για merged_data_no_td	32
4.6	Ιστόγραμμα χαρακτηριστικών 16-19 ανά κλάση για merged_data_no_td	32
4.7	Ιστόγραμμα χαρακτηριστικών 20-22 ανά κλάση για merged_data_no_td	33
4.8	Ιστογράμματα χαρακτηριστικών 0-3 ανά κλάση για merged_data2 .	33
4.9	Ιστόγραμμα χαρακτηριστικών 4-7 ανά κλάση για merged_data2 . .	34
4.10	Ιστόγραμμα χαρακτηριστικών 8-11 ανά κλάση για merged_data2 . .	34
5.1	Τα 20 καλύτερα μοντέλα με βάση το σετ επικύρωσης από όλες τις προσεγγίσεις με τις αποδόσεις τους στο σετ επικύρωσης για το merged_data_no_td	41
5.2	Τα 20 καλύτερα μοντέλα με βάση το σετ επικύρωσης από όλες τις προσεγγίσεις με τις αποδόσεις τους στο σετ ελέγχου για το merged_data_no_td	41
5.3	Τα 20 καλύτερα μοντέλα με βάση το σετ επικύρωσης από όλες τις προσεγγίσεις με τις αποδόσεις τους στο σετ επικύρωσης για το merged_data2	42
5.4	Τα 20 καλύτερα μοντέλα με βάση το σετ επικύρωσης από όλες τις προσεγγίσεις με τις αποδόσεις τους στο σετ ελέγχου για το merged_data2	42
5.5	Τα 20 καλύτερα μοντέλα για το merged_data_no_td χωρίς το χαρακτηριστικό Severity	49
5.6	Confusion Matrix και καμπύλες ROC των 3 καλύτερων μοντέλων για το merged_data_no_td χωρίς το χαρακτηριστικό Severity	50
5.7	Το καλύτερο μοντέλο για το merged_data_no_td χωρίς το χαρακτηριστικό Severity	51

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

5.8 Τα 20 καλύτερα μοντέλα για το merged_data_no_td με το χαρακτηριστικό Severity	53
5.9 Confusion Matrix και καμπύλες ROC των 3 καλύτερων μοντέλων για το merged_data_no_td με το χαρακτηριστικό Severity	54
5.10 Το καλύτερο μοντέλο για το merged_data_no_td με το χαρακτηριστικό Severity	55
5.11 Στατιστικά μετρικών απόδοσης μοντέλων μεθόδου 2 για pos_neg_ratio = 1	57
5.12 Στατιστικά μετρικών απόδοσης μοντέλων μεθόδου 2 για pos_neg_ratio = 0.8	57
5.13 Στατιστικά μετρικών απόδοσης μοντέλων μεθόδου 2 για pos_neg_ratio = 0.5	58
5.14 Στατιστικά μετρικών απόδοσης μοντέλων μεθόδου 2 για pos_neg_ratio = 0.25	58
5.15 Στατιστικά μετρικών απόδοσης μοντέλων μεθόδου 2 για pos_neg_ratio = 0.2	59
5.16 Στατιστικά μετρικών απόδοσης μοντέλων μεθόδου 2 για pos_neg_ratio = 0.15	59
5.17 Τα 20 καλύτερα μοντέλα για το merged_data_no_td χωρίς το χαρακτηριστικό Severity για ασθενείς με Διαβήτη	62
5.18 Confusion Matrix και καμπύλες ROC των 3 καλύτερων μοντέλων για το merged_data_no_td χωρίς το χαρακτηριστικό Severity για ασθενείς με Διαβήτη	63
5.19 Το καλύτερο μοντέλο για το merged_data_no_td χωρίς το χαρακτηριστικό Severity για ασθενείς με Διαβήτη	64
5.20 Τα 20 καλύτερα μοντέλα για το merged_data_no_td χωρίς το χαρακτηριστικό Severity για ασθενείς χωρίς Διαβήτη	66
5.21 Confusion Matrix και καμπύλες ROC των 3 καλύτερων μοντέλων για το merged_data_no_td χωρίς το χαρακτηριστικό Severity για ασθενείς χωρίς Διαβήτη	67
5.22 Το καλύτερο μοντέλο για το merged_data_no_td χωρίς το χαρακτηριστικό Severity για ασθενείς χωρίς Διαβήτη	68
5.23 Τα 20 καλύτερα μοντέλα για το merged_data_no_td με το χαρακτηριστικό Severity για ασθενείς με Διαβήτη	70
5.24 Confusion Matrix και καμπύλες ROC των 3 καλύτερων μοντέλων για το merged_data_no_td με το χαρακτηριστικό Severity για ασθενείς με Διαβήτη	71
5.25 Το καλύτερο μοντέλο για το merged_data_no_td με το χαρακτηριστικό Severity για ασθενείς με Διαβήτη	72
5.26 Τα 20 καλύτερα μοντέλα για το merged_data_no_td με το χαρακτηριστικό Severity για ασθενείς χωρίς Διαβήτη	74
5.27 Confusion Matrix και καμπύλες ROC των 3 καλύτερων μοντέλων για το merged_data_no_td με το χαρακτηριστικό Severity για ασθενείς χωρίς Διαβήτη	75
5.28 Το καλύτερο μοντέλο για το merged_data_no_td με το χαρακτηριστικό Severity για ασθενείς χωρίς Διαβήτη	76

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

6.1	Σελίδα εφαρμογής με τα δεδομένα εισόδου που μπορεί να δώσει ο χρήστης - Τμήμα 1	81
6.2	Σελίδα εφαρμογής με τα δεδομένα εισόδου που μπορεί να δώσει ο χρήστης - Τμήμα 2	82
6.3	Σελίδα εφαρμογής με συμπληρωμένα δεδομένα εισόδου για τον ασθενή 672 - Τμήμα 1	83
6.4	Σελίδα εφαρμογής με συμπληρωμένα δεδομένα εισόδου για τον ασθενή 672 - Τμήμα 2	84
6.5	Λειτουργία κουμπιών "Show input passed" και "Make prediction" .	85

Κατάλογος πινάκων

5.1	Επισκόπηση των κλάσεων των πινάκων δεδομένων	36
5.2	Μετρικές απόδοσης για αξιολόγηση μοντέλων	39
5.3	Κατώφλια μετρικών απόδοσης για φιλτράρισμα μοντέλων για επιλογή καλύτερης προσέγγισης βελτιστοποίησης υπερπαραμέτρων	40
5.4	Επισκόπηση των κλάσεων του πίνακα δεδομένων merged_data_no_td ανάλογα με το υποσύνολο των ασθενών	45
5.5	Κατώφλια μετρικών απόδοσης για φιλτράρισμα μοντέλων για μέθοδο 1 για merged_data_no_td χωρίς το χαρακτηριστικό Severity	48
5.6	Κατώφλια μετρικών απόδοσης για φιλτράρισμα μοντέλων για μέθοδο 1 για merged_data_no_td με το χαρακτηριστικό Severity	52
5.7	Κατώφλια μετρικών απόδοσης για φιλτράρισμα μοντέλων για μέθοδο 3 για merged_data_no_td χωρίς το χαρακτηριστικό Severity για ασθενείς με Διαβήτη	61
5.8	Κατώφλια μετρικών απόδοσης για φιλτράρισμα μοντέλων για μέθοδο 3 για merged_data_no_td χωρίς το χαρακτηριστικό Severity για ασθενείς χωρίς Διαβήτη	65
5.9	Κατώφλια μετρικών απόδοσης για φιλτράρισμα μοντέλων για μέθοδο 3 για merged_data_no_td με το χαρακτηριστικό Severity για ασθενείς με Διαβήτη	69
5.10	Κατώφλια μετρικών απόδοσης για φιλτράρισμα μοντέλων για μέθοδο 3 για merged_data_no_td με το χαρακτηριστικό Severity για ασθενείς χωρίς Διαβήτη	73
5.11	Τα καλύτερα μοντέλα που επιλέχτηκαν από τις μεθόδους 1 και 3	77
5.12	Μοντέλο 225 vs Μοντέλο 939	78
5.13	Μοντέλο 281 vs Μοντέλο 451	78
5.14	Μοντέλο 207 vs Μοντέλο 1601	79

Λέξεις κλειδιά

Παρακάτω παρατίθενται ορισμένες από τις πιο συχνά χρησιμοποιούμενες λέξεις κλειδιά:

- gen_data: Δημογραφικά δεδομένα ασθενών
- lab_data: Αιματολογικά δεδομένα ασθενών
- sensor_data: Συμβατικά δεδομένα ασθενών
- merged_data_no_td: Πίνακας δεδομένων που περιέχει δημογραφικά, αιματολογικά και συμβατικά δεδομένα
- merged_data2: Πίνακας δεδομένων που περιέχει δημογραφικά και συμβατικά δεδομένα
- Severity: Δυαδικό χαρακτηριστικό σοβαρής νόσησης ασθενή
- Diabetes: Δυαδικό χαρακτηριστικό που καθορίζει αν ένας ασθενής έχει διαβήτη ή όχι
- th: Κατώφλι για την αφαίρεση ασθενών που έχουν κενές τιμές στα χαρακτηριστικά τους

1

Εισαγωγή

1.1 ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

Η γενική ιδέα του προβλήματος που πάμε να λύσουμε στη συγκεκριμένη διπλωματική είναι αν μπορεί να κατασκευαστεί κάποιο μοντέλο μηχανικής μάθησης χρησιμοποιώντας κάποια εύκολα για συλλογή δεδομένα από νοσούντες της COVID-19, το οποίο θα δίνει καλές εκτιμήσεις για το ρίσκο θνητότητας των ασθενών.

Σαφώς, το πρόβλημα αυτό γεννάει διάφορα ερωτήματα, όπως:

- Ποια είναι τα κατάλληλα δεδομένα για αυτόν τον σκοπό?
- Είναι εφικτή η κατασκευή τέτοιων μοντέλων? Αν ναι, τότε η κατασκευή τέτοιων μοντέλων πώς συγκρίνεται με την κρίση κάποιου γιατρού? Υπάρχει δηλαδή κάποια πρακτική σημασία στην κατασκευή τέτοιων μοντέλων?
- Με ποιον τρόπο θα κατασκευαστούν αυτά τα μοντέλα?

Εμείς θα εστιάσουμε κυρίως στο κατά πόσο μπορούν να φτιαχτούν τέτοια μοντέλα δοκιμάζοντας διάφορες μεθόδους κατασκευής με σκοπό να λάβουμε καλές εκτιμήσεις για το ρίσκο θνητότητας χρησιμοποιώντας δημογραφικά, αιματολογικά και συμβατικά δεδομένα όπως μέτρηση θερμοκρασίας σώματος, υψηλής πίεσης, κλπ. Επίσης, δίνουμε έμφαση στην ευκολία της συλλογής των δεδομένων και της αξιοποίησής τους με εύκολο τρόπο. Συγκεκριμένα, για έναν ασθενή της COVID-19 ο οποίος εισέρχεται στο νοσοκομείο για νοσηλεία, θέλουμε να εκτιμήσουμε το ρίσκο θνητότητας μονάχα από τα πρώτα δεδομένα που θα συλλεχθούν από αυτόν. Οι λόγοι που επιλέγεται αυτή η προσέγγιση είναι οι ακόλουθοι:

- Με αυτόν τον απλό τρόπο μπορούμε να λάβουμε μια αίσθηση του πόσο καλά μοντέλα μπορούν να φτιαχτούν αν χρησιμοποιούσαμε ακόμα πιο πολλά ή σύνθετα δεδομένα αλλά και δεδομένα συσσωρευτικού τύπου όπως μέσος όρος

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

θερμοκρασίας σώματος μέσα σε 1 εβδομάδα, κτλπ. Εξετάζουμε δηλαδή την επιλυσιμότητα του προβλήματος και τι περιθώρια ανάπτυξης υπάρχουν.

- Ο σημαντικότερος όμως λόγος είναι γιατί θέλουμε να λάβουμε μια έγκαιρη εκτίμηση για το ρίσκο θνητότητας των ασθενών που νοσηλεύονται χωρίς να απαιτείται καθημερινή συλλογή δεδομένων, έτσι ώστε να υπάρχει πρόγνωση της κατάστασης που θα βρεθούν οι ασθενείς, το οποίο συνδέεται άμεσα με τον υλικό εξοπλισμό που θα χρειαστεί το νοσοκομείο στο άμεσο μέλλον και στην περίπτωση έλλειψης εξοπλισμού, να μπορούν να γίνονται καλύτερες αποφάσεις για την επιλογή των περιθαλπτώμενων ασθενών.

1.2 ΣΚΟΠΟΣ - ΣΥΝΕΙΣΦΟΡΑ ΤΗΣ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Όπως αναφέρθηκε και παραπάνω, ο σκοπός της διπλωματικής αυτής είναι η κατασκευή μοντέλων μηχανικής μάθησης τα οποία χρησιμοποιώντας δημογραφικά, αιματολογικά και συμβατικά δεδομένα όπως μέτρηση θερμοκρασίας σώματος για ασθενείς της COVID-19, θα μπορούν να κάνουν καλές εκτιμήσεις για το ρίσκο θνητότητας των ασθενών αυτών.

Για τον σκοπό αυτόν έγιναν πολλές διαφορετικές υλοποιήσεις με σκοπό την εύρεση των καλύτερων δυνατών μοντέλων χρησιμοποιώντας τη βιβλιοθήκη XGBoost. Συγκεκριμένα, δοκιμάστηκαν τα παρακάτω:

- Έλεγχος απόδοσης μοντέλων που κατασκευάζονται με δημογραφικά και συμβατικά δεδομένα.
- Έλεγχος απόδοσης μοντέλων που κατασκευάζονται με δημογραφικά, αιματολογικά και συμβατικά δεδομένα.
- Έλεγχος απόδοσης μοντέλων αν χωριστούν οι ασθενείς σε δύο κατηγορίες έχοντας ή όχι μια δυαδική ιδιότητα, όπως αν έχουν Διαβήτη.
- Έλεγχος απόδοσης μοντέλων χρησιμοποιώντας τεχνικές βελτιστοποίησης των υπερπαραμέτρων του μοντέλου.
- Έλεγχος απόδοσης μοντέλων χρησιμοποιώντας τεχνικές μείωσης της ανισορροπίας των δεδομένων.
- Έλεγχος απόδοσης μοντέλων χρησιμοποιώντας όμοια υπο-μοντέλα εκπαιδευμένα σε λιγότερα δεδομένα τα οποία απαρτίζουν ένα υπερ-μοντέλο, οι προβλέψεις του οποίου αποτελούν τον μέσο όρο των προβλέψεων των επιμέρους υπο-μοντέλων.
- Έλεγχος απόδοσης μοντέλων χρησιμοποιώντας διάφορες μετρικές αξιολόγησης στο σετ επικύρωσης για την παύση της εκπαίδευσης εκεί που το μοντέλο αποδίδει καλύτερα.

1.3 ΔΙΑΡΘΡΩΣΗ ΤΗΣ ΑΝΑΦΟΡΑΣ

Η διάρθρωση της παρούσας διπλωματικής εργασίας είναι η εξής:

- **Κεφάλαιο 2:** Γίνεται ανασκόπηση της ερευνητικής δραστηριότητας στον συγκεκριμένο τομέα.
- **Κεφάλαιο 3:** Αναλύεται το θεωρητικό υπόβαθρο που χρησιμοποιήθηκε.
- **Κεφάλαιο 4:** Παρουσιάζεται η διαδικασία κατασκευής των δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση μοντέλων, καθώς και οι κατανομές των χαρακτηριστικών.
- **Κεφάλαιο 5:** Παρουσιάζονται οι προσεγγίσεις για τη βελτιστοποίηση των υπερπαραμέτρων των μοντέλων, διάφορες μέθοδοι κατασκευής μοντέλων, τα αποτελέσματα των μεθόδων αυτών, παρουσίαση των καλύτερων μοντέλων και σύγκριση αποτελεσμάτων μεταξύ μοντέλων.
- **Κεφάλαιο 6:** Παρουσιάζεται μια απλή διαδικτυακή εφαρμογή που κατασκευάστηκε που χρησιμοποιεί ένα μοντέλο για πρόβλεψη ρίσκου θνητότητας ασθενή του οποίου τα δεδομένα εισάγει ο χρήστης.
- **Κεφάλαιο 7:** Παρουσιάζονται τα συμπεράσματα της ανάλυσης καθώς και πιθανές προεκτάσεις, νέες ιδέες για την προσέγγιση του προβλήματος.

2

Επισκόπηση της Ερευνητικής Περιοχής

Στο κεφάλαιο αυτό γίνεται μία σύντομη αναφορά σε ήδη υπάρχουσες υλοποιήσεις στο κομμάτι της πρόβλεψης του ρίσκου θνητότητας ασθενών της COVID-19. Καθώς η διπλωματική πραγματεύεται ένα ζήτημα πολύ πρόσφατο, κατά την έναρξή της δεν βρέθηκαν άλλες υλοποιήσεις, πέραν της υλοποίησης του COVIDAnalytics.io¹[1].

Η ομάδα του COVIDAnalytics.io προσπάθησε να λύσει αρκιβώς αυτό το πρόβλημα χρησιμοποιώντας δημογραφικά, αιματολογικά και συμβατικά δεδομένα, όπως μέτρηση θερμοκρασίας σώματος. Τα δεδομένα που χρησιμοποιήσε η συγκεκριμένη ομάδα προήλθαν από τις παρακάτω πηγές:

- Νοσοκομεία από την πόλη Κρεμόνα², της Ιταλίας.
- Την ομάδα ομάδας νοσοκομείων HM Hospitales³ της Ισπανίας με 15 γενικά νοσοκομεία και 21 κλινικά κέντρα που καλύπτουν περιοχές της Μαδρίτης, της Καλικίας και της Λεόν.
- Το νοσοκομειακό δίκτυο Hartford HealthCare⁴, το οποίο εξυπηρετεί ασθενείς στο Κονέκτικατ της Αμερικής.

Χρησιμοποιώντας τα δεδομένα αυτά, κατασκεύασαν ένα μοντέλο πάνω σε 2781 ασθενείς που νοσηλεύτηκαν λόγω της νόσου COVID-19, εκ των οποίων το 25% πέθανε. Για την κατασκευή του μοντέλου τους χρησιμοποίησαν τη βιβλιοθήκη XGBoost και, έφτιαξαν μια απλή διαδικτυακή εφαρμογή όπου μπορούσε ο χρήστης να συμπληρώσει τα σχετικά δεδομένα που χρησιμοποιούσε το μοντέλο για κάποιον που

¹COVIDAnalytics.io, https://www.COVIDAnalytics.io/mortality_calculator

²Νοσοκομεία Κρεμόνας, <https://www.asst-cremona.it/>

³HM Hospitales, <https://www.hmhospitales.com/>

⁴Hartford HealthCare, <https://hartfordhealthcare.org/>

νοσεί από COVID-19 και να του δωθεί μια πρόβλεψη για το ρίσκο θνητότητας. Τα δεδομένα που μπορούσε να εισάγει ο χρήστης ήταν μετρήσεις μια ημέρας μονάχα. Το μοντέλο τους είχε μετρική απόδοσης auc = 0.9 χρησιμοποιώντας δημογραφικά, αιματολογικά και συμβατικά δεδομένα και auc = 0.83 χρησιμοποιώντας δημογραφικά και συμβατικά δεδομένα μόνο. Δυστυχώς, δεν υπήρχε αναφορά κάποιας άλλης μετρικής αξιολόγησης για τα συγκεκριμένα μοντέλα.

Δεδομένης της συγκεκριμένης υλοποίησης, η οποία είχε ξεκάθαρο στόχο και δομή, στη συγκεκριμένη διπλωματική έγινε η προσπάθεια να φτάσουμε την απόδοση του συγκεκριμένου μοντέλου του COVIDAnalytics.io. Έτσι, χρησιμοποιήσαμε τα ίδια δεδομένα όπου ήταν εφικτό, δηλαδή τα ίδια μεγέθη για τα δημογραφικά, αιματολογικά και συμβατικά δεδομένα. Επίσης, τα δεδομένα που χρησιμοποιήσαμε λήφθηκαν μονάχα από μια μέρα για κάθε ασθενή όπως και στο COVIDAnalytics.io. Κατά αυτόν τον τρόπο θα μπορούσαμε να κάνουμε άμεσες συγκρίσεις χρησιμοποιώντας τα δικά μας μοντέλα και το μοντέλο του COVIDAnalytics.io.

Να σημειωθεί ότι η συγκεκριμένη ομάδα, έκανε υλοποιήσεις και για άλλα αντίστοιχα προβλήματα, όπως:

- Πρόβλεψη ρίσκου νόσησης από COVID-19 χρησιμοποιώντας την ίδιο λογική με την πρόβλεψη ρίσκου θνητότητας.
- Εκτίμηση πορείας του επιδημιολογικού κύματος της COVID-19.
- Εκτίμηση ανάγκης εξοπλισμού αναπνευστικής υποστήριξης για συγκεκριμένα νοσοκομεία χρησιμοποιώντας δημογραφικά δεδομένα για το επιδημιακό φορτίο της COVID-19.

Δυστυχώς, η ιστοσελίδα της συγκεκριμένης ομάδας κατέβηκε στις 7 Οκτώβρη και δεν μπόρεσαν να γίνουν συγκρίσεις για συγκεκριμένους ασθενείς με το μοντέλο που είχε αναρτηθεί εκεί.

3

Θεωρητικό Υπόβαθρο

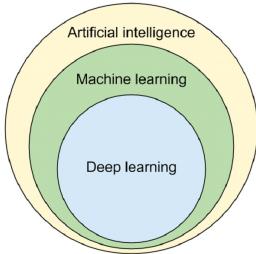
3.1 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Η μηχανική μάθηση (**Machine Learning - ML**)⁵ είναι η μελέτη υπολογιστικών αλγορίθμων οι οποίοι μπορούν να βελτιώνονται αυτόματα μέσω απόκτησης εμπειρίας και χρήσης δεδομένων[2]. Θεωρείται τμήμα της τεχνητής νοημοσύνης (**Artificial Intelligence - AI**). Οι αλγόριθμοι μηχανικής μάθησης κατασκευάζουν ένα μοντέλο χρησιμοποιώντας ένα δείγμα δεδομένων, τα **δεδομένα εκπαίδευσης**, με σκοπό να μπορέσουν να κάνουν προβλέψεις ή να παίρνουν αποφάσεις πάνω σε δεδομένα στα οποία δεν έχουν προγραμματιστεί προηγουμένως. Οι αλγόριθμοι μηχανικής μάθησης βρίσκουν χρησιμότητα σε ένα μεγάλο εύρος εφαρμογών, όπως είναι η ιατρική, η αναγνώριση φωνής και η υπολογιστική όραση (**Computer Vision**). Σκοπός των αλγορίθμων αυτών είναι να λύσουν προβλήματα τα οποία δεν επιδέχονται λύσεις με συμβατικούς αλγορίθμους. Ένας λειτουργικός ορισμός της μηχανικής μάθησης, όπως ορίστηκε από τον Tom M. Mitchell είναι: Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από την εμπειρία E σε σχέση με μια ομάδα στόχων T και μετρικής απόδοσης P αν η απόδοση στους στόχους T , όπως μετριέται από την P , βελτιώνεται με την εμπειρία E .[2] Την σήμερον ημέρα, η μηχανική μάθηση έχει δύο σκοπούς: Ο ένας είναι η ταξινόμηση δεδομένων με βάση μοντέλα που έχουν αναπτυχθεί κι ο άλλος είναι να κάνει προβλέψεις για το μέλλον με βάση αυτά τα μοντέλα.

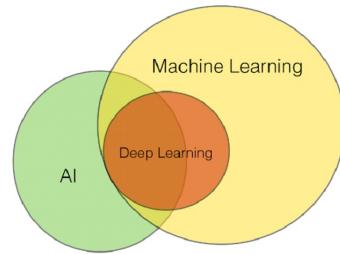
Ο όρος μηχανική μάθηση χρησιμοποιήθηκε πρώτη φορά το 1959 από τον Arthur Samuel και ξεκίνησε με στόχο την επίτευξη τεχνητής νοημοσύνης δίνοντας έμφαση

⁵Machine Learning, https://en.wikipedia.org/wiki/Machine_learning

σε συμβολικές μεθόδους, λογικές και βασισμένες σε γνώση προσεγγίσεις. Άρχισε να θεωρείται ξεχωριστός τομέας και να αναπτύσσεται στη δεκατία του 90, όπου ο νέος στόχος ήταν η αντιμετώπιση επιλύσιμων προβλημάτων πρακτικής φύσεως. Έτσι, απομακρύνθηκε από τις συμβολικές μεθόδους που είχε κληρονομήσει από την τεχνητή νοημοσύνη και κινήθηκε προς μεθόδους και μοντέλα βασισμένα στην στατιστική και την θεωρία πιθανοτήτων[3]. Ακόμα και σήμερα βέβαια, πολλές πηγές θεωρούν ότι η μηχανική μάθηση παραμένει υποσύνολο της τεχνητής νοημοσύνης[4][5], ενώ άλλες θεωρούν ότι μόνο ένα ”έξυπνο” υποσύνολο της μηχανικής μάθησης είναι τιμήμα της τεχνητής νοημοσύνης. Ο Judea Pearl έκανε το διαχωρισμό, ορίζοντας ότι στη μηχανική μάθηση το μοντέλο μαθαίνει και κάνει προβλέψεις βασισμένο σε παθητικές παρατηρήσεις, ενώ στην τεχνητή νοημοσύνη υπονοείται ότι υπάρχει κάποιος πράκτορας ο οποίος αλληλεπιδρά με το περιβάλλον για να μάθει και να δράσει ώστε να μεγιστοποιήσει τις πιθανότητές του να επιτύχει αποτελεσματικά τους στόχους του.



Σχήμα 3.1: Η μηχανική μάθηση ως υποσύνολο του AI



Σχήμα 3.2: Η μηχανική μάθηση και το AI είναι ξεχωριστοί τομείς με επικάλυψη

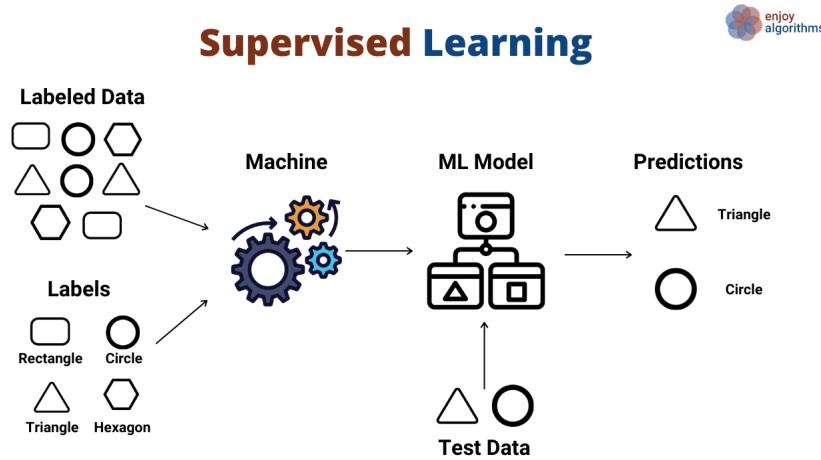
Η μηχανική μάθηση έχει άμεση σχέση με τη βελτιστοποίηση, καθώς πολλά προβλήματα δομούνται με την ελαχιστοποίηση κάποιας συνάρτησης κόστους πάνω στα δεδομένα εκπαίδευσης. Οι συναρήτσεις κόστους εκφράζουν την ασυμφωνία των προβλέψεων του μοντέλου πάνω στα δεδομένα εκπαίδευσης και των πραγματικών εμφανίσεων. Η διαφορά μεταξύ βελτιστοποίησης και μηχανικής μάθησης ανάγεται στο ότι η μηχανική μάθηση έχει σαν στόχο τη γενίκευση. Οι αλγόριθμοι βελτιστοποίησης ελαχιστοποιούν το κόστος πάνω στα δεδομένα εκπαίδευσης, ενώ η μηχανική μάθηση ασχολείται με την ελαχιστοποίηση του κόστους πάνω σε άγνωστα δεδομένα. Επιπρόσθετα, η μηχανική μάθηση συνδέεται άμεσα με τη στατιστική, άλλα διαφέροντα σε σχέση με τη βασικό τους στόχο: η στατιστική προσπαθεί να κάνει γενικεύσεις για έναν πληθυσμό από ένα δείγμα του, ενώ η μηχανική μάθηση εντοπίζει γενικά προβλέψιμα μοτίβα πάνω στον πληθυσμό από το δείγμα.[6]

3.1.1 Supervised Learning

Οι αλγόριθμοι εποπτευόμενης μάθησης (**Supervised Learning - SL**)[7] δημιουργούν ένα μαθηματικό μοντέλο ενός συνόλου δεδομένων που περιέχει τόσο τις εισόδους όσο και τις επιθυμητές εξόδους.[8] Τα δεδομένα είναι γνωστά ως δεδομένα εκπαίδευσης και αποτελούνται από ένα σύνολο παραδειγμάτων εκπαίδευσης.

ΚΕΦΑΛΑΙΟ 3. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

Κάθε παράδειγμα εκπαίδευσης έχει μία ή περισσότερες εισόδους και την επιθυμητή έξοδο, γνωστή και ως εποπτικό σύμα (**supervisory signal**). Στο μαθηματικό μοντέλο, κάθε παράδειγμα εκπαίδευσης αντιπροσωπεύεται από έναν πίνακα ή διάνυσμα, μερικές φορές ονομάζεται διάνυσμα χαρακτηριστικών και τα δεδομένα εκπαίδευσης αντιπροσωπεύονται από έναν πίνακα. Μέσω επαναληπτικής βελτιστοποίησης μιας συνάρτησης στόχου, οι αλγόριθμοι εποπτευόμενης μάθησης μαθαίνουν μια συνάρτηση που μπορεί να χρησιμοποιηθεί για την πρόβλεψη της εξόδου που σχετίζεται με νέες εισόδους. Μια βέλτιστη συνάρτηση θα επιτρέψει στον αλγόριθμο να καθορίσει σωστά την έξοδο για εισόδους που δεν ήταν μέρος των δεδομένων εκπαίδευσης.[9][10] Ένας αλγόριθμος που βελτιώνει την ακρίβεια των εξόδων ή των προβλέψεων του με την πάροδο του χρόνου λέγεται ότι έχει μάθει να εκτελεί αυτήν την εργασία.



Σχήμα 3.3: Εποπτευόμενη Μάθηση

3.1.2 Random Forest

Τα τυχαία δάση ή τα δάση τυχαίων αποφάσεων είναι μια μέθοδος εκμάθησης για ταξινόμηση, παλινδρόμηση κι άλλες εργασίες που λειτουργεί με την κατασκευή πολλών δέντρων αποφάσεων κατά την προπόνηση. Για εργασίες ταξινόμησης, η έξοδος του τυχαίου δάσους είναι η κλάση που επιλέγεται από τα περισσότερα δέντρα. Για εργασίες παλινδρόμησης, επιστρέφεται η μέση ή μέση πρόβλεψη των επιμέρους δέντρων. Τα δάση τυχαίων αποφάσεων διορθώνουν τη συνήθεια των δέντρων αποφάσεων να προσαρμόζονται υπερβολικά στο σετ εκπαίδευσής τους. Τα τυχαία δάση γενικά ξεπερνούν τα δέντρα αποφάσεων, αλλά η ακρίβειά τους είναι χαμηλότερη από τα δέντρα με gradient boosting. Ωστόσο, τα χαρακτηριστικά δεδομένων μπορούν να επηρεάσουν την απόδοσή τους.[11][12]

Τα δέντρα αποφάσεων είναι μια δημοφιλής μέθοδος για διάφορες εργασίες μηχανικής μάθησης, αλλά είναι επιρρεπή σε υπερεκπαίδευση όσο αυξάνεται το βάθος τους. Μαθαίνουν δηλαδή πολύ καλά το σετ εκπαίδευσης, έχοντας χαμηλή προκατάληψη, αλλά αποτυγχάνουν στο να κάνουν ταξινόμηση σε νέα δεδομένα, έχοντας υψηλή διακύμανση. Αυτό το φαινόμενο είναι γνωστό ως bias-variance tradeoff[13] στον χώρο της μηχανικής μάθησης. Τα τυχαία δάση είναι ένας τρόπος να γίνει ταξινόμηση σαν μέσος όρος των αποφάσεων πολλαπλών δέντρων απόφασης, που εκπαιδεύονται σε διαφορετικά μέρη του ίδιου εκπαιδευτικού σετ, με στόχο τη μείωση της διακύμανσης. Αυτό έρχεται σε βάρος μιας μικρής αύξησης της προκατάληψης και κάποια απώλεια ερμηνείας, αλλά γενικά ενισχύει σημαντικά την απόδοση στο τελικό μοντέλο. Τα δάση λαμβάνουν υπόψην την ομαδική εργασία πολλών δέντρων βελτιώνοντας έτσι την απόδοση ενός μόνο τυχαίου δέντρου. Αν και δεν είναι αρκετά παρόμοια, τα δάση δίνουν τα αποτελέσματα μιας σταυροειδούς επικύρωσης K-fold.

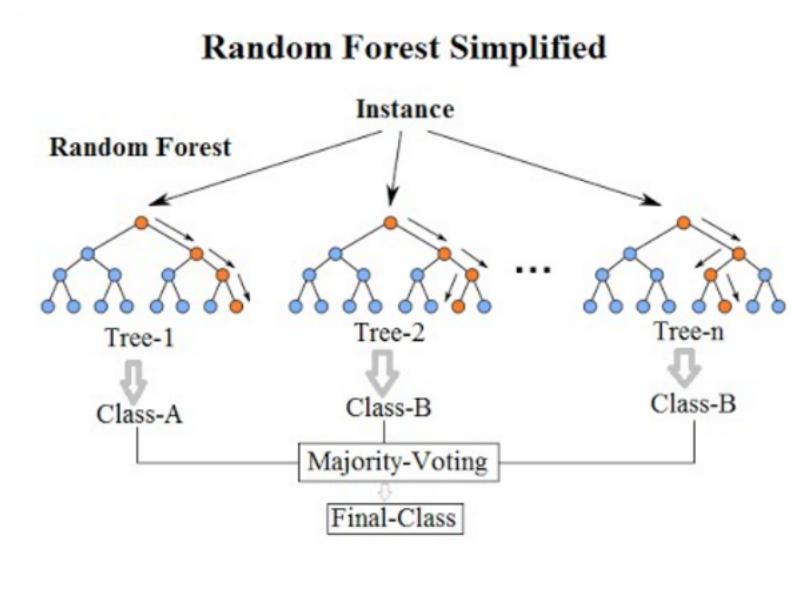
Ο αλγόριθμος εκπαίδευσης των τυχαίων δασών εφαρμόζει μια γενική τεχνική bootstrap aggregating ή bagging πάνω σε αλγορίθμους κατασκευής δέντρων, ως εξής:

Δεδομένου ενός σετ εκπαίδευσης με εισόδους $X = x_1, \dots, x_n$ κι εξόδους $Y = y_1, \dots, y_n$, κάνοντας bagging B φορές γίνεται επιλογή τυχαίων δειγμάτων με αντικατάσταση δημιουργώντας ένα δέντρο για κάθε δείγμα.

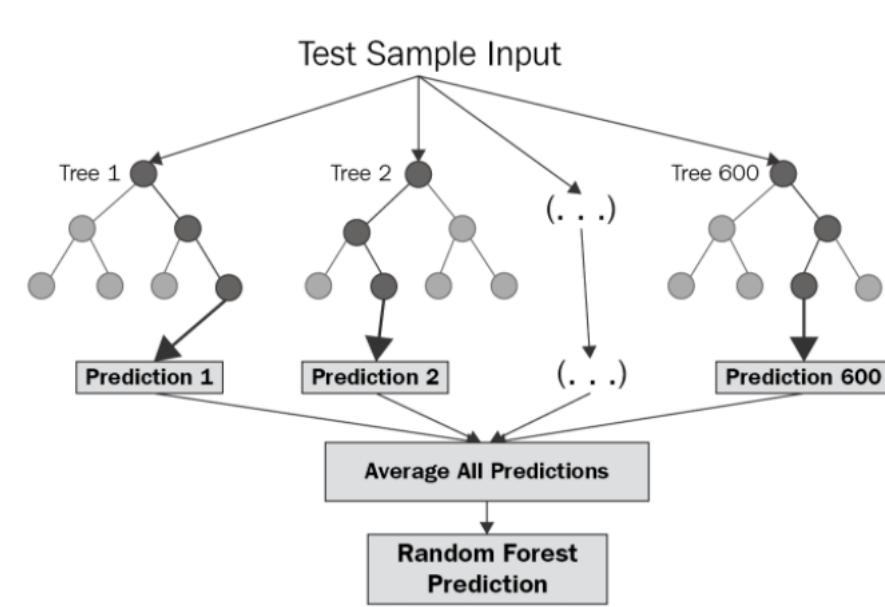
Συγκεκριμένα, για $b = 1, \dots, B$:

1. Δειγματοληψία με αντικατάσταση n παραδειγμάτων εκπαίδευσης από τα X, Y και δημιουργία του αντίστοιχου δείγματος X_b, Y_b .
2. Εκπαίδευση ενός δέντρου f_b πάνω στο δείγμα X_b, Y_b

Αφού εκπαιδευτούν όλα τα δέντρα, η πρόβλεψη πάνω σε άγνωστα δεδομένα γίνεται υπολογίζοντας τη μέση πρόβλεψη από όλα τα δέντρα f_b που κατασκευάστηκαν.



Σχήμα 3.4: Τυχαία Δάση με ταξινόμηση



Σχήμα 3.5: Τυχαία Δάση με παλινδρόμηση

3.1.3 Gradient Boosting

To gradient boosting[14] είναι μια τεχνική μηχανικής μάθησης για παλινδρόμηση, ταξινόμηση και άλλες εργασίες, η οποία παράγει ένα μοντέλο πρόβλεψης με τη μορφή ενός συνόλου ασθενών μοντέλων πρόβλεψης, συνήθως δένδρων αποφάσεων. Τα gradient boosted trees συνήθως έχουν καλύτερη επίδοση σε σχέση με τα τυχαία δάση. Η διαφορά με τα τυχαία δάση, είναι ότι στα gradient boosted trees τα δέντρα κατασκευάζονται το ένα μετά το άλλο με ένα γενικευμένο τρόπο επτρέποντας την χρήση οποιασδήποτε αυθαίρετης διαφορίσιμης συνάρτησης κόστους. Ο γενικός αλγόριθμος των gradient boosted trees φαίνεται παρακάτω:

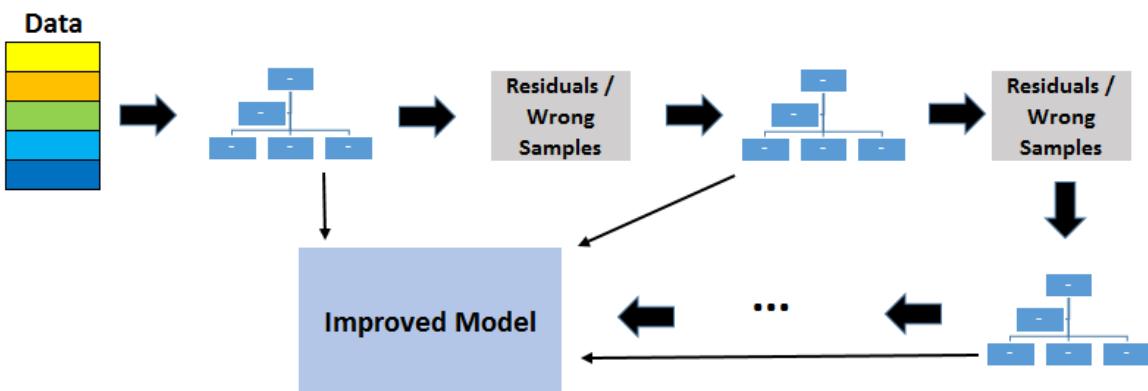
Έστω το σετ δεδομένων εκπαίδευσης $\{(x_i + y_i)\}_{i=1}^n$, μια διαφορίσιμη συνάρτηση κόστους $L(y, F(x))$ κι ο αριθμός επαναλήψεων M . Τότε για την κατασκευή των gradient boosted trees έχουμε:

1. Αρχικοποίηση του μοντέλου με μια σταθερά: $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^n L(y_i, \gamma)$.
2. Για $m = 1$ έως M :
 - (α') Ύπολογισμός των ψευδο-υπολοίπων:

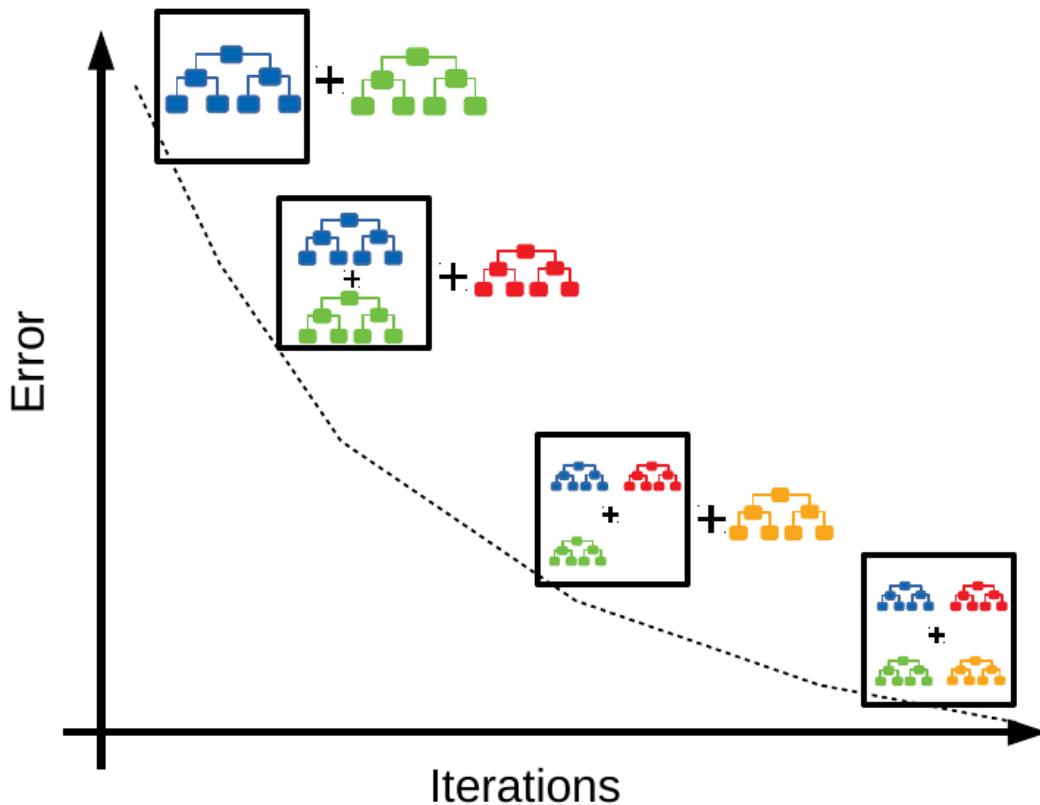
$$r_{im} = -[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}]_{F(x)=F_{m-1}(x)} \text{ για } i = 1, \dots, n.$$
 - (β') Κατασκευή ενός ασθενούς μοντέλου $h_m(x)$, εκπαιδευόντας το με το σετ δεδομένων εκπαίδευσης $\{(x_i + r_{im})\}_{i=1}^n$.
 - (γ') Ύπολογισμός του πολλαπλασιαστή γ_m λύνοντας το μονοδιάστατο πρόβλημα βελτιστοποίησης:

$$\gamma_m = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^n L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)).$$
 - (δ') Ανανέωση του μοντέλου:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$
3. Επιστροφή τελικού μοντέλου $F_M(x)$.



Σχήμα 3.6: Διαδικασία κατασκευής των gradient boosted trees



Σχήμα 3.7: Επίδοση των gradient boosted trees ανά επανάληψη

3.1.4 Binary Logistic Regression

Η λογιστική παλινδρόμηση (**logistic regression**)^[15] είναι ένα στατιστικό μοντέλο που στη βασική του μορφή χρησιμοποιεί μια λογιστική συνάρτηση για να μοντελοποιήσει μια δυαδική εξαρτημένη μεταβλητή. Μαθηματικά, ένα δυαδικό λογιστικό μοντέλο έχει μια εξαρτημένη μεταβλητή με δύο πιθανές τιμές, όπως το επιτυχία/αποτυχία που αντιπροσωπεύεται από μια δείκτρια μεταβλητή, με τιμές ενδείξεων "0" και "1". Στο λογιστικό μοντέλο, οι λογαριθμικές πιθανότητες, δηλαδή ο λογάριθμος των πιθανοτήτων, για την τιμή που φέρει την ένδειξη "1" είναι ένας γραμμικός συνδυασμός μίας ή περισσοτέρων ανεξάρτητων μεταβλητών (προγνωστικές μεταβλητές). Οι ανεξάρτητες μεταβλητές μπορούν να είναι μια δυαδική μεταβλητή (δύο κλάσεις, κωδικοποιημένες από μια δείκτρια μεταβλητή) ή μια συνεχής μεταβλητή (οποιαδήποτε πραγματική τιμή). Η αντίστοιχη πιθανότητα της τιμής με την ένδειξη "1" μπορεί να κυμαίνεται μεταξύ 0 (σίγουρα η τιμή "0") και 1 (σίγουρα η τιμή "1"), εξ'ου και η χρήση των τιμών "0" και "1". Η συνάρτηση που μετατρέπει τις λογαριθμικές πιθανότητες σε πραγματική πιθανότητα είναι η λογιστική συνάρτηση. Η μονάδα μέτρησης για την κλίμακα λογαριθμικών πιθανοτήτων ονομάζεται logit, που αντιστοιχεί σε ένα logistic unit.

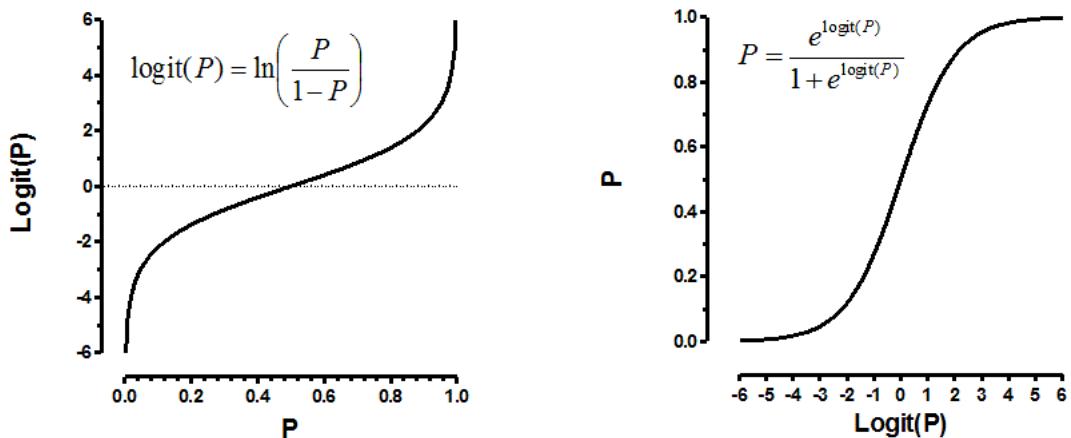
Το μοντέλο λογιστικής παλινδρόμησης δίνει μια πιθανότητα σαν έξοδο με βάση

μια συγκεκριμένη είσοδο και συνεπώς δεν εκτελεί στατιστική ταξινόμηση. Παρ' όλα αυτά, μπορεί να χρησιμοποιηθεί για την ταξινόμηση, επιλέγοντας μια τιμή αποκοπής (συνήθως 0.5) και ταξινομώντας τις εισόδους με πιθανότητα μεγαλύτερη από την αποκοπή ως μία κατηγορία κι εκείνες με πιθανότητα κάτω από την αποκοπή ως την άλλη κατηγορία.

Μαθηματικά, ένα λογιστικό μοντέλο μπορεί να διατυπωθεί ως εξής:
 Έστω ότι έχουμε τις ανεξάρτητες μεταβλητές $X = x_1, \dots, x_n$ και μια δυαδική εξαρτημένη μεταβλητή εξόδου $Y = 0$ ή 1. Θεωρούμε ότι $p = P(Y = 1)$ κι υποθέτουμε μια γραμμική σχέση μεταξύ των ανεξάρτητων μεταβλητών X και των λογαριθμικών πιθανοτήτων του γεγονότος $Y = 1$. Ετσι, αν l είναι η λογαριθμική πιθανότητα, b είναι η βάση του λογαρίθμου και β_i οι παράμετροι του μοντέλου, τότε έχουμε:

$$\begin{aligned} l &= \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m \\ \iff & \frac{p}{1-p} = b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m} \\ \iff & p = \frac{b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m}}{b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m} + 1} \\ \iff & p = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m)}} \\ \iff & p = S_b(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m) \end{aligned}$$

Όπου S_b είναι η σιγμοειδής συνάρτηση με βάση b , η οποία είναι συνήθως το e .



Σχήμα 3.8: Η λογιστική συνάρτηση με βάση e κι η αντίστροφή της που είναι η πιθανότητα p

3.2 XGBoost

Το **XGBoost**[16][17] αποτελεί ακρώνυμο για το "Extreme Gradient Boosting", όπου ο όρος ήρθε από το paper Greedy Function Approximation: A Gradient Boosting Machine, Friedman[14]. Το XGBoost αποτελεί μια βιβλιοθήκη ανοιχτού λογισμικού, η οποία παρέχει ένα κανονικοποιημένο framework για gradient boosting υλοποιήσεις μηχανικής μάθησης[18][19]. Έγινε πολύ δημοφιλές από ομάδες που το χρησιμοποίησαν και κατάφεραν να νικήσουν διαγωνισμούς μηχανικής μάθησης.[20] Κάποια από τα χαρακτηριστικά που κάνουν το XGBoost να ξεχωρίζει σε σχέση με άλλους gradient boosting αλγορίθμους είναι:

- Έξυπνος τρόπος εισαγωγής ποινής σε δέντρα
- Αναλογική μείωση των φύλλων του δέντρου
- Newton Boosting
- Περισσότερες δυνατότητες για τυχαιοποίηση
- Μπορεί να υλοποιηθεί για απλά και διανεμημένα συστήματα, καθώς και για υπολογισμούς με εξωτερική μνήμη
- Αυτόματη επιλογή χαρακτηριστικών

Δεδομένου ενός σετ εκπαίδευσης, για την κατασκευή ενός μοντέλου μηχανικής μάθησης ψάχνουμε να βρούμε τις παραμέτρους Θ του μοντέλου ελαχιστοποιώντας μια συνάρτηση κόστους (**objective function**). Κάθε τέτοια συνάρτηση $Obj(\Theta)$ μπορεί να οριστεί με 2 όρους, ως εξής:

$$Obj(\Theta) = L(\Theta) + \Omega(\Theta)$$

Όπου ο όρος L είναι η συνάρτηση κόστους κατά την εκπαίδευση και η Ω αποτελεί έναν όρο κανονικοποίησης. Η συνάρτηση κόστους μετράει πόσο καλές προβλέψεις κάνει το μοντέλο πάνω στα δεδομένα εκπαίδευσης. Ο όρος κανονικοποίησης ελέγχει την συνθετότητα του μοντέλου με σκοπό την αποφυγή της υπερεκπαίδευσης. Τα μοντέλα του XGBoost προσπαθούν να βρουν μια ισορροπία μεταξύ αυτών των 2 όρων, κάτι που είναι γνωστό στη μηχανική μάθηση ως το bias-variance tradeoff[13].

Για όλα τα μοντέλα που κατασκευάζουμε χρησιμοποιούμε τη συνάρτηση κόστους της δυαδικής λογιστικής παλινδρόμησης[21]. Η μαθητιματική της μορφή φαίνεται παρακάτω:

$$L(\Theta) = - \sum_{n=1}^N (y_n \ln s_n + (1-y_n) \ln(1-s_n)), \text{ όπου } s_n = \sigma(\Theta^T x_n) \text{ και } \sigma(t) = \frac{1}{1+\exp(-t)} = P(\omega_1|x), \text{ όπου } P(\omega_1|x) \text{ η εκτιμώμενη πιθανότητα του μοντέλου να ανήκει η κάθε εγγραφή } x \text{ στην κλάση } \omega_1, \text{ όπου } \omega_1 \text{ και } \omega_2 \text{ οι δύο κλάσεις του προβλήματος.}$$

Τα μοντέλα που προκύπτουν από το XGBoost είναι (**Classification and Regression Trees (CART)**)[22][23], δηλαδή κάνουν ταξινόμηση κι επιστρέφουν και μια τιμή

σαν regression score σε κάθε φύλλο του δέντρου. Επίσης, συνδυάζουν τη δομή των Random Forest και των Gradient Boosted δέντρων κι έχουν σαν σκοπό να είναι απλά και αποδοτικά. Μαθηματικά, ο αλγόριθμος του XGBoost φαίνεται παρακάτω:

Έστω το σετ δεδομένων εκπαίδευσης $\{(x_i + y_i)\}_{i=1}^n$, μια διαφορίσιμη συνάρτηση κόστους $L(y, F(x))$, ο αριθμός ασθενών μοντέλων M κι ο ρυθμός εκμάθησης α . Τότε για την κατασκευή του XGBoost μοντέλου έχουμε:

$$1. \text{ Αρχικοποίηση του μοντέλου με μια σταθερά: } \hat{f}_{(0)}(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^N L(y_i, \theta).$$

2. Για $m = 1$ έως M :

(α') Υπολογισμός των gradients και των hessians:

$$\hat{g}_m(x_i) = [\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}]_{f(x)=\hat{f}_{m-1}(x)}$$

$$\hat{h}_m(x_i) = [\frac{\partial^2 L(y_i, f(x_i))}{\partial^2 f(x_i)}]_{f(x)=\hat{f}_{m-1}(x)}$$

για $i = 1, \dots, N$.

(β') Κατασκευή του ασθενούς μοντέλου $\hat{f}_m(x)$, εκπαιδευόντας το με το σετ δεδομένων εκπαίδευσης $\{(x_i, -\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)})\}_{i=1}^N$ λύνοντας το μονοδιάστατο πρόβλημα βελτιστοποίησης:

$$\hat{\phi}_m = \underset{\phi \in \Phi}{\operatorname{argmin}} \sum_{i=1}^N \frac{1}{2} \hat{h}_m(x_i) [-\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} - \phi(x_i)]^2.$$

$$\hat{f}_m(x) = \alpha \hat{\phi}_m(x)$$

(γ') Ανανέωση του μοντέλου:

$$\hat{f}_m(x) = \hat{f}_{m-1}(x) + \hat{f}_m(x).$$

$$3. \text{ Επιστροφή τελικού μοντέλου } \hat{f}(x) = \hat{f}_M(x) = \sum_{m=0}^M \hat{f}_m(x).$$

3.2.1 XGBoost Parameters

Παρακάτω παρουσιάζονται οι βασικές παράμετροι που παρέχονται από τη βιβλιοθήκη του XGBoost⁶ με σκοπό την γενική ρύθμιση των μοντέλων αλλά και την βελτιστοποίησή τους. Οι παρακάτω παράμετροι θα χωριστούν σε κατηγορίες ανάλογα με τον σκοπό που έχουν:

Ρυθμιστικές Παράμετροι:

- **booster:** Καθορίζει το είδος των μοντέλων που φτιάχνονται. Στην περίπτωσή μας χρησιμοποιούμε boosted trees οπότε δώθηκε η τιμή gbtree.

⁶XGBoost documentation, <https://xgboost.readthedocs.io/en/latest/index.html>

ΚΕΦΑΛΑΙΟ 3. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

- **base_score:** Η αρχική τιμή πρόβλεψης που δίνεται για κάθε εγγραφή. Ορίζεται και ως γενική προκατάληψη (global bias) και χρησιμοποιήθηκε η προκαθορισμένη τιμή που είναι 0.5.
- **num_boost_round:** Ο μέγιστος αριθμός των επαναλήψεων κατά την εκπαίδευση. Χρησιμοποιήθηκε η τιμή 1000.
- **early_stopping_rounds:** Ο μέγιστος αριθμός επαναλήψεων που πρέπει να γίνουν για να σταματήσει η εκπαίδευση, εφόσον δεν υπάρχει βελτίωση στο σετ επικύρωσης. Δώθηκε η τιμή 50.

Γενικές Παράμετροι:

- **max_depth:** Μέγιστο βάθος δέντρου. Η αύξηση αυτής της τιμής κάνει το μντέλο πιο σύνθετο και πιο πιθανό να υπερεκπαιδευτεί.
Διάστημα πιθανών τιμών: $[0, \infty]$
- **min_child_weight:** Το ελάχιστο άθροισμα των βαρών εγγραφών (hessian) που απαιτείται για να γίνει διαχωρισμός ενός κόμβου και να αποκτήσει παιδιά. Όσο αυξάνεται η τιμή αυτή, τόσο πιο συντηρητικός γίνεται ο αλγόριθμος.
Διάστημα πιθανών τιμών: $[0, \infty]$
- **eta:** Ο ρυθμός εκπαίδευσης. Όσο πιο μικρή η τιμή, τόσο πιο συντηρητικός γίνεται ο αλγόριθμος.
Διάστημα πιθανών τιμών: $[0, 1]$
- **subsample:** Ο λόγος υποδειγματοληψίας των εγγραφών εκπαίδευσης πριν ξεκινήσει η εκπαίδευση των δέντρων σε κάθε γύρο. Χαμηλότερες τιμές μειώνουν την πιθανότητα υπερεκπαίδευσης.
Διάστημα πιθανών τιμών: $[0, 1]$
- **colsample_by_tree:** Ο λόγος υποδειγματοληψίας των χαρακτηριστικών για την κατασκευή κάθε δέντρου. Χαμηλότερες τιμές μειώνουν την πιθανότητα υπερεκπαίδευσης.
Διάστημα πιθανών τιμών: $[0, 1]$
- **scale_pos_weight:** Ελέγχει το βάρος που δίνεται στις εγγραφές της θετικής κλάσης σε σχέση με την αρνητική. Χρήσιμη παράμετρος για ασύμμετρα δεδομένα.
Διάστημα πιθανών τιμών: $[0, \infty]$

Παράμετροι κανονικοποίησης:

- **min_split_loss (gamma):** Η ελάχιστη μείωση της συνάρτησης κόστους που απαιτείται ώστε να γίνει διαχωρισμός ενός κόμβου και να αποκτήσει παιδιά. Όσο πιο ψηλή είναι η τιμή, τόσο πιο συντηρητικός γίνεται ο αλγόριθμος.
Διάστημα πιθανών τιμών: $[0, \infty]$
- **lambda:** L2 κανονικοποίηση (Ridge Regression) των βαρών της συνάρτησης κόστους. Όσο πιο ψηλή είναι η τιμή, τόσο πιο συντηρητικός γίνεται ο αλγόριθμος.
Διάστημα πιθανών τιμών: $[0, \infty]$

- **alpha:** L1 κανονικοποίηση (Lasso Regression) των βαρών της συνάρτησης κόστους. Όσο πιο ψηλή είναι η τιμή, τόσο πιο συντηρητικός γίνεται ο αλγόριθμος. Διάστημα πιθανών τιμών: $[0, \infty]$

3.2.2 Μετρικές Αξιολόγησης

Κατά τη διαδικασία εκπαίδευσης ενός XGBoost μοντέλου χρησιμοποιείται μια ή πολλές μετρικές αξιολόγησης σε σχέση με το σετ επικύρωσης. Ο ρόλος τους είναι καθοδηγητικός, καθώς σε κάθε boosting επανάληψη γίνεται υπολογισμός της εκάστοτε μετρικής στο σετ επικύρωσης και το καλύτερο μοντέλο επιλέγεται με βάση το γύρο εκπαίδευσης όπου η μετρική αξιολόγησης για το σετ επικύρωσης φτάνει τη μεγαλύτερη τιμή. Να τονιστεί ότι η μετρική αξιολόγησης κι η συνάρτηση κόστους είναι διαφορετικές έννοιες καθώς η συνάρτηση κόστους επιδέχεται βελτιστοποίηση κατά την κατασκευή των μοντέλων, ενώ η μετρική αξιολόγησης δείχνει την επίδοση του μοντέλου στο σετ επικύρωσης σε κάθε γύρο και σταματάει την εκπαίδευση εκεί που το μοντέλο τα πηγαίνει καλύτερα. Έχει δηλαδή στόχο την παύση της εκπαίδευσης. Ωστόσο, για να έχουμε μια εποπτική εικόνα του πώς εξελίσσεται το μοντέλο μας ανά γύρο, μπορούμε να χρησιμοποιήσουμε μια μετρική αξιολόγησης για το σετ εκπαίδευσης και το σετ επικύρωσης για να παράξουμε ένα διάγραμμα μάθησης. Η εκπαίδευση συνεχίζεται μέχρι να ικανοποιηθεί η συνθήκη που θέτει η παράμετρος *early stopping rounds*. Η βιβλιοθήκη XGBoost περιέχει κάποιες έτοιμες συναρτήσεις για τη χρήση συγκεκριμένων μετρικών αξιολόγησης, αλλά για σκοπούς πληρότητας και καλύτερης διερεύνησης υλοποιήθηκαν και κάποιες επιπλέον. Οι μετρικές αξιολόγησης που χρησιμοποιήθηκαν, ανά κατηγορία, παρουσιάζονται παρακάτω:

Έτοιμες Μετρικές Αξιολόγησης:

- **error:** Ορίζεται ως ο λόγος $\frac{\text{wrong cases}}{\text{all cases}}$. Μπορεί να ειδωθεί κι ως $1 - \text{accuracy}$, όπου $\text{accuracy} = \frac{\text{correct cases}}{\text{all cases}}$. Όσο πιο χαμηλή τιμή παίρνει, τόσο το καλύτερο.
- **logloss:** Ορίζεται ως $\text{LogLoss} = -\frac{1}{N} \sum_{i=1}^N [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$, όπου N ο αριθμός των δειγμάτων, y_i η κλάση στην οποία ανήκει το i -στό δείγμα και p_i η πιθανότητα να ανήκει στη θετική κλάση το i -στό δείγμα. Όσο πιο χαμηλή τιμή έχει τόσο λιγότερες λάθος ταξινομήσεις γίνονται.
- **auc:** Ορίζεται ως την επιφάνεια κάτω από την καμπύλη ROC (Receiver operating characteristic)[24]. Όσο πιο υψηλή τιμή παίρνει τόσο πιο "δυαδικά" κάνει ταξινόμηση το μοντέλο, δηλαδή οι πιθανότητες τείνουν προς το 0 ή 1 για την αρνητική και θετική κλάση αντίστοιχα.
- **aucpr:** Ορίζεται ως την επιφάνεια κάτω από την Precision-Recall[25] καμπύλη. Όσο πιο υψηλή τιμή παίρνει τόσο πιο "δυαδικά" κάνει ταξινόμηση το μοντέλο, δηλαδή οι πιθανότητες τείνουν προς το 0 ή 1 για την αρνητική και θετική κλάση αντίστοιχα. Ουσιαστικά η επιφάνεια αυτή ταυτίζεται με το μέσο Precision για διαφορετικά κατώφλια πιθανότητας για ταξινόμηση στη θετική ή αρνητική κλάση.

- **map:** Αποτελεί μια παραλλαγή της μετρικής aucpr.

Υλοποιημένες Μετρικές Αξιολόγησης:

- **F1-score:** Είναι ο αρμονικός μέσος όρος των μετρικών Precision και Recall κι ορίζεται ως $F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$.
- **Balanced Accuracy:** Ορίζεται ως ο μέσος όρος του Sensitivity \equiv True Positive Rate (TPR) \equiv Recall $= \frac{TP}{TP+FN}$ και του Specificity \equiv True Negative Rate (TNR) $\equiv 1 - \text{False Positive Rate (FPR)} = \frac{TN}{FP+TN}$, δηλαδή $\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}$. [26]
- **Precision:** Ορίζεται ως ο λόγος $\frac{TP}{TP+FP}$.
- **Recall:** Ορίζεται ως ο λόγος $\frac{TP}{TP+FN}$.
- **MCC:** Σημαίνει Matthews Correlation Coefficient κι είναι μια μετρική η οποία εκτιμάει την ποιότητα ενός δυαδικού ταξινομητή. Ορίζεται ως ο λόγος $\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$. [27]

3.3 SMOTE

Η τεχνική SMOTE ορίζεται ως Synthetic Minority Over-sampling TTechnique[28]. Η συγκεκριμένη τεχνική βρίσκει εφαρμογή σε προβλήματα μηχανικής μάθησης όπου υπάρχει μεγάλη ανισορροπία μεταξύ του αριθμού των εγγραφών ανά κλάση[29]. Η ανισορροπία των κλάσεων σε ένα δυαδικό πρόβλημα ταξινόμησης μετριέται από τον λόγο $\frac{\text{positive class cases}}{\text{negative class cases}}$. Όσο αυτός ο λόγος απομακρύνεται από την μονάδα, τόσο αυξάνεται η ανισορροπία των κλάσεων. Συνήθως θεωρούμε ως θετική κλάση εκείνη η οποία έχει τις λιγότερες εγγραφές κι ως αρνητική κλάση εκείνη η οποία έχει τις περισσότερες. Με αυτήν την παραδοχή η θετική κλάση ορίζεται ως κλάση μειονότητας κι η αρνητική κλάση ως κλάση πλειονότητας. Η βασική λειτουργία του SMOTE είναι να δημιουργήσει συνθετικά δεδομένα για την κλάση μειονότητας κι η αρνητική κλάση ως κλάση πλειονότητας. Η βασική λειτουργία του SMOTE είναι να δημιουργήσει συνθετικά δεδομένα για την κλάση μειονότητας από τα ήδη υπάρχοντα, με σκοπό την μείωση της ανισορροπίας μεταξύ των 2 κλάσεων, ευελπιστώντας ότι θα βελτιωθεί κι η απόδοση του μοντέλου. Για τη δημιουργία των συνθετικών δειγμάτων χρησιμοποιείται ο αλγόριθμος kNN (k Nearest Neighbours)[30] πάνω στον χώρο των χαρακτηριστικών. Η γενική διαδικασία που ακολουθείται για αυτόν τον σκοπό μπορεί να οριστεί τμηματικά ως εξής:

1. Απομόνωση των εγγραφών της κλάσης μειονότητας.
2. Ορισμός των k κοντινότερων γειτόνων που θα χρησιμοποιηθούν για τη δημιουργία των συνθετικών δειγμάτων.
3. Ορισμός του νέου επιθυμητού λόγου $\frac{\text{positive class cases}}{\text{negative class cases}}$. Ο νέος λόγος καθορίζει πόσες νέες συνθετικές εγγραφές θα παραχθούν και κατά συνέπεια πόσες συνθετικές εγγραφές θα παραχθούν ανά υπάρχουσα εγγραφή κι αν θα παραχθεί συνθετική εγγραφή για κάθε υπάρχουσα εγγραφή.

4. Μέχρι να παραχθούν όλες οι ζητούμενες συνθετικές εγγραφές:

- Επιλογή μιας τυχαίας εγγραφής i της κλάσης μειονότητας κι εύρεση των k κοντινότερων γειτόνων για την εγγραφή i .

- Επιλογή κάποιου τυχαίου γείτονα j από τους k κοντινότερους γείτονες της εγγραφής i κι υπολογισμός της απόστασης:

$$dif(i, j) = feature_vector(i) - feature_vector(j)$$

Όπου $feature_vector(i)$ είναι το διάνυσμα χαρακτηριστικών της εγγραφής i και $feature_vector(j)$ είναι το διάνυσμα χαρακτηριστικών του γείτονα j της εγγραφής i .

- Δημιουργία της συνθετικής εγγραφής l , δηλαδή δημιουργία του διανύσματος χαρακτηριστικών της εγγραφής l , ως εξής:

$$feature_vector(l) = feature_vector(i) + T \cdot dif(i, j)$$

Όπου το T παίρνει μια τυχαία τιμή μέσα στο διάστημα $(0, 1)$.

Για $T = 1$ έχουμε $feature_vector(l) = feature_vector(j)$ και για $T = 0$ έχουμε $feature_vector(l) = feature_vector(i)$.

Κατά τον τρόπο αυτό, με τη χρήση της τεχνικής SMOTE επεκτείνεται ο χώρος των δεδομένων με σκοπό να δημιουργηθούν μεγαλύτερες και λιγότερο εξειδικευμένες περιοχές αποφάσεων για κοντινά δεδομένα των 2 κλάσεων, πράγμα που διευκολύνει την ταξινόμηση των δεδομένων.

Μια επιπλέον πρακτική κατά τη χρήση της τεχνικής SMOTE είναι η υποδειγματοληφία των δεδομένων της κλάσης πλειονότητας με σκοπό την περαιτέρω μείωση της ανισορροπίας των κλάσεων. Κατά τη διαδικασία αυτή βέβαια μπορεί να χαθεί πληροφορία από την κλάση πλειονότητας, οπότε για να έχει εφαρμογή η συγκεκριμένη πρακτική πρέπει τα δεδομένα να είναι ικανοποιητικά σε αριθμό, ώστε να μην υπάρξει σημαντική μείωση στην πληροφορία που χάνεται.

Στο συγκεκριμένο πρόβλημα που πάμε να λύσουμε, για τη χρήση της τεχνικής SMOTE ορίστηκαν 3 παράμετροι για τις απαραίτητες ρυθμίσεις, οι οποίες παρουσιάζονται παρακάτω:

Παράμετροι για χρήση της τεχνικής SMOTE:

- **smote_ratio:** Εδώ ορίζεται ο λόγος $\frac{\text{positive class cases}}{\text{negative class cases}}$ που ζητείται να επιτευχθεί μετά από τη χρήση της τεχνικής SMOTE στην κλάση μειονότητας.
- **neighbours:** Εδώ ορίζεται ο αριθμός των κοντινότερων γειτόνων k που θα χρησιμοποιηθούν κατά την εφαρμογή της τεχνικής SMOTE.
- **class_ratio:** Εδώ ορίζεται ο λόγος $\frac{\text{positive class cases}}{\text{negative class cases}}$ που ζητείται να επιτευχθεί, αφού χρησιμοποιηθεί η τεχνική SMOTE, κάνοντας υποδειγματοληφία στην κλάση πλειονότητας.

Είναι σημαντικό να τονίσουμε ότι για τη χρήση της τεχνικής SMOTE, είναι απαραίτητη η ανυπαρξία κενών τιμών σε όλα τα χαρακτηριστικά των δεδομένων. Κατά συνέπεια, πριν την εφαρμογή της πρέπει να γίνει μια συμπλήρωση των κενών τιμών με κάποια μέθοδο. Στο συγκεκριμένο πρόβλημα, χρησιμοποιήθηκε μια έτοιμη μέθοδος εκτίμησης των κενών τιμών από τις τιμές των υπόλοιπων χαρακτηριστικών.

ΚΕΦΑΛΑΙΟ 3. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

Είναι σαφές ότι με αυτή τη διαδικασία, εισάγονται σφάλματα στο σετ δεδομένων και κατά συνέπεια επηρεάζεται το τελικό αποτέλεσμα της ταξινόμησης. Για το λόγο αυτό είναι σημαντικό τα δεδομένα να έχουν ακεραιότητα, αξιοπιστία και να είναι σημαντικά σε αριθμό.

3.4 CROSS VALIDATION

To Cross Validation είναι μια τεχνική επικύρωσης ενός μοντέλου για την αξιολόγηση του τρόπου με τον οποίο οι προβλέψεις του, που είναι προϊόντα στατιστικής ανάλυσης γενικεύονται σε ένα ανεξάρτητο σύνολο δεδομένων[31]. Χρησιμοποιείται κυρίως σε ρυθμίσεις όπου ο στόχος είναι η πρόβλεψη και κάποιος θέλει να εκτιμήσει πόσο ακριβές θα αποδώσει ένα μοντέλο πρόβλεψης σε πραγματικά δεδομένα. Σε ένα πρόβλημα πρόβλεψης, ένα μοντέλο εκπαιδεύεται με τα δεδομένα εκπαίδευσης κι η απόδοσή του ελέγχεται πάνω σε ένα σύνολο άγνωστων δεδομένων ή δεδομένων που εμφανίστηκαν για πρώτη φορά (δεδομένα επικύρωσης ή ελέγχου). Ο στόχος της διασταυρούμενης επικύρωσης είναι να δοκιμάσει την ικανότητα του μοντέλου να προβλέπει νέα δεδομένα που δεν χρησιμοποιήθηκαν κατά την κατασκευή του, προκειμένου να επισημάνει προβλήματα όπως η υπερεκπαίδευση (overfitting) ή η προκατάληψη επιλογής (selection bias) και να δώσει μια αντικειμενικότερη εικόνα για το πώς το μοντέλο γενικεύει σε ένα ανεξάρτητο σύνολο δεδομένων που θα προέκυπτε σε ένα πραγματικό πρόβλημα.

3.4.1 k-fold Cross Validation

Για τη χρήση της τεχνικής k-fold Cross Validation[32], το αρχικό σύνολο δεδομένων χωρίζεται τυχαία σε k ίσου μεγέθους υποδείγματα. Από τα k υποδείγματα, ένα μόνο υπόδειγμα χρησιμοποιείται ως το σετ επικύρωσης για τον έλεγχο του μοντέλου και τα υπόλοιπα $k - 1$ υποδείγματα χρησιμοποιούνται ως το σετ εκπαίδευσης. Η κατηγοριοποίηση των υποδειγμάτων σε σετ εκπαίδευσης και επικύρωσης γίνεται με κυκλικό τρόπο k φορές, μέχρις ότου κάθε ένα από τα k υποδείγματα έχει χρησιμοποιηθεί ακριβώς μια φορά ως σετ επικύρωσης. Σε κάθε μια από αυτές τις k επαναλήψεις, υπολογίζεται κάποια μετρική απόδοσης του μοντέλου. Σαν τελική απόδοση του μοντέλου από όλες τις επαναλήψεις ορίζεται ο μέσος όρος των αποδόσεων από όλες τις επιμέρους k επαναλήψεις[33]. Το πλεονέκτημα αυτής της μεθόδου έναντι μιας μορφής επαναλαμβανόμενης τυχαίας υποδειγματοληψίας είναι ότι όλες οι παρατηρήσεις χρησιμοποιούνται τόσο για εκπαίδευση όσο και για επικύρωση, και κάθε παρατήρηση χρησιμοποιείται για επικύρωση ακριβώς μία φορά. Συνήθως επιλέγεται η τιμή 10 για το k , αλλά παραμένει προσδιορίσιμη παράμετρος ανάλογα με το πρόβλημα.

4

Κατασκευή δεδομένων

4.1 ΚΑΤΑΣΚΕΥΗ ΔΕΔΟΜΕΝΩΝ

Για το συγκεκριμένο πρόβλημα χρησιμοποιήθηκαν δεδομένα από τα **HM Hospitales**, που είναι ομάδα νοσοκομείων της Ισπανίας. Τα δεδομένα αυτά δώθηκαν στην Διεθνή Επιστημονική Κοινότητα με σκοπό την καλύτερη κατανόηση της νόσου Covid-19 και την καλύτερη αντιμετώπιση της πανδημίας. Τα δεδομένα περιέχουν ανώνυμες πληροφορίες ασθενών που νοσηλεύτηκαν σε κάποιο νοσοκομείο της παραπάνω αλυσίδας και είναι χωρισμένα σε διαφορετικά αρχεία ανάλογα με το είδος της πληροφορίας που περιέχουν. Αξίζει να σημειωθεί ότι στην πρωταρχική τους μορφή τα δεδομένα ήταν στην ισπανική γλώσσα κι έγινε μετάφρασή τους στην αγγλική δια χειρός. Επίσης, τα δεδομένα αυτά πέρασαν αρκετά στάδια τροποποίησης μέχρις ότου να ξεκινήσει η προεπεξεργασία τους για την κατασκευή μοντέλων. Θα αναλυθούν τα στάδια με τη σειρά στη συνέχεια.

4.1.1 Αρχική μορφή δεδομένων

Το κάθε αρχείο έχει τη δομή πίνακα, όπου η κάθε στήλη αντιπροσωπεύει έναν ασθενή κι η κάθε στήλη περιέχει ένα είδος πληροφορίας. Στην αρχική τους μορφή, οι κατηγορίες των αρχείων παρουσιάζονται παρακάτω, καθώς και οι στήλες που χρησιμοποιήθηκαν ανά κατηγορία:

- **Δημιογραφικά Δεδομένα (gen_data):** Τα δεδομένα αυτά περιέχουν γενικές δημιογραφικές πληροφορίες για τον κάθε ασθενή. Οι στήλες που χρησιμοποιο-

ΚΕΦΑΛΑΙΟ 4. ΚΑΤΑΣΚΕΥΗ ΔΕΔΟΜΕΝΩΝ

ήθηκαν είναι οι παρακάτω:

- **Age:** Η ηλικία του ασθενή
 - **Sex:** Το φύλο του ασθενή
 - **ID:** Ο κωδικός του ασθενή
 - **Date of Admission as Inpatient:** Η ημερομηνία που ξεκίνησε η νοσηλεία του ασθενή
 - **Reason for discharge as Inpatient:** Ο λόγος που σταμάτησε η νοσηλεία του ασθενή
 - **COVID diagnosis during admission:** Η διάγνωση του ασθενή για τη νόσο Covid-19 κατά την εισαγωγή του ασθενή για νοσηλεία
 - **Date of entry to the ICU:** Η ημερομηνία που έγινε εισαγωγή του ασθενή σε μονάδα εντατικής θεραπείας
 - **Days spent at ICU:** Οι ημέρες που πέρασε ο ασθενής στη μονάδα εντατικής θεραπείας
- **Δεδομένα βασικών μετρήσεων (sensor_data):** Τα δεδομένα αυτά περιέχουν μετρήσεις σταθερών που πραγματοποιούνται με απλούς αισθητήρες, όπως θερμοκρασία σώματος, κορεσμός οξυγόνου, κ.ο.κ. Οι στήλες που χρησιμοποιήθηκαν είναι οι παρακάτω:
 - **ID:** Ο κωδικός του ασθενή
 - **Constant record date:** Η ημερομηνία που έγιναν οι μετρήσεις
 - **Maximum blood pressure value:** Μέτρηση της υψηλής πίεσης
 - **Minimum blood pressure value:** Μέτρηση της χαμηλής πίεσης
 - **Temperature value:** Μέτρηση της θερμοκρασίας σώματος
 - **Heart rate value:** Μέτρηση των καρδιακών παλμών
 - **Oxygen saturation value:** Μέτρηση του κορεσμού οξυγόνου - **Δεδομένα αιματολογικών μετρήσεων (lab_data):** Τα δεδομένα αυτά περιέχουν μετρήσεις από αιματολογικές εξετάσεις. Οι στήλες που χρησιμοποιήθηκαν είναι οι παρακάτω:
 - **ID:** Ο κωδικός του ασθενή
 - **Test date:** Η ημερομηνία που έγιναν οι μετρήσεις
 - **Test Value:** Η τιμή της μέτρησης
 - **Measurement Units:** Οι μονάδες μέτρησης του μεγέθους που μετρήθηκε
 - **Lab Test Name:** Το μέγεθος που μετρήθηκε. Εδώ πέρα από την αιματολογικά μεγέθη που χρησιμοποιήθηκαν από την αιματολογική επιλογή των covidanalytics.io. Τα μεγέθη που επιλέχθηκαν φαίνονται παρακάτω:
 - * **Leukocytes ($10^3/L$):** Αριθμός λευκοκυττάρων ανά μL
 - * **Creatinine (mg/dL):** Μάζα κρεατινίνης ανά dL

- * **Aspartate Aminotransferase (U/L):** Ασπαρτική αμινοτρανσφεράση σε μονάδες ανά L
 - * **Sodium (mmol/L):** Συγκέντρωση νατρίου σε mmol/L
 - * **Prothrombin Time (s):** Χρόνος προθρομβίνης σε sec
 - * **Platelet Count ($10^3/L$):** Αριθμός αιμοπεταλίων ανά μL
 - * **Blood Glucose (mg/dL):** Πυκνότητα γλυκόζης αίματος σε mg/dL
 - * **Mean Corpuscular Hemoglobin (pg):** Μέση μάσα αιμοσφαιρίνης ανά ερυθρό αιμοσφαιρίο
 - * **C-Reactive Protein (mg/L):** Πυκνότητα C-αντιδρώσας πρωτεΐνης σε mg/L
 - * **Hemoglobin (g/dL):** Πυκνότητα αιμοσφαιρίνης σε g/dL
 - * **Alanine Aminotransferase (U/L):** Μονάδες αμινοτρασφεράσης της αλανίνης ανά L
 - * **Potassium (mmol/L):** Συγκέντρωση καλίου σε mmol/L
- **Δεδομένα διάγνωσης (ICD10_data):** Τα δεδομένα αυτά περιέχουν πληροφορίες για άλλες συννοσηρότητες των ασθενών πέραν της Covid-19 κωδικοποιημένα με βάση το πρωτόκολλο ICD-10. Οι στήλες που χρησιμοποιήθηκαν είναι οι παρακάτω:
 - **ID:** Ο κωδικός του ασθενή
 - **DIA_0i:** Η i-στή διάγνωση του ασθενή, όπου $i \in [1, 19]$
 - **POAD_0i:** Επιβεβαίωση της i-στής διάγνωσης κατά την εισαγωγή του ασθενή για νοσηλεία, όπου $i \in [1, 19]$

4.1.2 Κατασκευή δεδομένων συννοσηροτήτων

Με βάση την υλοποίηση του covidanalytics.io για την κατασκευή των μοντέλων χρησιμοποιήθηκαν ως χαρακτηριστικά και 4 γενικές κατηγορίες συννοσηροτήτων. Οι κατηγορίες αυτές είναι οι: Διαβήτης, Καρδιακή αρρυθμία, Χρόνια νεφρική νόσος και Στεφανιαία αθηροσκλήρωση. Αυτά τα χαρακτηριστικά λαμβάνουν σαφώς δυαδικές τιμές και στην περίπτωσή μας κατασκευάστηκαν με βάση τα δεδομένα διάγνωσης (ICD10_data). Ο τρόπος κατασκευής ήταν ο εξής:

- Αρχικοποίηση του πίνακα συννοσηροτήτων **comorbidity_data** με στήλες ID, Cardiac dysrhythmias, Chronic Kidney Disease, Coronary atherosclerosis, Diabetes με όλες τις στήλες συννοσηροτήτων μηδενισμένες και τη στήλη ID συμπληρωμένη με τα IDs των ασθενών από το ICD10_data.
- Έπειτα από έλεγχο μέσω στης ιστοσελίδας <https://www.icd10data.com/> βρέθηκαν οι κωδικοί που αντιστοιχούν σε μορφές των 4 συννοσηροτήτων που μας ενδιαφέρουν, οι οποίες για κάθε νόσο είναι:
 - Cardiac dysrhythmias: ICD10 code in [I49.0, I49.9]
 - Chronic Kidney Disease: ICD10 code in [N18.0, N18.9]
 - Coronary atherosclerosis: ICD10 code in [I25.0, I25.9]

- Diabetes: ICD10 code in [E08.0, E13.9]

Έτσι, για κάθε ασθενή έγινε έλεγχος στις στήλες DIA_0i για να διαπιστωθεί αν έχουν κάποιες από αυτές τις συννοσηρότητες κι η επιβεβαίωση ότι της είχαν κατά την εισαγωγή για νοσηλεία γινόταν μέσω των αντίστοιχων στηλών POAD_0i. Όποιος ασθενής είχε κάποια από τις 4 συννοσηρότες σημαδευόταν στην αντίστοιχη στήλη του πίνακα comorbidity_data με 1.

4.1.3 Κατασκευή δεδομένων σοβαρά νοσούντων

Έγινε μια προσπάθεια να φτιαχτεί ένα χαρακτηριστικό από τα δεδομένα, το οποίο θα δήλωνε τη σοβαρότητα της νόσησης από Covid-19. Ο τρόπος που κατασκευάστηκε αυτό το χαρακτηριστικό ήταν αρκετά απλός. Όσοι ασθενείς είχαν μπει σε μονάδες εντατικής θεραπείας (ICU) θεωρήθηκε ότι νόσησαν σοβαρά, οπότε δηλιμουργήθηκε ο πίνακας **severity_data** ο οποίος είχε τις στήλες ID και Severity. Η στήλη ID περιείχε τα IDs των ασθενών που υπήρχαν στον πίνακα gen_data. Το Severity θεωρήθηκε ως δυαδική μεταβλητή, οπότε για κάθε ασθενή, χρησιμοποιήθηκε η στήλη Days spent at ICU από τον πίνακα gen_data για να ελεγχθεί αν νόσησαν σοβαρά. Όσοι ασθενείς ήταν για τουλάχιστον 1 μέρα με βάση τη στήλη αυτή, θεωρήθηκαν ως σοβαρά νοσούντες και σημαδεύτηκαν με 1 στη στήλη Severity, ενώ οι υπόλοιποι σημαδεύτηκαν με 0.

4.1.4 Ενοποίηση δεδομένων

Αφού έγινε λήψη των δεδομένων στην αρχική τους μορφή με βάση τις πληροφορίες που χρειάζονται, το επόμενο βήμα ήταν να γίνει μια τροποποίηση στην μορφή των δεδομένων, ώστε να μπορέσει να γίνει ενοποίησή τους σε ένα μόνο πίνακα. Η διαδικασία που ακολουθήθηκε για το σκοπό αυτό ήταν η εξής:

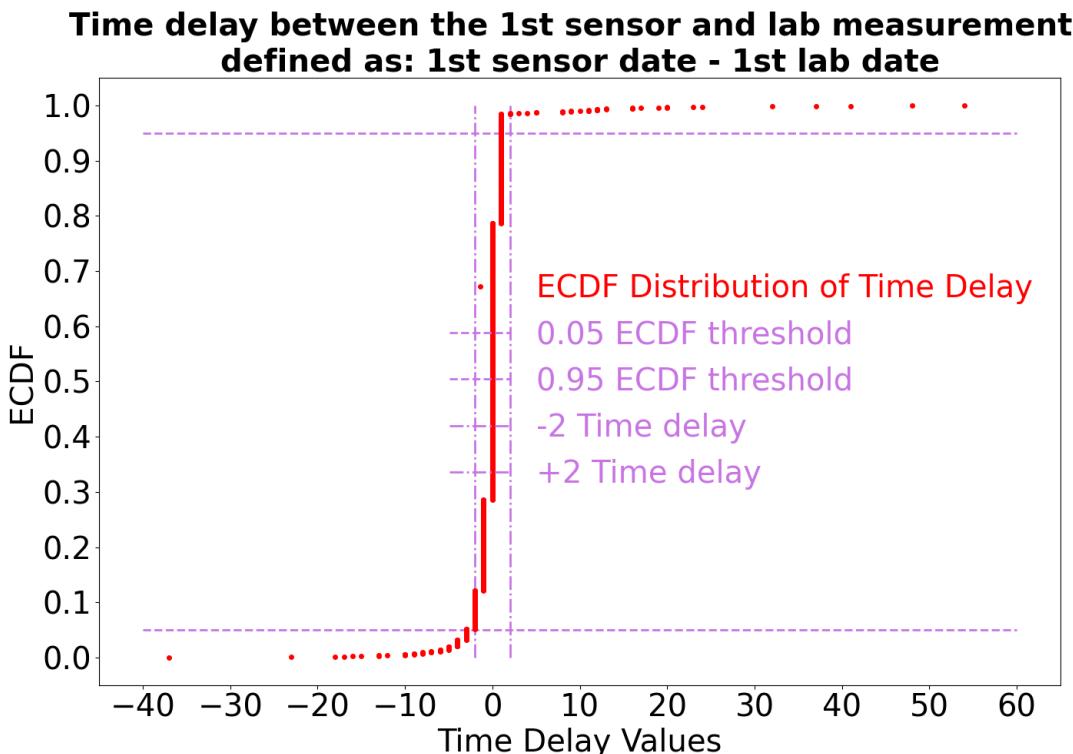
1. Τροποποίηση των ημερομηνιών στις στήλες Date of Admission as Inpatient (gen_data), Date of entry to the ICU (gen_data), Constant record date (sensor_data), Test Date (lab_data) ώστε να είναι στο ίδιο format και να μπορεί να γίνει επεξεργασία μεταξύ πινάκων.
2. Αρχικοποίηση των ακόλουθων πινάκων δεδομένων:
 - **merged_data_no_td**: Πίνακας δεδομένων μετά από ενοποίηση των επιμέρους πινάκων gen_data, sensor_data και lab_data. Η κατάληξη no_td υποδηλώνει no time delay μεταξύ των μετρήσεων από τον πίνακα sensor_data και lab_data. Για την κατασκευή του πίνακα merged_data_no_td χρησιμοποιήθηκαν μονάχα ασθενείς που είχαν δεδομένα και για τους 3 επιμέρους πίνακες gen_data, sensor_data και lab_data. Η διασταύρωση αυτή έγινε με βάση τη στήλη ID από κάθε πίνακα.
 - **merged_data2**: Πίνακας δεδομένων μετά από ενοποίηση των επιμέρους πινάκων gen_data και sensor_data. Ουσιαστικά δεν περιέχονται δεδομένα από αιμοτολογικές εξετάσεις. Για την κατασκευή του πίνακα merged_data2

χρησιμοποιήθηκαν μονάχα ασθενείς που είχαν δεδομένα και για τους 2 επιμέρους πίνακες gen_data, sensor_data. Η διασταύρωση αυτή έγινε με βάση τη στήλη ID από κάθε πίνακα.

- Για τον πίνακα δεδομένων merged_data_no_td, για κάθε ασθενή έγινε εύρεση της πρώτης και της τελευταίας ημερομηνίας που έγινε συλλογή δεδομένων βασικών μετρήσεων (sensor_data) και δεδομένων αιμοτολογικών μετρήσεων (lab_data) αντίστοιχα. Η κάθε ημερομηνία ελέγχθηκε ώστε να αναφέρεται είτε στην μέρα εισαγωγής του ασθενή για νοσηλεία είτε μετά από τη μέρα αυτή. Έπειτα, για να βεβαιωθεί ότι οι μετρήσεις στα sensor_data και στα lab_data αναφέρονται στις ίδιες ημερομηνίες, έγινε για κάθε ασθενή έλεγχος επικάλυψης θεωρώντας τα παράθυρα [ημερομηνία 1ης μέτρησης, ημερομηνία τελευταίας μέτρησης] για τα sensor_data και τα lab_data αντίστοιχα. Έπειτα, για τους ασθενείς που παρουσίαζαν επικάλυψη στα παράθυρα αυτά υπολογίστηκε η χρονική μετατόπιση:

time delay = date of 1st sensor measurement - date of 1st lab measurement.

Η κατανομή του time delay φαίνεται παρακάτω:



Σχήμα 4.1: Εμπειρική κατανομή πιθανότητας της χρονικής μετατόπισης πρώτων μετρήσεων των ασθενών

Έπειτα επιλέχθηκαν οι ασθενείς που είχαν time delay = 0, καθώς οι μετρήσεις αυτές είναι εκείνες που μας ενδιαφέρουν για την επίλυση του συγκεκριμένου προβλήματος. Η διαδικασία αυτή θα μπορούσε να απλοποιηθεί εφόσον έγινε επιλογή των ασθενών αυτών, αλλά προτιμήθηκε ώστε να υπάρξει μια εποπτική

ΚΕΦΑΛΑΙΟ 4. ΚΑΤΑΣΚΕΥΗ ΔΕΔΟΜΕΝΩΝ

εικόνα για τις ημερομηνίες των πρώτων μετρήσεων, αλλά και για περίπτωση εναλλακτικής υλοποίησης στο μέλλον.

Κατά αντιστοιχία, για τον πίνακα δεδομένων merged_data2 χρησιμοποιήθηκαν οι πρώτες μετρήσεις των δεδομένων βασικών μετρήσεων (sensor_data), οι οποίες έγιναν είτε την μέρα εισαγωγής του ασθενή για νοσηλεία είτε μετά τη μέρα αυτή.

4. Για τους εναπομείναντες ασθενείς στους πίνακες merged_data_no_td και merged_data2, αποθηκεύτηκαν οι τιμές των χαρακτηριστικών που μας ενδιαφέρουν από τους επιμέρους πίνακες δεδομένων gen_data, sensor_data και lab_data.
5. Οι πίνακες comorbidity_data και severity_data ενοποιήθηκαν με τους αντίστοιχους πίνακες merged_data_no_td και merged_data2 κρατώντας μόνο τους ασθενείς που υπήρχαν και στους 3 πίνακες κάθε φορά χρησιμοποιώντας τη στήλη ID του κάθε πίνακα.
6. Για τους τελικούς πίνακες merged_data_no_td και merged_data2 έγιναν τα εξής βήματα:
 - Οι τιμές Male και Female της στήλης Sex μετατράπηκαν σε 1 και 0 αντίστοιχα.
 - Οι τιμές Positive και Suspected Covid της στήλης COVID diagnosis during admission μετατράπηκαν σε 0 και 1 αντίστοιχα.
 - Διαγραφή ασθενών που έχουν τιμή 1 στη στήλη COVID diagnosis during admission.
 - Από τις πιθανές τιμές της στήλης Reason for discharge as Inpatient κρατήθηκαν ως σχετικές για το πρόβλημα μόνο οι Sent Home και Death, οι οποίες μετατράπηκαν σε 0 και 1 αντίστοιχα. Η στήλη Reason for discharge as Inpatient πήρε πλέον τη μορφή της δυαδικής κλάσης που μας ενδιαφέρει για το ρίσκο θνητότητας, καθώς το Sent Home μπορεί να ερμηνευθεί ως Alive και το Death ως Death.
 - Διαγραφή των στηλών ID, COVID diagnosis during admission, Date of entry to the ICU και Days spent at ICU που πλέον δεν χρειάζονται ώστε να μείνει μόνο η κλάση και τα χαρακτηριστικά.

Έτσι στην τελική μορφή τους, τα ενοποιημένα δεδομένα, δηλαδή οι πίνακες merged_data_no_td και merged_data2 περιέχουν τα παρακάτω χαρακτηριστικά:

merged_data_no_td

- **Age:** Ακέραια μεταβλητή που δηλώνει την ηλικία του ασθενή
- **Sex:** Δυαδική μεταβλητή που δηλώνει το φύλο του ασθενή, όπου το 1 αντιστοιχεί σε αρσενικό και το 0 σε θηλυκό
- **Reason for discharge as Inpatient:** Η δυαδική κλάση θνητότητας που πάιρνει τιμές 0 (Ζωντανός) και 1 (Νεκρός)
- **Maximum blood pressure value:** Συνεχής μεταβλητή που περιέχει τη μέτρηση της υψηλής πίεσης

- **Minimum blood pressure value:** Συνεχής μεταβλητή που περιέχει τη μέτρηση της χαμηλής πίεσης
- **Temperature value:** Συνεχής μεταβλητή που περιέχει τη μέτρηση της θερμοκρασίας σώματος
- **Heart rate value:** Συνεχής μεταβλητή που περιέχει τη μέτρηση των καρδιακών παλμών
- **Oxygen saturation value:** Συνεχής μεταβλητή που περιέχει τη μέτρηση του κορεσμού οξυγόνου
- **Leukocytes ($10^3/L$):** Συνεχής μεταβλητή που περιέχει τη μέτρηση του αριθμού λευκοκυττάρων ανά μL
- **Creatinine (mg/dL):** Συνεχής μεταβλητή που περιέχει τη μέτρηση της μάζας κρεατινίνης ανά dL
- **Aspartate Aminotransferase (U/L):** Συνεχής μεταβλητή που περιέχει τη μέτρηση της ασπαρτικής αμινοτρανσφεράσης σε μονάδες ανά L
- **Sodium (mmol/L):** Συνεχής μεταβλητή που περιέχει τη μέτρηση της συγκέντρωσης νατρίου σε mmol/L
- **Prothrombin Time (s):** Συνεχής μεταβλητή που περιέχει τη μέτρηση του χρόνου προθρομβίνης σε sec
- **Platelet Count ($10^3/L$):** Συνεχής μεταβλητή που περιέχει τη μέτρηση του αριθμού αιμοπεταλίων ανά μL
- **Blood Glucose (mg/dL):** Συνεχής μεταβλητή που περιέχει τη μέτρηση της πυκνότητας γλυκόζης αίματος σε mg/dL
- **Mean Corpuscular Hemoglobin (pg):** Συνεχής μεταβλητή που περιέχει τη μέτρηση της μέσης μάζας αιμοσφαιρίνης ανά ερυθρό αιμοσφαιρίο
- **C-Reactive Protein (mg/L):** Συνεχής μεταβλητή που περιέχει τη μέτρηση της πυκνότητας C-αντιδρώσας πρωτεΐνης σε mg/L
- **Hemoglobin (g/dL):** Συνεχής μεταβλητή που περιέχει τη μέτρηση της πυκνότητας αιμοσφαιρίνης σε g/dL
- **Alanine Aminotransferase (U/L):** Συνεχής μεταβλητή που περιέχει τη μέτρηση των μονάδων αμινοτρασφεράσης της αλανίνης ανά L
- **Potassium (mmol/L):** Συνεχής μεταβλητή που περιέχει τη μέτρηση της συγκέντρωσης καλίου σε mmol/L
- **Cardiac dysrhythmias:** Δυαδική μεταβλητή που δηλώνει αν ο ασθενής πάσχει παράλληλα από κάποια μορφή καρδιακής αρρυθμίας
- **Chronic Kidney Disease:** Δυαδική μεταβλητή που δηλώνει αν ο ασθενής πάσχει παράλληλα από κάποια μορφή χρόνιας νεφρικής νόσου
- **Coronary atherosclerosis:** Δυαδική μεταβλητή που δηλώνει αν ο ασθενής πάσχει παράλληλα από κάποια μορφή στεφανιαίας αθηροσκλήρωσης

ΚΕΦΑΛΑΙΟ 4. ΚΑΤΑΣΚΕΥΗ ΔΕΔΟΜΕΝΩΝ

- **Diabetes:** Δυαδική μεταβλητή που δηλώνει αν ο ασθενής πάσχει παράλληλα από κάποια μορφή Διαβήτη
- **Severity:** Δυαδική μεταβλητή που δηλώνει αν ο ασθενής νοσεί σοβαρά από Covid-19

merged_data2

- **Age:** Ακέραια μεταβλητή που δηλώνει την ηλικία του ασθενή
- **Sex:** Δυαδική μεταβλητή που δηλώνει το φύλο του ασθενή, όπου το 1 αντιστοιχεί σε αρσενικό και το 0 σε θηλυκό
- **Reason for discharge as Inpatient:** Η δυαδική κλάση θυητότητας που πάιρνει τιμές 0 (Ζωντανός) και 1 (Νεκρός)
- **Maximum blood pressure value:** Συνεχής μεταβλητή που περιέχει τη μέτρηση της υψηλής πίεσης
- **Minimum blood pressure value:** Συνεχής μεταβλητή που περιέχει τη μέτρηση της χαμηλής πίεσης
- **Temperature value:** Συνεχής μεταβλητή που περιέχει τη μέτρηση της θερμοκρασίας σώματος
- **Heart rate value:** Συνεχής μεταβλητή που περιέχει τη μέτρηση των καρδιακών παλμών
- **Oxygen saturation value:** Συνεχής μεταβλητή που περιέχει τη μέτρηση του κορεσμού οξυγόνου
- **Cardiac dysrhythmias:** Δυαδική μεταβλητή που δηλώνει αν ο ασθενής πάσχει παράλληλα από κάποια μορφή καρδιακής αρρυθμίας
- **Chronic Kidney Disease:** Δυαδική μεταβλητή που δηλώνει αν ο ασθενής πάσχει παράλληλα από κάποια μορφή χρόνιας νεφρικής νόσου
- **Coronary atherosclerosis:** Δυαδική μεταβλητή που δηλώνει αν ο ασθενής πάσχει παράλληλα από κάποια μορφή στεφανιαίας αθηροσκλήρωσης
- **Diabetes:** Δυαδική μεταβλητή που δηλώνει αν ο ασθενής πάσχει παράλληλα από κάποια μορφή Διαβήτη
- **Severity:** Δυαδική μεταβλητή που δηλώνει αν ο ασθενής νοσεί σοβαρά από Covid-19

4.2 ΚΑΘΑΡΙΣΜΟΣ ΔΕΔΟΜΕΝΩΝ

Προτού ξεκινήσει η εκπαίδευση των μοντέλων, πρέπει πρώτα να βεβαιωθούμε ότι οι τιμές που έχουν τα δεδομένα των πινάκων merged_data_no_td και merged_data2 είναι στη σωστή μορφή και να κάνουμε οποιαδήποτε τροποποίηση χρειάζεται σε σχέση με τις τιμές των χαρακτηριστικών. Αξίζει να σημειωθεί ότι οι διορθώσεις που

αναφέρονται παρακάτω δεν άλλαξαν τη μορφή των πινάκων merged_data_no_td και merged_data2 όπως παρουσιάστηκαν στην τελική τους μορφή στην προηγούμενη υποενότητα. Χάριν απλότητας οι στήλες που αφαιρέθηκαν μετά την κατασκευή των πινάκων αυτών δεν αναφέρθηκαν καθόλου, ώστε να παρουσιαστεί συνοπτικά η διαδικασία του καθαρισμού χωρίς να υπάρχει παραπάνω σύγχυση στον αναγνώστη.

4.2.1 Διαδικασία Καθαρισμού Δεδομένων

Η διαδικασία που ακολουθήθηκε για τον καθαρισμό των δεδομένων είναι η εξής:

- Διόρθωση τίτλων χαρακτηριστικών για τις περιπτώσεις που δεν έγινε σωστή μετάφραση λόγω καδικοποιητή της ισπανικής γλώσσας κατά την μετατροπή στην αγγλική γλώσσα.
- Διόρθωση τιμών που αντιστοιχούν σε κενές εγγραφές, αλλά λανθασμένα έχουν μηδενιστεί κατά την δημιουργία των δεδομένων. Παραδείγματος χάριν, για το χαρακτηριστικό Maximum blood pressure value υπήρχαν πολλές τιμές 0 που προφανώς δεν είναι λογικές τιμές για το συγκεκριμένο μέγεθος κι αντικαταστάθηκαν με NaN τιμές.
- Διαγραφή χαρακτηριστικών τα οποία είχαν κενές τιμές για τουλάχιστον 50% των εγγραφών. Για παράδειγμα η στήλη Nitrogen, που δεν αναφέρθηκε προηγουμένως έπασχε από αυτό το πρόβλημα κι αφαιρέθηκε.
- Διαγραφή χαρακτηριστικών τα οποία είχαν ίδιες τιμές για τουλάχιστον 50% των εγγραφών. Για τον συγκεκριμένο έλεγχο αγνοήθηκαν οι στήλες Reason for discharge as Inpatient, Sex, Severity, Diabetes, Coronary atherosclerosis, Cardiac dysrhythmia και Chronic Kidney Disease για προφανείς λόγους.
- Αντικατάσταση των ακραίων τιμών των χαρακτηριστικών με κενές τιμές. Η αναγνώριση των ακραίων τιμών έγινε με 2 τρόπους:
 - Επιτρεπτές τιμές θεωρήθηκαν μόνο όσες βρίσκονταν στο διάστημα τιμών [Q1 - 1.5IQR, Q3 + 1.5IQR], όπου Q1 το πρώτο τεταρτημόριο, Q3 το τρίτο τεταρτημόριο και IQR το ενδοτεταρτημοριακό εύρος.
 - Ελέγχοντας τα ιστογράμματα των χαρακτηριστικών για παράλογες τιμές ή μη σχετικές με το πρόβλημα. Κατά αυτόν τον τρόπο για παράδειγμα διαπιστώθηκε ότι υπήρχαν εγγραφές ανηλίκων ($Age < 18$) μεταξύ των ασθενών, οι οποίες μετά αφαιρέθηκαν. Τα ιστογράμματα αυτά χωρίστηκαν ανά κλάση (Ζωντανός ή Νεκρός) ώστε να δωθεί επίσης μια εποπτική εικόνα για την κατανομή των δεδομένων στις κλάσεις αυτές. Αυτά τα ιστογράμματα θα παρουσιαστούν για κάθε χαρακτηριστικό στη συνέχεια μετά τον καθαρισμό.
- Αφαιρέθηκαν ασθενείς που είχαν κενή τιμή στην κλάση Reason for discharge as Inpatient.
- Αφαιρέθηκαν ασθενείς που είχαν συμπληρωμένες τιμές για λιγότερο από το κατώφλι $th\%$ των συνολικών χαρακτηριστικών. Συγκεκριμένα, ανά πίνακα

ΚΕΦΑΛΑΙΟ 4. ΚΑΤΑΣΚΕΥΗ ΔΕΔΟΜΕΝΩΝ

δεδομένων:

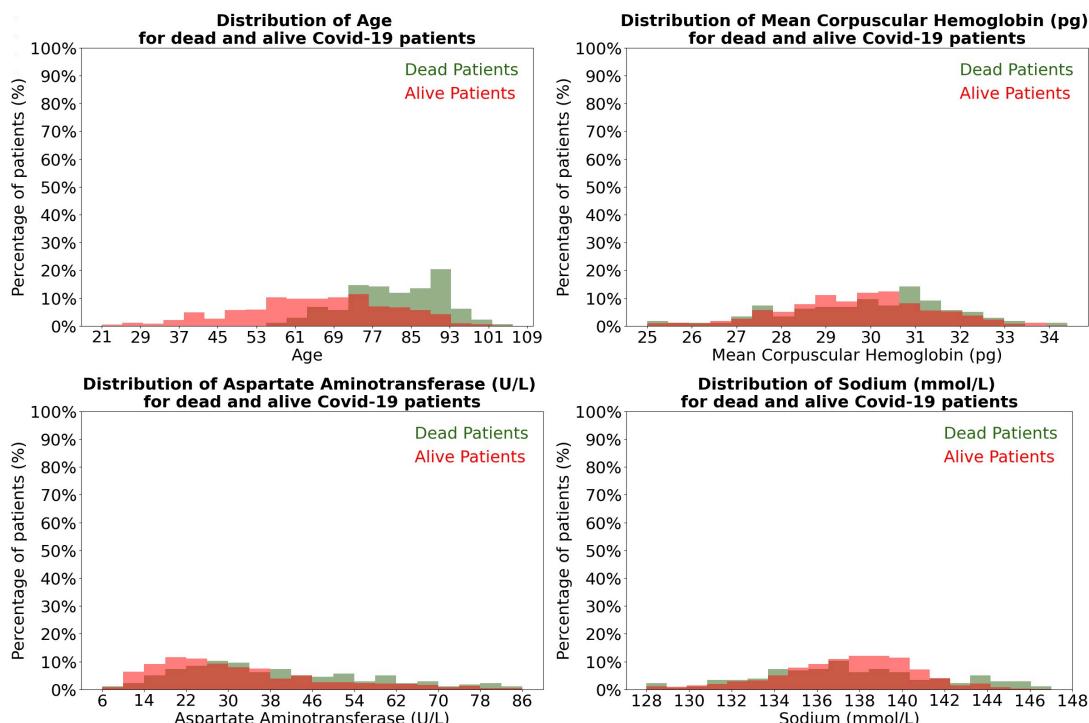
- **merged_data_no_td**: Για τον συγκεκριμένο πίνακα δοκιμάστηκαν οι τιμές 80%, 85% και 90% ελέγχοντας τον αριθμό των ασθενών που έμεναν, το λόγο των κλάσεων και την απόδοση μοντέλων που κατασκευάστηκαν και χρίθηκε ως καλύτερο κατώφλι η τιμή 85%.
- **merged_data2**: Για τον συγκεκριμένο πίνακα δοκιμάστηκαν οι τιμές 90% και 99%. Τα αποτελέσματα για αυτόν τον πίνακα θα παρουσιαστούν μετέπειτα.

Αφού οι πίνακες δεδομένων ”πέρασαν” τους παραπάνω ελέγχους, τα δεδομένα ήταν έτοιμα για εκπαίδευση.

4.2.2 Οπτικοποίηση των τελικών δεδομένων

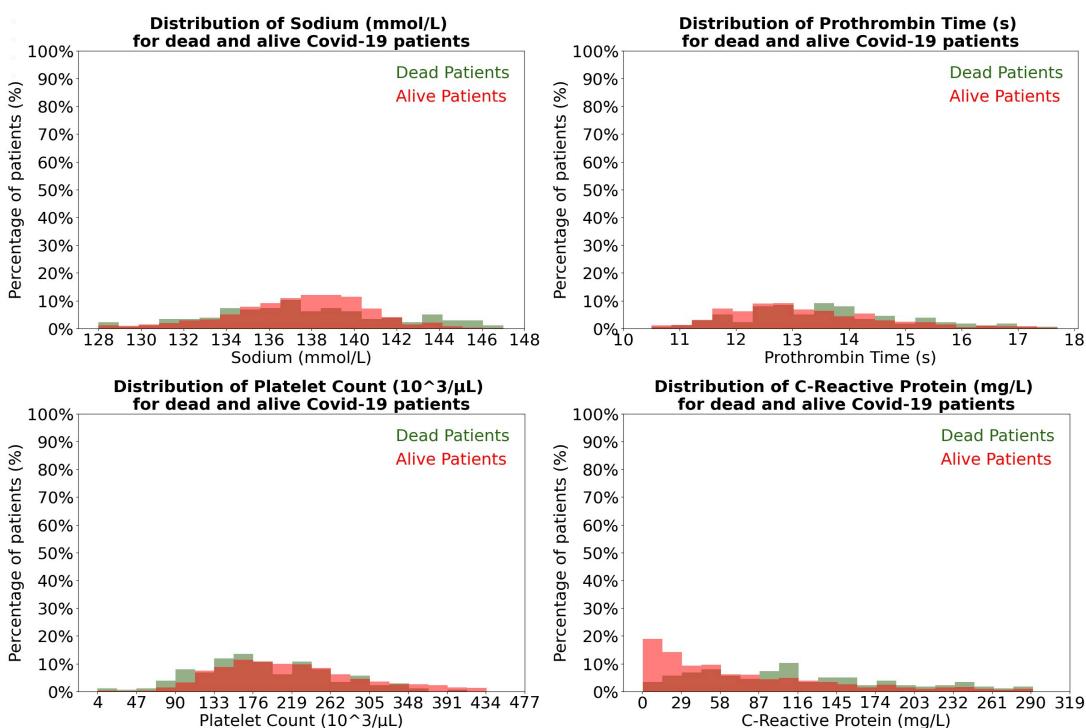
Σε αυτήν την υποενότητα παρουσιάζονται τα ιστογράμματα των χαρακτηριστικών ξεχωριστά για τους 2 πίνακες merged_data_no_td και merged_data2. Για κάθε χαρακτηριστικό υπάρχουν 2 ιστογράμματα: ένα για τους ζωντανούς ασθενείς κι ένα για τους νεκρούς με ετικέτες Alive και Dead αντίστοιχα. Επίσης, για κάθε πίνακα ορίζεται το κατώφλι th.

merged_data_no_td με th = 85%

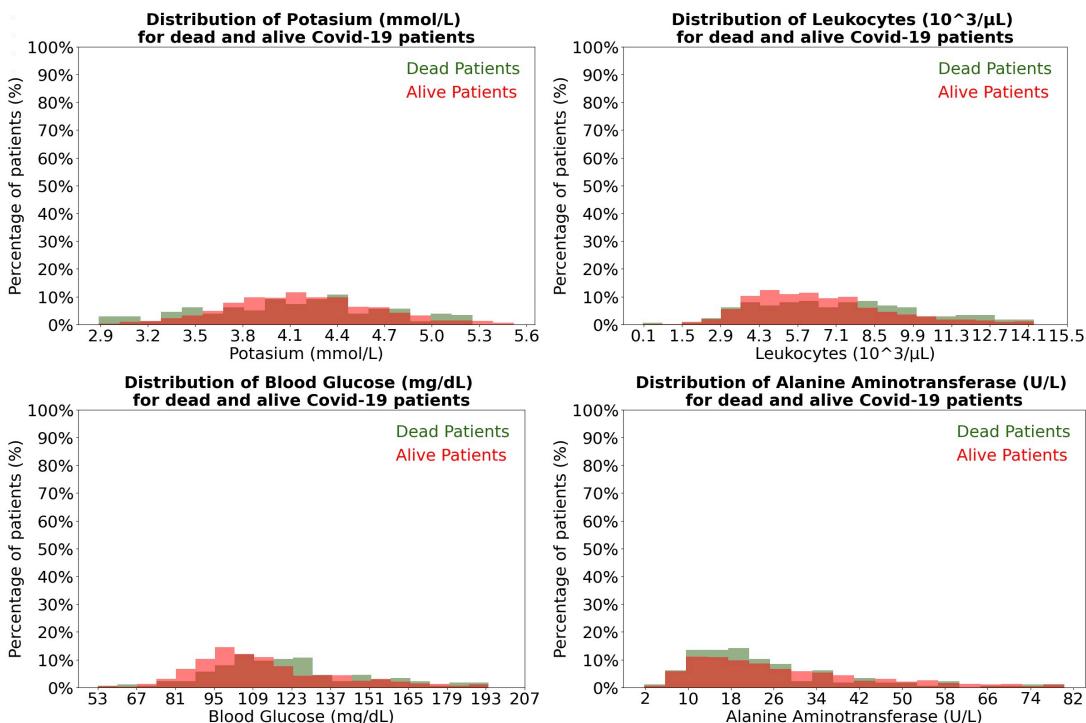


Σχήμα 4.2: Ιστογράμματα χαρακτηριστικών 0-3 ανά κλάση για merged_data_no_td

4.2. ΚΑΘΑΡΙΣΜΟΣ ΔΕΔΟΜΕΝΩΝ

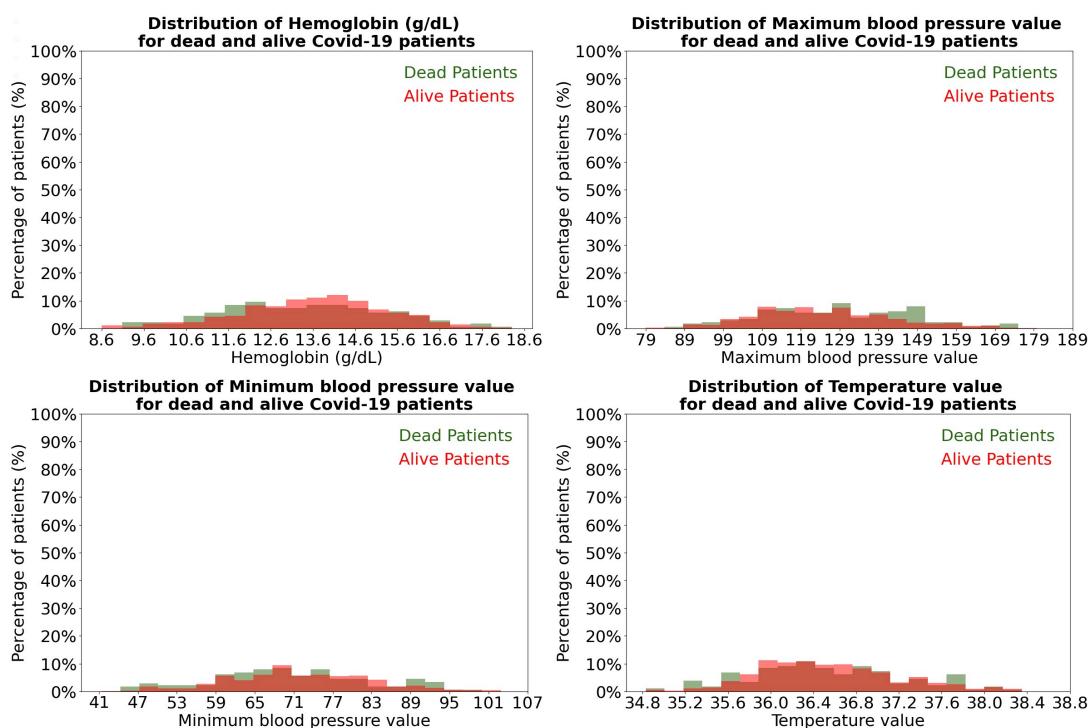


Σχήμα 4.3: Ιστόγραμμα χαρακτηριστικών 4-7 ανά κλάση για merged_data_no_td

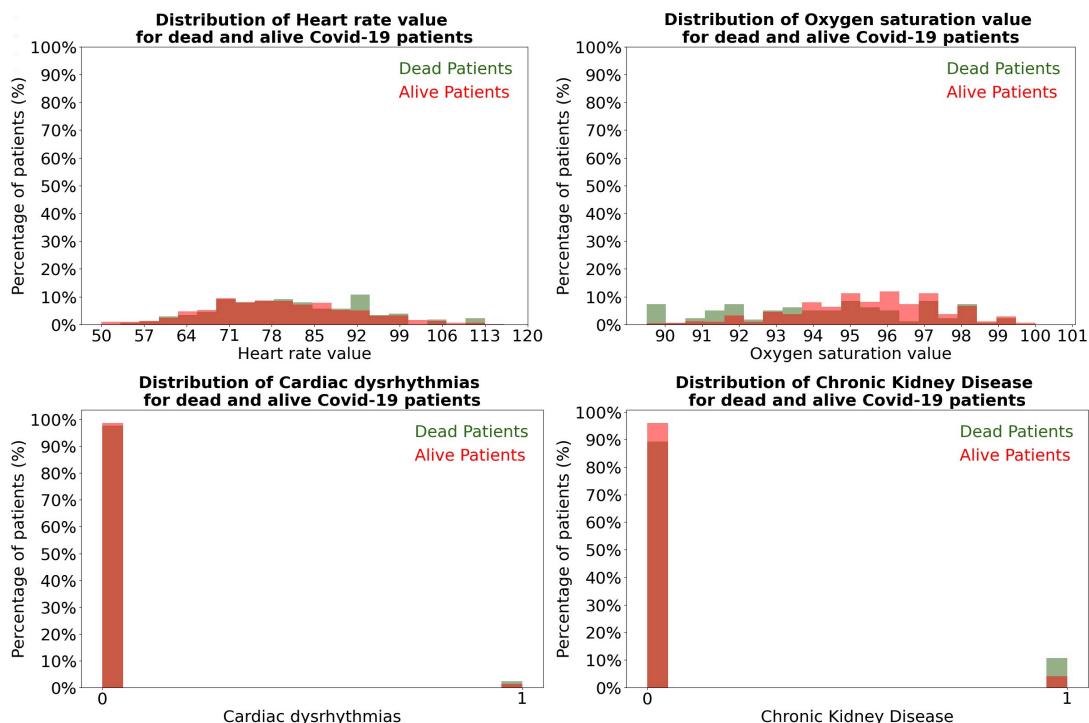


Σχήμα 4.4: Ιστόγραμμα χαρακτηριστικών 8-11 ανά κλάση για merged_data_no_td

ΚΕΦΑΛΑΙΟ 4. ΚΑΤΑΣΚΕΥΗ ΔΕΔΟΜΕΝΩΝ

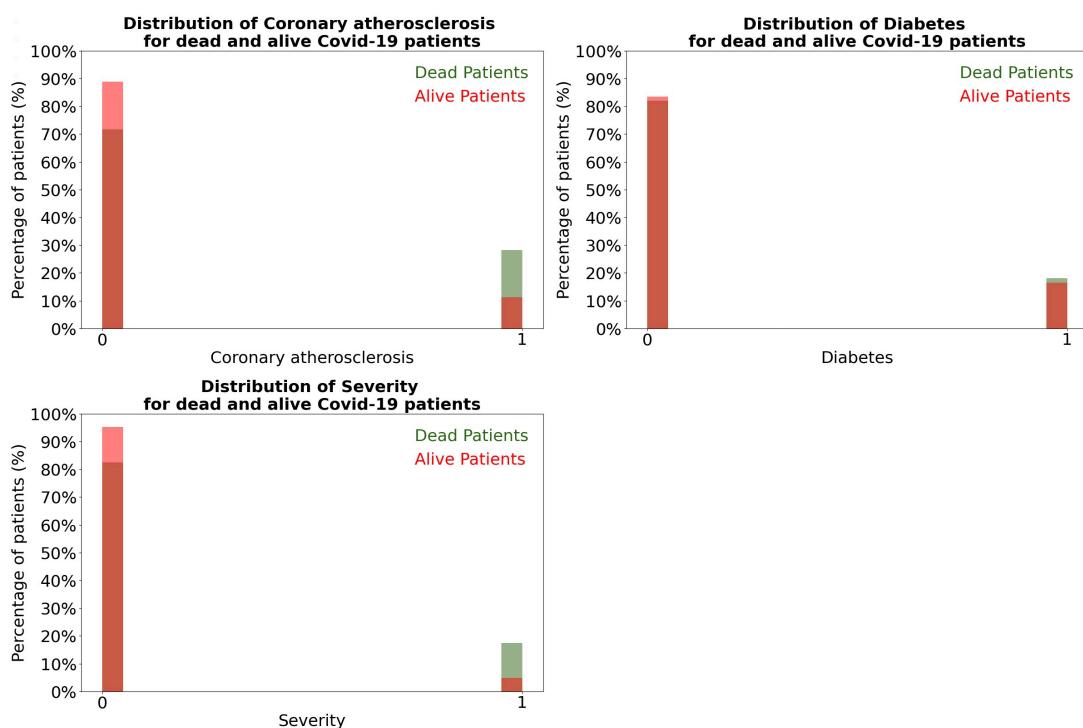


Σχήμα 4.5: Ιστόγραμμα χαρακτηριστικών 12-15 ανά κλάση για merged_data_no_td



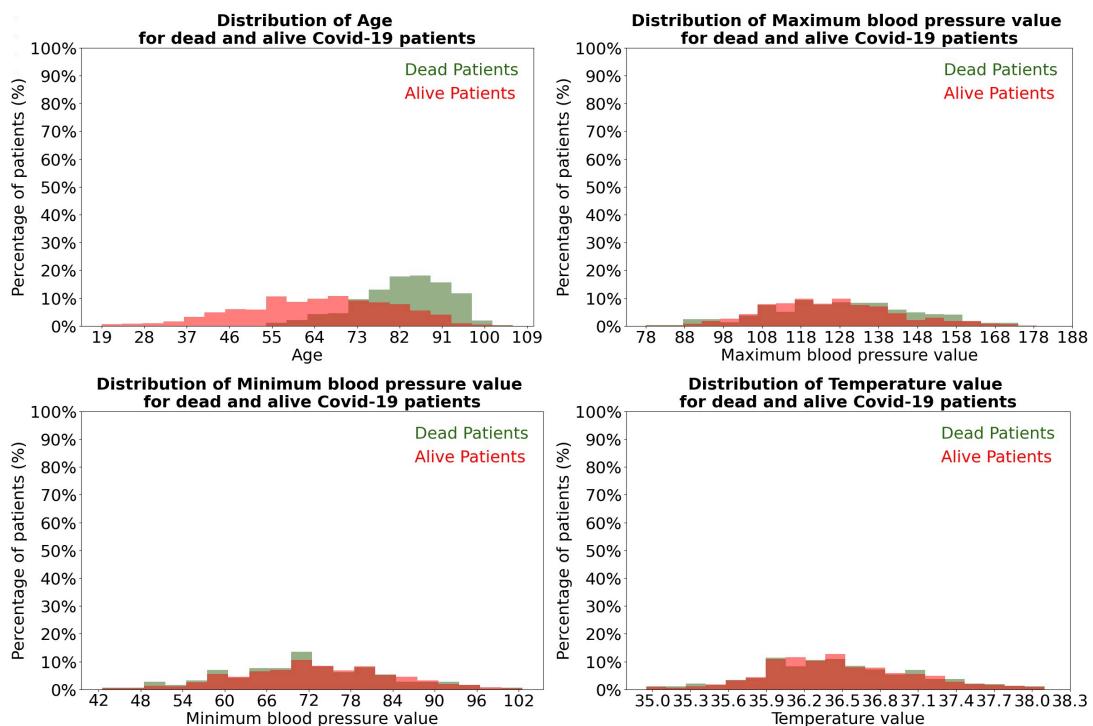
Σχήμα 4.6: Ιστόγραμμα χαρακτηριστικών 16-19 ανά κλάση για merged_data_no_td

4.2. ΚΑΘΑΡΙΣΜΟΣ ΔΕΔΟΜΕΝΩΝ



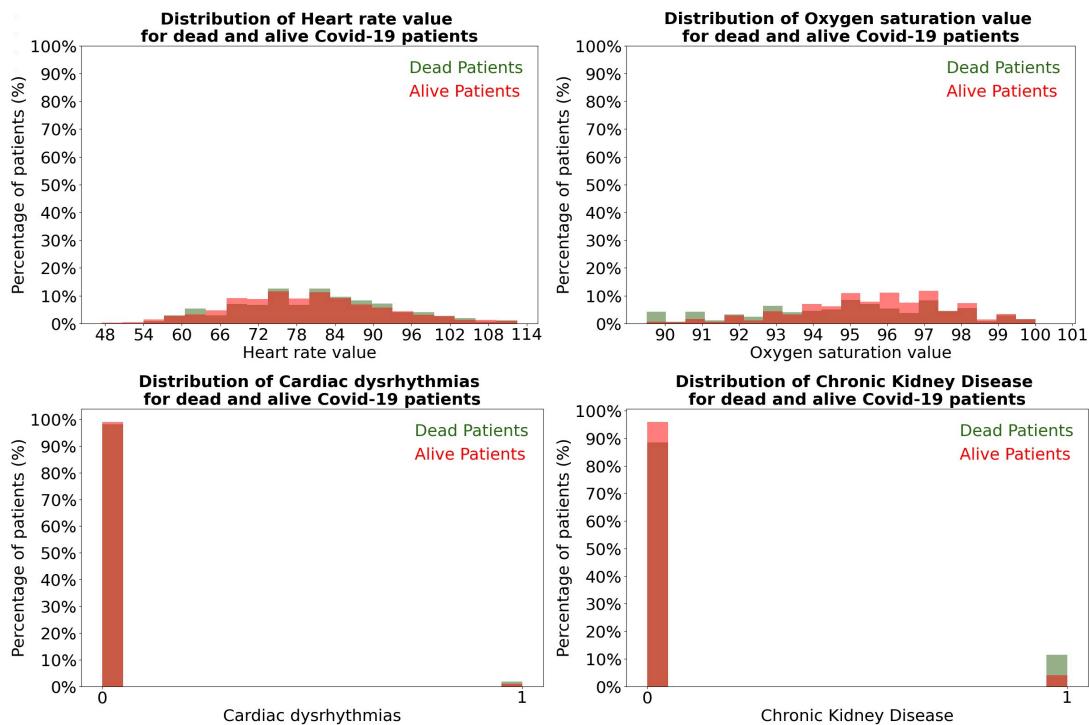
Σχήμα 4.7: Ιστογραμμα χαρακτηριστικών 20-22 ανά κλάση για merged_data_no_td

merged_data2 με th = 99%

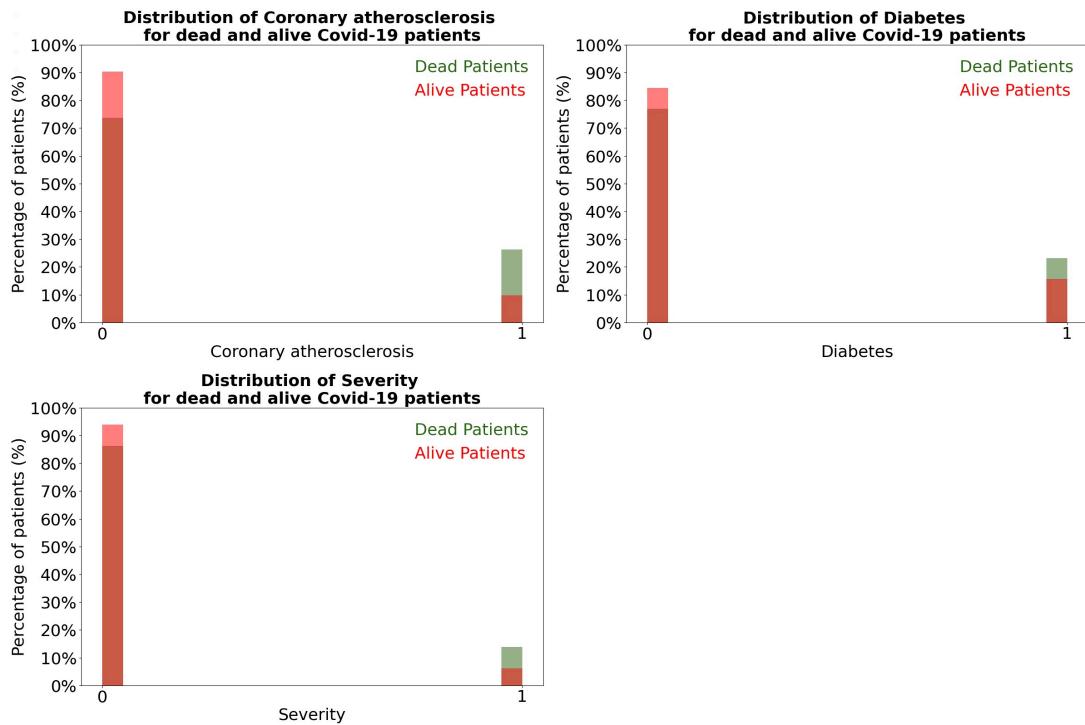


Σχήμα 4.8: Ιστογράμματα χαρακτηριστικών 0-3 ανά κλάση για merged_data2

ΚΕΦΑΛΑΙΟ 4. ΚΑΤΑΣΚΕΥΗ ΔΕΔΟΜΕΝΩΝ



Σχήμα 4.9: Ιστόγραμμα χαρακτηριστικών 4-7 ανά κλάση για merged_data2



Σχήμα 4.10: Ιστόγραμμα χαρακτηριστικών 8-11 ανά κλάση για merged_data2

5

Κατασκευή Μοντέλων κι Αποτελέσματα

5.1 ΕΚΠΑΙΔΕΥΣΗ ΜΟΝΤΕΛΩΝ

Όλα τα μοντέλα που κατασκευάστηκαν ήταν XGBoost μοντέλα ταξινομητών με objective function Binary Logistic Regression. Οι λόγοι που επιλέχθηκε το XGBoost ήταν:

- Τα μοντέλα του covidanalytics.io είναι XGBoost μοντέλα και θέλαμε να κάνουμε σύγκριση των αποτελεσμάτων.
- Η βιβλιοθήκη του XGBoost είναι πολύ εύχρηστη, δέχεται προγραμματισμό από το χρήστη κι επίσης παράγει πολύ καλά μοντέλα.

Ο λόγος που επιλέχθηκε ως objective function το Binary Logistic Regression ήταν το γεγονός ότι θέλαμε ο ταξινομητής μας να επιστρέψει ένα ρίσκο θνητότητας και συνεπώς μια πιθανότητα θνητότητας. Έτσι κρίθηκε ως πιο κατάλληλη συνάρτηση κόστους αυτή του Binary Logistic Regression, η οποία χρησιμοποιήθηκε και από το covidanalytics.io.

Επιπλέον, για την κατασκευή όλων των μοντέλων, εκτός αν δηλωθεί κάτι αλλο ρητά για συγκεκριμένα μοντέλα, χρησιμοποιήθηκε stratified learning με διαχωρισμό των δεδομένων σε 80% για το σετ εκπαίδευσης, 10% για το σετ επικύρωσης και 10% για το σετ ελέγχου. Ο διαχωρισμός των δεδομένων στα σετ αυτά ήταν ίδιος για όλους τους ελέγχους, εκτός αν δηλωθεί κάτι άλλο ρητά. Για κάθε μοντέλο XGBoost που κατασκευάζεται χρησιμοποιείται μια μετρική αξιολόγησης η οποία υπολογίζεται σε κάθε επανάληψη για την εκτίμηση της απόδοσης του μοντέλου πάνω στο

ΚΕΦΑΛΑΙΟ 5. ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΩΝ ΚΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

σετ επικύρωσης. Η εκπαίδευση σταματάει στην επανάληψη που η μετρική αξιολόγησης είχε την υψηλότερη τιμή στο σετ επικύρωσης. Αυτή η μετρική αξιολόγησης έχει στην ουσία το ρόλο της εποπτείας της απόδοσης του μοντέλου ανά επανάληψη και γι' αυτό το λόγο για την κατασκευή των καμπυλών μάθησης χρησιμοποιούμε τη μετρική αξιολόγησης που χρησιμοποιήθηκε τόσο στο σετ επικύρωσης όσο και στο σετ εκπαίδευσης. Οι μετρικές αξιολόγησης που χρησιμοποιήθηκαν για την εκπαίδευση αναλύονται διεξοδικά στην υποενότητα 3.2.2. Αναφέρονται ονομαστικά εδώ για υπενθύμιση και για λόγους συνέπειας:

Μετρικές Αξιολόγησης:

- **error**
- **logloss**
- **auc**
- **aucpr**
- **map**
- **F1-score**
- **Balanced Accuracy**
- **Precision**
- **Recall**
- **MCC**

Για την κατασκευή των μοντέλων τα δεδομένα χωρίστηκαν σε 4 κατηγορίες με σκοπό η καθεμία να εξεταστεί ξεχωριστά. Οι κατηγορίες αυτές ήταν:

- Ο πίνακας δεδομένων merged_data_no_td χωρίς το χαρακτηριστικό Severity
- Ο πίνακας δεδομένων merged_data_no_td με το χαρακτηριστικό Severity
- Ο πίνακας δεδομένων merged_data2 χωρίς το χαρακτηριστικό Severity
- Ο πίνακας δεδομένων merged_data2 με το χαρακτηριστικό Severity

Πίνακας 5.1: Επισκόπηση των κλάσεων των πινάκων δεδομένων

Πίνακας Δεδομένων	Κατώφλι th	Ασθενείς	Ζωντανοί	Νεκροί	Νεκροί Ζωντανοί
merged_data_no_td	85%	1348	1171	177	15.1%
merged_data2	99%	2401	2082	319	15.3%

Ο λόγος που εξετάσαμε ως διαφορετική περίπτωση το γεγονός να περιλαμβάνεται το χαρακτηριστικό Severity στα δεδομένα εκπαίδευσης ήταν επειδή το χαρακτηριστικό αυτό κανονικά δεν είναι γνωστό a priori, αλλά πρέπει να γίνει εκτίμησή του. Θέλαμε όμως να δούμε αν συνεισφέρει θεωρητικά έστω στην απόδοση των μοντέλων. Επομένως, κανονικά για να χρησιμοποιηθεί αυτό το χαρακτηριστικό θα πρέπει να προηγείται ένας δυαδικός ταξινομητής για το χαρακτηριστικό Severity.

5.2. ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΥΠΕΡΠΑΡΑΜΕΤΡΩΝ

Έγιναν κάποιες απόπειρες να δημιουργηθεί ένας τέτοιος ταξινομητής αλλά δεν απέδωσαν καρπούς λόγω έλλειψης δεδομένων και πιθανόν λόγω του ότι ένας δυαδικός ταξινομητής ίσως να μην είναι η κατάλληλη προσέγγιση για το συγκεκριμένο χαρακτηριστικό. Μια καλύτερη προσέγγιση θα ήταν να εκτιμηθούν οι μέρες παραμονής σε μονάδες εντατικής θεραπείας, οπότε το χαρακτηριστικό Severity από δυαδικές να λαμβάνει ακέραιες τιμές. Λόγω περιορισμένου χρόνου όμως κι έλλειψης δεδομένων για ασθενείς που μπήκαν σε μονάδες εντατικής θεραπείας δεν έγινε κάποια ουσιαστική προσπάθεια πάνω στην κατεύθυνση αυτή. Υπενθυμίζεται ότι το Severity είναι ένα χαρακτηριστικό το οποίο δηλώνει σοβαρή νόσηση, η οποία ορίζεται ως την είσοδο ενός ασθενή σε μονάδα εντατικής θεραπείας για τουλάχιστον 1 μέρα. Όταν όμως λαμβάνουμε δεδομένα για έναν ασθενή για να εκτιμήσουμε την πιθανότητα να πεθάνει, δεν γνωρίζουμε ακόμα αν θα μπει σε μονάδα εντατικής θεραπείας. Τα δεδομένα που λαμβάνουμε, λαμβάνονται σχετικά κοντά με την εισαγωγή του στο νοσοκομείο για νοσηλεία και δεν γίνεται έλεγχος αν τα δεδομένα αυτά λήφθηκαν μετά την εισαγωγή του σε μονάδες εντατικής θεραπείας. Υποτίθεται ο σκοπός των μοντέλων που φτιάχνουμε είναι να μας παρέχουν πληροφορία πρόβλεψης θνητότητας με δεδομένα που συλλέγονται μια φορά και μια μέρα για έναν ασθενή. Θα μπορούσαμε σαφώς να ορίσουμε μια νέα κατεύθυνση, όπου κατασκευάζουμε μοντέλα για άτομα που είναι αποκλειστικά σε μονάδες εντατικής θεραπείας και χρησιμοποιούμε μετρήσεις αφού μπουν στις μονάδες αυτές ή να χρησιμοποιούμε ολότελα συσσωρευτικές μετρήσεις διαφόρων ημερών αντί μετρήσεις μιας μέρας μόνο. Δεν ήταν αυτή όμως η κατεύθυνση της συγκεκριμένης διπλωματικής. Τέτοια ζητήματα θα συζητηθούν πιο εκτενώς στο κεφάλαιο των προεκτάσεων.

5.2 ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΥΠΕΡΠΑΡΑΜΕΤΡΩΝ

Αφού τα δεδομένα χωρίστηκαν σε σετ εκπαίδευσης, επικύρωσης κι ελέγχου όπως ορίστηκαν παραπάνω μπορούσε να ξεκινήσει η εκπαίδευση. Για την κατασκευή των μοντέλων που ακολουθούν θέσαμε boosting_rounds = 1000 και stopping_rounds = 20. Υπενθυμίζεται ότι η κλάση που ορίστηκε ήταν η στήλη Reason for discharge as Inpatient και οι υπόλοιπες στήλες αποτελούν τα χαρακτηριστικά. Στο αρχικό στάδιο της εκπαίδευσης δοκιμάστηκαν 6 βασικές προσεγγίσεις για την βελτιστοποίηση των υπερπαραμέτρων του μοντέλου για να δωθεί μια αρχική κατεύθυνση σχετικά με το ποιες υπερπαράμετροι έχουν τη μεγαλύτερη επίδραση. Οι ορισμοί των υπερπαραμέτρων αυτών δίνονται στην υποενότητα 3.2.1. Οι διαφορετικές προσεγγίσεις αυτές ορίζονται παρακάτω παρουσιάζοντας και τα αντίστοιχα αποτελέσματα ξεχωριστά για τους πίνακες δεδομένων merged_data_no_td με th = 85% και merged_data2 με th = 99% χωρίς το χαρακτηριστικό Severity.

1. **Προσέγγιση 1:** Βελτιστοποίηση των γενικών υπερπαραμέτρων χωρίς τη συμπλήρωση των κενών τιμών των χαρακτηριστικών.

Για την βελτιστοποίηση των γενικών υπερπαραμέτρων εκτελέστηκε η ακόλουθη τυχαία αναζήτηση πλέγματος για όλες τις προσεγγίσεις, επιλέγοντας ένα τυχαίο 10% των σημείων του παρακάτω πλέγματος όπως γράφτηκε στην

python:

```
random_grid_params = [
(max_depth, min_child_weight, eta, subsample, colsample_bytree, scale_pos_weight)
for max_depth in [6, 7, 8, 9, 10]
for min_child_weight in [1, 2, 3, 4, 5]
for eta in [0.05, 0.1]
for subsample in [0.6, 0.8, 1]
for colsample_bytree in [0.6, 0.8, 1]
for scale_pos_weight in [1, 1.5, 2, 2.5]
]
```

Για κάθε σημείο που επιλέχθηκε από το παραπάνω πλέγμα, κατασκευάστηκαν 10 μοντέλα, ένα μοντέλο για κάθε μια από τις μετρικές αξιολόγησης που ορίστηκαν, κατασκευάζοντας συνολικά 1800 μοντέλα.

- Προσέγγιση 2:** Συμπλήρωση των κενών τιμών των χαρακτηριστικών για κάθε σετ ξεχωριστά, χρησιμοποιώντας τη συνάρτηση Iterative Imputer από τη βιβλιοθήκη του scikitlearn με τις προκαθορισμένες ρυθμίσεις. Βελτιστοποίηση των γενικών υπερπαραμέτρων.

Για την βελτιστοποίηση των γενικών υπερπαραμέτρων εκτελέστηκε η ίδια διαδικασία όπως και στην προσέγγιση 1 κατασκευάζοντας συνολικά 1800 μοντέλα.

- Προσέγγιση 3:** Ύποδειγματοληψία της κλάσης πλειονότητας. Βελτιστοποίηση των γενικών υπερπαραμέτρων και της παραμέτρου class_ratio από τις παραμέτρους SMOTE, χωρίς τη συμπλήρωση των κενών τιμών των χαρακτηριστικών.

Για την βελτιστοποίηση των γενικών υπερπαραμέτρων εκτελέστηκε η ίδια διαδικασία όπως και στην προσέγγιση 1, ενώ για την παράμετρο class_ratio δώθηκαν οι τιμές [0.35, 0.4, 0.45, 0.5]. Για κάθε τιμή της παραμέτρου class_ratio κατασκευάστηκαν 1800 μοντέλα με βάση το πλέγμα των γενικών υπερπαραμέτρων. Συνολικά κατασκευάστηκαν 7200 μοντέλα για την προσέγγιση αυτή.

- Προσέγγιση 4:** Συμπλήρωση των κενών τιμών των χαρακτηριστικών για κάθε σετ ξεχωριστά, χρησιμοποιώντας τη συνάρτηση Iterative Imputer από τη βιβλιοθήκη του scikitlearn με τις προκαθορισμένες ρυθμίσεις. Ύποδειγματοληψία της κλάσης πλειονότητας. Βελτιστοποίηση των γενικών υπερπαραμέτρων και της παραμέτρου class_ratio από τις παραμέτρους SMOTE.

Για την βελτιστοποίηση των γενικών υπερπαραμέτρων εκτελέστηκε η ίδια διαδικασία όπως και στην προσέγγιση 3 κατασκευάζοντας συνολικά 7200 μοντέλα.

- Προσέγγιση 5:** Συμπλήρωση των κενών τιμών των χαρακτηριστικών για κάθε σετ ξεχωριστά, χρησιμοποιώντας τη συνάρτηση Iterative Imputer από τη βιβλιοθήκη του scikitlearn με τις προκαθορισμένες ρυθμίσεις. Κατασκευή συνθετικών δειγμάτων της κλάσης μειονότητας χρησιμοποιώντας την τεχνική SMOTE

5.2. ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΥΠΕΡΠΑΡΑΜΕΤΡΩΝ

που ορίστηκε στην υποενότητα 3.3. Βελτιστοποίηση των γενικών υπερπαραμέτρων και των παραμέτρων SMOTE εκτός από την παράμετρο class_ratio.

Για την βελτιστοποίηση των γενικών υπερπαραμέτρων εκτελέστηκε η ίδια διαδικασία όπως και στην προσέγγιση 1, ενώ για τις παραμέτρους SMOTE φτιάχτηκε το πλέγμα:

```
smote_grid = [  
    (smote_ratio, smote_neighbours)  
    for smote_ratio in [0.2, 0.25, 0.3]  
    for smote_neighbours in [5, 8]  
]
```

Για κάθε σημείο του πλέγματος smote_grid κατασκευάστηκαν 1800 μοντέλα, οπότε συνολικά κατασκευάστηκαν 9000 μοντέλα.

6. **Προσέγγιση 6:** Συμπλήρωση των κενών τιμών των χαρακτηριστικών για κάθε σετ ξεχωριστά, χρησιμοποιώντας τη συνάρτηση Iterative Imputer από τη βιβλιοθήκη του scikitlearn με τις προκαθορισμένες ρυθμίσεις. Κατασκευή συνθετικών δειγμάτων της κλάσης μειονότητας χρησιμοποιώντας την τεχνική SMOTE που ορίστηκε στην υποενότητα 3.3. Υποδειγματοληψία της κλάσης πλειονότητας. Βελτιστοποίηση των γενικών υπερπαραμέτρων και όλων των παραμέτρων SMOTE.

Για την βελτιστοποίηση των γενικών υπερπαραμέτρων εκτελέστηκε η ίδια διαδικασία όπως και στην προσέγγιση 1, ενώ για τις παραμέτρους SMOTE φτιάχτηκε το πλέγμα:

```
smote_grid = [  
    (smote_ratio, smote_neighbours, class_ratio)  
    for smote_ratio in [0.2, 0.25, 0.3]  
    for smote_neighbours in [5, 8]  
    for class_ratio in [0.35, 0.4, 0.45]  
]
```

Για κάθε σημείο του πλέγματος smote_grid κατασκευάστηκαν 1800 μοντέλα, οπότε συνολικά κατασκευάστηκαν 32400 μοντέλα.

Αθροίζοντας τα μοντέλα όλων των προσεγγίσεων, κατασκευάστηκαν συνολικά 59400 μοντέλα για κάθε πίνακα δεδομένων. Για κάθε μοντέλο που κατασκευάστηκε έγινε υπολογισμός των παρακάτω μετρικών απόδοσης, τόσο στο σετ επικύρωσης όσο και στο σετ ελέγχου:

Πίνακας 5.2: Μετρικές απόδοσης για αξιολόγηση μοντέλων

Precision	Average Precision	F1-score	MCC	Recall	AUC
-----------	-------------------	----------	-----	--------	-----

ΚΕΦΑΛΑΙΟ 5. ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΩΝ ΚΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

Για κάθε μοντέλο που κατασκευάστηκε αποθηκεύτηκαν σε ένα πίνακα τα ακόλουθα:

- Το ID του μοντέλου
- Η μετρική αξιολόγησης που χρησιμοποιήθηκε κατά την εκπαίδευση
- Η τιμή της κάθε μετρικής απόδοσης που αναφέρθηκε παραπάνω ξεχωριστά για το σετ επικύρωσης κι ελέγχου
- Ο αριθμός των επαναλήψεων που χρειάστηκαν για να εκπαιδευτεί το μοντέλο
- Οι τιμές των υπερπαραμέτρων που χρησιμοποιήθηκαν

Έτσι, ήμασταν σε θέση να συγκρίνουμε αυτές τις προσεγγίσεις μεταξύ τους ταυτόχρονα αφού αποθηκεύσαμε σε έναν πίνακα πληροφορίες για όλα τα μοντέλα κάθε προσέγγισης. Ο τρόπος που έγινε η αξιολόγηση ήταν ο εξής:

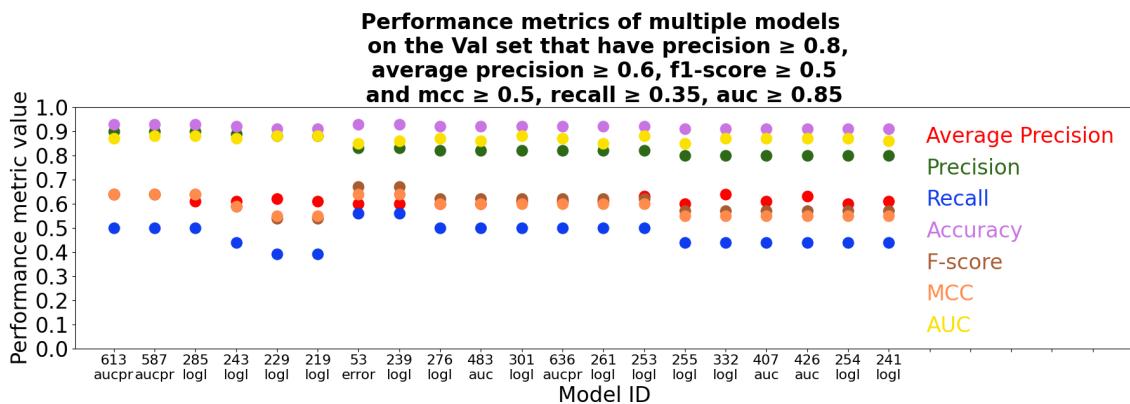
- Φιλτράρισμα των μοντέλων ορίζοντας ένα κατώφλι για κάθε μετρική απόδοσης για το σετ επικύρωσης. Τα κατώφλια που επιλέχθηκαν ήταν:

Πίνακας 5.3: Κατώφλια μετρικών απόδοσης για φιλτράρισμα μοντέλων για επιλογή καλύτερης προσέγγισης βελτιστοποίησης υπερπαραμέτρων

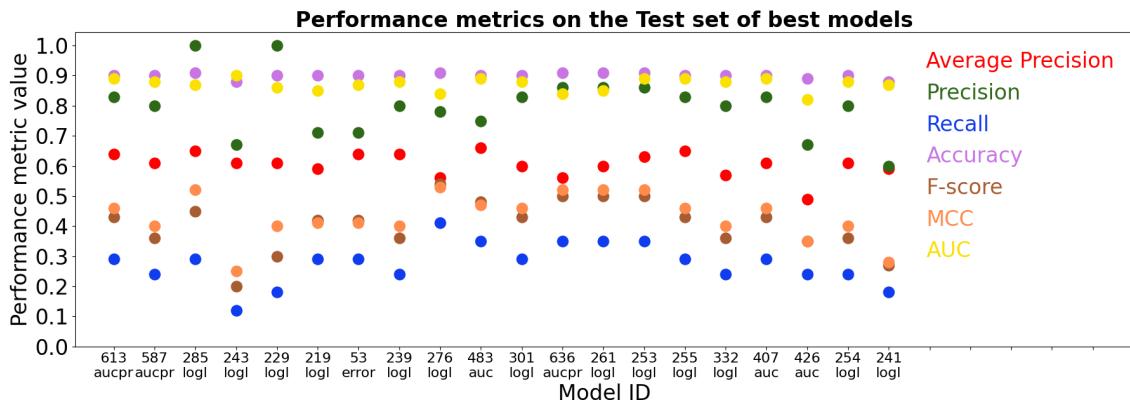
Data	Precision	Average Precision	F1-score	MCC	Recall	AUC
merged_data_no_td	0.8	0.6	0.5	0.5	0.35	0.85
merged_data2	0.45	0.4	0.5	0.4	0.5	0.7

- Επιλογή των 20 καλύτερων από τα εναπομείναντα μοντέλα για κάθε πίνακα δεδομένων με βάση την υψηλότερη τιμή της μετρικής Precision, η οποία είναι κι η πιο σχετική με το πρόβλημα. Δοκιμάστηκε κι η μετρική MCC που σχετίζεται με την ποιότητα του μοντέλου αλλά η απόδοση έπεφτε πολύ στο σετ ελέγχου σε σχέση με το σετ επικύρωσης.
- Δημιουργία διαγράμματος, όπου στον x άξονα υπάρχει το Model ID μαζί με την μετρική αξιολόγησης που χρησιμοποιήθηκε κατά την εκπαίδευση και στον y άξονα είναι οι τιμές των μετρικών απόδοσης. Για κάθε μετρική απόδοσης επιλέχθηκε διαφορετικό χρώμα ώστε να ξεχωρίζουν οπτικά. Το διάγραμμα αυτό φτιάχτηκε για το σετ επικύρωσης αλλά και για το σετ ελέγχου και παρουσιάζεται παρακάτω πρώτα για τον πίνακα merged_data_no_td κι έπειτα για τον πίνακα merged_data2:

5.2. ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΥΠΕΡΠΑΡΑΜΕΤΡΩΝ



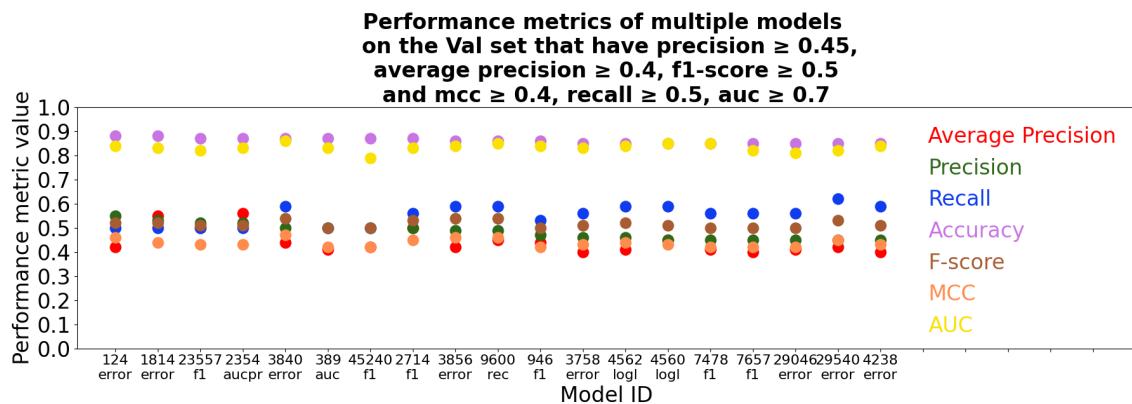
Σχήμα 5.1: Τα 20 καλύτερα μοντέλα με βάση το σετ επικύρωσης από όλες τις προσεγγίσεις με τις αποδόσεις τους στο σετ επικύρωσης για το merged_data_no_td



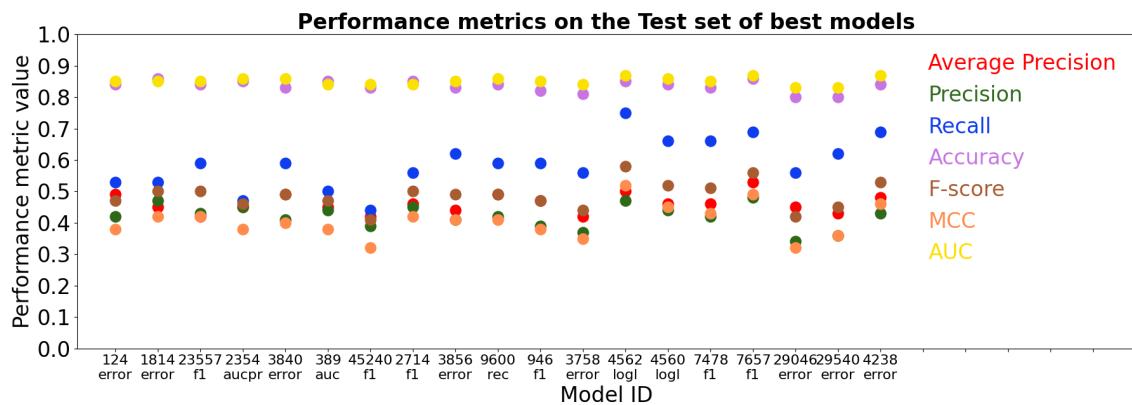
Σχήμα 5.2: Τα 20 καλύτερα μοντέλα με βάση το σετ επικύρωσης από όλες τις προσεγγίσεις με τις αποδόσεις τους στο σετ ελέγχου για το merged_data_no_td

Όπως βλέπουμε από τα παραπάνω 2 διαγράμματα όλα τα model IDs έχουν τιμή ≥ 1800 , επομένως προέκυψαν όλα από την 1η προσέγγιση για τον πίνακα δεδομένων merged_data_no_td. Συνεπώς, αυτή είναι που επιλέχθηκε για τη συνέχεια των πειραμάτων. Να σημειωθεί ότι για κατώφλι $th = 80\%$ προέκυψαν μοντέλα κι από τη 2η προσέγγιση, ενώ για κατώφλι $th = 90\%$ προέκυψαν μοντέλα μόνο από την 1η. Για κανένα κατώφλι δεν προέκυψαν στα καλύτερα μοντέλα κάποια από τις προσεγγίσεις 3, 4, 5 και 6. Αυτό πιθανόν οφείλεται στο γεγονός ότι δεν είχαμε πολλά δεδομένα. Η ανάλυση αυτή έγινε και για τα κατώφλια 80% και 90% αλλά δεν παρουσιάζεται γιατί οι αποδόσεις των μοντέλων ήταν χαμηλότερες σε σχέση με τη χρήση του κατωφλίου $th = 85\%$.

ΚΕΦΑΛΑΙΟ 5. ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΩΝ ΚΙ ΑΠΟΤΕΛΕΣΜΑΤΑ



Σχήμα 5.3: Τα 20 καλύτερα μοντέλα με βάση το σετ επικύρωσης από όλες τις προσεγγίσεις με τις αποδόσεις τους στο σετ επικύρωσης για το merged_data2



Σχήμα 5.4: Τα 20 καλύτερα μοντέλα με βάση το σετ επικύρωσης από όλες τις προσεγγίσεις με τις αποδόσεις τους στο σετ ελέγχου για το merged_data2

Όπως βλέπουμε από τα παραπάνω 2 διαγράμματα φαίνεται ότι προκύπτουν μοντέλα από όλες τις προσεγγίσεις αφού υπάρχουν Model IDs με τιμή $\geq 59400 - 32400 = 27000$. Ωστόσο, οι αποδόσεις των μοντέλων αυτών δεν είναι ιδιαίτερα καλές και δεδομένου ότι παρατηρούμε πως δεν υπάρχουν μεγάλες διαφορές στις αποδόσεις των μοντέλων μεταξύ των διαφορετικών προσεγγίσεων θα επιλέξουμε την απλούστερη προσέγγιση που είναι η προσέγγιση 1. Παρόλα αυτά, αξίζει να σημειωθεί ότι φαίνεται να υπάρχει αρκετά καλή γενίκευση από το σετ επικύρωσης προς το σετ ελέγχου. Η ανάλυση αυτή έγινε και για το κατώφλι $th = 90\%$ αλλά παραλήφθηκε λόγω του ότι λαμβάναμε χειρότερα αποτελέσματα.

Συνεπώς, και για τους 2 πίνακες δεδομένων merged_data_no_td και merged_data2 θα χρησιμοποιήσουμε την προσέγγιση 1 για τη βελτιστοποίηση των υπερπαραμέτρων γενικεύοντας και για τις περιπτώσεις που λαμβάνεται υπόψη και το χαρακτηριστικό Severity.

5.3 ΜΕΘΟΔΟΙ ΚΑΤΑΣΚΕΥΗΣ ΚΑΛΥΤΕΡΩΝ ΜΟΝΤΕΛΩΝ

Χρησιμοποιώντας την προσέγγιση 1 για την βελτιστοποίηση των υπερπαραμέτρων από την προηγούμενη υποενότητα θα κατασκευάσουμε με διάφορους τρόπους μοντέλα. Συνεπώς, θα βελτιστοποιήσουμε την επιλογή των γενικών υπερπαραμέτρων μόνο στα μοντέλα που θα φτιάξουμε και θα κρατήσουμε τα καλύτερα μοντέλα ανά μέθοδο κι έπειτα θα επιλέξουμε το συνολικά καλύτερο. Για την κατασκευή των μοντέλων που ακολουθούν θέσαμε `boosting_rounds = 1000` και `stopping_rounds = 50`. Επίσης, επειδή τα μοντέλα που φτιάχτηκαν με τον πίνακα δεδομένων `merged_data2` ήταν σημαντικά χειρότερα ως προς τη μετρική Precision που μας ενδιαφέρει περισσότερο, θα τα αγνοήσουμε. Η ανάλυσή μας λοιπόν θα επικεντρωθεί στον πίνακα δεδομένων `merged_data_no_td` με το χαρακτηριστικό Severity και χωρίς. Υπενθυμίζεται ότι η κλάση που ορίστηκε ήταν η στήλη Reason for discharge as Inpatient και οι υπόλοιπες στήλες αποτελούν τα χαρακτηριστικά. Παρακάτω θα παρουσιαστούν οι διαφορετικές μέθοδοι που χρησιμοποιήθηκαν για την κατασκευή των καλύτερων μοντέλων κι έπειτα θα παρουσιαστούν τα σχετικά αποτελέματα για τον πίνακα δεδομένων `merged_data_no_td` με και χωρίς το χαρακτηριστικό Severity ξεχωριστά.

Για κάθε μοντέλο που κατασκευάστηκε με τις παρακάτω μεθόδους, έγινε υπολογισμός των ίδιων μετρικών απόδοσης όπως ορίστηκαν στην υποενότητα 4.3.1, τόσο στο σετ επικύρωσης όσο και στο σετ ελέγχου. Έπειτα, για κάθε μοντέλο που κατασκευάστηκε αποθηκεύτηκαν πάλι σε έναν πίνακα όλες οι σχετικές πληροφορίες που μας ενδιαφέρουν όπως ορίστηκαν στην υποενότητα 4.3.1 ώστε να μπορέσουμε να συγκρίνουμε όλα τα μοντέλα μεταξύ τους ανά μέθοδο και να βρούμε τα καλύτερα. Παρουσιάζονται λοιπόν οι μέθοδοι:

- Μέθοδος 1:** Χρήση όλων των ασθενών από τον πίνακα `merged_data_no_td` κι εκτέλεση της ίδιας διαδικασίας με την προσέγγιση 1 όπως ορίστηκε στην υποενότητα 4.3.1 χρησιμοποιώντας αυτή τη φορά μεγαλύτερο πλέγμα.
- Μέθοδος 2:** Κρατώντας το σετ επικύρωσης (10% των συνολικών δεδομένων) και ελέγχου (10% των συνολικών δεδομένων) σταθερό, κάναμε το εξής:
 - Χωρισμός του σετ εκπαίδευσης (80% των συνολικών δεδομένων) σε θετική (Alive patients) κι αρνητική κλάση (Dead patients).
 - Δημιουργία υποσετ εκπαίδευσης τα οποία έχουν λόγο κλάσεων $\frac{\text{Positive Cases}}{\text{Negative Cases}} = \text{pos_neg_ratio}$. Κάθε υποσετ εκπαίδευσης περιέχει όλους τους ασθενείς της θετικής κλάσης (Dead patients) του αρχικού σετ εκπαίδευσης και ασθενείς της αρνητικής κλάσης (Alive patients) όσους ορίζονται από το λόγο `pos_neg_ratio`. Αν δηλαδή οι νεκροί ασθενείς σε ένα υποσετ εκπαίδευσης είναι N , τότε οι ζωντανοί θα είναι $\frac{N}{\text{pos_neg_ratio}}$. Οι ασθενείς της αρνητικής κλάσης λαμβάνονται με τυχαία δειγματοληψία χωρίς αντικατάσταση από το αρχικό σετ εκπαίδευσης, ώστε κάθε υποσετ εκπαίδευσης να έχει μοναδικούς ασθενείς για την αρνητική κλάση. Αν για το τελευταίο υποσετ εκπαίδευσης δεν υπάρχουν αρκετοί μοναδικοί ασθενείς στο αρχικό σετ εκπαίδευσης, τότε για τον σχηματισμό του λαμ-

βάνονται τυχαία χωρίς αντικατάσταση και μη μοναδικοί ασθενείς από το αρχικό σετ εκπαίδευσης μέχρι να συμπληρωθεί ο απαραίτητος αριθμός.

- Για κάθε υποσετ εκπαίδευσης που κατασκευάστηκε γίνεται εκπαίδευση ενός XGBoost υπό-μοντέλου χρησιμοποιώντας το ίδιο σετ επικύρωσης για όλα τα υποσετ. Να τονιστεί εδώ ότι πλέον το σετ εκπαίδευσης με το σετ επικύρωσης δεν έχουν τον ίδιο λόγο κλάσεων. Ο λόγος που επιλέχθηκε αυτή η τακτική ήταν για να μπορέσει να γίνει σύγκριση πάνω στην ίδια βάση μεταξύ των μεθόδων, καθώς η αξιολόγηση των μοντέλων γίνεται με βάση την απόδοση στο σετ επικύρωσης και στη "μεταφορά" της απόδοσης αυτής στο σετ ελέγχου. Συνεπώς, κρίθηκε λογικό τα σετ επικύρωσης κι ελέγχου να είναι ακριβώς τα ίδια για κάθε μέθοδο.
- Για κάθε υπό-μοντέλο που κατασκευάστηκε έγινε υπολογισμός των πιθανοτήτων ταξινόμησης για το σετ επικύρωσης και ελέγχου. Η τελική πιθανότητα ταξινόμησης προκύπτει από τον μέσο όρο των επιμέρους πιθανοτήτων των υπό-μοντέλων. Κατά αυτό τον τρόπο λοιπόν κατασκευάζουμε ένα τελικό μοντέλο το οποίο απαρτίζεται από ξεχωριστά υπό-μοντέλα εκπαιδευμένα πάνω σε υποσετ εκπαίδευσης με ορισμένο από εμάς λόγο κλάσεων λαμβάνοντας υπόψην όλα τα δεδομένα εκπαίδευσης, αποφεύγοντας έτσι κάποια μορφή υποδειγματοληψίας.

Να σημειωθεί ότι για την κατασκευή του κάθε μοντέλου έγινε βελτιστοποίηση των γενικών υπερπαραμέτρων με τυχαία αναζήτηση πλέγματος.

3. **Μέθοδος 3:** Χωρισμός του συνόλου δεδομένων σε 2 υποσύνολα, όπου το ένα υποσύνολο περιέχει ασθενείς που έχουν μια συγκεκριμένη ιδιότητα, ενώ το άλλο περιέχει ασθενείς που δεν την έχουν. Χωρίσαμε τα δεδομένα μας δηλαδή εισάγοντας μια συνθήκη. Οι ιδιότητες που δοκιμάστηκαν έπρεπε να είναι δυαδικές μεταβλητές κι επιλέχθηκαν οι ακόλουθες: Diabetes, Coronary atherosclerosis και Severity. Δεν επιλέχθηκαν τα χαρακτηριστικά Cardiac dysrhythmias και Chronic Kidney Disease γιατί δεν υπήρχαν σε πολλούς ασθενείς. Για τον πίνακα δεδομένων merged_data_no_td χωρίς το χαρακτηριστικό Severity έγινε διαχωρισμός μόνο με βάση το χαρακτηριστικό Diabetes. Ο λόγος που δεν έγινε διαχωρισμός και με το χαρακτηριστικό Coronary atherosclerosis ήταν επειδή από τους ελέγχους που κάναμε δεν ήταν πολύ αποτελεσματικός στον πίνακα merged_data_no_td με το χαρακτηριστικό Severity. Μετά τον διαχωρισμό του συνόλου δεδομένων σε 2 υποσύνολα, για κάθε υποσύνολο έγινε διαχωρισμός σε σετ εκπαίδευσης (80%), σετ επικύρωσης (10%) και σετ ελέγχου (10%) με stratified τρόπο ανά υποσύνολο. Έπειτα, για κάθε υποσύνολο έγινε εκπαίδευση ενός XGBoost μοντέλου βελτιστοποιώντας τις γενικές υπερπαραμέτρους με τυχαία αναζήτηση πλέγματος. Έτσι, η πρόβλεψη θνητότητας για έναν καινούργιο ασθενή γίνεται από διαφορετικό μοντέλο ανάλογα με το αν έχει ή οχι μια συγκεκριμένη ιδιότητα. Παρατίθενται παρακάτω οι κατανομές των κλάσεων ανάλογα με την ιδιότητα που επιλέχθηκε:

5.3. ΜΕΘΟΔΟΙ ΚΑΤΑΣΚΕΥΗΣ ΚΑΛΥΤΕΡΩΝ ΜΟΝΤΕΛΩΝ

Πίνακας 5.4: Επισκόπηση των αλάσεων του πίνακα δεδομένων merged_data_no_td ανάλογα με το υποσύνολο των ασθενών

Είδος συνόλου	Ασθενείς	Ζωντανοί	Νεκροί	$\frac{\text{Νεκροί}}{\text{Ζωντανοί}}$
Όλοι οι ασθενείς	1348	1171	177	15.1%
Ασθενείς με Diabetes = 1	225	193	32	16.6%
Ασθενείς με Diabetes = 0	1123	978	145	14.8%
Ασθενείς με Severity = 1	87	56	31	55.4%
Ασθενείς με Severity = 0	1261	1115	146	13.1%
Ασθενείς με Coronary atherosclerosis = 1	181	131	50	38.2%
Ασθενείς με Coronary atherosclerosis = 0	1167	1040	127	12.2%

Στην υποενότητα αυτή θα ορίσουμε επίσης τους τρόπους με τους οποίους έγινε η αξιολόγηση των μοντέλων ώστε να κρατήσουμε τα καλύτερα και πιο αξιόπιστα ανά μέθοδο. Συγκεκριμένα:

Μέθοδοι 1 και 3:

- Συλλογή όλων των μοντέλων που κατασκευάστηκαν και φιλτράρισμά τους ορίζοντας ένα κατώφλι για κάθε μετρική απόδοσης τόσο για το σετ επικύρωσης όσο και για το σετ ελέγχου. Έτσι κρατάμε μοντέλα που ξέρουμε ότι κάνουν καλή γενίκευση κι επίσης τα πάνε καλά στο σετ ελέγχου.
- Κρατάμε το σετ ελέγχου (10% των δεδομένων) στην άκρη. Ενώνουμε το σετ εκπαίδευσης (80% των δεδομένων) με το σετ επικύρωσης (10% των δεδομένων) κι όριζουμε ένα προσωρινό σετ (90% των δεδομένων). Χρησιμοποιώντας αυτό το προσωρινό σετ, δημιουργούμε 9 stratified ίσου μεγέθους υποσέτ (10% των δεδομένων το καθένα) κι εκτελούμε 9-fold cross validation για κάθε μοντέλο που έχουμε κρατήσει ξεχωριστά, χρησιμοποιώντας 1 υποσέτ για επικύρωση και τα άλλα 8 για εκπαίδευση κάθε φορά. Για κάθε fold υπολογίζουμε τις μετρικές απόδοσης του μοντέλου στο σετ επικύρωσής του και στο γενικό σετ ελέγχου. Αφού ολοκληρωθούν τα folds, τότε παίρνουμε την μέση απόδοση του μοντέλου ως τον μέσο όρο των μετρικών αποδόσεων στα σετ επικύρωσης και στο σετ ελέγχου.
- Κρατάμε τα καλύτερα 20 μοντέλα με βάση τη μέση απόδοσή τους στο σετ επικύρωσης από το 9-fold cross validation στη μετρική MCC. Δηλαδή, τα μοντέλα με τις 20 υψηλότερες τιμές στο μέσο MCC είναι αυτά που επιλέγονται. Για το κριτήριο επιλογής των 20 καλύτερων μοντέλων σε αυτό το σημείο δοκιμάστηκαν επίσης τα ακόλουθα:
 - Υψηλότερη μέση τιμή στη μετρική απόδοσης Precision
 - Χαμηλότερη τιμή στο SEM της μετρικής Precision
 - Χαμηλότερη τιμή στο SEM της μετρικής MCC
 - Υψηλότερη τιμή στο λόγο $\frac{\text{Μέση τιμή Precision}}{\text{SEM για Precision}}$
 - Υψηλότερη τιμή στο λόγο $\frac{\text{Μέση τιμή MCC}}{\text{SEM για MCC}}$

ΚΕΦΑΛΑΙΟ 5. ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΩΝ ΚΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

Η υψηλότερη μέση τιμή στη μετρική MCC όμως έδινε τα καλύτερα αποτελέσματα οπότε αυτό ήταν το κριτήριο που επιλέχθηκε.

4. Δημιουργία διαγράμματος, όπου στον x άξονα υπάρχει το Model ID μαζί με την μετρική αξιολόγησης που χρησιμοποιήθηκε κατά την εκπαίδευση και στον y άξονα είναι οι μέσες τιμές των μετρικών απόδοσης από το 9-fold cross validation μαζί με το διάστημα εμπιστοσύνης [mean - 1.96SEM, mean + 1.96SEM] ζωγραφισμένο με μια γραμμή πάνω και κάτω από τη μέση τιμή της κάθε μετρικής απόδοσης. Για κάθε μετρική απόδοσης επιλέχθηκε διαφορετικό χρώμα ώστε να ξεχωρίζουν οπτικά. Το διάγραμμα αυτό φτιάχτηκε για το σετ επικύρωσης αλλά και για το σετ ελέγχου.
5. Δημιουργία αντίστοιχου διαγράμματος με το παραπάνω, όπου παρουσιάζονται τα ίδια μοντέλα αλλά με τις αποδόσεις τους στα σετ δεδομένων όπως είχαν αρχικά χωριστεί κατά την δημιουργία τους. Χρήση δηλαδή των αποδόσεων των μοντέλων που χρησιμοποιήθηκαν στο βήμα (1).
6. Κοιτώντας τα παραπάνω 4 διαγράμματα επιλογή των 3 καλύτερων μοντέλων έπειτα από οπτική διερεύνηση με βάση τα ακόλουθα κριτήρια:
 - Τα μοντέλα παρουσιάζουν καλή γενίκευση από το σετ επικύρωσης προς το σετ ελέγχου και στα διαγράμματα του 9-fold cross validation και στα αντίστοιχα διαγράμματα των αποθηκευμένων μοντέλων.
 - Τα μοντέλα έχουν καλό Precision και σχετικά καλό MCC.
 - Τα μοντέλα δεν έχουν μεγάλη απόκλιση στις αποδόσεις τους στα διάγραμματα του 9-fold cross validation και στα αντίστοιχα διαγράμματα των αποθηκευμένων μοντέλων.
7. Για τα 3 καλύτερα μοντέλα που επιλέχθηκαν, παρουσίαση των διαγραμμάτων:
 - Confusion Matrix
 - ROC Curve
8. Επιλογή του καλύτερου μοντέλου με βάση τα παραπάνω 2 διαγράμματα και παρουσίαση των διαγραμμάτων:
 - Learning Curve
 - Confusion Matrix
 - ROC Curve
 - Feature Importance Bar Plot

Μέθοδος 2:

Για την συγκεκριμένη μέθοδο πρέπει πρώτα να διαπιστώσουμε κατά πόσο παράγει καλά αποτελέσματα κι αν όντως παράγει, τότε να δοκιμάσουμε τον ίδιο τρόπο αξιολόγησης με τις μεθόδους 1 και 3. Για αυτό το σκοπό έγιναν τα παρακάτω:

1. Κατασκευή μοντέλων με βελτιστοποίηση υπερπαραμέτρων όπως στην πρόσγειση 1 χρησιμοποιώντας τυχαία αναζήτηση πλέγματος δοκιμάζοντας διάφορες τιμές για το λόγο pos_neg_ratio.

5.3. ΜΕΘΟΔΟΙ ΚΑΤΑΣΚΕΥΗΣ ΚΑΛΥΤΕΡΩΝ ΜΟΝΤΕΛΩΝ

2. Οπτικοποίηση γενικής απόδοσης βλέποντας κάποια στατιστικά για τις τιμές των μετρικών απόδοσης των μοντέλων, όπως μέση τιμή, 1o τεταρτημόριο, κτλπ. Θα παρουσιαστούν αυτά πλήρως μαζί με τα αποτελέσματα στην επόμενη υποενότητα.
3. Επιλογή καλύτερης τιμής για το λόγο pos_neg_ratio.
4. Σχολιασμός γενικής απόδοσης σε σχέση με τις άλλες μεθόδους.

5.4 ΑΠΟΤΕΛΕΣΜΑΤΑ ΜΕΘΟΔΟΥ 1

Τα αποτελέσματα της μεθόδου αυτής θα παρουσιαστούν ξεχωριστά για τον πίνακα δεδομένων merged_data_no_td με και χωρίς το χαρακτηριστικό Severity.

5.4.1 merged_data_no_td χωρίς το χαρακτηριστικό Severity

Για την βελτιστοποίηση των γενικών υπερπαραμέτρων εκτελέστηκε τυχαία αναζήτηση πλέγματος, επιλέγοντας ένα τυχαίο 20% των σημείων του παρακάτω πλέγματος:

```
random_grid_params = [  
    (max_depth, min_child_weight, eta, subsample, colsample_bytree, scale_pos_weight)  
    for max_depth in [5, 6, 7, 8, 9, 10]  
    for min_child_weight in [1, 2, 3, 4, 5, 6]  
    for eta in [0.05, 0.1]  
    for subsample in [0.5, 0.6, 0.7, 0.8, 0.9, 1]  
    for colsample_bytree in [0.5, 0.6, 0.7, 0.8, 0.9, 1]  
    for scale_pos_weight in [1, 1.5, 2, 2.5]  
]
```

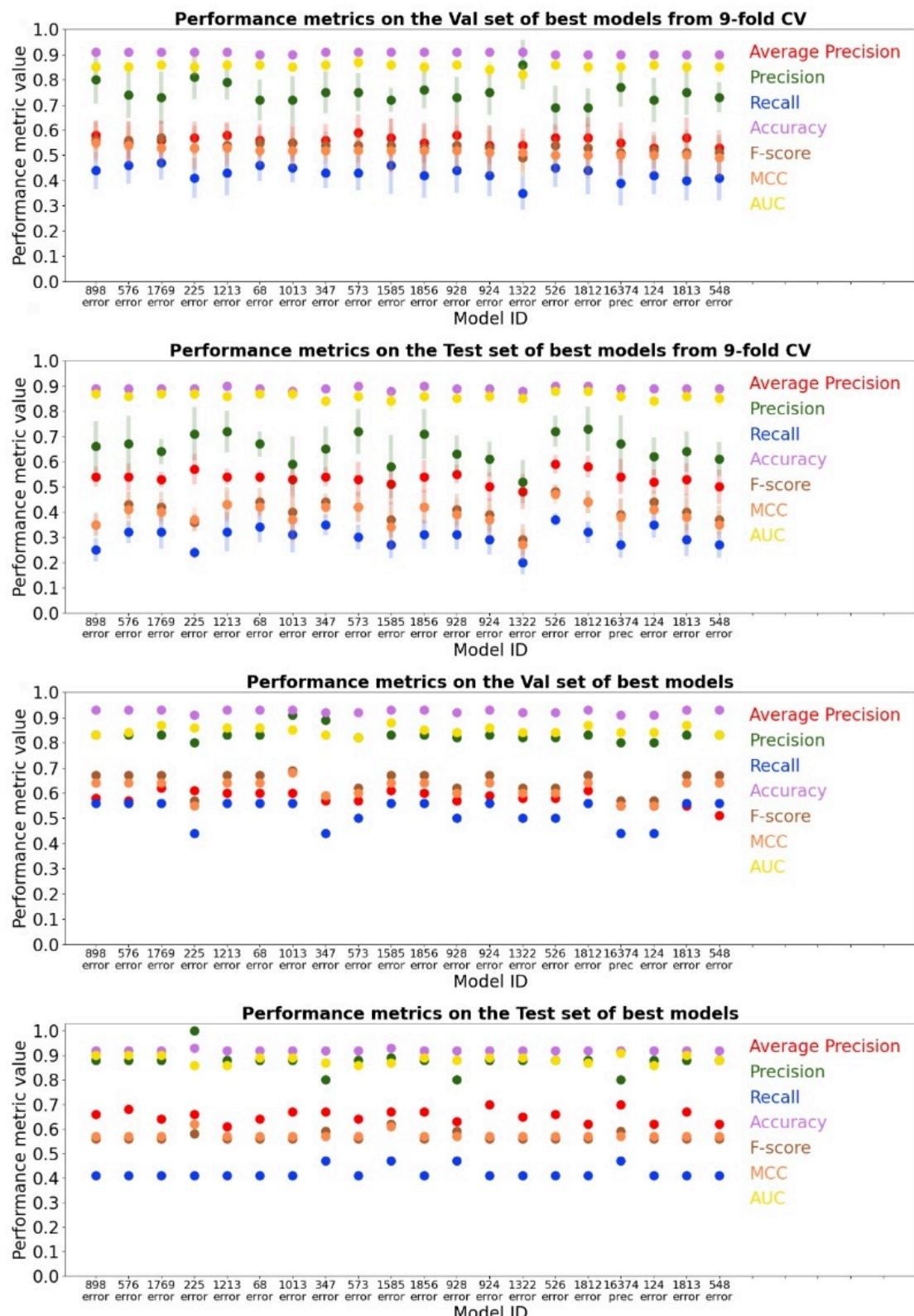
Για κάθε σημείο που επιλέχθηκε από το παραπάνω πλέγμα, κατασκευάστηκαν 10 μοντέλα, ένα μοντέλο για κάθε μια από τις μετρικές αξιολόγησης που ορίστηκαν, κατασκευάζοντας συνολικά 20740 μοντέλα.

Για το φιλτράρισμα των μοντέλων με βάση κατώφλια που εφαρμόστηκαν στο σετ επικύρωσης κι ελέγχου για τις μετρικές απόδοσης, επιλέχθηκαν οι παραπάνω τιμές:

Πίνακας 5.5: Κατώφλια μετρικών απόδοσης για φιλτράρισμα μοντέλων για μέθοδο 1 για merged_data_no_td χωρίς το χαρακτηριστικό Severity

Precision	Average Precision	F1-score	MCC	Recall	AUC
0.8	0.5	0.5	0.5	0.4	0.7

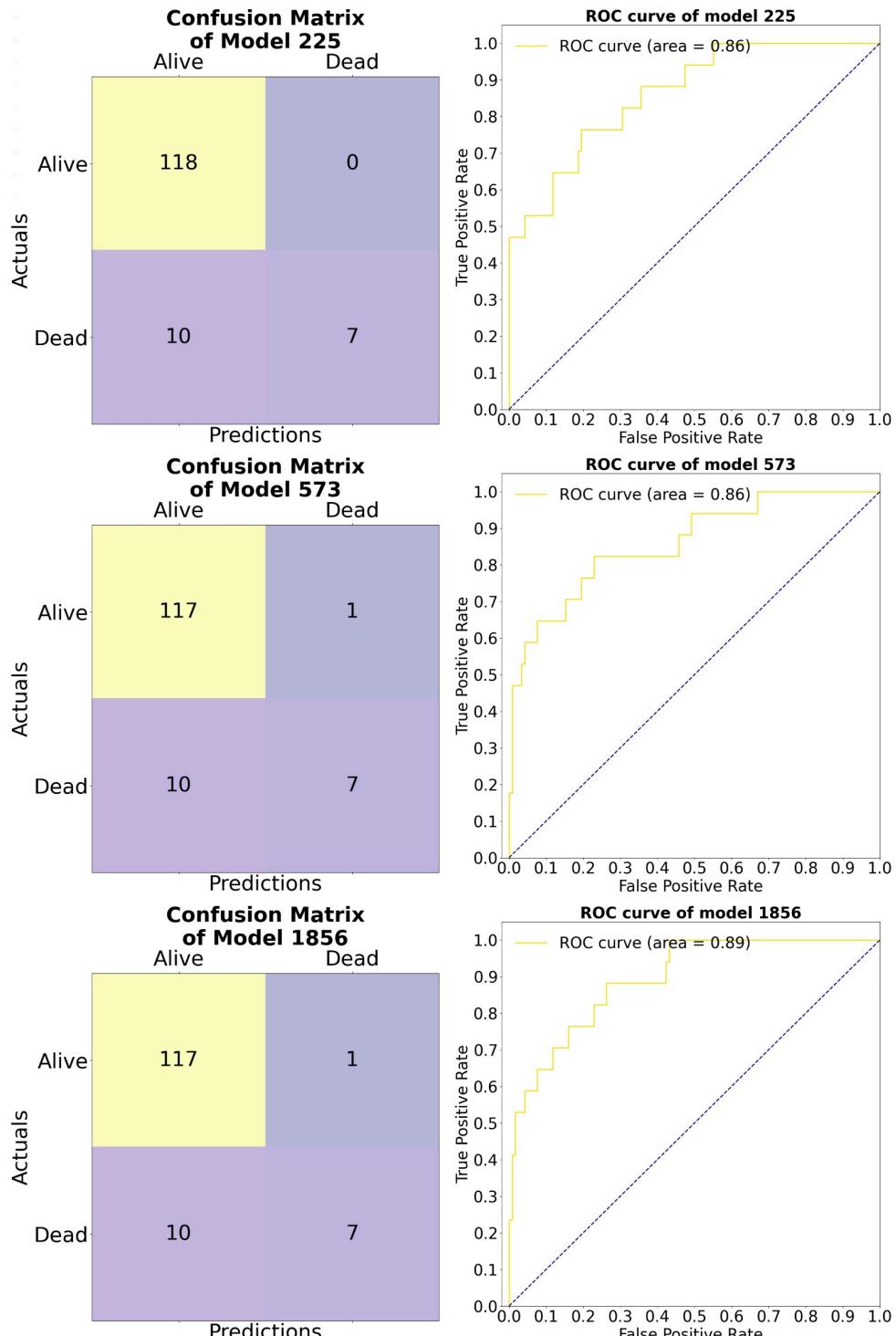
Από τα εναπομείναντα μοντέλα, τα 20 καλύτερα που έμειναν μετά την κατάταξή τους με βάση τη μέση τιμή της μετρικής MCC στο σετ επικύρωσης από το 9-fold cross validation παρουσιάζονται παρακάτω πρώτα για την απόδοσή τους στο 9-fold cross validation κι έπειτα για την απόδοσή τους στο αρχικό σετ δεδομένων.



Σχήμα 5.5: Τα 20 καλύτερα μοντέλα για το merged_data_no_td χωρίς το χαρακτηριστικό Severity

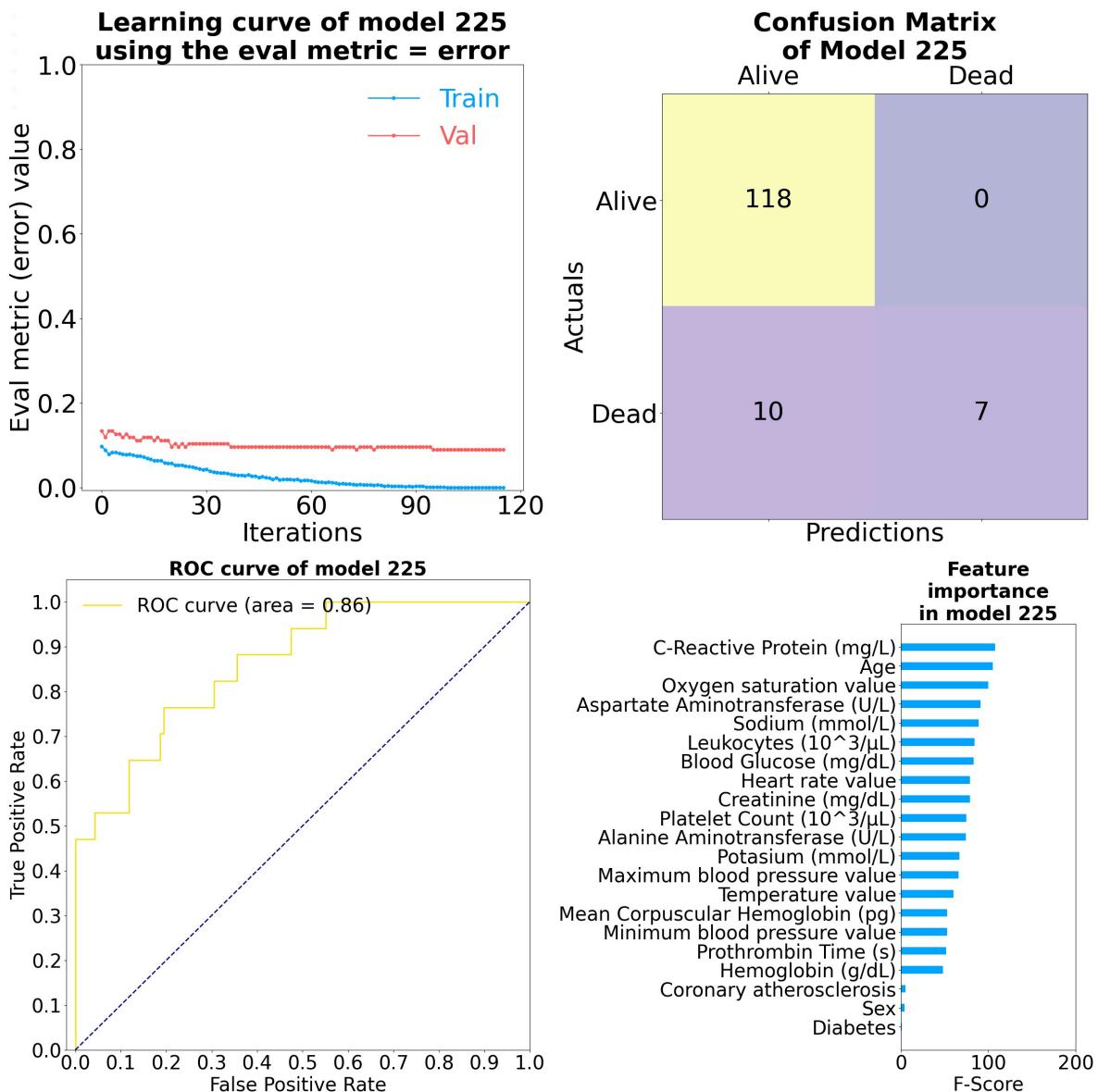
ΚΕΦΑΛΑΙΟ 5. ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΩΝ ΚΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

Κοιτώντας τα παραπάνω 4 διαγράμματα επιλέχτηκαν τα μοντέλα 225, 573 και 1856 ως τα καλύτερα και παρουσιάζονται παρακάτω:



Σχήμα 5.6: Confusion Matrix και καμπύλες ROC των 3 καλύτερων μοντέλων για το merged_data_no_td χωρίς το χαρακτηριστικό Severity

Κοιτώντας λοιπόν τα καλύτερα 3 μοντέλα, θεωρούμε ως καλύτερο, το μοντέλο 225 το οποίο έχει το καλύτερο Precision, το οποίο εν προκειμένω δεν κάνει ποτέ λάθος όταν πιθανολογήσει έναν ασθενή ως νεκρό. Το μοντέλο 225 παρουσιάζεται παρακάτω:



Σχήμα 5.7: Το καλύτερο μοντέλο για το merged_data_no_td χωρίς το χαρακτηριστικό Severity

5.4.2 merged_data_no_td με το χαρακτηριστικό Severity

Για την βελτιστοποίηση των γενικών υπερπαραμέτρων εκτελέστηκε τυχαία αναζήτηση πλέγματος, επιλέγοντας ένα τυχαίο 10% των σημείων του παρακάτω πλέγματος:

```
random_grid_params = [  
    (max_depth, min_child_weight, eta, subsample, colsample_bytree, scale_pos_weight)  
    for max_depth in [5, 6, 7, 8, 9, 10]  
    for min_child_weight in [1, 2, 3, 4, 5, 6]  
    for eta in [0.05]  
    for subsample in [0.5, 0.6, 0.7, 0.8, 0.9, 1]  
    for colsample_bytree in [0.5, 0.6, 0.7, 0.8, 0.9, 1]  
    for scale_pos_weight in [1, 2, 3, 4]  
]
```

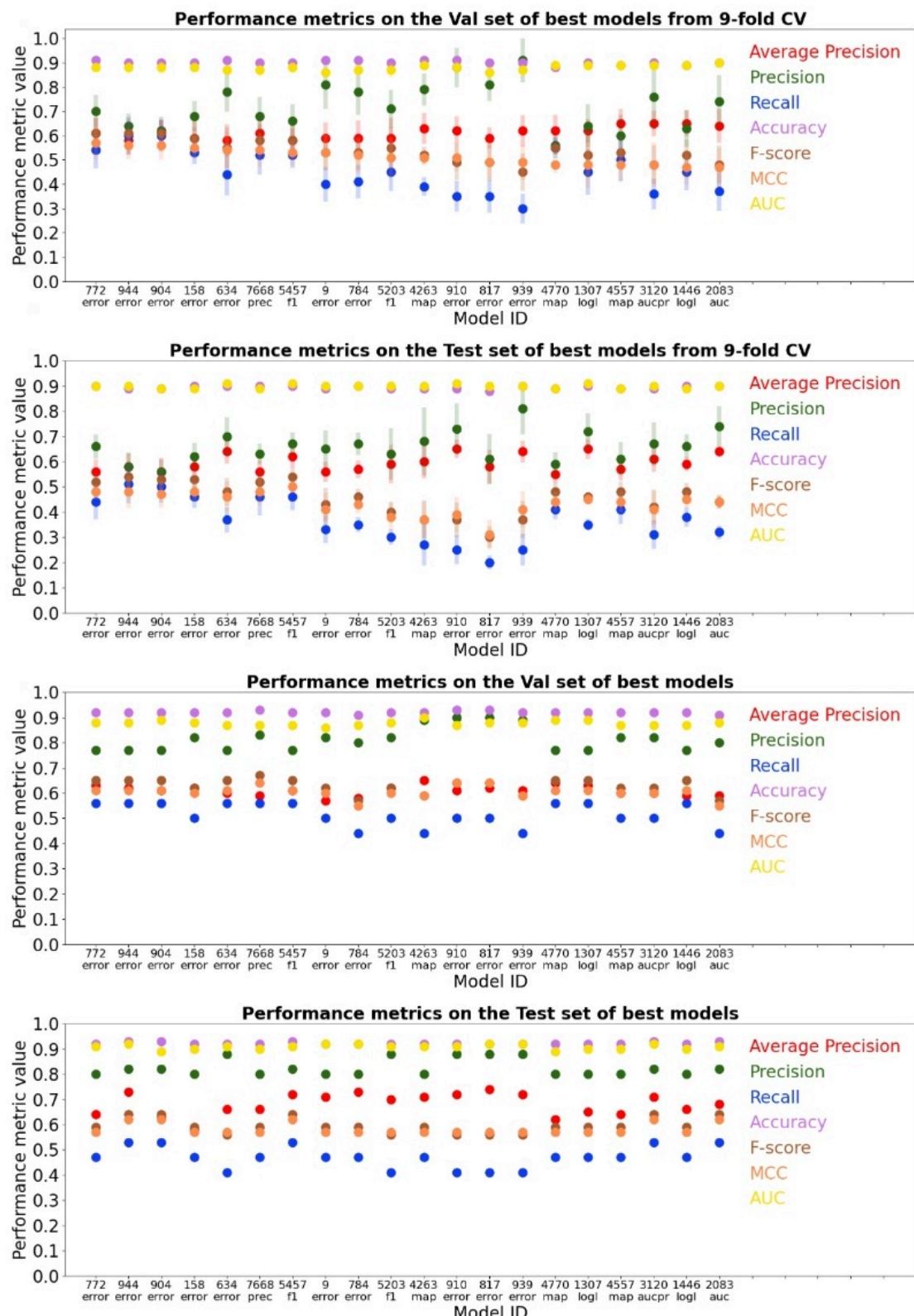
Για κάθε σημείο που επιλέχθηκε από το παραπάνω πλέγμα, κατασκευάστηκαν 10 μοντέλα, ένα μοντέλο για κάθε μια από τις μετρικές αξιολόγησης που ορίστηκαν, κατασκευάζοντας συνολικά 5184 μοντέλα.

Για το φιλτράρισμα των μοντέλων με βάση κατώφλια που εφαρμόστηκαν στο σετ επικύρωσης κι ελέγχου για τις μετρικές απόδοσης, επιλέχθηκαν οι παραπάνω τιμές:

Πίνακας 5.6: Κατώφλια μετρικών απόδοσης για φιλτράρισμα μοντέλων για μέθοδο 1 για merged_data_no_td με το χαρακτηριστικό Severity

Precision	Average Precision	F1-score	MCC	Recall	AUC
0.76	0.5	0.5	0.55	0.4	0.8

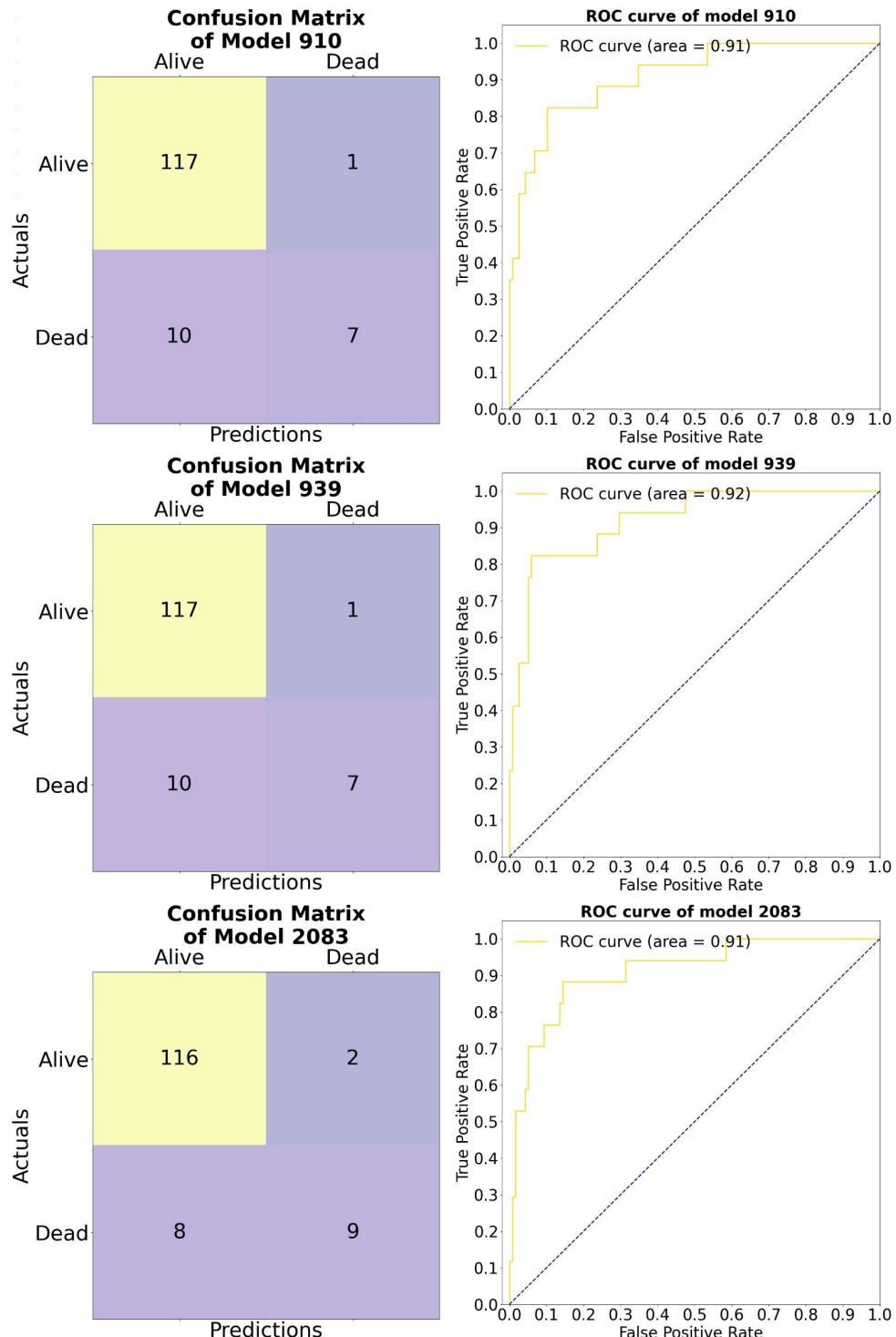
Από τα εναπομείναντα μοντέλα, τα 20 καλύτερα που έμειναν μετά την κατάταξή τους με βάση τη μέση τιμή της μετρικής MCC στο σετ επικύρωσης από το 9-fold cross validation παρουσιάζονται παρακάτω πρώτα για την απόδοσή τους στο 9-fold cross validation κι έπειτα για την απόδοσή τους στο αρχικό σετ δεδομένων.



Σχήμα 5.8: Τα 20 καλύτερα μοντέλα για το merged_data_no_td με το χαρακτηριστικό Severity

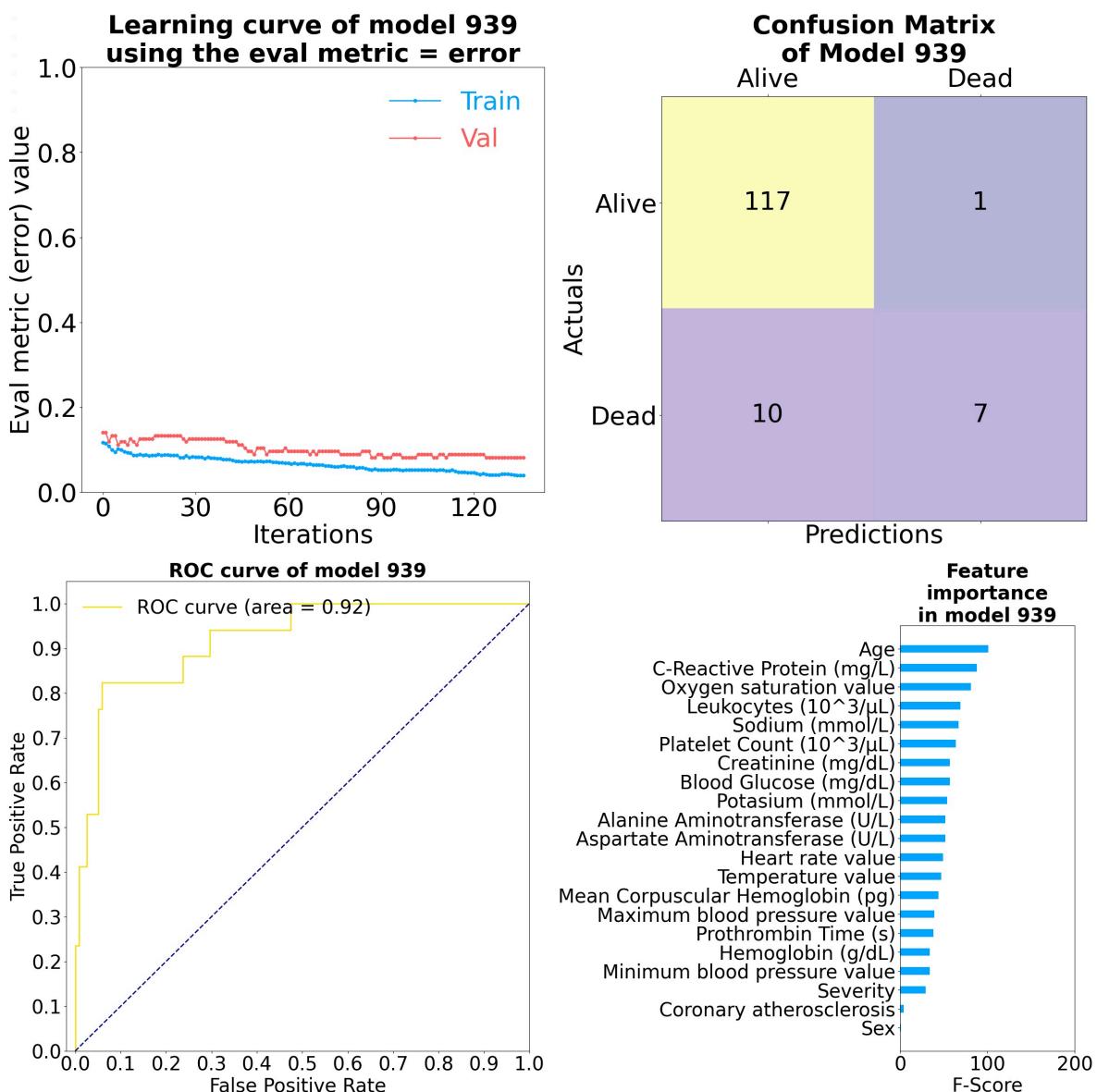
ΚΕΦΑΛΑΙΟ 5. ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΩΝ ΚΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

Κοιτώντας τα παραπάνω 4 διαγράμματα επιλέχτηκαν τα μοντέλα 910, 939 και 2083 ως τα καλύτερα και παρουσιάζονται παρακάτω:



Σχήμα 5.9: Confusion Matrix και καμπύλες ROC των 3 καλύτερων μοντέλων για το merged_data_no_td με το χαρακτηριστικό Severity

Κοιτώντας λοιπόν τα καλύτερα 3 μοντέλα, θεωρούμε ως καλύτερο το μοντέλο 939 το οποίο έχει το υψηλότερο Precision μαζί με το μοντέλο 910, αλλά έχει απειροελάχιστα υψηλότερο auc. Παρ'όλα αυτά, τα μοντέλα που κατασκευάστηκαν χωρίς το χαρακτηριστικό Severity είναι καλύτερα ως προς τη μετρική Precision. Τα συγκεκριμένα μοντέλα φαίνεται έχασαν από Precision και κέρδισαν σε auc. Το μοντέλο 939 παρουσιάζεται παρακάτω:



Σχήμα 5.10: Το καλύτερο μοντέλο για το merged_data_no_td με το χαρακτηριστικό Severity

5.5 ΑΠΟΤΕΛΕΣΜΑΤΑ ΜΕΘΟΔΟΥ 2

Για τη δοκιμή της συγκεκριμένης μεθόδου δώθηκαν διάφορες τιμές για τον λόγο pos_neg_ratio και θα παρουσιαστούν τα αποτελέσματα για την κάθε τιμή ξεχωριστά. Όσο μειώνεται ο λόγος αυτός, τόσο μειώνονται και τα συνολικά υπομοντέλα που κατασκευάζονται καθώς πλησιάζουμε προς τον πραγματικό λόγο τον κλάσεων. Η δοκιμή έγινε στον πίνακα δεδομένων merged_data_no_td με το χαρακτηριστικό Severity. Σημειώνεται ότι ο πραγματικός λόγος κλάσεων στα αρχικά δεδομένα είναι περίπου 0.152. Έτσι, για κάθε τιμή που δώθηκε στο λόγο pos_neg_ratio κατασκευάσαμε μοντέλα βελτιστοποιώντας τις γενικές υπερπαραμέτρους με τυχαία αναζήτηση πλέγματος, επιλέγοντας ένα τυχαίο 10% των σημείων του παρακάτω πλέγματος:

```
random_grid_params = [  
(max_depth, min_child_weight, eta, subsample, colsample_bytree, scale_pos_weight)  
for max_depth in [5, 6, 7, 8, 9, 10]  
for min_child_weight in [1, 2, 3, 4, 5, 6]  
for eta in [0.05]  
for subsample in [0.5, 0.6, 0.7, 0.8, 0.9, 1]  
for colsample_bytree in [0.5, 0.6, 0.7, 0.8, 0.9, 1]  
for scale_pos_weight in [1, 2, 3, 4]  
]
```

Για κάθε σημείο που επιλέχθηκε από το παραπάνω πλέγμα, κατασκευάστηκαν 10 μοντέλα, ένα μοντέλο για κάθε μια από τις μετρικές αξιολόγησης που ορίστηκαν, κατασκευάζοντας συνολικά 5184 μοντέλα.

Για κάθε λόγο pos_neg_ratio θα παρουσιαστούν τα στατιστικά των μετρικών απόδοσης από όλα τα μοντέλα ώστε να λάβουμε μια αρχική εκτίμηση για την απόδοση των μοντέλων ώστε να αποφανθούμε εν τέλει αν η συγκεκριμένη μέθοδος ευδοκιμεί στο συγκεκριμένο πρόβλημα.

5.5. ΑΠΟΤΕΛΕΣΜΑΤΑ ΜΕΘΟΔΟΥ 2

pos_neg_ratio = 1 - Κατασκευή 7 υπομοντέλων

	Val - Best Eval Metric Score	Val - Average Precision	Val - Precision	Val - Accuracy	Val - Recall	Val - F-score	Val - MCC	Val - AUC	Best Round
count	2364.00000	2364.00000	2364.00000	2364.00000	2364.00000	2364.00000	2364.00000	2364.00000	2364.00000
mean	0.60183	0.58168	0.26487	0.64761	0.86531	0.40206	0.33532	0.85193	62.45153
std	0.22267	0.04585	0.04015	0.08655	0.09215	0.04450	0.05921	0.02366	43.61107
min	0.22434	0.33000	0.14000	0.20000	0.61000	0.25000	0.10000	0.73000	1.00000
25%	0.39225	0.56000	0.24000	0.61000	0.78000	0.38000	0.30000	0.84000	23.96429
50%	0.58949	0.59000	0.27000	0.66000	0.89000	0.41000	0.34000	0.86000	61.71429
75%	0.83903	0.61000	0.29000	0.71000	0.94000	0.43000	0.38000	0.87000	90.71429
max	1.00000	0.68000	0.38000	0.79000	1.00000	0.52000	0.48000	0.90000	214.42857
	Test - Best Eval Metric Score	Test - Average Precision	Test - Precision	Test - Accuracy	Test - Recall	Test - F-score	Test - MCC	Test - AUC	Best Round
count	2364.00000	2364.00000	2364.00000	2364.00000	2364.00000	2364.00000	2364.00000	2364.00000	2364.00000
mean	0.60183	0.57815	0.32753	0.73333	0.91597	0.47698	0.43980	0.89991	62.45153
std	0.22267	0.05523	0.06845	0.08982	0.04461	0.07027	0.07368	0.01564	43.61107
min	0.22434	0.34654	0.13821	0.21481	0.70588	0.24286	0.11856	0.79212	1.00000
25%	0.39225	0.54820	0.28302	0.69630	0.88235	0.43243	0.39881	0.89482	23.96429
50%	0.58949	0.58980	0.32000	0.74074	0.94118	0.47619	0.44107	0.90379	61.71429
75%	0.83903	0.61904	0.37209	0.80000	0.94118	0.52459	0.48942	0.90977	90.71429
max	1.00000	0.71010	0.50000	0.87407	1.00000	0.64000	0.61536	0.92772	214.42857

Σχήμα 5.11: Στατιστικά μετρικών απόδοσης μοντέλων μεθόδου 2 για pos_neg_ratio = 1

pos_neg_ratio = 0.8 - Κατασκευή 6 υπομοντέλων

	Val - Best Eval Metric Score	Val - Average Precision	Val - Precision	Val - Accuracy	Val - Recall	Val - F-score	Val - MCC	Val - AUC	Best Round
count	77.00000	77.00000	77.00000	77.00000	77.00000	77.00000	77.00000	77.00000	77.00000
mean	0.58641	0.60623	0.30857	0.72364	0.80805	0.44234	0.37286	0.87052	69.54329
std	0.21456	0.03628	0.04129	0.06341	0.09092	0.04245	0.05127	0.01547	41.30915
min	0.20370	0.46000	0.19000	0.47000	0.61000	0.31000	0.21000	0.82000	2.83333
25%	0.43013	0.59000	0.28000	0.68000	0.72000	0.42000	0.35000	0.86000	44.00000
50%	0.58670	0.62000	0.30000	0.73000	0.83000	0.44000	0.38000	0.87000	67.00000
75%	0.78383	0.63000	0.34000	0.78000	0.89000	0.47000	0.40000	0.88000	100.83333
max	0.92593	0.65000	0.38000	0.80000	0.94000	0.53000	0.47000	0.89000	178.16667
	Test - Best Eval Metric Score	Test - Average Precision	Test - Precision	Test - Accuracy	Test - Recall	Test - F-score	Test - MCC	Test - AUC	Best Round
count	77.00000	77.00000	77.00000	77.00000	77.00000	77.00000	77.00000	77.00000	77.00000
mean	0.58641	0.60141	0.40564	0.80981	0.88312	0.54935	0.51164	0.90909	69.54329
std	0.21456	0.05282	0.07693	0.05907	0.05269	0.06554	0.06311	0.01196	41.30915
min	0.20370	0.43978	0.23188	0.60000	0.70588	0.37209	0.32655	0.87587	2.83333
25%	0.43013	0.58271	0.34043	0.76296	0.82353	0.50000	0.46436	0.90229	44.00000
50%	0.58670	0.61094	0.38095	0.80000	0.88235	0.54237	0.50727	0.91226	67.00000
75%	0.78383	0.63648	0.48276	0.86667	0.94118	0.60870	0.56257	0.91725	100.83333
max	0.92593	0.68187	0.53571	0.88889	0.94118	0.66667	0.63185	0.92772	178.16667

Σχήμα 5.12: Στατιστικά μετρικών απόδοσης μοντέλων μεθόδου 2 για pos_neg_ratio = 0.8

ΚΕΦΑΛΑΙΟ 5. ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΩΝ ΚΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

pos_neg_ratio = 0.5 - Κατασκευή 4 υπομοντέλων

	Val - Best Eval Metric Score	Val - Average Precision	Val - Precision	Val - Accuracy	Val - Recall	Val - F-score	Val - MCC	Val - AUC	Best Round
count	71.00000	71.00000	71.00000	71.00000	71.00000	71.00000	71.00000	71.00000	71.00000
mean	0.55205	0.61761	0.36761	0.79859	0.64873	0.46408	0.37718	0.87479	68.81338
std	0.21886	0.03690	0.05630	0.03681	0.06845	0.03404	0.03986	0.01602	57.27652
min	0.12222	0.49000	0.29000	0.73000	0.56000	0.40000	0.30000	0.83000	1.50000
25%	0.40449	0.60000	0.32000	0.77000	0.61000	0.44000	0.35000	0.87000	22.50000
50%	0.58472	0.62000	0.35000	0.79000	0.67000	0.45000	0.37000	0.88000	49.75000
75%	0.76629	0.64500	0.41000	0.83500	0.67000	0.49000	0.41000	0.89000	108.75000
max	0.88782	0.68000	0.50000	0.87000	0.83000	0.54000	0.46000	0.91000	214.00000
	Val - Best Eval Metric Score	Test - Average Precision	Test - Precision	Test - Accuracy	Test - Recall	Test - F-score	Test - MCC	Test - AUC	Best Round
count	71.00000	71.00000	71.00000	71.00000	71.00000	71.00000	71.00000	71.00000	71.00000
mean	0.55205	0.59555	0.50198	0.87126	0.73322	0.59013	0.53490	0.90760	68.81338
std	0.21886	0.04429	0.06003	0.02296	0.09158	0.04458	0.05123	0.01223	57.27652
min	0.12222	0.44209	0.35000	0.78519	0.47059	0.45714	0.37657	0.86939	1.50000
25%	0.40449	0.57512	0.45491	0.85926	0.70588	0.56168	0.50160	0.90105	22.50000
50%	0.58472	0.60273	0.50000	0.87407	0.76471	0.59459	0.53968	0.90977	49.75000
75%	0.76629	0.62602	0.54167	0.88889	0.82353	0.62857	0.57497	0.91600	108.75000
max	0.88782	0.66606	0.65000	0.91852	0.88235	0.70270	0.65876	0.92722	214.00000

Σχήμα 5.13: Στατιστικά μετρικά απόδοσης μοντέλων μεθόδου 2 για pos_neg_ratio = 0.5

pos_neg_ratio = 0.25 - Κατασκευή 2 υπομοντέλων

	Val - Best Eval Metric Score	Val - Average Precision	Val - Precision	Val - Accuracy	Val - Recall	Val - F-score	Val - MCC	Val - AUC	Best Round
count	63.00000	63.00000	63.00000	63.00000	63.00000	63.00000	63.00000	63.00000	63.00000
mean	0.52911	0.58190	0.46698	0.84841	0.60270	0.51810	0.44111	0.87127	89.46825
std	0.24583	0.06032	0.08286	0.03153	0.06842	0.04439	0.05397	0.02400	90.09430
min	0.09630	0.37000	0.31000	0.76000	0.50000	0.41000	0.31000	0.80000	1.00000
25%	0.29611	0.55000	0.41500	0.83500	0.56000	0.49000	0.41000	0.86000	15.75000
50%	0.58785	0.60000	0.48000	0.86000	0.56000	0.51000	0.43000	0.88000	57.50000
75%	0.69444	0.62000	0.50000	0.87000	0.61000	0.54500	0.47000	0.89000	141.00000
max	0.89530	0.65000	0.71000	0.91000	0.83000	0.63000	0.58000	0.90000	302.50000
	Val - Best Eval Metric Score	Test - Average Precision	Test - Precision	Test - Accuracy	Test - Recall	Test - F-score	Test - MCC	Test - AUC	Best Round
count	63.00000	63.00000	63.00000	63.00000	63.00000	63.00000	63.00000	63.00000	63.00000
mean	0.52911	0.59773	0.55819	0.88736	0.61625	0.57935	0.52010	0.90162	89.46825
std	0.24583	0.05445	0.06226	0.02014	0.09080	0.04454	0.05176	0.00989	90.09430
min	0.09630	0.46502	0.30556	0.77037	0.47059	0.41509	0.32650	0.87014	1.00000
25%	0.29611	0.55607	0.53452	0.88519	0.58824	0.55903	0.50521	0.89531	15.75000
50%	0.58785	0.59502	0.55556	0.88889	0.58824	0.58824	0.52891	0.90229	57.50000
75%	0.69444	0.64599	0.58824	0.89630	0.64706	0.61111	0.55263	0.90852	141.00000
max	0.89530	0.71564	0.69231	0.91111	0.88235	0.68293	0.64109	0.92074	302.50000

Σχήμα 5.14: Στατιστικά μετρικά απόδοσης μοντέλων μεθόδου 2 για pos_neg_ratio = 0.25

5.5. ΑΠΟΤΕΛΕΣΜΑΤΑ ΜΕΘΟΔΟΥ 2

pos_neg_ratio = 0.2 - Κατασκευή 2 υπομοντέλων

	Val - Best Eval Metric Score	Val - Average Precision	Val - Precision	Val - Accuracy	Val - Recall	Val - F-score	Val - MCC	Val - AUC	Best Round
count	57.00000	57.00000	57.00000	57.00000	57.00000	57.00000	57.00000	57.00000	57.00000
mean	0.50546	0.57947	0.57474	0.87544	0.55000	0.55105	0.48456	0.85772	99.08772
std	0.25654	0.05146	0.14024	0.03892	0.03449	0.06710	0.08691	0.02934	90.25515
min	0.08148	0.38000	0.29000	0.76000	0.44000	0.38000	0.27000	0.77000	1.00000
25%	0.28031	0.56000	0.50000	0.87000	0.56000	0.53000	0.45000	0.85000	24.50000
50%	0.58333	0.59000	0.56000	0.88000	0.56000	0.56000	0.49000	0.86000	77.50000
75%	0.63889	0.62000	0.67000	0.90000	0.56000	0.61000	0.55000	0.88000	180.00000
max	0.88628	0.66000	0.89000	0.92000	0.61000	0.65000	0.61000	0.90000	360.00000
	Val - Best Eval Metric Score	Test - Average Precision	Test - Precision	Test - Accuracy	Test - Recall	Test - F-score	Test - MCC	Test - AUC	Best Round
count	57.00000	57.00000	57.00000	57.00000	57.00000	57.00000	57.00000	57.00000	57.00000
mean	0.50546	0.61212	0.60024	0.89201	0.52116	0.54350	0.49341	0.90350	99.08772
std	0.25654	0.05457	0.08874	0.01546	0.11892	0.06562	0.06241	0.01670	90.25515
min	0.08148	0.41463	0.41379	0.83704	0.23529	0.34783	0.35151	0.83001	1.00000
25%	0.28031	0.59951	0.54167	0.88148	0.47059	0.50000	0.45676	0.90229	24.50000
50%	0.58333	0.62064	0.58824	0.89630	0.52941	0.55172	0.50521	0.90603	77.50000
75%	0.63889	0.63767	0.66667	0.90370	0.58824	0.5824	0.53306	0.91226	180.00000
max	0.88628	0.71532	0.81818	0.92593	0.76471	0.65000	0.62147	0.93021	360.00000

Σχήμα 5.15: Στατιστικά μετρικά απόδοσης μοντέλων μεθόδου 2 για pos_neg_ratio = 0.2

pos_neg_ratio = 0.15 - Κατασκευή 1 μοντέλου

	Val - Best Eval Metric Score	Val - Average Precision	Val - Precision	Val - Accuracy	Val - Recall	Val - F-score	Val - MCC	Val - AUC	Best Round
count	52.00000	52.00000	52.00000	52.00000	52.00000	52.00000	52.00000	52.00000	52.00000
mean	0.50907	0.58846	0.66365	0.88923	0.49462	0.54404	0.50250	0.86442	94.51923
std	0.29070	0.07840	0.15135	0.03463	0.11117	0.08541	0.08422	0.03654	87.35986
min	0.06667	0.32000	0.27000	0.73000	0.22000	0.33000	0.26000	0.75000	1.00000
25%	0.25880	0.57000	0.59000	0.89000	0.44000	0.52000	0.46750	0.84750	29.25000
50%	0.58543	0.61000	0.67000	0.90000	0.56000	0.55500	0.51000	0.87500	70.50000
75%	0.72922	0.63000	0.77250	0.90000	0.56000	0.59000	0.55000	0.89000	130.50000
max	1.00000	0.68000	1.00000	0.93000	0.67000	0.69000	0.68000	0.92000	352.00000
	Val - Best Eval Metric Score	Test - Average Precision	Test - Precision	Test - Accuracy	Test - Recall	Test - F-score	Test - MCC	Test - AUC	Best Round
count	52.00000	52.00000	52.00000	52.00000	52.00000	52.00000	52.00000	52.00000	52.00000
mean	0.50907	0.57863	0.60948	0.88561	0.36312	0.43387	0.40256	0.88878	94.51923
std	0.29070	0.10100	0.16343	0.02284	0.13470	0.11754	0.11444	0.02803	87.35986
min	0.06667	0.31809	0.29630	0.79259	0.11765	0.17391	0.13483	0.79536	1.00000
25%	0.25880	0.54387	0.51786	0.87963	0.29412	0.34483	0.30558	0.87861	29.25000
50%	0.58543	0.58932	0.58824	0.88889	0.38235	0.45658	0.43416	0.89432	70.50000
75%	0.72922	0.64870	0.69423	0.90370	0.47059	0.51932	0.49405	0.90753	130.50000
max	1.00000	0.77680	1.00000	0.91852	0.64706	0.62069	0.58755	0.92921	352.00000

Σχήμα 5.16: Στατιστικά μετρικά απόδοσης μοντέλων μεθόδου 2 για pos_neg_ratio = 0.15

ΚΕΦΑΛΑΙΟ 5. ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΩΝ ΚΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

Εξετάζοντας τις παραπάνω εικόνες από πάνω προς τα κάτω δίνοντας έμφαση στις μετρικές απόδοσης Precision, Recall, MCC και Accuracy παρατηρούμε τα εξής:

- Όσο αυξάνεται το pos_neg_ratio παίρνουμε πιο υψηλές τιμές για το Recall.
- Όσο αυξάνεται το pos_neg_ratio παίρνουμε πιο χαμηλές τιμές για το Precision.
- Όσο αυξάνεται το pos_neg_ratio παίρνουμε πιο χαμηλές τιμές για το Accuracy.
- Όταν κατασκευάζονται τουλάχιστον 4 μοντέλα παίρνουμε πιο χαμηλές τιμές αλλά παρόμοιες τιμές για το MCC στο σετ επικύρωσης. Αφού μειώσουμε τον αριθμό των μοντέλων κάτω από 4, τότε βελτιώνεται αρκετά το MCC στο σετ επικύρωσης. Το σετ ελέγχου δεν επηρεάζεται έντονα.
- Όταν χρησιμοποιούμε τουλάχιστον 2 υπομοντέλα παίρνουμε μέγιστο Precision στο σετ ελέγχου ίσο με περίπου 0.82 για pos_neg_ratio = 0.2.
- Το Precision με το Recall φαίνεται να λειτουργούν σαν ισοζύγιο. Αυξάνοντας το ένα μειώνεται το άλλο κι αντιστρόφως.

Με βάση τις παραπάνω παρατηρήσεις όσο πιο χαμηλό είναι το pos_neg_ratio τόσο πιο καλά είναι τα μοντέλα μας από άποψη Precision, Accuracy και MCC. Τα υψηλά pos_neg_ratio μας δίνουν καλύτερο Recall αλλά θυσιάζουμε τις υπόλοιπες μετρικές απόδοσης, εκ των οποίων η πιο σημαντική για το πρόβλημα που πάμε να λύσουμε, είναι το Precision. Με βάση αυτόν το συλλογισμό την καλύτερη απόδοση την έχουμε για pos_neg_ratio = 0.15. Για το συγκεκριμένο λόγο pos_neg_ratio όμως έχουμε 1 μοντέλο, το οποίο προσεγγιστικά έχει κατασκευαστεί από τα αρχικά μας δεδομένα με μια μικρή υποδειγματοληψία στους αρνητικούς ασθενείς, δηλαδή αντιπροσωπεύει σχεδόν ίδια μοντέλα με αυτά της μεθόδου 1. Υπενθυμίζεται ότι ο πραγματικός λόγος κλάσεων στα αρχικά δεδομένα είναι περίπου 0.152. Συνεπώς, χρησιμοποιώντας το pos_neg_ratio = 0.15 συγχρίνουμε έμμεσα την μέθοδο 1 με την μέθοδο 2 και φαίνεται πως η μέθοδος 1 υπερτερεί της μεθόδου 2. Έτσι, η μέθοδος 2 απορρίπτεται και δεν θα χρησιμοποιηθεί για περεταίρω ελέγχους

5.6 ΑΠΟΤΕΛΕΣΜΑΤΑ ΜΕΘΟΔΟΥ 3

Όπως αναφέρθηκε και στην υποενότητα 4.3.2 για τον πίνακα δεδομένων merged_data_no_td με το χαρακτηριστικό Severity χρησιμοποιήθηκαν ως ιδιότητες τα χαρακτηριστικά Diabetes, Coronary atherosclerosis και Severity. Για τον ίδιο πίνακα δεδομένων χωρίς το χαρακτηριστικό Severity χρησιμοποιήθηκε σαν ιδιότητα μόνο το χαρακτηριστικό Diabetes. Για λόγους εξοικονόμησης χώρου κι άσκοπης κούρασης του αναγνώστη θα παρουσιαστούν μόνο οι διαχωρισμοί των δεδομένων που έγιναν χρησιμοποιώντας το χαρακτηριστικό Diabetes και για τις 2 περιπτώσεις του merged_data_no_td με και χωρίς το χαρακτηριστικό Severity ξεχωριστά. Ο λόγος που δεν παρουσιάζονται οι διαχωρισμοί με βάση τα χαρακτηριστικά Coronary atherosclerosis και Severity είναι επειδή δεν δημιουργήθηκαν αποτελεσματικότερα μοντέλα.

5.6.1 merged_data_no_td χωρίς το χαρακτηριστικό Severity για Diabetes = 1

Για την βελτιστοποίηση των γενικών υπερπαραμέτρων εκτελέστηκε τυχαία αναζήτηση πλέγματος, επιλέγοντας ένα τυχαίο 10% των σημείων του παρακάτω πλέγματος:

```
random_grid_params = [
(max_depth, min_child_weight, eta, subsample, colsample_bytree, scale_pos_weight)
for max_depth in [5, 6, 7, 8, 9, 10]
for min_child_weight in [1, 2, 3, 4, 5, 6]
for eta in [0.05]
for subsample in [0.5, 0.6, 0.7, 0.8, 0.9, 1]
for colsample_bytree in [0.5, 0.6, 0.7, 0.8, 0.9, 1]
for scale_pos_weight in [1, 2, 3, 4]
]
```

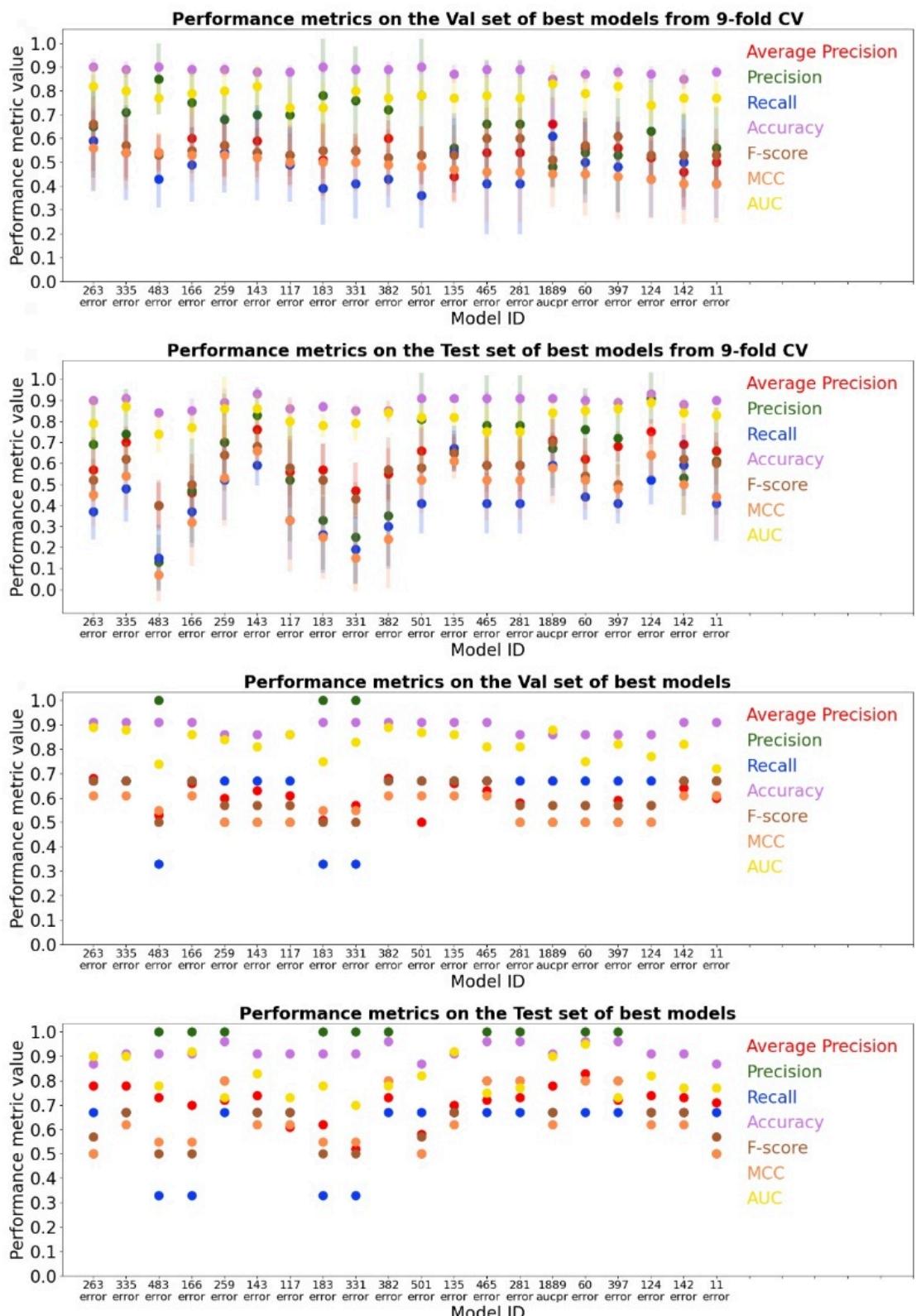
Για κάθε σημείο που επιλέχθηκε από το παραπάνω πλέγμα, κατασκευάστηκαν 10 μοντέλα, ένα μοντέλο για κάθε μια από τις μετρικές αξιολόγησης που ορίστηκαν, κατασκευάζοντας συνολικά 5184 μοντέλα.

Για το φιλτράρισμα των μοντέλων με βάση κατώφλια που εφαρμόστηκαν στο σετ επικύρωσης κι ελέγχου για τις μετρικές απόδοσης, επιλέχθηκαν οι παραπάνω τιμές:

Πίνακας 5.7: Κατώφλια μετρικών απόδοσης για φιλτράρισμα μοντέλων για μέθοδο 3 για merged_data_no_td χωρίς το χαρακτηριστικό Severity για ασθενείς με Διαβήτη

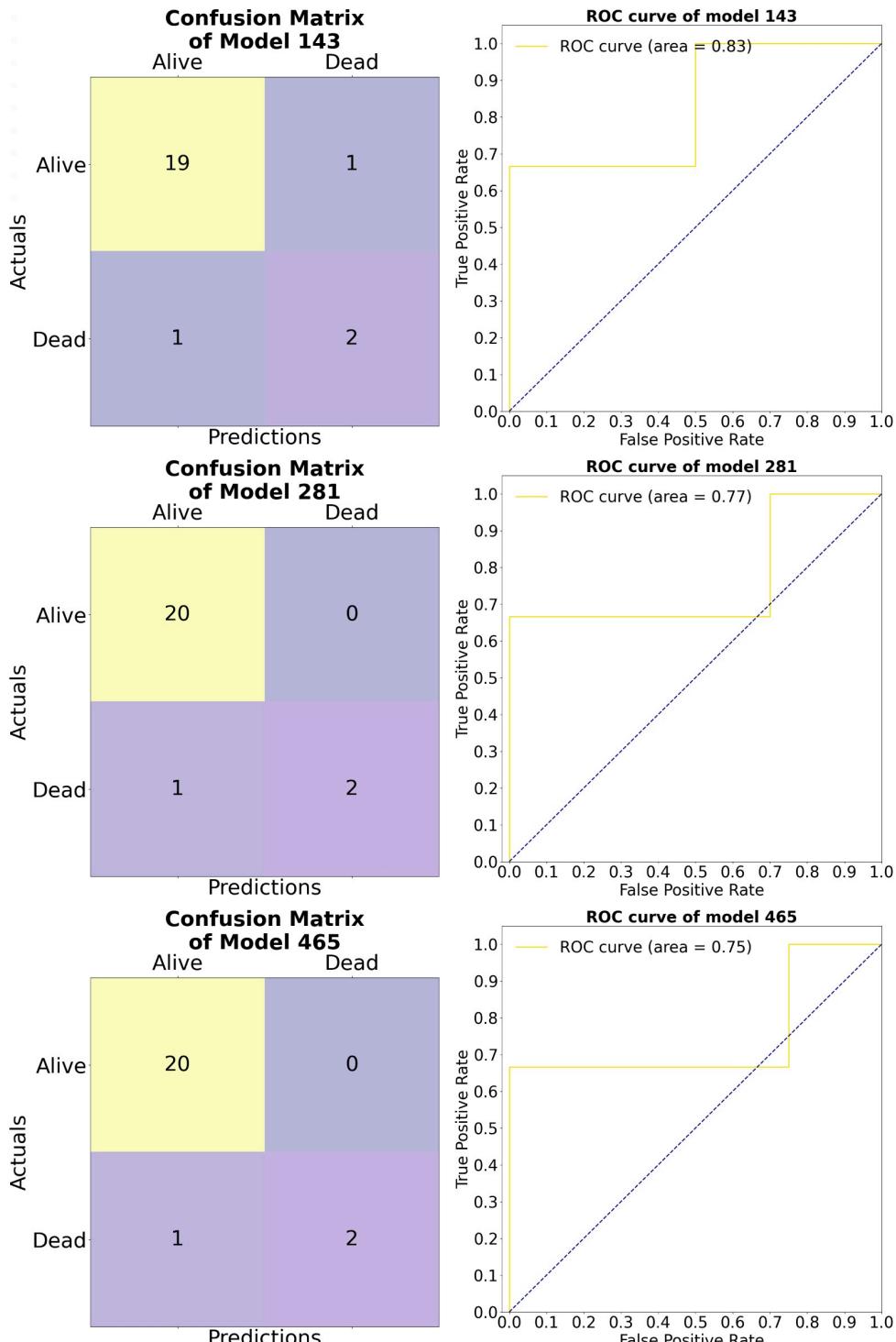
Precision	Average Precision	F1-score	MCC	Recall	AUC
0.5	0.5	0.5	0.4	0.3	0.7

Από τα εναπομείναντα μοντέλα, τα 20 καλύτερα που έμειναν μετά την κατάταξή τους με βάση τη μέση τιμή της μετρικής MCC στο σετ επικύρωσης από το 9-fold cross validation παρουσιάζονται παρακάτω πρώτα για την απόδοσή τους στο 9-fold cross validation κι έπειτα για την απόδοσή τους στο αρχικό σετ δεδομένων.



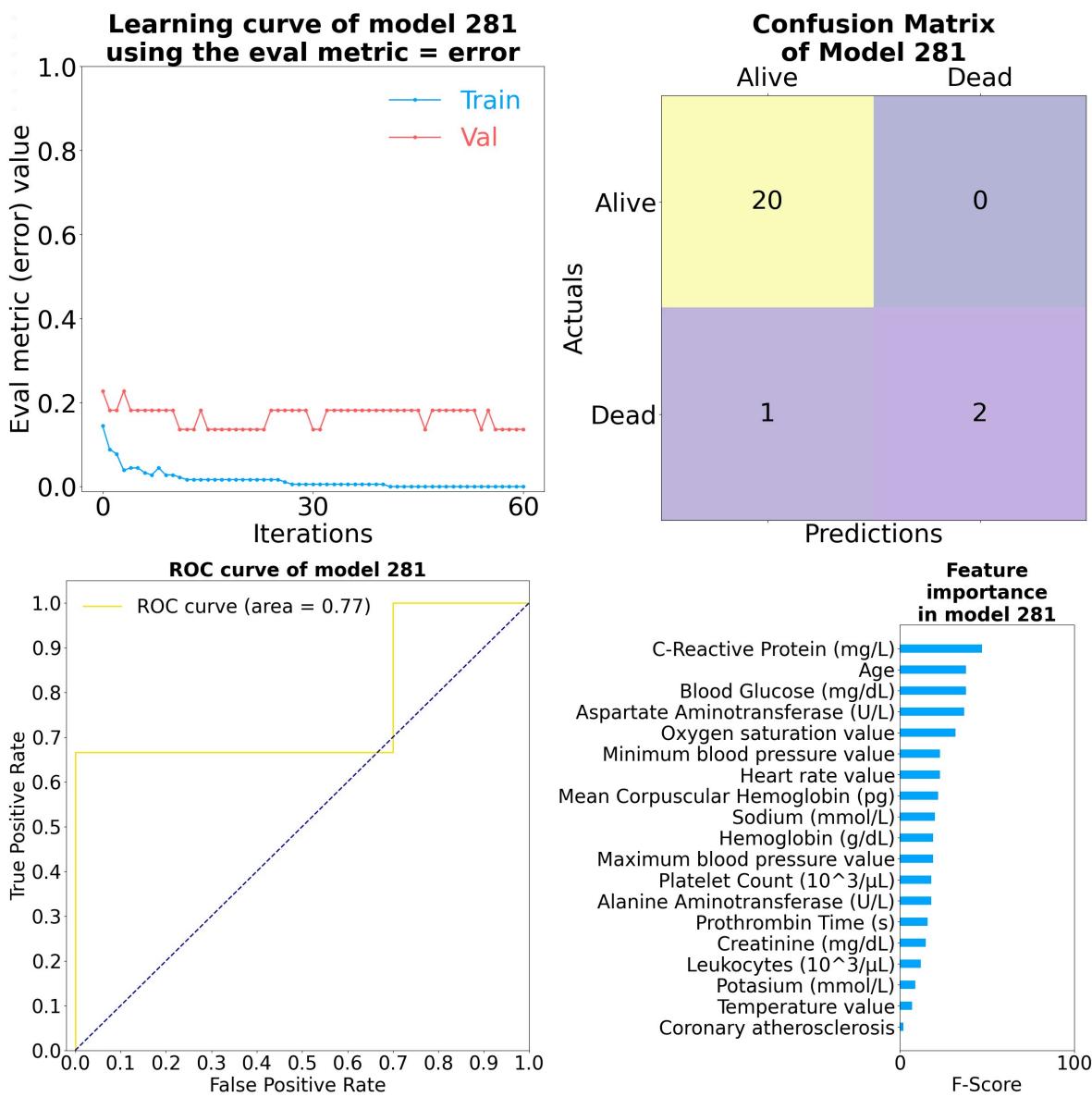
Σχήμα 5.17: Τα 20 καλύτερα μοντέλα για το merged_data_no_td χωρίς το χαρακτηριστικό Severity για ασθενείς με Διαβήτη

Κοιτώντας τα παραπάνω 4 διαγράμματα επιλέχτηκαν τα μοντέλα 143, 281, 465 ως τα καλύτερα και παρουσιάζονται παρακάτω.



Σχήμα 5.18: Confusion Matrix και καμπύλες ROC των 3 καλύτερων μοντέλων για το merged_data_no_td χωρίς το χαρακτηριστικό Severity για ασθενείς με Διαβήτη

Κοιτώντας λοιπόν τα καλύτερα 3 μοντέλα, θεωρούμε ως καλύτερο το μοντέλο 281 το οποίο έχει το υψηλότερο Precision μαζί με το μοντέλο 465, αλλά έχει απειροελάχιστα υψηλότερο auc. Το μοντέλο 281 παρουσιάζεται παρακάτω:



Σχήμα 5.19: Το καλύτερο μοντέλο για το merged_data_no_td χωρίς το χαρακτηριστικό Severity για ασθενείς με Διαβήτη

5.6.2 merged_data_no_td χωρίς το χαρακτηριστικό Severity για Diabetes = 0

Για την βελτιστοποίηση των γενικών υπερπαραμέτρων εκτελέστηκε τυχαία αναζήτηση πλέγματος, επιλέγοντας ένα τυχαίο 10% των σημείων του παρακάτω πλέγματος:

```
random_grid_params = [
(max_depth, min_child_weight, eta, subsample, colsample_bytree, scale_pos_weight)
for max_depth in [5, 6, 7, 8, 9, 10]
for min_child_weight in [1, 2, 3, 4, 5, 6]
for eta in [0.05]
for subsample in [0.5, 0.6, 0.7, 0.8, 0.9, 1]
for colsample_bytree in [0.5, 0.6, 0.7, 0.8, 0.9, 1]
for scale_pos_weight in [1, 2, 3, 4]
]
```

Για κάθε σημείο που επιλέχθηκε από το παραπάνω πλέγμα, κατασκευάστηκαν 10 μοντέλα, ένα μοντέλο για κάθε μια από τις μετρικές αξιολόγησης που ορίστηκαν, κατασκευάζοντας συνολικά 5184 μοντέλα.

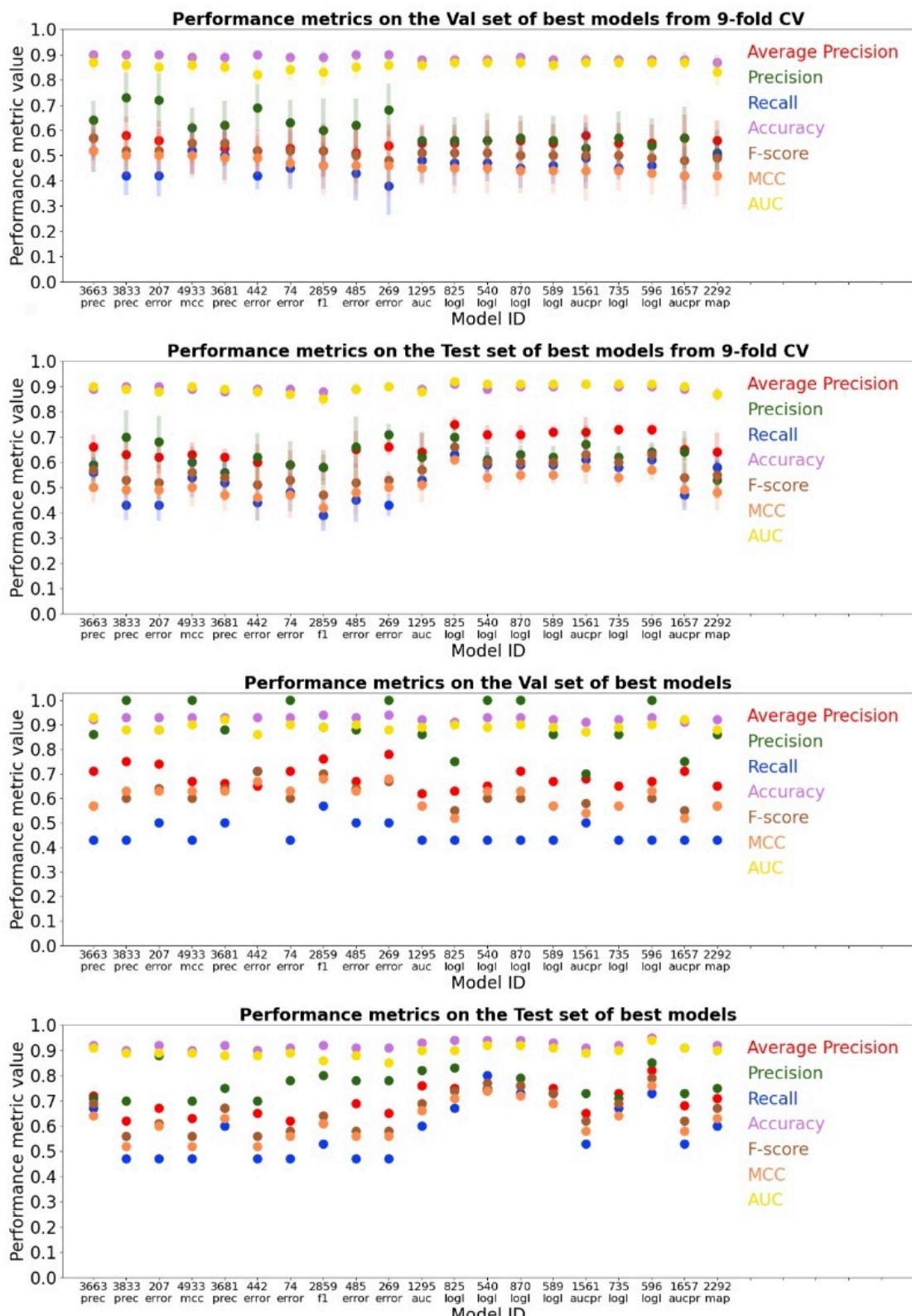
Για το φιλτράρισμα των μοντέλων με βάση κατώφλια που εφαρμόστηκαν στο σετ επικύρωσης κι ελέγχου για τις μετρικές απόδοσης, επιλέχθηκαν οι παραπάνω τιμές:

Πίνακας 5.8: Κατώφλια μετρικών απόδοσης για φιλτράρισμα μοντέλων για μέθοδο 3 για merged_data_no_td χωρίς το χαρακτηριστικό Severity για ασθενείς χωρίς Διαβήτη

Precision	Average Precision	F1-score	MCC	Recall	AUC
0.7	0.6	0.5	0.5	0.4	0.8

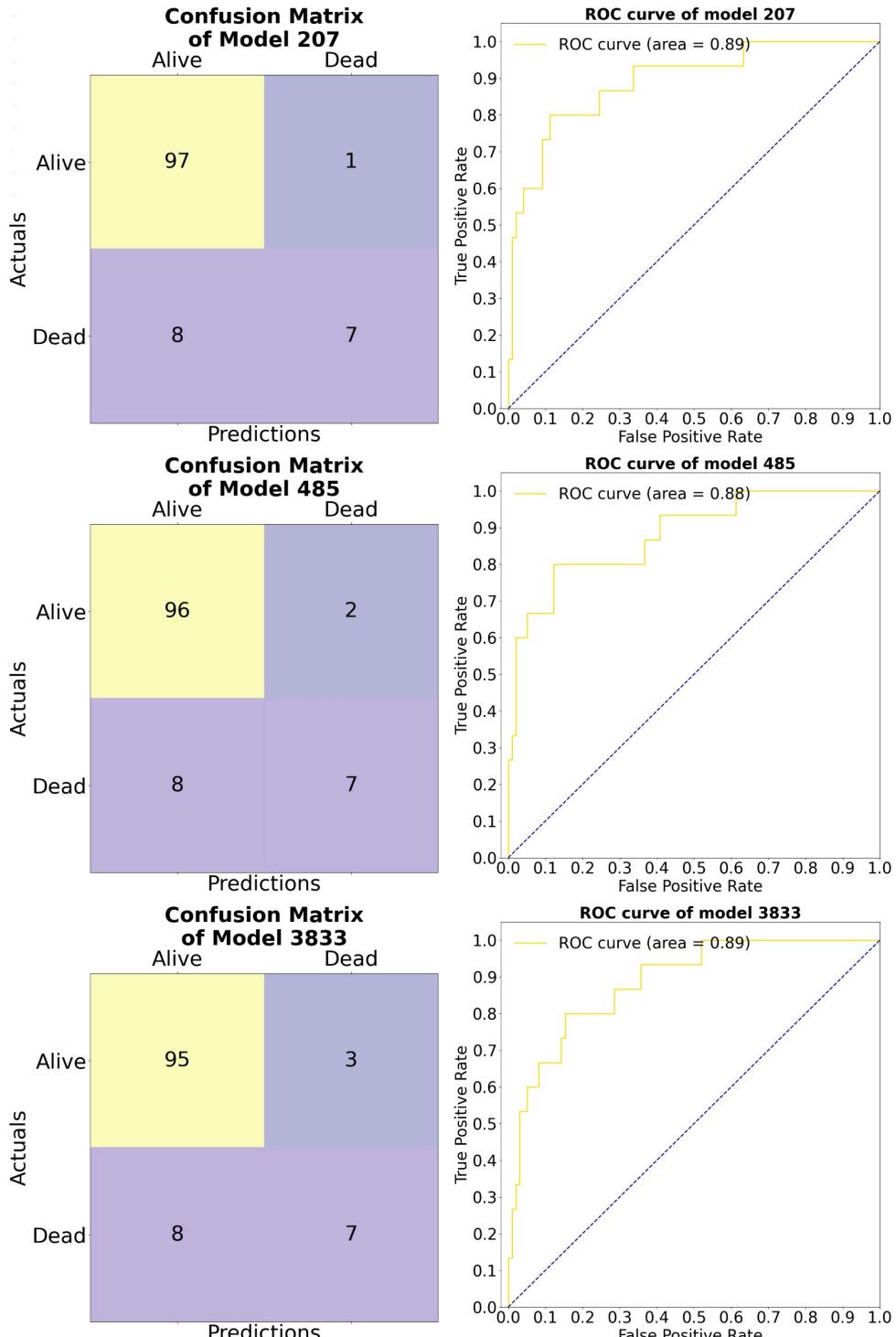
Από τα εναπομείναντα μοντέλα, τα 20 καλύτερα που έμειναν μετά την κατάταξή τους με βάση τη μέση τιμή της μετρικής MCC στο σετ επικύρωσης από το 9-fold cross validation παρουσιάζονται παρακάτω πρώτα για την απόδοσή τους στο 9-fold cross validation κι έπειτα για την απόδοσή τους στο αρχικό σετ δεδομένων.

ΚΕΦΑΛΑΙΟ 5. ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΩΝ ΚΙ ΑΠΟΤΕΛΕΣΜΑΤΑ



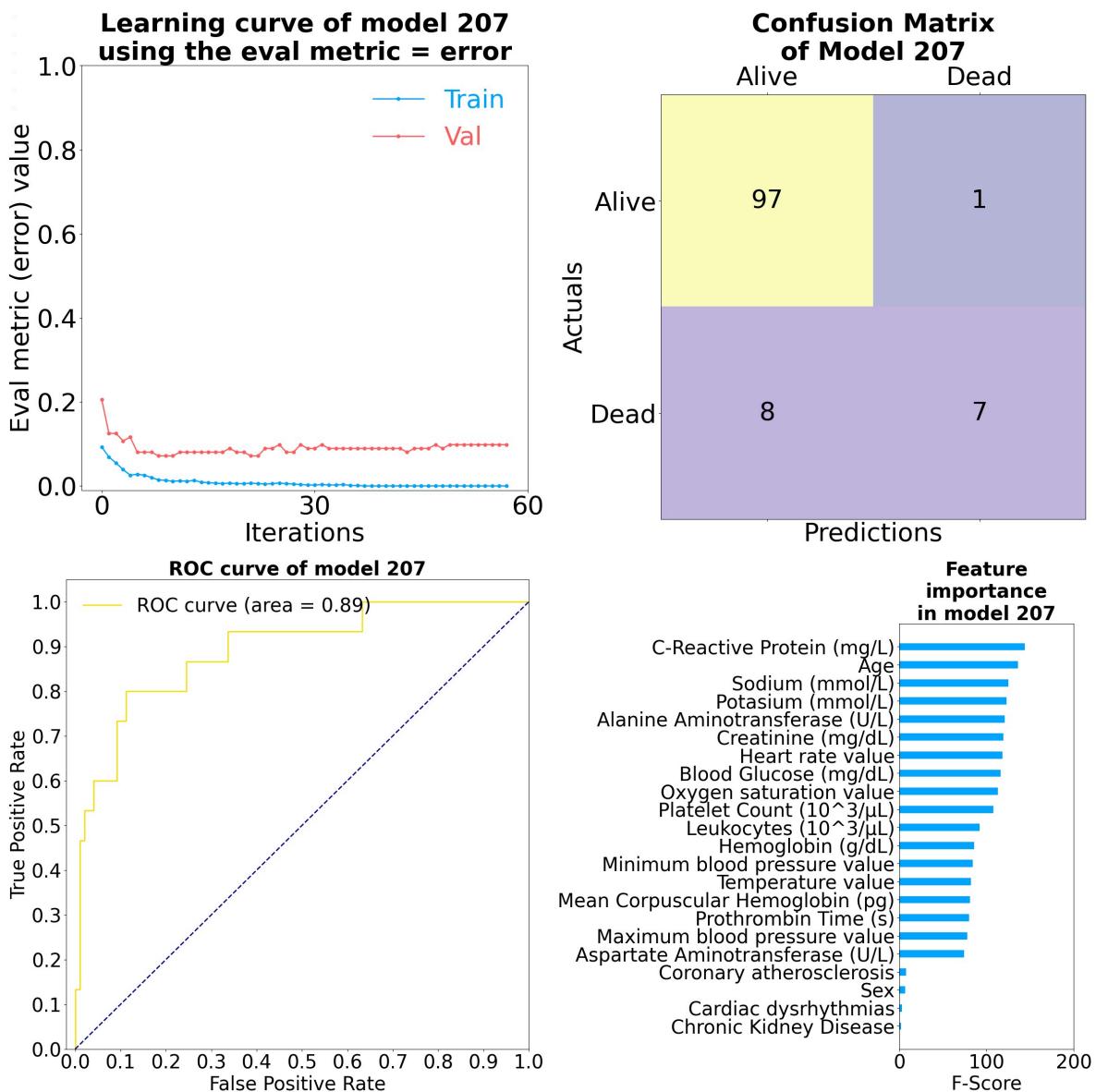
Σχήμα 5.20: Τα 20 καλύτερα μοντέλα για το merged_data_no_td χωρίς το χαρακτηριστικό Severity για ασθενείς χωρίς Διαβήτη

Κοιτώντας τα παραπάνω 4 διαγράμματα επιλέχτηκαν τα μοντέλα 3833, 207, 485 ως τα καλύτερα και παρουσιάζονται παρακάτω.



Σχήμα 5.21: Confusion Matrix και καμπύλες ROC των 3 καλύτερων μοντέλων για το merged_data_no_td χωρίς το χαρακτηριστικό Severity για ασθενείς χωρίς Διαβήτη

Κοιτώντας λοιπόν τα καλύτερα 3 μοντέλα, θεωρούμε ως καλύτερο το μοντέλο 207 το οποίο έχει το υψηλότερο Precision. Το μοντέλο 207 παρουσιάζεται παρακάτω:



Σχήμα 5.22: Το καλύτερο μοντέλο για το merged_data_no_td χωρίς το χαρακτηριστικό Severity για ασθενείς χωρίς Διαβήτη

5.6.3 merged_data_no_td με το χαρακτηριστικό Severity για Diabetes = 1

Για την βελτιστοποίηση των γενικών υπερπαραμέτρων εκτελέστηκε τυχαία αναζήτηση πλέγματος, επιλέγοντας ένα τυχαίο 10% των σημείων του παρακάτω πλέγματος:

```
random_grid_params = [
(max_depth, min_child_weight, eta, subsample, colsample_bytree, scale_pos_weight)
for max_depth in [5, 6, 7, 8, 9, 10]
for min_child_weight in [1, 2, 3, 4, 5, 6]
for eta in [0.05]
for subsample in [0.5, 0.6, 0.7, 0.8, 0.9, 1]
for colsample_bytree in [0.5, 0.6, 0.7, 0.8, 0.9, 1]
for scale_pos_weight in [1, 2, 3, 4]
]
```

Για κάθε σημείο που επιλέχθηκε από το παραπάνω πλέγμα, κατασκευάστηκαν 10 μοντέλα, ένα μοντέλο για κάθε μια από τις μετρικές αξιολόγησης που ορίστηκαν, κατασκευάζοντας συνολικά 5184 μοντέλα.

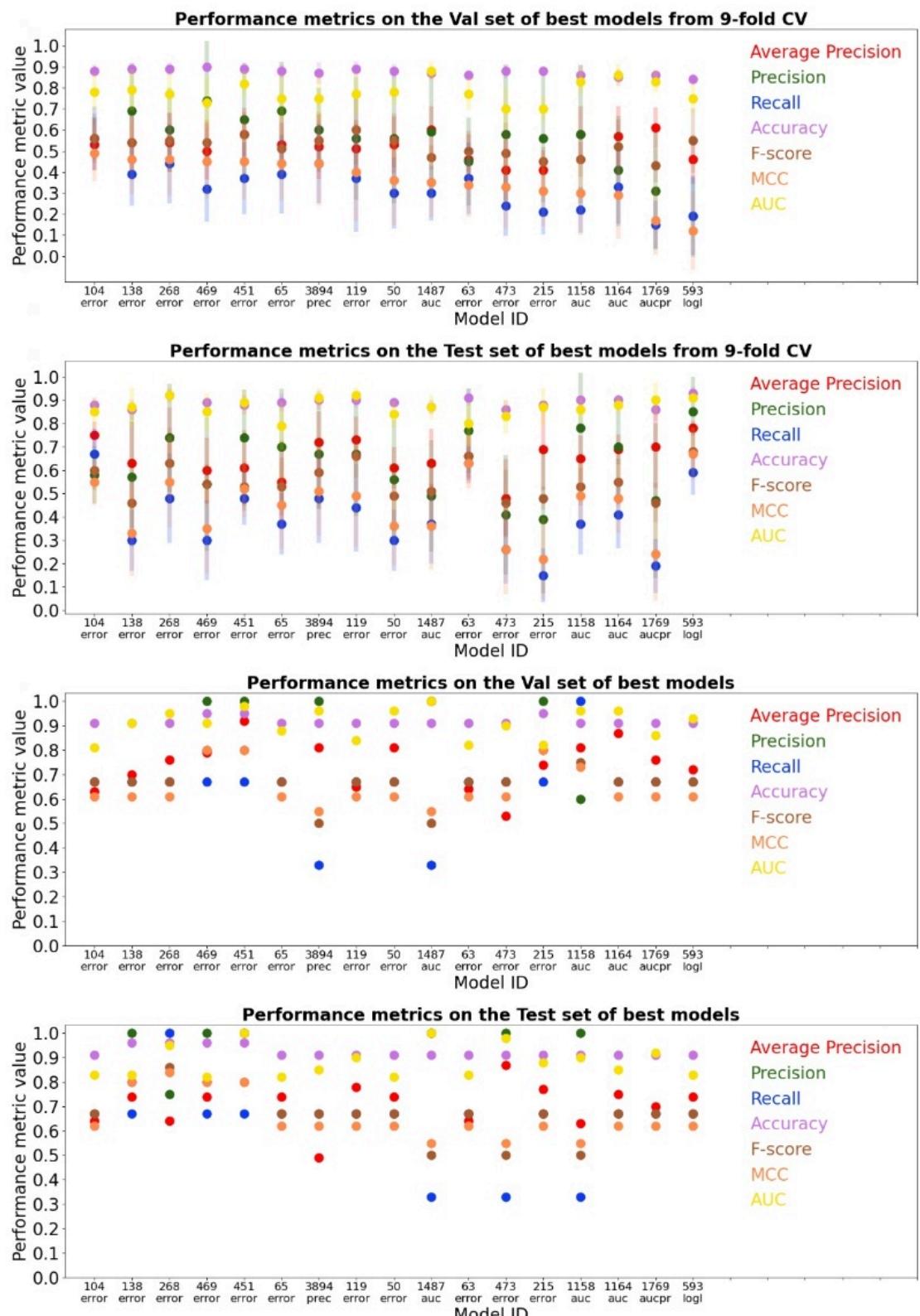
Για το φιλτράρισμα των μοντέλων με βάση κατώφλια που εφαρμόστηκαν στο σετ επικύρωσης κι ελέγχου για τις μετρικές απόδοσης, επιλέχθηκαν οι παραπάνω τιμές:

Πίνακας 5.9: Κατώφλια μετρικών απόδοσης για φιλτράρισμα μοντέλων για μέθοδο 3 για merged_data_no_td με το χαρακτηριστικό Severity για ασθενείς με Διαβήτη

Precision	Average Precision	F1-score	MCC	Recall	AUC
0.6	0.4	0.5	0.5	0.3	0.8

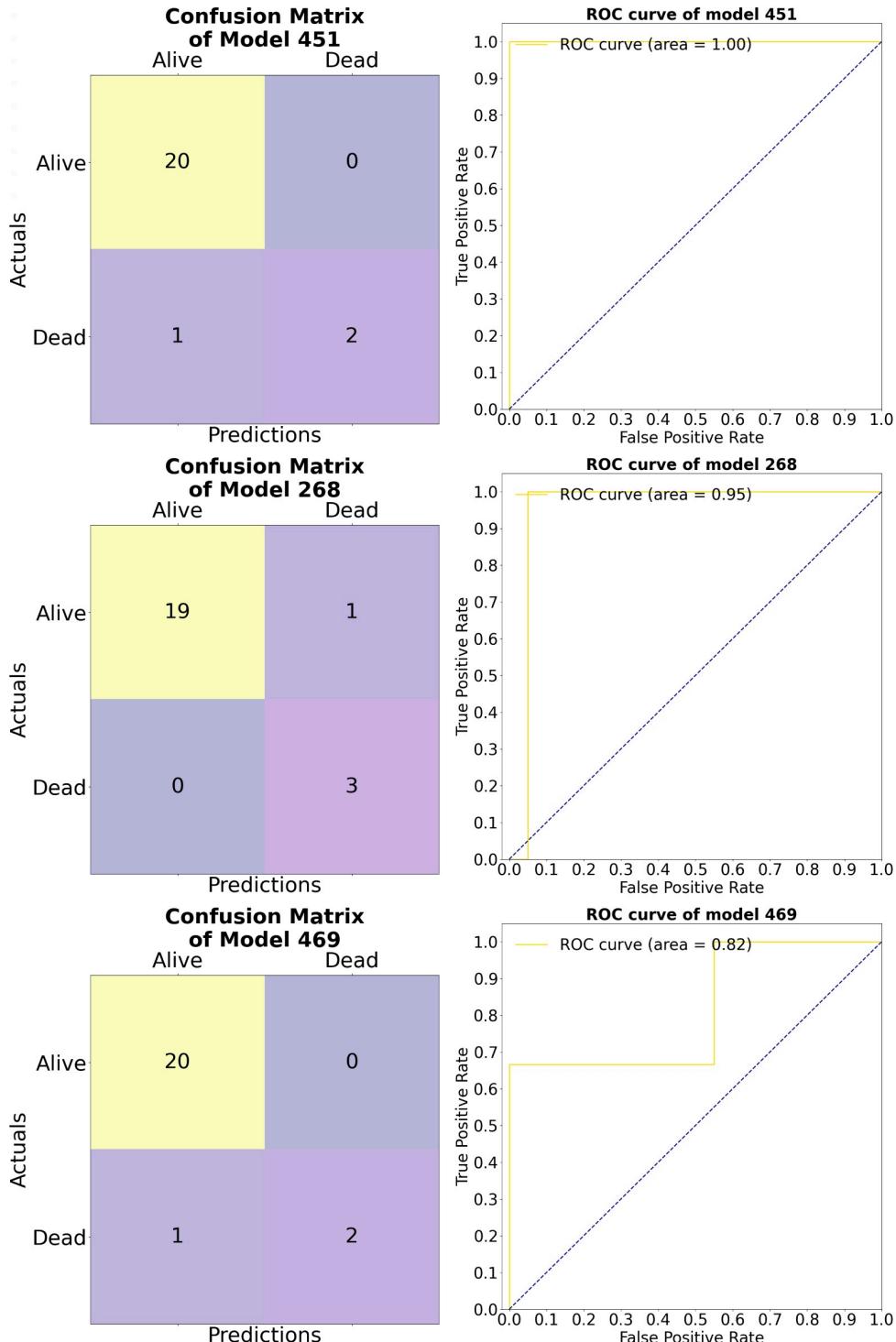
Από τα εναπομείναντα μοντέλα, τα 20 καλύτερα που έμειναν μετά την κατάταξή τους με βάση τη μέση τιμή της μετρικής MCC στο σετ επικύρωσης από το 9-fold cross validation παρουσιάζονται παρακάτω πρώτα για την απόδοσή τους στο 9-fold cross validation κι έπειτα για την απόδοσή τους στο αρχικό σετ δεδομένων.

ΚΕΦΑΛΑΙΟ 5. ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΩΝ ΚΙ ΑΠΟΤΕΛΕΣΜΑΤΑ



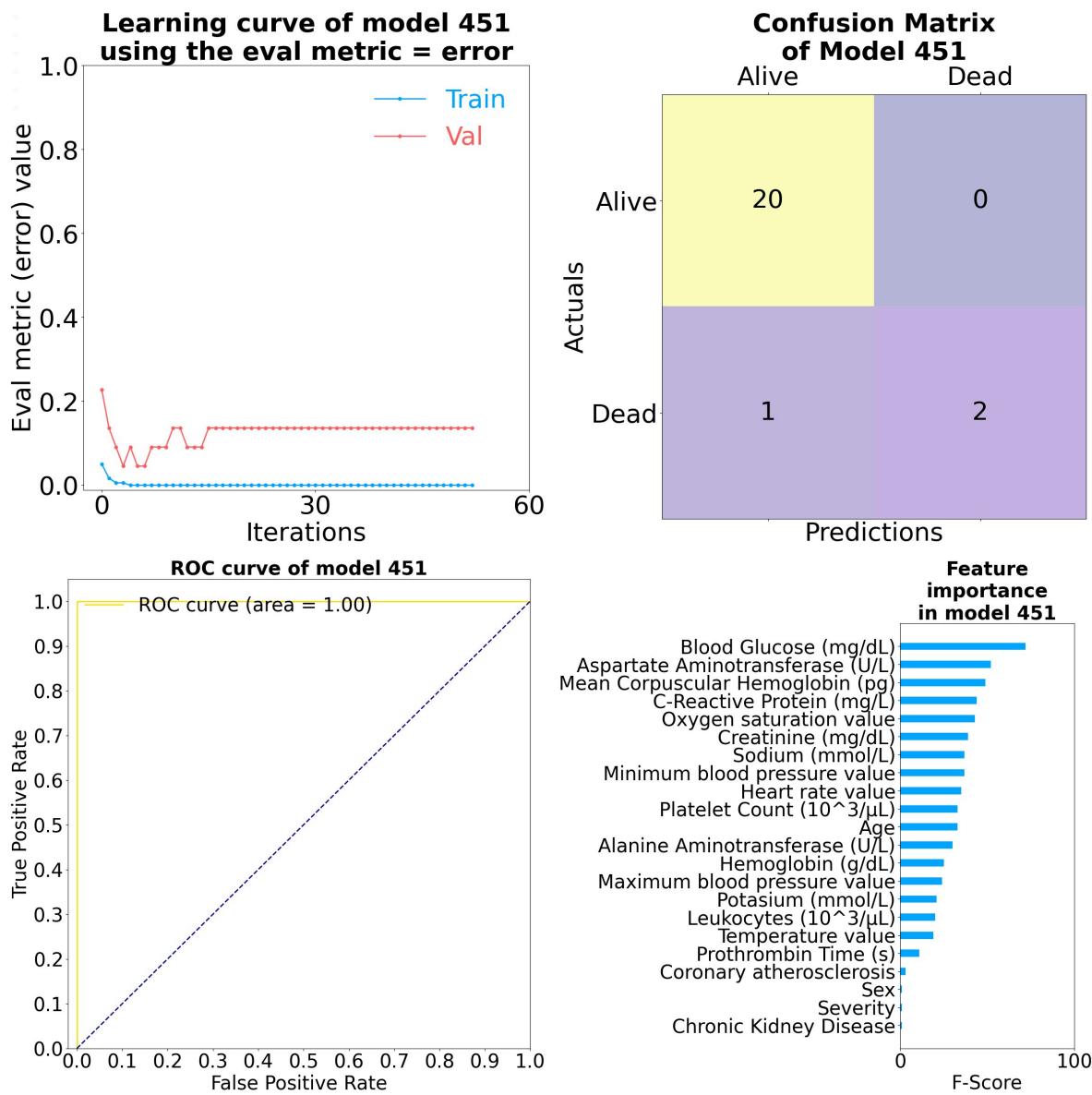
Σχήμα 5.23: Τα 20 καλύτερα μοντέλα για το merged_data_no_td με το χαρακτηριστικό Severity για ασθενείς με Διαβήτη

Κοιτώντας τα παραπάνω 4 διαγράμματα επιλέχτηκαν τα μοντέλα 451, 268 και 469 ως τα καλύτερα και παρουσιάζονται παρακάτω.



Σχήμα 5.24: Confusion Matrix και καμπύλες ROC των 3 καλύτερων μοντέλων για το merged_data_no_td με το χαρακτηριστικό Severity για ασθενείς με Διαβήτη

Κοιτώντας λοιπόν τα καλύτερα 3 μοντέλα, θεωρούμε ως καλύτερο το μοντέλο 451 το οποίο έχει το υψηλότερο Precision μαζί με το μοντέλο 469, αλλά έχει υψηλότερο auc. Το μοντέλο 451 παρουσιάζεται παρακάτω:



Σχήμα 5.25: Το καλύτερο μοντέλο για το merged_data_no_td με το χαρακτηριστικό Severity για ασθενείς με Διαβήτη

5.6.4 merged_data_no_td με το χαρακτηριστικό Severity για Diabetes = 0

Για την βελτιστοποίηση των γενικών υπερπαραμέτρων εκτελέστηκε τυχαία αναζήτηση πλέγματος, επιλέγοντας ένα τυχαίο 10% των σημείων του παρακάτω πλέγματος:

```
random_grid_params = [
(max_depth, min_child_weight, eta, subsample, colsample_bytree, scale_pos_weight)
for max_depth in [5, 6, 7, 8, 9, 10]
for min_child_weight in [1, 2, 3, 4, 5, 6]
for eta in [0.05]
for subsample in [0.5, 0.6, 0.7, 0.8, 0.9, 1]
for colsample_bytree in [0.5, 0.6, 0.7, 0.8, 0.9, 1]
for scale_pos_weight in [1, 2, 3, 4]
]
```

Για κάθε σημείο που επιλέχθηκε από το παραπάνω πλέγμα, κατασκευάστηκαν 10 μοντέλα, ένα μοντέλο για κάθε μια από τις μετρικές αξιολόγησης που ορίστηκαν, κατασκευάζοντας συνολικά 5184 μοντέλα.

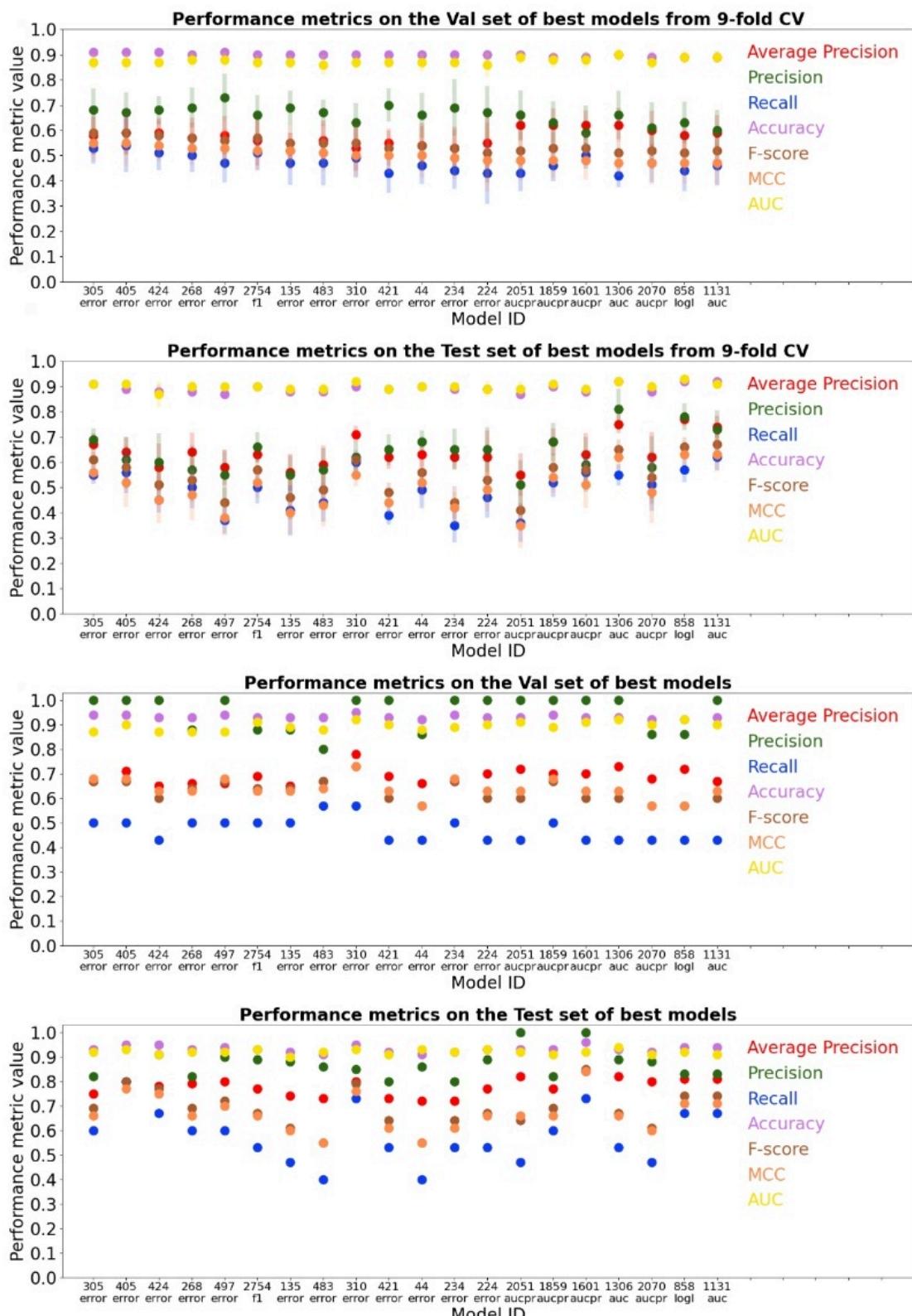
Για το φιλτράρισμα των μοντέλων με βάση κατώφλια που εφαρμόστηκαν στο σετ επικύρωσης κι ελέγχου για τις μετρικές απόδοσης, επιλέχθηκαν οι παραπάνω τιμές:

Πίνακας 5.10: Κατώφλια μετρικών απόδοσης για φιλτράρισμα μοντέλων για μέθοδο 3 για merged_data_no_td με το χαρακτηριστικό Severity για ασθενείς χωρίς Διαβήτη

Precision	Average Precision	F1-score	MCC	Recall	AUC
0.8	0.4	0.5	0.5	0.3	0.8

Από τα εναπομείναντα μοντέλα, τα 20 καλύτερα που έμειναν μετά την κατάταξή τους με βάση τη μέση τιμή της μετρικής MCC στο σετ επικύρωσης από το 9-fold cross validation παρουσιάζονται παρακάτω πρώτα για την απόδοσή τους στο 9-fold cross validation κι έπειτα για την απόδοσή τους στο αρχικό σετ δεδομένων.

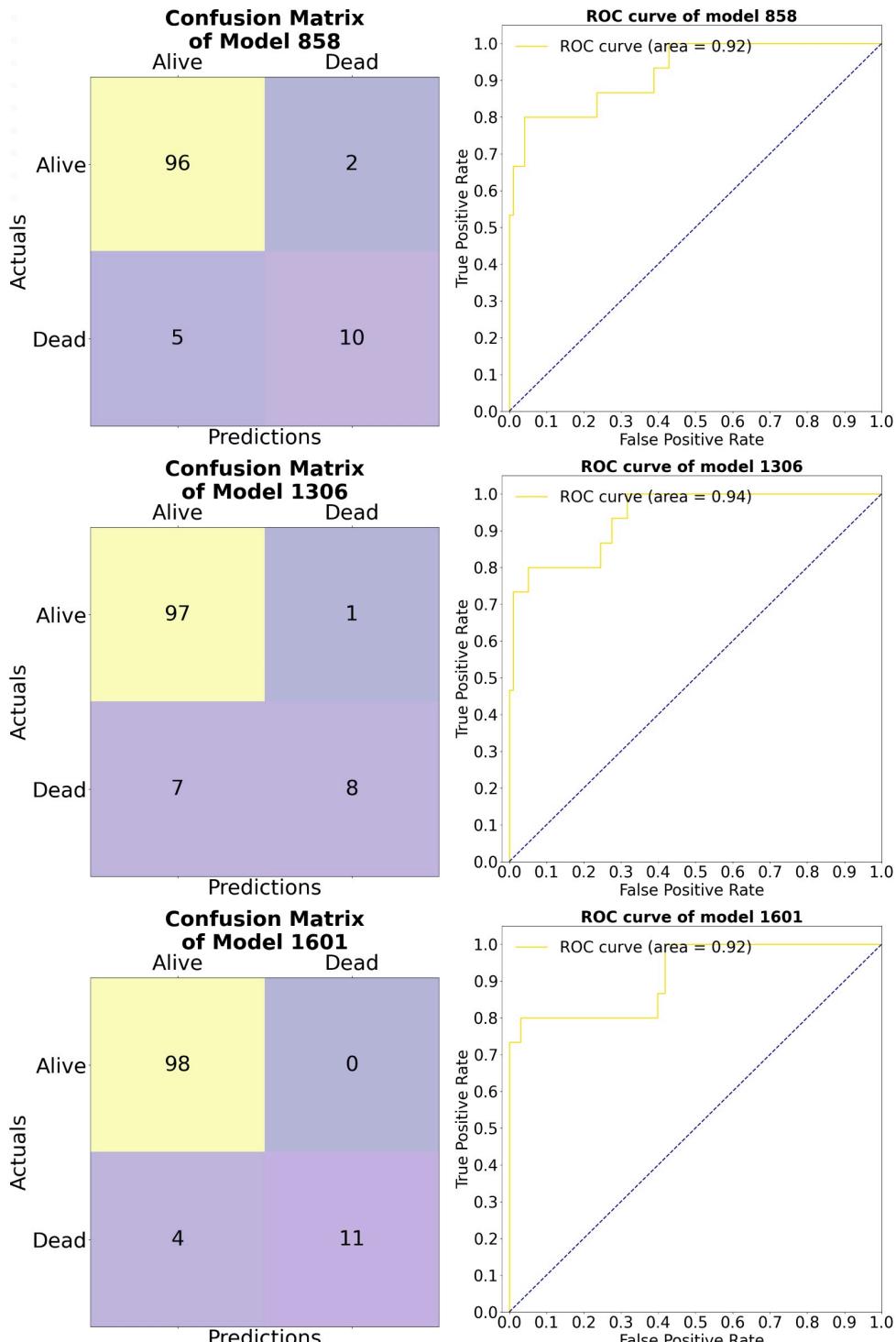
ΚΕΦΑΛΑΙΟ 5. ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΩΝ ΚΙ ΑΠΟΤΕΛΕΣΜΑΤΑ



Σχήμα 5.26: Τα 20 καλύτερα μοντέλα για το merged_data_no_td με το χαρακτηριστικό Severity για ασθενείς χωρίς Διαβήτη

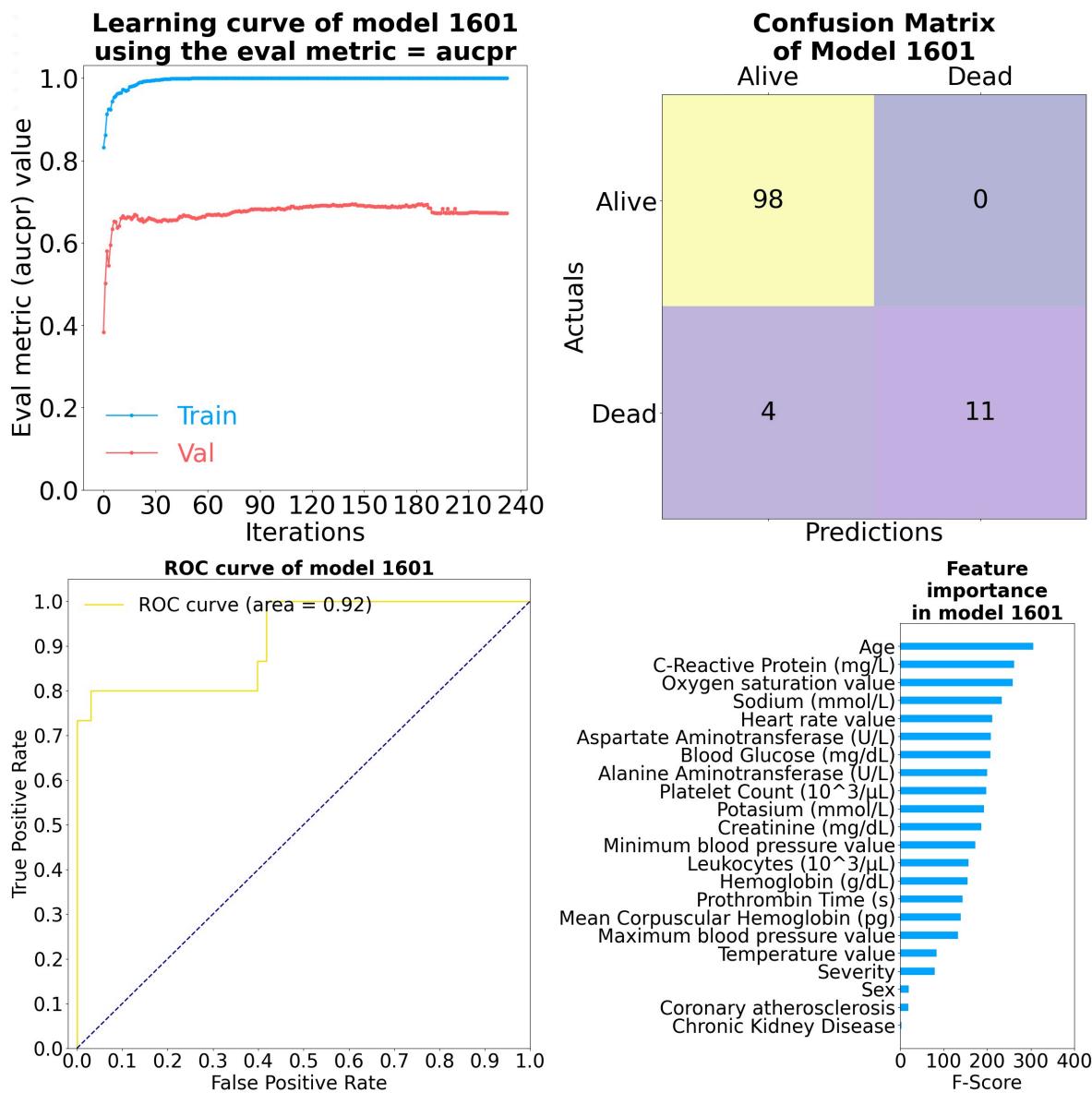
5.6. ΑΠΟΤΕΛΕΣΜΑΤΑ ΜΕΘΟΔΟΥ 3

Κοιτώντας τα παραπάνω 4 διαγράμματα επιλέχτηκαν τα μοντέλα 1306, 858 και 1601 ως τα καλύτερα και παρουσιάζονται παρακάτω:



Σχήμα 5.27: Confusion Matrix και καμπύλες ROC των 3 καλύτερων μοντέλων για το merged_data_no_td με το χαρακτηριστικό Severity για ασθενείς χωρίς Διαβήτη

Κοιτώντας λοιπόν τα καλύτερα 3 μοντέλα, θεωρούμε ως καλύτερο το μοντέλο 1601 το οποίο έχει το υψηλότερο Precision. Το μοντέλο 1601 παρουσιάζεται παρακάτω:



Σχήμα 5.28: Το καλύτερο μοντέλο για το merged_data_no_td με το χαρακτηριστικό Severity για ασθενείς χωρίς Διαβήτη

5.7 ΣΥΓΚΡΙΣΗ ΜΟΝΤΕΛΩΝ

Έχοντας δημιουργήσει εκατοντάδες χιλιάδες μοντέλα με πολλούς διαφορετικούς τρόπους υλοποίησης καταλήξαμε εν τέλει στα τελικά μας μοντέλα για τις μεθόδους 1 και 3. Παραθέτουμε αυτά τα μοντέλα παρακάτω για λόγους εποπτείας:

Πίνακας 5.11: Τα καλύτερα μοντέλα που επιλέχτηκαν από τις μεθόδους 1 και 3

Data	Ασθενείς	Severity	Method	Model ID
merged_data_no_td	Όλοι	Όχι	Μέθοδος 1	225
merged_data_no_td	Όλοι	Ναι	Μέθοδος 1	939
merged_data_no_td	Διαβητικοί μόνο	Όχι	Μέθοδος 3	281
merged_data_no_td	Μη διαβητικοί μόνο	Όχι	Μέθοδος 3	207
merged_data_no_td	Διαβητικοί μόνο	Ναι	Μέθοδος 3	451
merged_data_no_td	Μη διαβητικοί μόνο	Ναι	Μέθοδος 3	1601

Για να δούμε πώς συμπεριφέρονται τα μοντέλα στην πράξη, κανονικά θα συγκρίνουμε τα μοντέλα που επιλέξαμε με το μοντέλο του covidanalytics.io χρησιμοποιώντας τα δεδομένα ελέγχου για τυχαίους ασθενείς. Η ιστοσελίδα του covidanalytics.io όμως δεν είναι πλέον διαθέσιμη και με βάση το Wayback Machine η τελευταία φορά που χρησιμποιήθηκε ήταν στις 7 Οκτώβρη 2021. Αναγκαστικά λοιπόν θα συγκρίνουμε τα μοντέλα μεταξύ τους μονάχα στα δεδομένα ελέγχου. Τα δεδομένα ελέγχου είναι τα ίδια για κάθε μοντέλο που φτιάχτηκε με την ίδια μέθοδο. Για τον σκοπό αυτό θα κάνουμε τις εξής συγκρίσεις:

- Μοντέλο 225 vs Μοντέλο 939
- Μοντέλο 281 vs Μοντέλο 451
- Μοντέλο 207 vs Μοντέλο 1601

Παραθέτουμε λοιπόν τις συγκρίσεις αυτές παρακάτω σε μορφή πινάκων. Υπενθυμίζουμε ότι για προβλέψεις μεγαλύτερες ή ίσες του 50% ο ασθενής προβλέπεται ότι θα πεθάνει και ταξινομείται ως νεκρός, ενώ για προβλέψεις μικρότερες του 50% ο ασθενής προβλέπεται ότι θα ζήσει και ταξινομείται ως ζωντανός.

ΚΕΦΑΛΑΙΟ 5. ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΩΝ ΚΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

Πίνακας 5.12: Μοντέλο 225 vs Μοντέλο 939

ID ασθενή	Κλάση ασθενή	Πρόβλεψη μοντέλου 225	Πρόβλεψη μοντέλου 939
672	Ζωντανός	2%	5%
471	Ζωντανός	6%	13%
869	Ζωντανός	1%	2%
323	Ζωντανός	24%	37%
688	Ζωντανός	0%	1%
1596	Νεκρός	8%	27%
382	Νεκρός	2%	4%
249	Νεκρός	43%	33%
1244	Νεκρός	57%	52%
173	Νεκρός	70%	67%

Παρατηρούμε ότι για τους ζωντανούς ασθενείς το μοντέλο 939 δίνει υψηλότερο ρίσκο θνητότητας σε σχέση με το μοντέλο 225, ενώ για τους νεκρούς ασθενείς συμβαίνει το ανάποδο. Με βάση αυτό το μικρό δείγμα λοιπόν, το οποίο σαφώς δεν είναι αντιπροσωπευτικό, το μοντέλο 225 φαίνεται να τα πηγαίνει καλύτερα. Προφανώς όμως για να αποφασιστεί το καλύτερο μοντέλο πρέπει να γίνουν πιο εκτενείς έλεγχοι. Εδώ κάνουμε απλώς μια αναφορά αναδεικνύωντας επίσης ότι το χαρακτηριστικό Severity δεν φαίνεται να βελτιώνει τις εκτιμήσεις στη συγκεκριμένη μέθοδο. Να σημειωθεί ότι το μοντέλο 225 δείξαμε ότι έχει μεγαλύτερο Precision από το μοντέλο 939, ενώ το μοντέλο 939 έχει υψηλότερο auc.

Πίνακας 5.13: Μοντέλο 281 vs Μοντέλο 451

ID ασθενή	Κλάση ασθενή	Πρόβλεψη μοντέλου 281	Πρόβλεψη μοντέλου 451
1228	Ζωντανός	28%	41%
999	Ζωντανός	32%	41%
337	Ζωντανός	36%	47%
693	Νεκρός	52%	49%
502	Νεκρός	32%	50%
859	Νεκρός	51%	51%

Παρατηρούμε ότι τόσο για τους ζωντανούς όσο και για τους νεκρούς ασθενείς το μοντέλο 451 δίνει υψηλότερο ρίσκο θνητότητας σε σχέση με το μοντέλο 281. Οι τιμές όμως που δίνονται είναι πάνω από 41% για το μοντέλο 451 και πάνω από το 28% για το μοντέλο 281. Συνεπώς, το μοντέλο 451 φαίνεται να προτιμάει να "γέρνει" τις προβλέψεις του προς τους νεκρούς κάτι το οποίο σε πραγματικές συνθήκες δεν είναι πολύ χρήσιμο. Να σημειωθεί ότι τα 2 μοντέλα έχουν το ίδιο Precision, αλλά το μοντέλο 451 έχει αρκετά υψηλότερο auc σε σχέση με το μοντέλο 281. Με βάση αυτό το μικρό δείγμα λοιπόν, το οποίο σαφώς δεν είναι αντιπροσωπευτικό, το μοντέλο 281 φαίνεται να είναι προτιμότερο καθώς "ξεκαθαρίζει" τους ασθενείς καλύτερα, ενώ το μοντέλο 451 οριακά τους μπερδεύει. Προφανώς όμως για να αποφασιστεί το καλύτερο μοντέλο πρέπει να γίνουν πιο εκτενείς έλεγχοι.

Πίνακας 5.14: Μοντέλο 207 vs Μοντέλο 1601

ID ασθενή	Κλάση ασθενή	Πρόβλεψη μοντέλου 207	Πρόβλεψη μοντέλου 1601
1227	Ζωντανός	39%	26%
805	Ζωντανός	41%	17%
91	Ζωντανός	32%	0%
856	Ζωντανός	42%	4%
903	Ζωντανός	37%	10%
358	Νεκρός	58%	99%
575	Νεκρός	45%	37%
990	Νεκρός	52%	71%
202	Νεκρός	43%	91%
185	Νεκρός	56%	75%

Παρατηρούμε ότι για τους ζωντανούς ασθενείς το μοντέλο 207 δίνει υψηλότερο ρίσκο θνητότητας σε σχέση με το μοντέλο 1601, ενώ για τους νεκρούς ασθενείς συμβαίνει το ανάποδο. Με βάση αυτό το μικρό δείγμα λοιπόν, το οποίο σαφώς δεν είναι αντιπροσωπευτικό, το μοντέλο 1601 φαίνεται να τα πηγαίνει καλύτερα. Προφανώς όμως για να αποφασιστεί το καλύτερο μοντέλο πρέπει να γίνουν πιο εκτενείς έλεγχοι. Εδώ κάνουμε απλώς μια αναφορά αναδεικνύωντας επίσης ότι το χαρακτηριστικό Severity φαίνεται να βελτιώνει τις εκτιμήσεις στη συγκεκριμένη μέθοδο. Να σημειωθεί ότι το μοντέλο 1601 δείχαμε ότι έχει μεγαλύτερο Precision και Recall από το μοντέλο 207 κι ελάχιστα μεγαλύτερο auc.

Από την συνολική εικόνα των συγκρίσεων το πιο αποτελεσματικό μοντέλο με βάση τις μετρικές απόδοσης θα ήταν ο συνδυασμός του μοντέλου 1601 για ασθενείς χωρίς διαβήτη και του μοντέλου 451 για ασθενείς με διαβήτη. Τα μοντέλα για ασθενείς με διαβήτη όμως εκπαιδεύτηκαν με ελάχιστα δεδομένα. Κοιτάζοντας τον πίνακα 5.4 βλέπουμε ότι μονάχα 225 ασθενείς από τους συνολικούς είχαν διαβήτη και συνεπώς τα μοντέλα που φτιάχτηκαν αν και τα πηγαίνουν καλά δεν έχουν μεγάλη αξιόπιστία. Αυτό φαίνεται κι από το γεγονός ότι το ρίσκο θνητότητας που επιστρέφει το μοντέλο 451 είναι πολύ κοντά στο 50% για θετικούς κι αρνητικούς ασθενείς. Για τον ίδιο λόγο δεν είναι τόσο αξιόπιστα τα μοντέλα 207 και 281. Φαίνεται πάντως πως αν είχαμε περισσότερα δεδομένα για διαβητικούς ασθενείς, τότε ένας τέτοιος διαχωρισμός των ασθενών σε διαβητικούς και μη θα έφερνε καλύτερα αποτελέσματα. Επίσης, θυμίζουμε ότι τα μοντέλα 939, 451 και 1601 έχουν κατασκευαστεί με το χαρακτηριστικό Severity, το οποίο κανονικά δεν είναι γνωστό αλλά πρέπει να εκτιμηθεί ξεχωριστά και δεν έχουμε κατασκευάσει κάποιο λειτουργικό εκτιμητή του συγκεκριμένου χαρακτηριστικού. Τα μοντέλα αυτά ήθελαν απλώς να αναδείξουν τη συνεισφορά ενός δικού μας τενχητού χαρακτηριστικού. Έτσι λοιπόν, το συνολικά πιο αξιόπιστο μοντέλο είναι το 225 το οποίο είναι κι αυτό που χρησιμοποιεί κι η διαδικτυακή εφαρμογή που παρουσιάζεται στη συνέχεια.

6

Web Application

Στην προσπάθεια αναπαραγωγής της δουλειάς του covidanalytics.io δημιουργήσαμε μια απλή διαδικτυακή εφαρμογή η οποία κάνει εκτίμηση του ρίσκου θνητότητας για κάποιον ασθενή. Συγκεκριμένα, ο χρήστης μπορεί να δώσει στην εφαρμογή ορισμένα δημογραφικά, γενικά και αιματολογικά δεδομένα για κάποιον πραγματικό ή υποθετικό ασθενή κι η εφαρμογή θα στείλει τα δεδομένα αυτά σε ένα αποθηκευμένο XGBoost μοντέλο από αυτά που κατασκευάσαμε και θα επιστρέψει το ρίσκο θνητότητας. Για την παρακάτω παρουσίαση χρησιμοποιήθηκε το μοντέλο 225 που κρίθηκε ως το καλύτερο μοντέλο της μεθόδου 1 για τον πίνακα δεδομένων merged_data_no_td χωρίς το χαρακτηριστικό Severity. Δείχνουμε παρακάτω πώς λειτουργεί η εφαρμογή με εικόνες.

Mortality Risk Calculator

Please fill the boxes below in order to make a prediction.

Age

- +

Sex

- +

Potassium (mmol/L)

- +

Creatinine (mg/dL)

- +

Prothrombin Time (s)

- +

Hemoglobin (g/dL)

- +

Aspartate Aminotransferase (U/L)

- +

Blood Glucose (mg/dL)

- +

Sodium (mmol/L)

- +

C-Reactive Protein (mg/L)

- +

Mean Corpuscular Hemoglobin (pg)

- +

Σχήμα 6.1: Σελίδα εφαρμογής με τα δεδομένα εισόδου που μπορεί να δώσει ο χρήστης - Τμήμα 1

ΚΕΦΑΛΑΙΟ 6. WEB APPLICATION

Alanine Aminotransferase (U/L)

-+

Platelet Count ($10^3/\mu\text{L}$)

-+

Leukocytes ($10^3/\mu\text{L}$)

-+

Maximum blood pressure value

-+

Minimum blood pressure value

-+

Temperature value

-+

Heart rate value

-+

Oxygen saturation value

-+

Cardiac dysrhythmias

-+

Chronic Kidney Disease

-+

Coronary atherosclerosis

-+

Diabetes

-+

Show input passed

Make prediction

Σχήμα 6.2: Σελίδα εφαρμογής με τα δεδομένα εισόδου που μπορεί να δώσει ο χρήστης - Τμήμα 2

Mortality Risk Calculator

Please fill the boxes below in order to make a prediction.

Age

88.00

- +

Sex

0.00

- +

Potassium (mmol/L)

4.23

- +

Creatinine (mg/dL)

1.04

- +

Prothrombin Time (s)

12.70

- +

Hemoglobin (g/dL)

11.80

- +

Aspartate Aminotransferase (U/L)

17.30

- +

Blood Glucose (mg/dL)

123.60

- +

Sodium (mmol/L)

136.40

- +

C-Reactive Protein (mg/L)

5.39

- +

Mean Corpuscular Hemoglobin (pg)

26.00

- +

Σχήμα 6.3: Σελίδα εφαρμογής με συμπληρωμένα δεδομένα εισόδου για τον ασθενή 672 - Τμήμα 1

ΚΕΦΑΛΑΙΟ 6. WEB APPLICATION

Alanine Aminotransferase (U/L)

11.30-+

Platelet Count ($10^3/\mu\text{L}$)

312.00-+

Leukocytes ($10^3/\mu\text{L}$)

9.00-+

Maximum blood pressure value

134.00-+

Minimum blood pressure value

63.00-+

Temperature value

35.86-+

Heart rate value

55.33-+

Oxygen saturation value

96.33-+

Cardiac dysrhythmias

0.00-+

Chronic Kidney Disease

0.00-+

Coronary atherosclerosis

0.00-+

Diabetes

0.00-+

Show input passed

Make prediction

Σχήμα 6.4: Σελίδα εφαρμογής με συμπληρωμένα δεδομένα εισόδου για τον ασθενή 672 - Τμήμα 2

	0
Age	88.0000
Sex	0.0000
Mean Corpuscular Hemoglobin (pg)	26.0000
Creatinine (mg/dL)	1.0400
Aspartate Aminotransferase (U/L)	17.3000
Sodium (mmol/L)	136.4000
Prothrombin Time (s)	12.7000
Platelet Count ($10^3/\mu\text{L}$)	0.0000
C-Reactive Protein (mg/L)	5.3900
Potassium (mmol/L)	4.2300
Leukocytes ($10^3/\mu\text{L}$)	0.0000
Blood Glucose (mg/dL)	123.6000
Alanine Aminotransferase (U/L)	11.3000
Hemoglobin (g/dL)	11.8000
Maximum blood pressure value	134.0000
Minimum blood pressure value	63.0000
Temperature value	35.8600
Heart rate value	55.3300
Oxygen saturation value	96.3300
Cardiac dysrhythmias	0.0000
Chronic Kidney Disease	0.0000
Coronary atherosclerosis	0.0000
Diabetes	0.0000
Platelet Count ($10^3/\mu\text{L}$)	312.0000
Leukocytes ($10^3/\mu\text{L}$)	9.0000

Make prediction

Mortality Risk score with input data given: 2 %

Σχήμα 6.5: Λειτουργία κουμπιών "Show input passed" και "Make prediction"

7

Συμπεράσματα και Προεκτάσεις

7.1 ΣΥΜΠΕΡΑΣΜΑΤΑ

Κοιτώντας όλα τα αποτελέσματα που συλλέξαμε μπορούμε να πούμε πως φαίνεται να είναι επιλύσιμο το πρόβλημα της πρόβλεψης ρίσκου θνητότητας ασθενών με COVID-19. Καταφέραμε να κατασκευάσουμε μάλιστα μοντέλα τα οποία έχουν αρκετά υψηλές τιμές στις μετρικές απόδοσής τους. Αυτό όμως δεν σημαίνει ότι αυτά τα μοντέλα είναι απαραίτητα αξιόπιστα λόγω των λίγων δεδομένων που είχαμε στη διάθεσή μας. Συγκεκριμένα, τα καλύτερά μας μοντέλα φτιάχτηκαν από 1348 ασθενείς του πίνακα δεδομένων merged_data_no_td, εκ των οποίων οι 1078 χρησιμοποιήθηκαν μονάχα για εκπαίδευση, οι 135 για επικύρωση και οι υπόλοιποι 135 για έλεγχο. Σαφώς, ένα τόσο μικρό δείγμα δεν μπορεί να είναι αντιπροσωπευτικό. Εφόσον θέλουμε να κάνουμε έναν έλεγχο αξιοπιστίας θα πρέπει να χρησιμοποιήσουμε δεδομένα από άλλες πηγές που το μοντέλο μας δεν έχει ξαναδεί και να δούμε πώς συμπεριφέρεται σε αυτά.

Επίσης, δεδομένου του ότι δεν έχουμε εικόνα του πώς ήταν η κατάσταση των ασθενών, τα ρίσκα θνητότητας των μοντέλων μας δεν είναι σίγουρο αν αντιστοιχούν σε πραγματικά ρίσκα θνητότητας. Με άλλα λόγια δεν ξέρουμε πρακτικά πώς διαφοροποιούνται ένας ασθενής με ρίσκο θνητότητας π.χ 49% κι ένας άλλος με ρίσκο θνητότητας 51%. Τα μοντέλα που φτιάχαμε κάνουν δυαδική ταξινόμηση με όριο απόφασης το 50% και συνεπώς ο ασθενής με ρίσκο θνητότητας 49% θα εκτιμηθεί ως ζωντανός, ενώ εκείνος με 51% θα εκτιμηθεί ως νεκρός, αλλά στην εικόνα τους δεν ξέρουμε πώς διαφοροποιούνται.

Σχετικά με την πρακτική χρήση των μοντέλων αυτών, σε καμία περίπτωση δεν

εγγυάται κανείς ότι θα κάνουν καλύτερες προβλέψεις από έναν γιατρό, καθώς δεν έχουμε τέτοια δεδομένα. Θα μπορούσαν βέβαια να ελεγχθούν οι προβλέψεις των μοντέλων με τις προβλέψεις των γιατρών σε πραγματικές συνθήκες, όχι για την συμβούλη των γιατρών από τα μοντέλα, αλλά για την εκτίμηση του ποιος από τους δύο κάνει καλύτερες προβλέψεις. Με άλλα λόγια, αν ένα νοσοκομείο έχει καταγραφή των προβλέψεων των γιατρών για το ρίσκο θνητότητας του κάθε ασθενή (έστω σε δυαδικό επίπεδο) και συγκριθεί αυτό με τις προβλέψεις των μοντέλων μας θα μπορέσουμε να πάρουμε μια εικόνα του πώς συγκρίνεται η απόδοσή τους σε σχέση με την εμπειρία των γιατρών. Μέχρι να γίνουν τέτοιοι έλεγχοι ωστόσο, δεν συνίσταται η χρήση των μοντέλων αυτών σε πραγματικές συνθήκες.

7.2 ΠΡΟΕΚΤΑΣΕΙΣ

Με σκοπό την κατασκευή καλύτερων μοντέλων για την επίλυση του προβλήματος πρόβλεψης του ρίσκου θνητότητας για νοσούντες της COVID-19 κι εξερεύνησης νέων προσεγγίσεων του προβλήματος, παρατίθενται παρακάτω μερικές ιδέες που θα μπορούσαν να δοκιμαστούν είτε ανεξάρτητα είτε συνδυαστικά μεταξύ τους:

- 1. Χρήση συσσωρευτικών δεδομένων:** Μια νέα κατεύθυνση που θα μπορούσε να δωθεί στη λύση του συγκεκριμένου προβλήματος θα ήταν η χρήση δεδομένων τα οποία συλλέγονται μέσα σε ένα διάστημα ημερών για τον κάθε ασθενή. Συγκεκριμένα, αν πάρουμε ως παράδειγμα το χαρακτηριστικό Temperature Value, θα μπορούσε για κάθε ασθενή να γίνεται μέτρηση του συγκεκριμένου μεγέθους καθημερινά για ένα εύλογο διάστημα ημερών, όπως μια εβδομάδα κι αξιοποίηση των επιμέρους μετρήσεων. Χρησιμοποιώντας αυτήν την προσέγγιση θα μπορούσε κάθε ασθενής να αντιπροσωπεύεται πλέον από κατανομές χαρακτηριστικών, αντί για μεμονωμένες τιμές. Κατά αυτόν τον τρόπο, θα είχαμε ουσιαστικά μια εποπτεία της πορείας της νόσου του κάθε ασθενή. Έτσι, θα μπορούσε η πορεία της νόσου ενός νέου ασθενή να συγκριθεί με τις πορείες των υπολοίπων και να μπορέσει να γίνει μια καλύτερη εκτίμηση για το αν ο συγκεκριμένος ασθενής έχει υψηλό ή χαμηλό ρίσκο θνητότητας. Πάνω σε αυτή τη λογική, θα μπορούσαν να χρησιμοποιηθούν κατανομές πιθανοτήτων για το κάθε χαρακτηριστικό σε συνδυασμό με κινούμενες μέσες τιμές και ρυθμοί αύξησης ή μείωσης των τιμών των χαρακτηριστικών ανά ημέρα. Έτσι, για κάθε χαρακτηριστικό θα μπορούσαμε να δημιουργήσουμε επιμέρους χαρακτηριστικά, καθένα εκ των οποίων θα έδινε μια νέα πληροφορία για το πρόβλημα. Πάνω σε αυτή τη βάση εκτιμώ ότι θα παίρναμε αρκετά καλύτερα αποτελέσματα. Τα μειονέκτηματα αυτής της μεθόδου έιναι ότι κάνει την υλοποίηση αρκετά πιο σύνθετη, απαιτεί μεγάλη αξιοπιστία, συνέπεια και λεπτομέρεια στα δεδομένα καθώς κι επίσης δεν θα μπορεί να δώσει γρήγορες εκτιμήσεις για την μελλοντική κατάσταση των ασθενών, πράγμα που δεν θα εξυπηρετούσε άμεσα το σκοπό των γρήγορων αποφάσεων επιλογής ασθενών για νοσηλεία.
- 2. Δημιουργία νέων χαρακτηριστικών:** Μια άλλη προσέγγιση θα η δημιουργία

ΚΕΦΑΛΑΙΟ 7. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΠΡΟΕΚΤΑΣΕΙΣ

καινούργιων χαρακτηριστικών μέσα από τα ήδη υπάρχοντα. Ένα παράδειγμα που παρουσιάστηκε για τη δημιουργία καινούργιου χαρακτηριστικού ήταν το χαρακτηριστικό Severity, το οποίο ορίστηκε ως μια δυαδική μεταβλητή που αντιστοιχούσε στο αν κάποιος ασθενής εισήχθηκε σε μονάδα εντατικής θεραπείας ή όχι. Θα μπορούσαν να κατασκευαστούν κι άλλα αντίστοιχα χαρακτηριστικά, καθώς και να δημιουργηθεί το χαρακτηριστικό Severity με διαφορετικό τρόπο.

3. **Συλλογή περισσότερων δεδομένων:** Ένας σημαντικός παράγοντας που θα βελτίωνε τα αποτελέσματα των μοντέλων θα ήταν η ύπαρξη περισσότερων δεδομένων τόσο από πλευράς ασθενών, όσο κι από πλευράς χαρακτηριστικών. Η χρήση μεγαλύτερου αριθμού ασθενών θα μας έδινε τη δυνατότητα να φτιάξουμε πιο αξιόπιστα μοντέλα τα οποία θα χρησιμοποιούσαν ένα πιο αντιπροσωπευτικό δείγμα του πληθυσμού που μελετάμε. Επίσης, με την ύπαρξη περισσότερων χαρακτηριστικών, όπως τιμές άλλων μεγεθών από αιματολογικές εξετάσεις θα μπορούσε να γίνει μια διερεύνηση των χαρακτηριστικών που φαίνεται να έχουν μεγαλύτερη επίδραση στις προβλέψεις του μοντέλου κι έπειτα να κρατηθούν εκείνα τα οποία παρουσιάζουν λιγότερη ετεροσυγέτιση με τα υπόλοιπα, αλλά κι αποτελούν τα εν τέλει πιο σημαντικά. Με τον τρόπο αυτό θα έχουμε στην ουσία καλύτερη "πρώτη ύλη" με την οποία κατασκευάζουμε τα μοντέλα μας.
4. **Εισαγωγή τοπικά περιγραφικών χαρακτηριστικών:** Όπως έχει αναφερθεί ήδη, πολλές φορές οι γιατροί των νοσοκομείων πρέπει να αποφασίσουν μεταξύ ασθενών που πρόκειται να νοσηλευθούν με βάση το ρίσκο θνητότητας του καθενός όταν το νοσοκομεία βρίσκονται σε υψηλό φόρτο. Αυτό όπως είναι αναμενόμενο έχει άμεση επίδραση στο ποιοι ασθενείς πεθαίνουν και ποιοι όχι. Είναι λογικό να φανταστούμε ότι πολλές φορές σε μια τέτοια περίπτωση δυαδικής επιλογής ασθενών, θα μπορούσαν να ζήσουν κι οι 2 ασθενείς, αλλά να έζησε μόνο ο ένας γιατί ο άλλος δεν μπόρεσε να νοσηλευτεί. Αυτή είναι μια μεταβλητή η οποία δεν παρουσιάζεται στα δεδομένα που μελετάμε κι έχει άμεση επίπτωση στις προβλέψεις μας. Έτσι λοιπόν, μια καλή ιδέα θα ήταν να υπάρχει ενα χαρακτηριστικό το οποίο θα περιγράφει τον νοσοκομειακό φόρτο κατά τη νοσηλεία του ασθενή, καθώς και την ύπαρξη εξοπλισμού αναπνευστικής υποστήριξης και μονάδων εντατικής θεραπείας. Είναι σημαντικό να τονίσουμε ότι ο νοσοκομειακός φόρτος δεν συνδέεται μονάχα με την δυνατότητα νοσηλείας ενός ασθενή, αλλά και με την ίδια την ψυχοσύνθεση και την ικανότητα των μελών του νοσοκομείου να μπορέσουν να ανταπεξέλθουν όσο πιο αποτελεσματικά μπορούν στις προκλήσεις που τους παρουσιάζονται. Πέραν του εξοπλισμού δηλαδή, η ζωή ενός ασθενή εξαρτάται από την ικανότητα του γιατρού που τον περιθάλπτει, που η ικανότητα σε καταστάσεις εξουθενωτικής πίεσης είναι λογικό να φθίνει.
5. **Εισαγωγή χαρακτηριστικών δυνατών θεραπειών:** Όπως είναι λογικό, η θνητότητα μιας νόσου έχει άμεση σχέση με την ικανότητα μας να την αντιμετωπίζουμε αποτελεσματικά, πράγμα που σχετίζεται με την ύπαρξη εξειδικευμένων φαρμάκων ή εμβολίων, ειδικού ιατρικού εξοπλισμού, κλπ. Συνεπώς, όταν κατασκευάζουμε ένα μοντέλο για την εκτίμηση του ρίσκου θνητότητας

ενός ασθενή της COVID-19, για να έχει πρακτική ισχύ στον χρόνο, θα πρέπει να είναι γνωστό κατά πόσο οι νέοι εξεταζόμενοι ασθενείς από το μοντέλο μας, βρίσκονται σε περίοδο ύπαρξης εμβολίου ή φαρμακευτικής αγωγής. Είναι σαφές ότι σε τέτοιες περιπτώσεις οι προβλέψεις των μοντέλων που έχουν εκπαιδευτεί σε προγενέστερες εποχές δύναται να έχουν μεγάλα σφάλματα.

6. **Δυναμική αυτοματοποιημένη κατασκευή μοντέλων:** Μια άλλη προσέγγιση θα ήταν η δημιουργία νέων μοντέλων με αυτοματοποιημένο τρόπο ανά χρονικά διαστήματα. Αυτή η προσέγγιση προσπαθεί να αντιμετωπίσει την δυναμική συμπεριφορά του φαινομένου που εξετάζουμε. Συγκεκριμένα, η διαθεσιμότητα σε φάρμακα, εμβόλια, ιατρικό προσωπικό, ο νοσοκομειακός φόρτος, η γνώση αντιμετώπισης της νόσου, η εμπειρία του προσωπικού, κ.ο.κ. είναι παράγοντες οι οποίες αλλάζουν συνεχώς κι έχουν άμεση επίδραση στο ρίσκο θνητότητας ενός ασθενή. Επειδή ακριβώς οι παράμετροι αυτοί είναι πολλές σε αριθμό, πολλές εκ των οποίων άγνωστες, άλλες μη μετρήσιμες, κ.ο.κ. θα μπορούσαμε να κατασκευάζουμε μοντέλα τα οποία θα εκπαιδεύονται πάνω σε δεδομένα συγκεκριμένης χρονικής περιόδου και θα ανανεώνονται με τη χρήση κάποιου χρονικού παραθύρου ανά διαστήματα. Για παράδειγμα, έστω ότι έχουμε δεδομένα 30 ημερών και κατασκευάζουμε το πρώτο μας μοντέλο. Το μοντέλο μας μπορεί να αντικατασταθεί την 40stή μέρα χρησιμοποιώντας τα δεδομένα της 10ης έως της 40stής μέρας για να κατασκευαστεί, κ.ο.κ. Με τον τρόπο αυτό πιθανόν να αντιμετωπίζαμε τη δυναμικότητα του προβλήματος και να είχαμε μια καλύτερη χρονική αξιοπιστία. Για την επίτευξη χωρικής αξιοπιστίας θα έπρεπε τα μοντέλα αυτά να κατασκευάζονται από τοπικά δεδομένα μόνο, δηλαδή από συγκεκριμένο νοσοκομείο ή πόλη ή χώρα. Έτσι θα λαμβάναμε επίσης υπόψην και τις ενδοπληθυσμιακές πιθανές διαφορές οι οποίες δεν είναι προφανείς στην πορεία μιας ασθένειας.

Βιβλιογραφία

- [1] Dimitris Bertsimas, Galit Lukin, Luca Mingardi, Omid Nohadani, Agni Orfanoudaki, Bartolomeo Stellato, Holly Wiberg, Sara Gonzalez-Garcia, Carlos Luis Parra-Calderon, Kenneth Robinson, et al. “*COVID-19 mortality risk assessment: An international multi-center study*“. PloS one, 15(12):e0243262, 2020.
- [2] Tom Mitchell. “*Machine learning*“. 1997.
- [3] Pat Langley. “*The changing science of machine learning*“, 2011.
- [4] Dr Michael J Garbade. “*Clearing the confusion: AI vs Machine learning vs Deep learning Differences*“. Towards Data Science, 14, 2018.
- [5] Eda Kavlakoglu. “*AI vs machine learning vs deep learning vs neural networks: what’s the difference*“. IBM. <https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deeplearning-vs-neural-networks> (accessed 10 January 2021). This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI, 10, 2020.
- [6] H IJ. “*Statistics versus machine learning*“. Nature methods, 15(4):233, 2018.
- [7] Bing Liu. “*Supervised learning*“. In “*Web data mining*“, pages 63–132. Springer, 2011.
- [8] Stuart Russell and Peter Norvig. “*Artificial intelligence: a modern approach*“. 2002.
- [9] Alexandru Niculescu-Mizil and Rich Caruana. “*Predicting good probabilities with supervised learning*“. In “*Proceedings of the 22nd international conference on Machine learning*“, pages 625–632, 2005.
- [10] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. “*Overview of supervised learning*“. In “*The elements of statistical learning*“, pages 9–41. Springer, 2009.
- [11] Thais Mayumi Oshiro, Pedro Santoro Perez, and José Augusto Baranauskas. “*How many trees in a random forest?*“. In “*International workshop on machine learning and data mining in pattern recognition*“, pages 154–168. Springer, 2012.
- [12] Gérard Biau and Erwan Scornet. “*A random forest guided tour*“. Test, 25(2):197–227, 2016.

- [13] Erica Briscoe and Jacob Feldman. “*Conceptual complexity and the bias/variance tradeoff*”. *Cognition*, 118(1):2–16, 2011.
- [14] Jerome H Friedman. “*Greedy function approximation: a gradient boosting machine*”. *Annals of statistics*, pages 1189–1232, 2001.
- [15] Juliana Tolles and William J Meurer. “*Logistic regression: relating patient characteristics to outcomes*”. *Jama*, 316(5):533–534, 2016.
- [16] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, et al. “*Xgboost: extreme gradient boosting*”. R package version 0.4-2, 1(4): 1–4, 2015.
- [17] Tianqi Chen and Carlos Guestrin. “*Xgboost: A scalable tree boosting system*”. In “*Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*”, pages 785–794, 2016.
- [18] Jason Brownlee. “*XGBoost With Python: Gradient Boosted Trees with XGBoost and Scikit-Learn*”. Machine Learning Mastery, 2016.
- [19] Tianqi Chen, Tong He, Michael Benesty, and Vadim Khotilovich. “*Package ‘xgboost’*”. R version, 90, 2019.
- [20] Didrik Nielsen. “*Tree boosting with xgboost—why does xgboost win every machine learning competition?*”. Master’s thesis, NTNU, 2016.
- [21] Sergios Theodoridis. “*Chapter 7 - Classification: a Tour of the Classics*”. In Sergios Theodoridis, editor, “*Machine Learning (Second Edition)*”, pages 301–350. Academic Press, second edition edition, 2020. ISBN 978-0-12-818803-3.
- [22] Wei-Yin Loh. “*Classification and regression trees*”. Wiley interdisciplinary reviews: data mining and knowledge discovery, 1(1):14–23, 2011.
- [23] Gretchen G Moisen. “*Classification and regression trees*”. In: Jørgensen, Sven Erik; Fath, Brian D. (Editor-in-Chief). Encyclopedia of Ecology, volume 1. Oxford, UK: Elsevier. p. 582–588., pages 582–588, 2008.
- [24] Jin Huang and Charles X Ling. “*Using AUC and accuracy in evaluating learning algorithms*”. *IEEE Transactions on knowledge and Data Engineering*, 17(3): 299–310, 2005.
- [25] Kendrick Boyd, Kevin H Eng, and C David Page. “*Area under the precision-recall curve: point estimates and confidence intervals*”. In “*Joint European conference on machine learning and knowledge discovery in databases*”, pages 451–466. Springer, 2013.
- [26] Vicente García, Ramón Alberto Mollineda, and José Salvador Sánchez. “*Index of balanced accuracy: A performance measure for skewed class distributions*”. In “*Iberian conference on pattern recognition and image analysis*”, pages 441–448. Springer, 2009.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- [27] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. “*Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric*“. PloS one, 12(6): e0177678, 2017.
- [28] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. “*SMOTE: synthetic minority over-sampling technique*“. Journal of artificial intelligence research, 16:321–357, 2002.
- [29] Alberto Fernández, Salvador García, Francisco Herrera, and Nitesh V Chawla. “*SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary*“. Journal of artificial intelligence research, 61:863–905, 2018.
- [30] Leif E Peterson. “*K-nearest neighbor*“. Scholarpedia, 4(2):1883, 2009.
- [31] Michael W Browne. “*Cross-validation methods*“. Journal of mathematical psychology, 44(1):108–132, 2000.
- [32] Davide Anguita, Luca Ghelardoni, Alessandro Ghio, Luca Oneto, and Sandro Ridella. “*The ‘K’in K-fold cross validation*“. In “*20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*“, pages 441–446. i6doc. com publ, 2012.
- [33] Tadayoshi Fushiki. “*Estimation of prediction error by using K-fold cross-validation*“. Statistics and Computing, 21(2):137–146, 2011.