ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Τμήμα Ψηφιακών Συστημάτων

Π.Μ.Σ. Μεγάλα Δεδομένα και Αναλυτική

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Πρόβλεψη επιπλοκών του εμφράγματος μυοκαρδίου με τη χρήση τεχνικών μηχανικής μάθησης

Καραβάς Μιχάλης

Επιβλέπων καθηγητής
Φιλιππάκης Μιχάλης

Πειραιάς, Ιούνιος 2022

Περιεχόμενα

1. ΕΙΣΑΓΩΓΗ

Το έμφραγμα μυοκαρδίου γνωστό και ως καρδιακή προσβολή αποτελεί την κύρια αιτία θανάτου παγκοσμίως και ευθύνεται για το 42% περίπου των συνολικών θανάτων από καρδιαγγειακή νόσο. Το έμφραγμα προκαλείται όταν οι περιοχές της καρδιάς δεν αιματώνονται σωστά, δε λαμβάνουν δηλαδή όσο αίμα απαιτείται προκειμένου να διεξάγεται ομαλά η λειτουργία της.[2] Η συχνότητα της ασθένειας (νοσηρότητα) παγκοσμίως είναι 195,3/100.000 για τους άντρες και 115/100.000 για τις γυναίκες, με τα συγκεκριμένα ποσοστά να μειώνονται κυρίως στις ανεπτυγμένες χώρες. Επίσης τις τελευταίες δεκαετίες, μειώθηκε και η θνητότητα και αρκετές από τις επιπλοκές που συνδέονται με το έμφραγμα του μυοκαρδίου. Η βελτίωση οφείλεται στην καλύτερη θεραπευτική αντιμετώπιση.

Η μηχανική μάθηση από την άλλη, είναι ένα αποτελεσματικό εργαλείο για την ανάλυση πολυσύνθετων ιατρικών συνόλων δεδομένων. Καθώς ο ρυθμός συλλογής τέτοιου είδους δεδομένων ολοένα και αυξάνεται γεννάται η δυνατότητα εφαρμογής τεχνικών μηχανικής μάθησης πάνω στα ιατρικά δεδομένα ως ένα ακόμα επικουρικό μέσο στα χέρια της ιατρικής επιστήμης.

Η μηχανική μάθηση υποβοηθά τον υπολογιστή να προσαρμόζεται σε νέες συνθήκες και να εξάγει συμπεράσματα βασισμένα σε ήδη υπάρχουσες παρατηρήσεις. Ένα μοντέλο θεωρείται ότι μαθαίνει όταν η μετρική αξιολόγησης του που έχουμε επιλέξει, βελτιώνεται καθώς του παρουσιάζονται περισσότερα παραδείγματα εκπαίδευσης. Στον τομέα της ιατρικής τα μοντέλα μηχανικής μάθησης δύναται να τροφοδοτηθούν με δεδομένα που μπορεί να προέρχονται από την πρωτογενή εξέταση ενός γιατρού έως ακτινογραφίες και μαγνητικές τομογραφίες. Το μοντέλο στη συνέχεια, καθώς του παρουσιάζονται τα παραδείγματα εκπαίδευσης, εξάγει γνώση αξιοποιώντας τα σημαντικότερα από τα χαρακτηριστικά του συνόλου δεδομένων. [10]

Η μηχανική μάθηση περιλαμβάνει μοντέλα επιβλεπόμενης και μη επιβλεπόμενης μάθησης. Η διαφοροποίηση τους έγκειται στην ύπαρξη ή μη ανατροφοδότησης της επιτυχίας ή αποτυχίας των προβλέψεων. Στην επιβλεπόμενη μάθηση το μοντέλο τροφοδοτείται με ένα σύνολο παραδειγμάτων και καλείται να εκπαιδευτεί ώστε να προβλέπει μια μεταβλητή στόχο. Μοντέλα επιβλεπόμενης μάθησης θεωρούνται στην πλειονότητα τους τα μοντέλα κατηγοριοποίησης (classification) και παλινδρόμησης (regression), ενώ μη επιβλεπόμενης μάθησης θεωρούνται οι αλγόριθμοι συσταδοποίησης (clustering) όπου τα μοντέλα καλούνται από μόνα τους να διακρίνουν τις διαφορετικές κλάσεις.

Τα ιατρικά σύνολα δεδομένων παρουσιάζουν ως επί το πλείστον την ιδιομορφία να μην είναι οι κλάσεις ενδιαφέροντος ισομερώς κατανεμημένες. Συνηθίζεται δηλαδή η αρνητική κλάση να είναι αρκετά πολυπληθέστερη από τη θετική. Αυτό δεν είναι αξιοπερίεργο καθώς αν σκεφτούμε ότι η κλάση ενδιαφέροντος είναι η

κατάληξη ή μη ενός ασθενούς, ή αν ένα δείγμα είναι θετικό ή αρνητικό σε εξέταση για νεοπλασία τότε στις περισσότερες περιπτώσεις το αποτέλεσμα θα είναι η μη κατάληξη του ασθενούς και η αρνητική απάντηση για τη νεοπλασία.

Αυτές οι συνθήκες του πραγματικού κόσμου δημιουργούν προσκόμματα στην ικανότητα των αλγορίθμων μηχανικής μάθησης να εκπαιδεύονται σωστά και να προβαίνουν σε ορθές προβλέψεις καθώς δεν υπάρχουν αρκετά παραδείγματα από τη θετική κλάση ώστε να εκπαιδευτούν πάνω σε αυτά. Παρ ’όλα αυτά έχουν αναπτυχθεί εξειδικευμένες τεχνικές οι οποίες θα παρουσιαστούν στη συγκεκριμένη εργασία ώστε κατά το δυνατόν να ξεπερνούνται αυτές οι δυσκολίες.

Στην παρούσα εργασία θα παρουσιαστεί μια πληθώρα τεχνικών μηχανικής μάθησης προκειμένου να προβλεφθεί η εξέλιξη της υγείας ασθενών που έχουν υποστεί έμφραγμα του μυοκαρδίου.

2. Το σύνολο δεδομένων

Το σύνολο δεδομένων [1] πάνω στο οποίο βασίστηκε η παρούσα εργασία έχει συλλεχθεί από το νοσοκομείο του Krasnoyarsk της Ρωσίας κατά τη χρονική περίοδο 1992-1995 και βρίσκεται διαθέσιμο στο αποθετήριο UCI στη διεύθυνση https://archive.ics.uci.edu/ml/datasets/Myocardial+infarction+complications.

2.1 Περιγραφή συνόλου δεδομένων

Το σύνολο δεδομένων αποτελείται από 1700 εγγραφές σε 124 στήλες οι οποίες περιλαμβάνουν τόσο συνεχή όσο και κατηγορικά χαρακτηριστικά.

1	Αναγνωριστικό ασθενούς	17	Ιστορικό εμμένουσας κολπικής μαρμαρυγής
2	Ηλικία	18	Ιστορικό κοιλιακής μαρμαρυγής
3	Φύλο	19	Ιστορικό παροξυσμικής κοιλιακής μαρμαρυγής
4	Πλήθος εμφραγμάτων στο ιστορικό	20	Ιστορικό κολποκοιλιακού αποκλεισμού 1ου βαθμού
5	Εμφάνιση στηθάγχης (χωρισμένη σε χρονικές κλάσεις)	21	Ιστορικό κολποκοιλιακού αποκλεισμού 3ου βαθμού
6	Κατηγορία στηθάγχης με βάση το τελευταίο έτος	22	Ιστορικό αποκλεισμού πρόσθιου αριστερού σκέλους
7	Στεφανιαία νόσος	23	Ιστορικό μερικού αποκλεισμού αριστερού σκέλους
8	Κληρονομικότητα στεφανιαίας νόσου	24	Ιστορικό πλήρους αποκλεισμού αριστερού σκέλους
9	Κατηγορία υπέρτασης μη οφειλόμενης σε παθολογικά αίτια	25	Ιστορικό μερικού αποκλεισμού δεξιού σκέλους
10	Υπέρταση οφειλόμενη σε παθολογικά αίτια	26	Ιστορικό πλήρους αποκλεισμού δεξιού σκέλους
11	Διάρκεια υπέρτασης (χωρισμένη σε χρονικές κλάσεις)	27	Ιστορικό διαβήτη
12	Ιστορικό καρδιακής ανεπάρκειας (χωρισμένη σε κλάσεις ανάλογα με την καρδιακή κοιλία)	28	Ιστορικό παχυσαρκίας
13	Ιστορικό αρρυθμίας	29	Ιστορικό θυρεοτοξίκωσης
14	Ιστορικό πρόωρης κολπικής συστολής	30	Ιστορικό χρόνιας βρογχίτιδας
15	Ιστορικό πρόωρης κοιλιακής συστολής	31	Ιστορικό αποφρακτικής χρόνιας βρογχίτιδας
16	Ιστορικό παροξυσμικής κολπικής μαρμαρυγής	32	Βρογχικό άσθμα στο ιστορικό

33	Ιστορικό χρόνιας πνευμονίας ιστορικό	49	Ύπαρξη δεξιού κοιλιακού εμφράγματος
34	Ιστορικό φυματίωσης	50	Φυσιολογική ένδειξη ΗΚΓ κατά την εισαγωγή στο νοσοκομείο
35	Συστολική πίεση κατά την εισαγωγή στα επείγοντα καρδιολογικά περιστατικά	51	Ένδειξη ΗΚΓ για κολπική μαρμαρυγή κατά την εισαγωγή στο νοσοκομείο
36	Διαστολική πίεση κατά την εισαγωγή στα επείγοντα καρδιολογικά περιστατικά	52	Ένδειξη ΗΚΓ για κολπική αρρυθμία κατά την εισαγωγή στο νοσοκομείο
37	Συστολική πίεση στην εντατική	53	Ύπαρξη ιδιοκοιλιακού ρυθμού στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο
38	Διαστολική πίεση στην εντατική	54	Φυσιολογική ένδειξη ΗΚΓ κατά την εισαγωγή στο νοσοκομείο με ρυθμό >90 παλμούς/λεπτό
39	Πνευμονικό οίδημα κατά την εισαγωγή στην εντατική	55	Φυσιολογική ένδειξη ΗΚΓ κατά την εισαγωγή στο νοσοκομείο με ρυθμό <60 παλμούς/λεπτό
40	Καρδιογενές σοκ κατά την εισαγωγή στην εντατική	56	Πρώιμες κολπικές έκτακτες συστολές στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο
41	Παροξυσμική κολπική μαρμαρυγή κατά την εισαγωγή στην εντατική	57	Συχνές πρώιμες κολπικές έκτακτες συστολές στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο
42	Υπερκοιλιακή ταχυκαρδία κατά την εισαγωγή στην εντατική	58	Πρώιμες κοιλιακές έκτακτες συστολές στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο
43	Κοιλιακή ταχυκαρδία κατά την εισαγωγή στην εντατική	59	Συχνές πρώιμες κοιλιακές έκτακτες συστολές στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο
44	Κοιλιακή μαρμαρυγή κατά την εισαγωγή στην εντατική	60	Παροξυσμική κολπική μαρμαρυγή στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο
45	Ύπαρξη πρόσθιου εμφράγματος μυοκαρδίου	61	Εμμένουσα κολπική μαρμαρυγή στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο
46	Ύπαρξη πλευρικού εμφράγματος μυοκαρδίου	62	Παροξυσμική υπερκοιλιακή ταχυκαρδία στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο
47	Ύπαρξη κατώτερου εμφράγματος μυοκαρδίου	63	Παροξυσμική κοιλιακή ταχυκαρδία στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο
48	Ύπαρξη οπίσθιου εμφράγματος μυοκαρδίου	64	Κοιλιακή ταχυκαρδία στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο

65	Φλεβοκομβοκολπικός αποκλεισμός στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο	80	Ινωδολυτική θεραπεία με Celiasum 500k IU
66	Κολποκοιλιακός αποκλεισμός 1ου βαθμού στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο	81	Ινωδολυτική θεραπεία με Celiasum 250k IU
67	Κολποκοιλιακός αποκλεισμός 2ου βαθμού-1ου τύπου στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο	82	Ινωδολυτική θεραπεία με Streptodecase 1.5m IU
68	Κολποκοιλιακός αποκλεισμός 2ου βαθμού-2ου τύπου στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο	83	Υποκαλιαιμία (< 4 mmol/L)
69	Κολποκοιλιακός αποκλεισμός 3ου βαθμού στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο	84	Κάλιο ορού αίματος
70	Αποκλεισμός πρόσθιου αριστερού σκέλους στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο	85	Αύξηση του νατρίου στον ορό αίματος >150 mmol/L
71	Αποκλεισμός οπίσθιου αριστερού σκέλους στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο	86	Νάτριο ορού αίματος
72	Μερικός αποκλεισμός αριστερού σκέλους στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο	87	AlAT ορού αίματος
73	Πλήρης αποκλεισμός αριστερού σκέλους στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο	88	AsAT ορού αίματος
74	Μερικός αποκλεισμός δεξιού σκέλους στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο	89	CPK ορού αίματος
75	Πλήρης αποκλεισμός δεξιού σκέλους στο ΗΚΓ κατά την εισαγωγή στο νοσοκομείο	90	Λευκά αιμοσφαίρια
76	Ινωδολυτική θεραπεία με Celiasum 750k IU	91	Ταχύτητα Καθίζησης Ερυθρών
77	Ινωδολυτική θεραπεία με Celiasum 1m IU	92	Χρόνος που μεσολάβησε από την έναρξη του επεισοδίου ως τη μεταφορά στο νοσοκομείο (χωρισμένη σε χρονικές κλάσεις)
78	Ινωδολυτική θεραπεία με Celiasum 3m IU	93	Επανεμφάνιση πόνων κατά τις πρώτες ώρες νοσηλείας
79	Ινωδολυτική θεραπεία με Streptase	94	Επανεμφάνιση πόνων κατά τη 2η ημέρα νοσηλείας

95	Επανεμφάνιση πόνων κατά την 3η μέρα νοσηλείας	110	Χορήγηση ακετυλοσαλικυλικού οξέως στην εντατική
96	Χορήγηση οπιούχων φαρμάκων στα επείγοντα καρδιολογικά περιστατικά	111	Χορήγηση Ticlid στην εντατική
97	Χορήγηση μη στεροειδών αντιφλεγμονώδων στα επείγοντα καρδιολογικά περιστατικά	112	Χορήγηση Trental στην εντατική
98	Χορήγηση λιδοκαίνης στα επείγοντα καρδιολογικά περιστατικά	113	Κολπική μαρμαρυγή (ως παρενέργεια)
99	Χορήγηση υγρών νιτρικών στην εντατική	114	Υπερκοιλιακή ταχυκαρδία (ως παρενέργεια)
100	Χορήγηση οπιούχων φαρμάκων στην εντατική κατά τις πρώτες ώρες νοσηλείας	115	Κοιλιακή ταχυκαρδία (ως παρενέργεια)
101	Χορήγηση οπιούχων φαρμάκων στην εντατική κατά τη 2η ημέρα νοσηλείας	116	Κοιλιακή μαρμαρυγή (ως παρενέργεια)
102	Χορήγηση οπιούχων φαρμάκων στην εντατική κατά τη 3η ημέρα νοσηλείας	117	Κολποκοιλιακός αποκλεισμός 3ου βαθμού (ως παρενέργεια)
103	Χορήγηση μη στεροειδών αντιφλεγμονώδων στην εντατική κατά τις πρώτες ώρες νοσηλείας	118	Πνευμονικό οίδημα (ως παρενέργεια)
104	Χορήγηση μη στεροειδών αντιφλεγμονώδων στην εντατική κατά τη 2η ημέρα νοσηλείας	119	Ρήξη μυοκαρδίου (ως παρενέργεια)
105	Χορήγηση μη στεροειδών αντιφλεγμονώδων στην εντατική κατά τη 3η ημέρα νοσηλείας	120	Σύνδρομο Dressler (ως παρενέργεια)
106	Χορήγηση λιδοκαίνης στην εντατική	121	Χρόνια καρδιακή ανεπάρκεια (ως παρενέργεια)
107	Χορήγηση Β- αναστολέων στην εντατική	122	Επανεμφάνιση εμφράγματος (ως παρενέργεια)μυοκαρδίου
108	Χορήγηση αναστολέων διαύλων ασβεστίου στην εντατική	123	Μετεμφραγματική στηθάγχη (ως παρενέργεια)
109	Χορήγηση αντιπηκτικών στην εντατική	124	Αιτία θανάτου (σε περίπτωση κατάληξης του ασθενούς)

Η κωδικοποίηση των κατηγορικών χαρακτηριστικών έγινε από τους δωρητές των δεδομένων.

2.2 Τι είναι η μηχανική μάθηση

Η μηχανική μάθηση συναποτελείται από 4 βασικά στοιχεία. Το πρώτο είναι ο αλγόριθμος μηχανικής μάθησης, το δεύτερο είναι τα δεδομένα με τα οποία αυτός θα τροφοδοτηθεί, τρίτο είναι η μετρική την οποία ο αλγόριθμος καλείται να βελτιστοποιήσει και τέταρτο είναι το μοντέλο το οποίο θα εξαχθεί. [4]

Αν εξεταστεί η μηχανική μάθηση από στατιστική σκοπιά τότε ο αλγόριθμος μηχανικής μάθησης καλείται να μάθει κάποια υποθετική συνάρτηση f τέτοια ώστε

Πρόβλεψη=f(Δεδομένα εισόδου)

όπου τα δεδομένα εισόδου αποτελούν τις ανεξάρτητες μεταβλητές και η πρόβλεψη την εξαρτημένη μεταβλητή.

Όταν λοιπόν εκπαιδευτεί ένας αλγόριθμος πάνω σε ένα σύνολο δεδομένων τότε καταλήγουμε με ένα μοντέλο και διαισθητικά θα μπορούσαμε αυτή τη διαδικασία να την αποτυπώσουμε με την παρακάτω εξίσωση:

Μοντέλο=Αλγόριθμος(Δεδομένα)

2.3 Ορισμός προβλημάτων μηχανικής μάθησης υπό εξέταση

Οι στήλες 2-112 αφορούν ποιοτικά και ποσοτικά χαρακτηριστικά του ασθενούς (φύλο, ηλικία κλπ), το ιατρικό του ιστορικό, κλινικές ενδείξεις που προηγήθηκαν του εμφράγματος (μαρμαρυγές, ταχυκαρδίες κλπ) και την ιατρική περίθαλψη που δέχτηκε ο ασθενής στο νοσοκομείο. Οι στήλες 113-123 αφορούν επιπλοκές που εγκαταστάθηκαν στον ασθενή μετά το έμφραγμα και η στήλη 124 αφορά την κατάληξη ή μη του ασθενούς.

Σύμφωνα με τους δωρητές, το σύνολο δεδομένων υπό εξέταση μπορεί να χρησιμοποιηθεί ως αντικείμενο μηχανικής μάθησης χρησιμοποιώντας τις στήλες 2-112 ως χαρακτηριστικά και κάθε μία από τις υπόλοιπες 12 ως εξαρτημένες μεταβλητές. Κατ' αυτόν τον τρόπο δημιουργείται μια πληθώρα προβλημάτων μηχανικής μάθησης που μπορούν να εξεταστούν, τα οποία γίνονται ακόμα περισσότερα αν εξετάζουμε την πρόβλεψη των παρενεργειών με βάση τις ενδείξεις του ασθενούς κατά την $1^η$,$2^η$ και $3^η$ ημέρα της νοσηλείας.

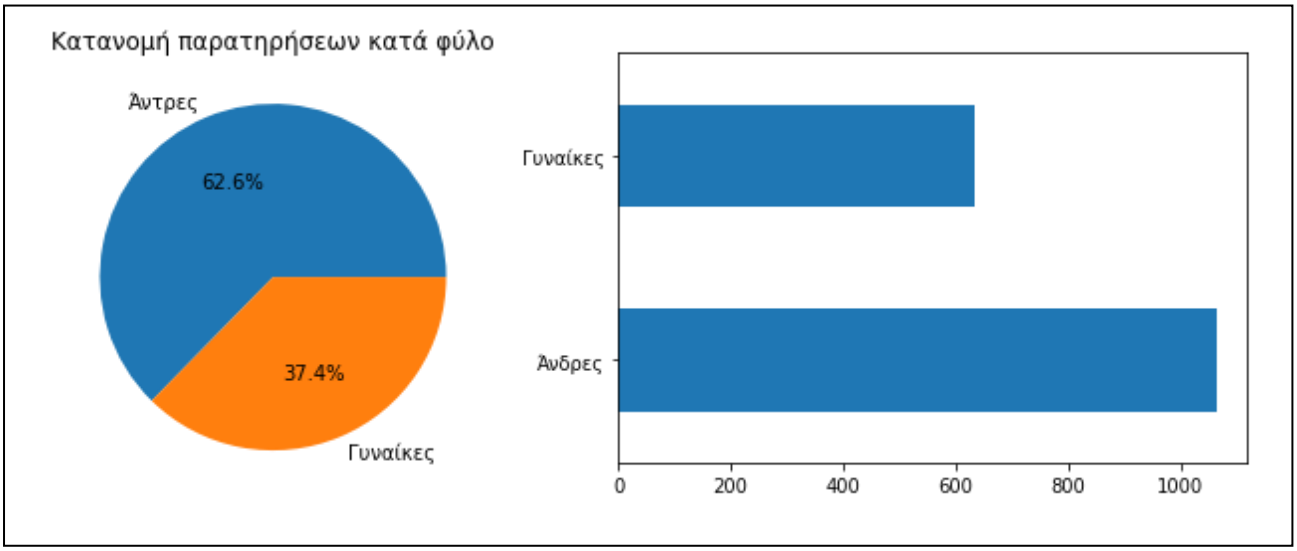
Αρχικά θα εξεταστεί ως εξαρτημένη μεταβλητή η κατάληξη ή μη του ασθενούς με την πρόβλεψη να γίνεται πάνω σε όλα τα διαθέσιμα δεδομένα στήλες 2-123, περιλαμβάνοντας δηλαδή και τις ενδεχόμενες μετεμφραγματικές επιπλοκές, καθώς είναι αυτή είναι η μεταβλητή με τη μεγαλύτερη σημασία από όλες. Στη συνέχεια οι προβλέψεις θα γίνουν πάνω στην ίδια μεταβλητή και με τη μεθοδολογία που προτείνουν οι δωρητές δηλαδή χρησιμοποιώντας διαθέσιμες ενδείξεις κατά την εισαγωγή στο νοσοκομείο, τις διαθέσιμες ενδείξεις ως την $1^η$ μέρα νοσηλείας, τις διαθέσιμες ενδείξεις ως την $2^η$ μέρα νοσηλείας και τις διαθέσιμες ενδείξεις ως την $3^η$ μέρα νοσηλείας. Τα δεδομένα που αφαιρούνται

κάθε φορά στα σενάρια 3,4 και 5 για να δημιουργήσουν ένα νέο σύνολο δεδομένων και ως εκ τούτου ένα νέο πρόβλημα μηχανικής μάθησης αφορούν την επανεμφάνιση των πόνων, τη χορήγηση οπιούχων και τη χορήγηση μη στεροειδών αντιφλεγμονωδών φαρμάκων.

2.4 Ανάλυση δεδομένων (EDA)

- **Φύλο**

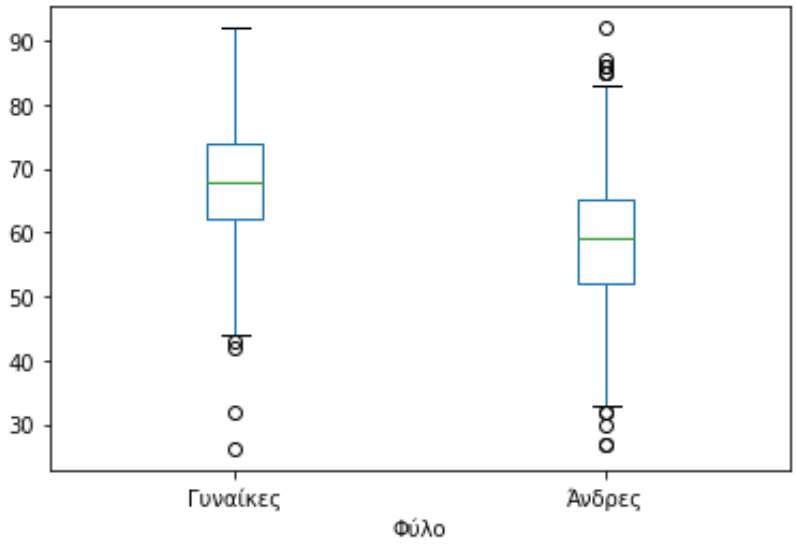
Από τις 1700 εγγραφές οι 1065 αφορούν άντρες και 635 γυναίκες.



Γράφημα 1

- **Ηλικία**

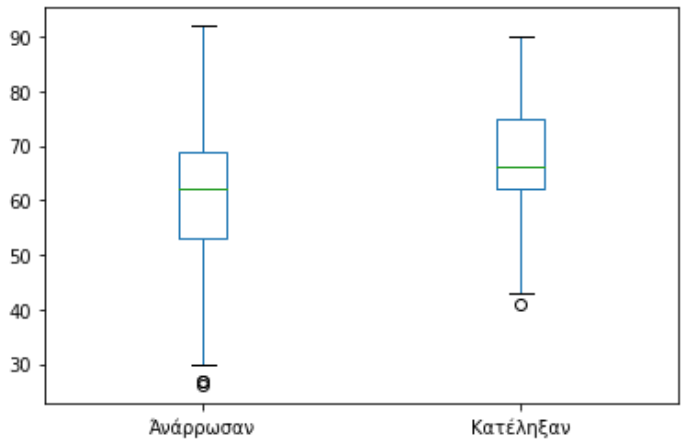
Η μέση ηλικία των ασθενών που περιλαμβάνονται στα δεδομένα είναι 61,8 έτη. Το εύρος κυμαίνεται από τα 26 ως τα 92 έτη. Η μέση ηλικία των ανδρών είναι τα 58,4 έτη με διάμεση τιμή 59 και των γυναικών 67,5 με διάμεση τιμή 68. Έκτροπες τιμές (outliers) παρουσιάζονται και στις 2 περιπτώσεις με τους άντρες να τις έχουν και στα 2 άκρα, ενώ η ηλικιακή κατανομή των γυναικών τις παρουσιάζει μόνο στο κάτω άκρο.



Θηκόγραμμα 1 ηλικιακής κατανομής ανά φύλο

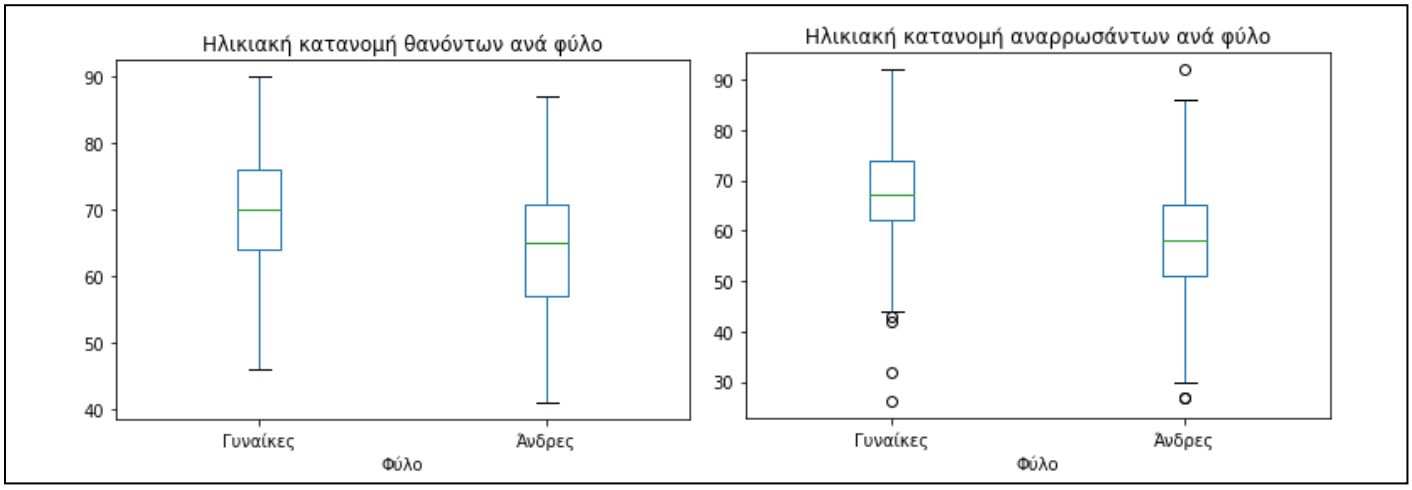
- **Κατάληξη ασθενούς (Μεταβλητή στόχος)**

Η μέση ηλικία θανόντων είναι τα 67,1 με διάμεσο τα 66 έτη ενώ η μέση ηλικία των αναρρωσάντων είναι τα 60,8 έτη με διάμεσο τα 62.



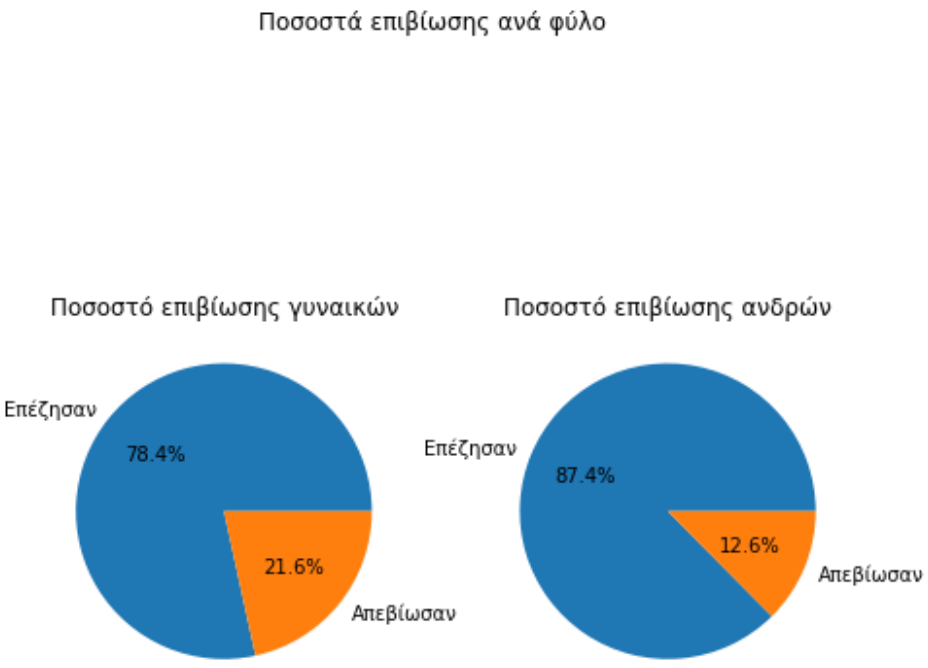
Θηκόγραμμα 2 Ηλικιακή κατανομή ως προς τη μεταβλητή στόχος

Εξετάζοντας δε τη μεταβλητή στόχος ως προς το φύλο έχουμε ότι η μέση ηλικία θανουσών γυναικών είναι τα 69,8 έτη με διάμεσο 70, ενώ η μέση ηλικία των θανόντων αντρών τα 64,3 με διάμεσο 65. Αντίστοιχα για τους αναρρώσαντες προκύπτει ότι για τις μεν γυναίκες η μέση ηλικία τους είναι τα 66,9 έτη με διάμεσο 67, για τους δε άντρες η μέση ηλικία είναι τα 57,5 έτη με διάμεση ηλικία τα 58 έτη.



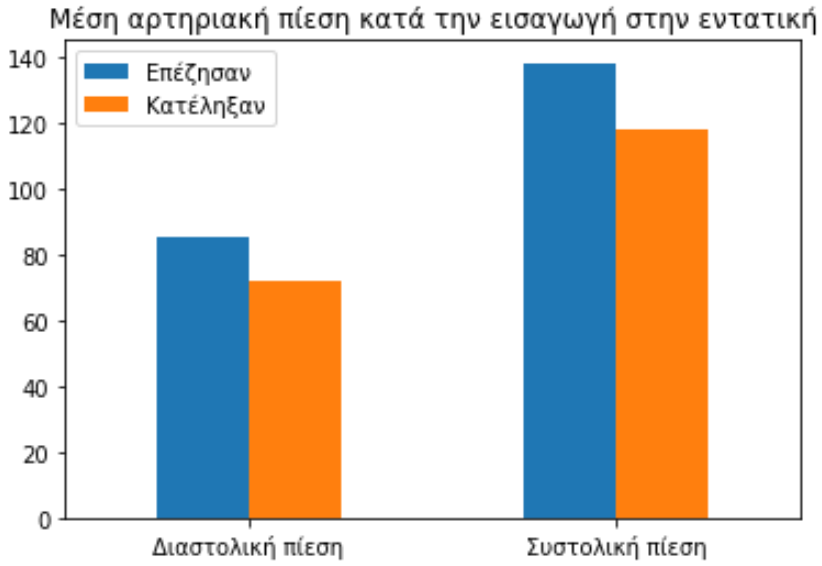
Θανατηφόρο κατάληξη συνολικά είχε το 15,9% (271) των περιστατικών, ενώ το 84,1% (1429) επιβίωσε.

Εξετάζοντας την κατάληξη των ασθενών με βάση το φύλο βλέπουμε πως για τους άνδρες το ποσοστό επιβίωσης ανέρχεται στο 87,4% ενώ για τις γυναίκες στο 78,4%.



Ένα άλλο ενδιαφέρον στοιχείο που προκύπτει είναι ότι κατά την εισαγωγή στη Μονάδα Εντατικής Θεραπείας οι ασθενείς που τελικά απεβίωσαν είχαν χαμηλότερη αρτηριακή πίεση -τόσο διαστολική όσο και συστολική- σε σχέση με

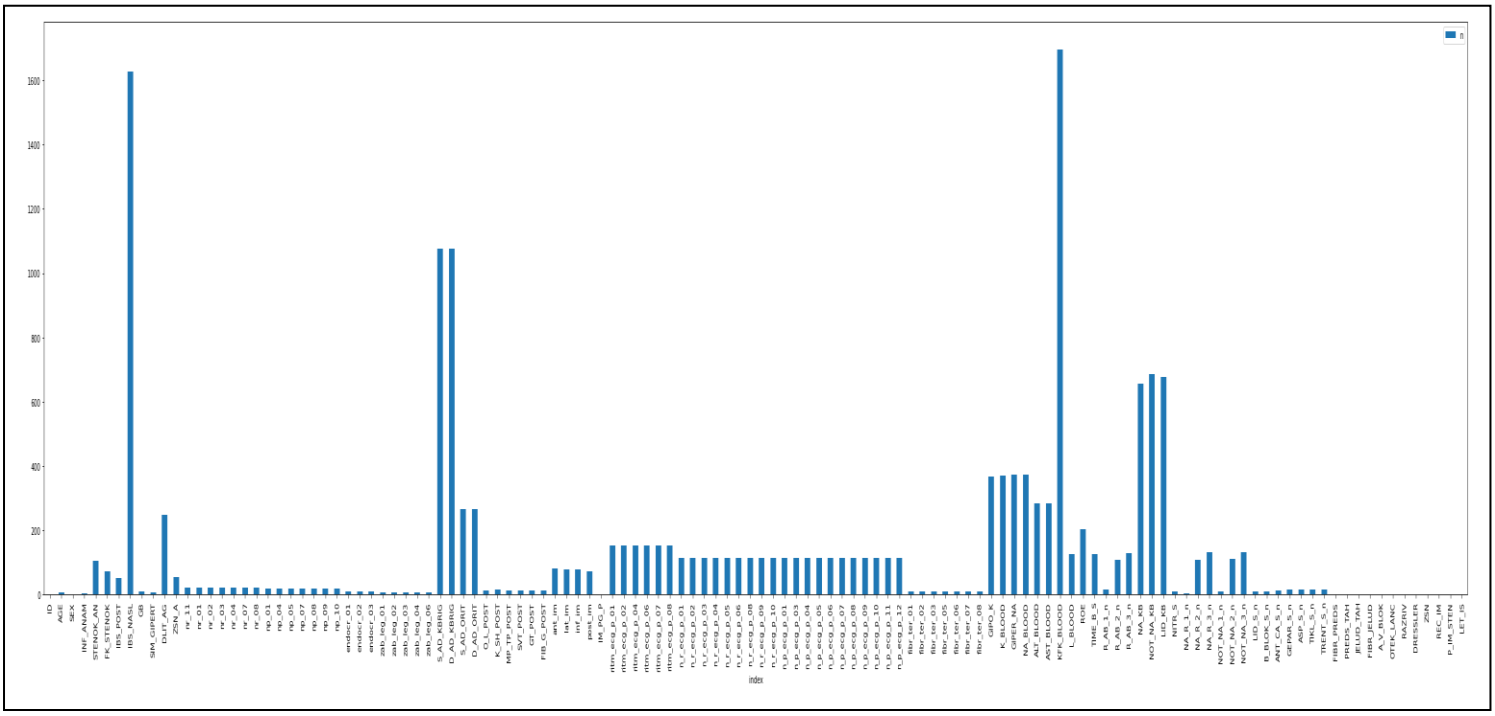
αυτούς που επέζησαν. Πιο συγκεκριμένα οι επιβιώσαντες είχαν 13,8 συστολική και 8,5 διαστολική, ενώ οι αποβιώσαντες είχαν 11,8 με 7,2 αντίστοιχα.



Από την εξέταση των στοιχείων προκύπτει πως οι άνδρες είναι σαφώς πιο επιρρεπείς στο να πάθουν έμφραγμα από τις γυναίκες (Γράφημα 1) ενώ το παθαίνουν και σε πολύ μικρότερη ηλικία από ότι οι γυναίκες (Θηκόγραμμα 1). Ιδιαίτερο ενδιαφέρον παρουσιάζει το τελευταίο γράφημα που δείχνει πως η θνητότητα μεταξύ των γυναικών είναι πολύ μεγαλύτερη σε σχέση με αυτή των αντρών.

2.5 Προεπεξεργασία δεδομένων

Το σύνολο δεδομένων παρουσιάζει ελλιπείς τιμές σε πολλές από τις στήλες του. Σε ορισμένες από αυτές το πλήθος των ελλιπών τιμών είναι τέτοιο που δε μπορούν να χρησιμοποιηθούν ώστε να τροφοδοτηθούν ως είσοδος στα μοντέλα μηχανικής μάθησης οπότε και θα παραλειφθούν. Σε περιπτώσεις όπου το πλήθος των ελλιπών τιμών δεν είναι απαγορευτικό για τη χρησιμοποίηση του συγκεκριμένου χαρακτηριστικού, τότε εφαρμόζεται για τα μεν κατηγορικά χαρακτηριστικά η αντικατάσταση των ελλιπών τιμών με την πιο κοινή τιμή της συγκεκριμένης στήλης, για τα δε συνεχή η αντικατάσταση με το μέσο της στήλης. Στο κάτωθι γράφημα εμφανίζεται το πλήθος των ελλιπών τιμών κάθε στήλης.



Ραβδόγραμμα απεικόνισης ελλιπών τιμών του συνόλου δεδομένων

3. Μεθοδολογία και μετρικές αξιολόγησης

3.1 Μεθοδολογίες αξιολόγησης αλγορίθμων

1) Holdout

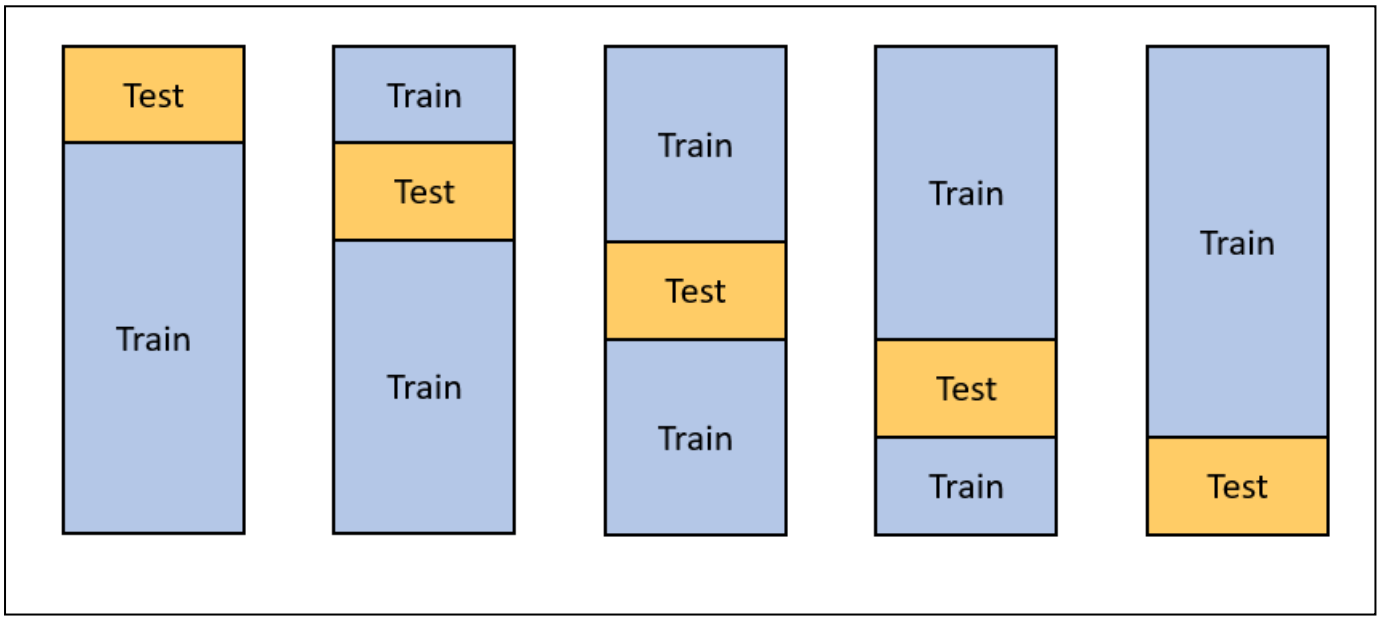
Αποτελεί την πιο απλή διαδικασία αξιολόγησης. Το σύνολο δεδομένων διαχωρίζεται σε 1 σύνολο εκπαίδευσης (train set) και 1 σύνολο ελέγχου (test set). Ο αλγόριθμος εκπαιδεύεται στο πρώτο σύνολο και η επίδοσή του αξιολογείται στο δεύτερο. Η μεθοδολογία αυτή υστερεί σε αξιοπιστία καθώς ο αρχικός διαχωρισμός είναι τυχαίος, γεγονός που μπορεί να οδηγήσει είτε σε υπερεκτίμηση των δυνατοτήτων του αλγορίθμου είτε σε υποεκτίμηση.

2) k-fold cross validation

Το k-fold cross validation είναι η πιο συνηθισμένη μέθοδος αξιολόγησης ενός αλγορίθμου. Πρόκειται για μία επαναληπτική διαδικασία η οποία, σε αντίθεση με το απλοϊκό train-test split χωρίζει το σύνολο δεδομένων σε k υποσύνολα (folds). Ο αλγόριθμος σε κάθε επανάληψη εκπαιδεύεται στα k-1 υποσύνολα και αξιολογείται στο εναπομείναν. Στο τέλος των k επαναλήψεων λαμβάνουμε το μέσο όρο των επιδόσεων του αλγορίθμου επί της μετρικής που του ορίσαμε. Η μεθοδολογία αυτή είναι αρτιότερη σε σχέση με το train-test split καθώς ο αλγόριθμος εκπαιδεύεται και αξιολογείται τελικά σε όλο το σύνολο δεδομένων. Το k-fold cross validation απαιτεί k φορές περισσότερο χρόνο από το train-test split. Σε περιπτώσεις όπου η μεταβλητή στόχος δεν είναι ισοκατανεμημένη μεταξύ των κλάσεων που την αποτελούν, τότε πρέπει να εφαρμόσουμε το λεγόμενο stratification δηλαδή σε κάθε επανάληψη να εξασφαλίσουμε ότι η αναλογία των κλάσεων που αποτελούν τη μεταβλητή στόχος παραμένει σταθερή τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο ελέγχου. Καθώς η εκπαίδευση και αξιολόγηση του μοντέλου λαμβάνει χώρα k φορές, στο τέλος προκειμένου να εξαχθεί μία τιμή για τη μετρική αξιολόγησης λαμβάνεται ο μέσος όρος των τιμών της μετρικής στα k folds.

Το cross validation χρησιμοποιείται επίσης προκειμένου να προσδιορίσουμε τις υπερπαραμέτρους ενός αλγορίθμου. Αυτό γίνεται μέσω μιας μεθόδου εξάντλησης η οποία ονομάζεται grid search. Στο grid search προσδιορίζουμε ένα εύρος τιμών για τις παραμέτρους που θέλουμε να βελτιστοποιήσουμε και έπειτα αξιολογούμε τον αλγόριθμο για κάθε δυνατό συνδυασμό αυτών των παραμέτρων μέσω k-fold cross validation. Από όλα τα μοντέλα που έτρεξαν μέσω αυτής της διαδικασίας κρατάμε αυτό το οποίο απέδωσε καλύτερα σε σχέση με τη μετρική αξιολόγησης που ορίσαμε. Το grid search είναι υπολογιστικά κοστοβόρο καθώς πραγματοποιεί k-fold cross validation σε όλους τους δυνατούς συνδυασμούς των παραμέτρων που του ορίζουμε. Έτσι λοιπόν αν θέλουμε να εξετάσουμε παραμέτρους οι οποίες λαμβάνουν διαφορετικές τιμές, τότε προκύπτουν

συνδυασμοί μοντέλων ίσοι με το γινόμενο των πληθαρίθμων κάθε παραμέτρου και τα μοντέλα που εξετάζονται τελικά είναι k φορές αυτό το πλήθος. Για παράδειγμα αν ψάχνω να βρω τις βέλτιστες τιμές ενός δένδρου απόφασης για τις παραμέτρους μέγιστο βάθος και μέγιστο πλήθος φύλλων, με την πρώτη να λαμβάνει 5 διακριτές τιμές και τη δεύτερη 6 τότε προκύπτουν 5*6=30 διαφορετικοί συνδυασμοί οι οποίοι μέσω του k-fold cross validation καταλήγουν σε 30*10=300 μοντέλα.



Οπτική αναπαράσταση ενός 5 fold cross validation (Πηγή www.analyticsvidhya.com)

3.2 Μετρικές αξιολόγησης

Οι αλγόριθμοι μηχανικής μάθησης που πραγματοποιούν ταξινόμηση (classifiers) εκπαιδεύονται πάνω στο σύνολο εκπαίδευσης που τους παρέχουμε και κατόπιν αξιολογούνται σε ένα σύνολο αξιολόγησης. Όταν οι κλάσεις περιέχουν ισοπληθή ή σχεδόν ισοπληθή παραδείγματα εκπαίδευσης τότε η εκπαίδευση είναι μια σχετικά απλή διαδικασία. Όταν όμως μία κλάση καταλαμβάνει τη μερίδα του λέοντος μέσα στο σύνολο δεδομένων τότε η εκπαίδευση του ταξινομητή γίνεται δυσκολότερη καθώς δεν υπάρχουν επαρκή παραδείγματα εκπαίδευσης από τις κλάσεις της μειονότητας. Στις περιπτώσεις αυτές θα πρέπει να χρησιμοποιούνται μετρικές αξιολόγησης οι οποίες λαμβάνουν υπ'οψιν τους αυτή την ανισομέρεια ανάμεσα στις διαφορετικές κλάσεις καθώς σε διαφορετική περίπτωση μπορεί η αξιολόγηση να καταλήξει σε εντελώς λανθασμένα συμπεράσματα για την απόδοση του ταξινομητή. Η ανάγκη δε αυτή γίνεται ακόμα μεγαλύτερη όταν η σωστή πρόβλεψη της κλάσης της μειονότητας αφορά κάτι πολύ σημαντικό όπως για παράδειγμα τη θετικότητα ενός τεστ σε ένα πολύ μολυσματικό ιό. Έτσι λοιπόν

θα χρησιμοποιηθούν περισσότερες της μιας μετρικές προκειμένου να αξιολογηθούν οι ταξινομητές και να αποδοθεί σφαιρικά η επίδοσή τους.

Πίνακας Σύγχυσης (confusion matrix)

Στα προβλήματα ταξινόμησης με βάση τα αποτελέσματα του αλγορίθμου μπορούμε στη συνέχεια να φτιάξουμε τον πίνακα σύγχυσης (confusion matrix), από τον οποίο μπορεί εύκολα να γίνει αντιληπτή η απόδοση του ταξινομητή σε κάθε κλάση.

		ΠΡΟΒΛΕΨΗ	
		0	1
ΠΡΑΓΜΑΤΙΚΗ	0	TN	FP
	1	FN	TP

Πάνω στον πίνακα σύγχυσης ορίζονται ορισμένα μεγέθη:

I) True Negatives TN το οποίο είναι το πλήθος των σωστών ταξινομήσεων της κλάσης 0

II) False Positives FP είναι το πλήθος των λανθασμένα ταξινομηθέντων στην κλάση 1

III) False Negatives FN είναι το πλήθος των λανθασμένα ταξινομηθέντων στην κλάση 0

IV) True Positives TP είναι το πλήθος των σωστών ταξινομήσεων της κλάσης 1

Α) Ακρίβεια (Accuracy)

Πρόκειται για την πιο κοινή και απλή μετρική αξιολόγησης ενός ταξινομητή. Ορίζεται ως ο λόγος των σωστών προβλέψεων προς το σύνολο όλων των προβλέψεων. Η ακρίβεια από μόνη της δεν είναι επαρκής μετρική αξιολόγησης ενός ταξινομητή στην περίπτωση των ανισοκατανεμημένων κλάσεων και αυτό μπορεί εύκολα να γίνει κατανοητό μέσα από ένα παράδειγμα. Έστω ότι έχουμε να κάνουμε με ένα δυαδικό πρόβλημα ταξινόμησης όπου η μία κλάση 0 καταλαμβάνει το 95% του συνόλου δεδομένων και η έτερη 1 το 5%. Αν ο ταξινομητής ταξινομήσει όλα τα παραδείγματα του συνόλου αξιολόγησης ότι ανήκουν στην κλάση 0 τότε θα έχει μεν επιτύχει ακρίβεια 95% το οποίο είναι ένα εξαιρετικό ποσοστό, όμως δε θα έχει προβλέψει σωστά ούτε μία εγγραφή της κλάσης 1! [3] Η ακρίβεια θα χρησιμοποιηθεί ως μετρική για την αξιολόγηση των μοντέλων προκειμένου να υπάρχει μια γενικότερη κατανόηση της επίδοσης κάθε ταξινομητή, αλλά χωρίς να είναι η μοναδική μετρική πάνω στην οποία θα αξιολογηθούν τα μοντέλα καθώς θα πρέπει υπάρχει ιδιαίτερη εποπτεία πάνω στις επιδόσεις των μοντέλων στην κλάση της μειονότητας.

Β) Ευαισθησία (Sensitivity/Recall)

Κατά σύμβαση όταν έχουμε να κάνουμε με ανισοκατανεμημένες κλάσεις αντιστοιχούμε το 0 στην πλειονοτική κλάση, η οποία καλείται και αρνητική, και το 1 στη μειονοτική η οποία καλείται και θετική. Το sensitivity ενός ταξινομητή έχει να κάνει με το λόγο True Positives προς το πλήθος όλων των εγγραφών που ανήκουν στην κλάση 1, πλήθος το οποίο προκύπτει από το άθροισμα των True Positives συν τα False Negatives .

$$Recall = \frac{TP}{(TP + FN)}$$

Καθώς αυτό το μέγεθος αφορά αποκλειστικά την κλάση 1 είναι πολύ σημαντικό να αξιολογείται και να λαμβάνεται υπ'οψιν. Όσο πιο υψηλό είναι το recall σημαίνει ότι τόσο πιο πιθανό είναι να έχουν ταξινομηθεί σωστά οι παρατηρήσεις της κλάσης 1.

Γ) Precision

Άλλη μία μετρική που αφορά τη θετική κλάση. Ορίζεται ως ο λόγος True Positives προς το σύνολο των εγγραφών που ταξινομήθηκαν στην κλάση 1, το οποίο σύνολο προκύπτει από το άθροισμα των True Positives συν τα False Positives.

$$Precision = \frac{TP}{(TP + FP)}$$

Όσο πιο υψηλό είναι το precision σημαίνει ότι από όσες εγγραφές έχουν ταξινομηθεί στην κλάση 1 τόσο πιο πιθανό είναι να έχουν ταξινομηθεί σωστά.

Δ) F1-score

Μεταξύ των μεγεθών precision και recall υφίσταται ένα tradeoff. Δηλαδή μειώνοντας το ένα αυξάνεται το άλλο και αντιστρόφως. Έτσι μπορεί ένα μοντέλο να παρουσιάζει εξαιρετικό precision και ταυτόχρονα το recall του να είναι πολύ χαμηλό. Αυτά τα 2 μεγέθη από μόνα τους δεν είναι ικανά να περιγράψουν την πλήρη απόδοση του μοντέλου. Καθώς λοιπόν υπάρχει αυτό το tradeoff εισάγεται ένα τρίτο μέγεθος το F1-score το οποίο είναι ο αρμονικός μέσος των precision και recall.

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Καμπύλες ROC & Precision-Recall

Οι ταξινομητές εκτός από το να αποδίδουν ταμπέλα (label) σε μια κλάση είναι δυνατόν και να αποδώσουν την πιθανότητα να ανήκει μία εγγραφή σε κάποια κλάση. Έτσι μπορεί να υποτεθεί ότι για υπολογιζόμενη πιθανότητα <0,5 η εγγραφή θα ταξινομείται ως αρνητική, ενώ αν η πιθανότητα είναι μεγαλύτερη ή ίση από 0,5 η εγγραφή θα ταξινομείται ως θετική. Το 0,5 είναι όμως ένα αυθαίρετο κατώφλι και στη θέση του θα μπορούσε να μπει οποιοσδήποτε άλλος αριθμός μεταξύ του 0 και του 1.

Η καμπύλη ROC είναι ένα γράφημα από το οποίο μπορούν να εξαχθούν συμπεράσματα για την απόδοση ενός μοντέλου χρησιμοποιώντας 2 μεγέθη. Την ευαισθησία ή διαφορετικά True Positive Rate και το False Positive Rate. Το False Positive Rate ορίζεται ως ο λόγος των False Positives προς το σύνολο των εγγραφών της κλάσης 0, δηλαδή το άθροισμα True Negatives συν False Positives. Στον άξονα χ μπαίνει το False Positive Rate και στον άξονα y το True Positive Rate. Τα σημεία της καμπύλης διαμορφώνονται για διαφορετικές τιμές του κατωφλίου πιθανοτήτων που διαχωρίζει τις 2 κλάσεις.

Στην περίπτωση που έχουμε ανισοκατανεμημένες κλάσεις τότε χρησιμοποιείται η καμπύλη Precision-Recall η οποία παράγεται με ανάλογο τρόπο με την καμπύλη ROC μόνο που ο άξονας χ αυτή τη φορά αντιπροσωπεύει το precision. Έτσι λοιπόν η συγκεκριμένη καμπύλη εστιάζει μόνο στη θετική κλάση.

Και για τις 2 καμπύλες κριτήριο απόδοσης ενός ταξινομητή αποτελεί το εμβαδό που σχηματίζεται κάτω από την καμπύλη (AUC). Όσο μεγαλύτερο είναι το εμβαδό τόσο καλύτερη η επίδοση του ταξινομητή.

3.3 Στάθμιση πιθανοτήτων

Πολλοί αλγόριθμοι μηχανικής μάθησης μπορούν να προβλέψουν πιθανότητες ή να εξάγουν πιθανοτικά αποτελέσματα σχετικά με το σε ποια κλάση ανήκει μια εγγραφή. Επί παραδείγματι η λογιστική παλινδρόμηση υπολογίζει ευθέως πιθανότητες, ενώ το SVM μπορεί και υπολογίζει αριθμούς που δύνανται να ερμηνευτούν ως πιθανότητες. Οι πιθανότητες προσφέρουν τον απαιτούμενο βαθμό λεπτομέρειας προκειμένου να αξιολογηθούν και να συγκριθούν μεταξύ τους διαφορετικά μοντέλα. Το πρόβλημα με τις πιθανότητες αυτές είναι πως συνήθως δεν είναι σταθμισμένες (calibrated). Έτσι κρίνεται σκόπιμο πριν την αξιολόγηση των μοντέλων να γίνεται στάθμιση των πιθανοτήτων αυτών. [3]

Για παράδειγμα έστω ότι ένα μοντέλο SVM επιστρέφει πιθανότητα 0,8 μια εγγραφή να ανήκει στην κλάση 1. Αν οι πιθανότητες είναι σταθμισμένες τότε αυτό σημαίνει ότι το 80% των εγγραφών που έλαβαν πιθανότητα 0,8 πράγματι θα ανήκουν στην κλάση 1. Αντίθετα αν οι πιθανότητες δεν είναι σταθμισμένες τότε στην κλάση 1 θα μπορούσε να ανήκει το 95% όσων εγγραφών η πιθανότητα υπολογίστηκε με 0,8 ότι ανήκει στην κλάση 1, άρα να έχει υπάρξει μια υποεκτίμηση, ή θα μπορούσε το 50% να ανήκει πράγματι στην κλάση 1 δηλαδή να έχει υπάρξει υπερεκτίμηση.

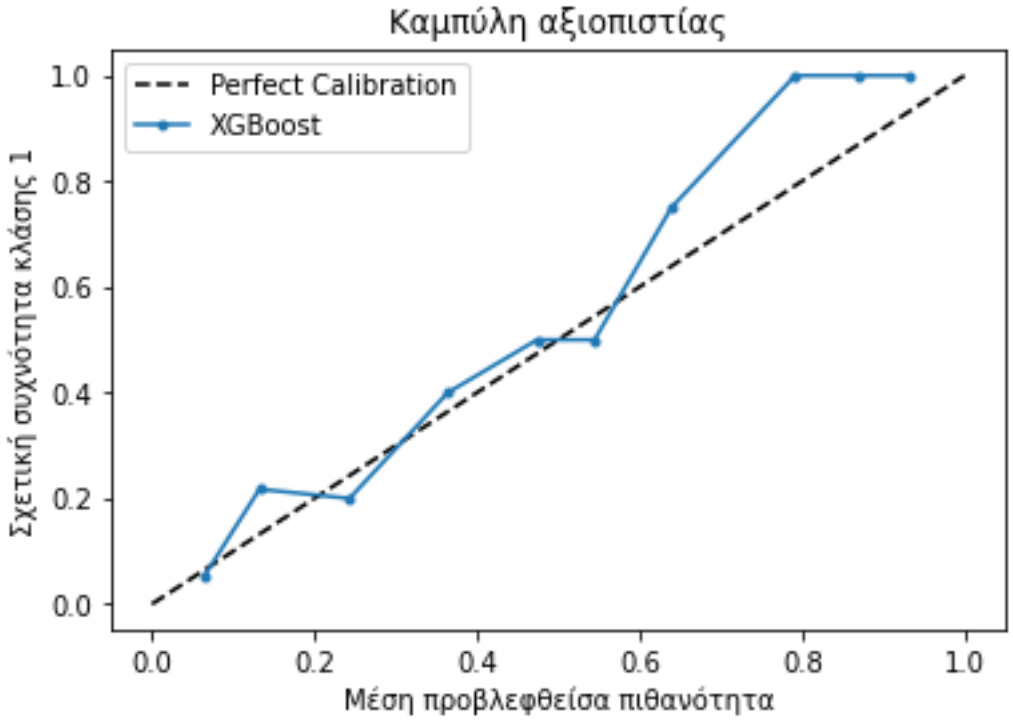
Η στάθμιση των πιθανοτήτων γίνεται μέσω της διόρθωσής της κλίμακας τους έτσι ώστε να ταιριάζει στην κατανομή των παρατηρήσεων στο σύνολο εκπαίδευσης. Σε αυτό το σημείο πρέπει να τονιστεί ότι η στάθμιση των πιθανοτήτων δε γίνεται ταυτόχρονα με την αξιολόγηση του μοντέλου καθώς αυτό θα συνιστούσε μεροληψία γιατί το ίδιο παράδειγμα εκπαίδευσης θα χρησιμοποιούνταν και για την εκπαίδευση του μοντέλου και για την στάθμιση.

Κατ’ επέκταση αν είχαμε να επιλέξουμε μεταξύ 2 μοντέλων εκ των οποίων το πρώτο αποδίδει 85% ακρίβεια με εμπιστοσύνη (confidence) 86% και το δεύτερο αποδίδει ακρίβεια πάλι 85%, αλλά με εμπιστοσύνη 99% τότε θα έπρεπε να επιλέξουμε το πρώτο καθώς φαίνεται πως οι πιθανότητες που αποδίδει (εμπιστοσύνη) είναι σταθμισμένες, ενώ το δεύτερο κάνει υπερεκτίμηση. [5]

Η στάθμιση μπορεί να γίνει μέσω 2 βασικών τεχνικών: Platt και isotonic regression. Η τεχνική Platt περνάει τα τύπου πιθανότητας αποτελέσματα που εξάγει το αρχικό μοντέλο σε ένα δεύτερο μοντέλο λογιστικής παλινδρόμησης. Τα μοντέλα λογιστικής παλινδρόμησης αποδίδουν σταθμισμένες πιθανότητες. Το isotonic regression από την άλλη χρησιμοποιεί γραμμική παλινδρόμηση με βάρη προκειμένου να επιτύχει την στάθμιση [3]. Για την εύρεση των βέλτιστων παραμέτρων στάθμισης δηλαδή ποιά είναι η καλύτερη μέθοδος μεταξύ των 2 και πόσες φορές χρειάζεται να γίνει cross validation χρησιμοποιήθηκε η εξαντλητική μέθοδος του gridsearch.

Για την οπτική επόπτευση της στάθμισης των πιθανοτήτων υπάρχει η καμπύλη αξιοπιστίας (reliability curve). Η καμπύλη αυτή έχει στον άξονα χ τις προβλεφθείσες από το μοντέλο πιθανότητες που αφορούν σε εγγραφές της θετικής κλάσης και στον άξονα y τις πραγματικές πιθανότητες (empirical probability) για τις αντίστοιχες εγγραφές ή αλλιώς τη σχετική συχνότητα της κλάσης 1. Στη διαγώνιο εμφανίζεται μια διακεκομμένη γραμμή η οποία υποδεικνύει την τέλεια στάθμιση. Κάτω από τη διαγώνιο το μοντέλο υπερεκτιμά τις πιθανότητες, ενώ πάνω από τη διαγώνιο υποεκτιμά.

Η στάθμιση πιθανοτήτων δεν είναι απαραίτητο βήμα κάθε φορά στα πλαίσια της αξιολόγησης ενός μοντέλου και η αξιοποίηση αυτού του εργαλείου εξαρτάται από τη σημαντικότητα του προβλεπόμενου γεγονότος. Υπάρχουν περιπτώσεις όπου οι μη σταθμισμένες πιθανότητες είναι ανεκτές όπως για παράδειγμα η πρόβλεψη εάν ένας καταναλωτής θα επαναλάβει αγορά από ένα κατάστημα. Στην υπό εξέταση περίπτωση όμως, όπου το προβλεπόμενο γεγονός αφορά την υγεία ασθενών, είναι σημαντικό να γνωρίζουμε ότι οι πιθανότητες των μοντέλων είναι σταθμισμένες.



4. Ταξινομητές και μεθοδολογίες εκπαίδευσης σε ανισοκατανεμημένες κλάσεις

4.1 Προετοιμασία δεδομένων

Ένα σημαντικό κομμάτι της επεξεργασίας ενός προβλήματος μηχανικής μάθησης έχει να κάνει με την διαμόρφωση του συνόλου δεδομένων σε μορφή τέτοια ώστε να είναι αξιοποιήσιμα από τους αλγόριθμους.

Στο παρόν πρόβλημα μηχανικής μάθησης εξετάζεται η κατάληξη ή μη του ασθενούς. Καθώς στα δεδομένα δεν παρέχεται ευθέως αυτή η πληροφορία, αλλά παρέχονται οι αιτίες θανάτου κωδικοποιημένες με διαφορετικά νούμερα (1-6) και με 0 η επιβίωση, προβαίνουμε εξ' αρχής στη μετατροπή της μεταβλητής αυτής σε δυαδική (binary) με το 0 να συνεχίζει να συμβολίζει την επιβίωση και το 1 να συμβολίζει την κατάληξη του ασθενούς.

Αρκετά από τα χαρακτηριστικά του συνόλου δεδομένων είναι αριθμητικά. Τα αριθμητικά δεδομένα επειδή συνήθως εκφράζουν μετρήσεις σε διαφορετικές μονάδες μέτρησης χρειάζεται να κανονικοποιηθούν προκειμένου να μη λάβει ο αλγόριθμος λανθασμένες πληροφορίες σχετικά με την βαρύτητα ενός χαρακτηριστικού σε σχέση με ένα άλλο. Ο μετασχηματισμός που φέρνει τα αριθμητικά δεδομένα στην ίδια κλίμακα είναι ο εξής:

$$Xnorm = \frac{X - Xmin}{Xmax - Xmin}$$

4.2 Ταξινομητές

Υπερπαράμετροι

Ο κάθε ταξινομητής κατά τη διάρκεια της εκπαίδευσης αυτό που κάνει είναι να προσαρμόζει κάποιες δικές του εσωτερικές παραμέτρους προκειμένου να βελτιστοποιήσει την απόδοση του ως προς τη μετρική που του έχει οριστεί. Πέρα από αυτές τις εσωτερικές παραμέτρους υπάρχει άλλο ένα σύνολο παραμέτρων που καλούνται υπερπαράμετροι και σε αντίθεση με τις εσωτερικές παραμέτρους ορίζονται από το χρήστη. Οι υπερπαράμετροι ελέγχουν τη διαδικασία εκπαίδευσης καθορίζοντας εν τέλει τις τιμές των εσωτερικών παραμέτρων του αλγορίθμου εξ' ου και η ονομασία υπέρ-παράμετροι, βρίσκονται δηλαδή ένα επίπεδο πάνω από τις εσωτερικές παραμέτρους. Οι υπερπαράμετροι αφορούν μόνο το στάδιο της εκπαίδευσης του αλγορίθμου και δεν αποτελούν κομμάτι του μοντέλου που προκύπτει.

Ensemble Learning

Η ensemble μάθηση είναι μια τεχνική της μηχανικής μάθησης όπου συνδυάζονται πολλαπλά μοντέλα (base/weak learners) προκειμένου να επιτευχθούν καλύτερες προβλέψεις. Η ensemble μάθηση περιλαμβάνει 2 μεγάλες κατηγορίες-στρατηγικές: το Boosting και το Bagging (Bootstrap Aggregating). Στο bagging εκπαιδεύονται πολλαπλά μοντέλα τα οποία στο τέλος καλούνται να ψηφίσουν προκειμένου να λάβει χώρα η πρόβλεψη. Κάθε μοντέλο μαθαίνει ανεξάρτητα από τα προηγούμενα. Αντίθετα στο boosting κάθε νέο μοντέλο ενημερώνεται για τα σφάλματα των προηγούμενων και η τελική πρόβλεψη είναι αποτέλεσμα του γραμμικού συνδυασμού των προβλέψεων όλων των μοντέλων. [7]

1. Logistic Regression

Η λογιστική παλινδρόμηση (logistic regression) είναι ένας αλγόριθμος της μηχανικής μάθησης που προέρχεται από τη στατιστική και είναι ιδανική για προβλήματα ταξινόμησης 2 κλάσεων. Ανήκει στα γενικευμένα γραμμικά μοντέλα (generalized linear models). Το όνομα της προέρχεται από τη λογιστική συνάρτηση η οποία καθώς παίρνει τιμές στο (0,1) χρησιμοποιείται για να αποδώσει αριθμούς που μπορούν να ερμηνευτούν ως πιθανότητες. [3]

Ο τύπος της λογιστικής συνάρτησης είναι

$$f(z) = \frac{e^z}{e^z + 1}$$

Καθώς η πρόβλεψη αφορά πιθανότητες μπορούμε να την εκφράσουμε ως εξής:

$$P(X)=P(Y=1|X) \quad (1)$$

Και η υπόθεση διαμορφώνεται ότι y=1 αν P(Y=1|X)>t για καθορισμένο t, το οποίο καλείται σύνορο απόφασης, διαφορετικά y=0.

Χρησιμοποιώντας το λογάριθμο του λόγου συμπληρωματικών πιθανοτήτων και την απλή γραμμική υπόθεση καθώς έχουμε να κάνουμε με γραμμικό μοντέλο, η σχέση (1) καταλήγει ως εξής:

$$\mathrm{P}(\mathrm{Y} = 1|\mathrm{X}) = \frac{e^{b_0+b_1*x}}{e^{b_0+b_1*x} + 1}$$

που αναπαριστά την υπόθεση της γραμμικής παλινδρόμησης.

Το logistic regression έχει την υπερπαράμετρο C η οποία είναι μια σταθερά κανονικοποίησης και είναι θετικός αριθμός. Η κανονικοποίηση παίζει το ρόλο της αποφυγής του overfitting του μοντέλου στα δεδομένα εκπαίδευσης μειώνοντας τη διακύμανση του μοντέλου.

Μία άλλη υπερπαράμετρος είναι το penalty. Το penalty επιβάλλει ένα κόστος στην αντικειμενική συνάρτηση και είναι ο τύπος της κανονικοποίησης που θα

εφαρμοστεί στο μοντέλο. Οι τιμές για την παράμετρο penalty l1 και l2 προέρχονται από τα μοντέλα lasso και ridge regression αντίστοιχα. Στο μεν πρώτο η αντικειμενική συνάρτηση έχει τύπο:

$$\sum(y_i-\sum x_{ij}b_j)^2 + \lambda\sum|b_j|$$

Ενώ στη ridge η αντικειμενική συνάρτηση έχει τύπο:

$$\sum(y_i-\sum x_{ij}b_j)^2 + \lambda\sum b_j^2$$

Όπου λ και στις δύο περιπτώσεις η σταθερά που καθορίζει τη βαρύτητα του penalty.

Τέλος υπάρχει και η υπερπαράμετρος solver που αφορά το ποιος θα είναι ο αλγόριθμος που θα επιλύσει το πρόβλημα βελτιστοποίησης, δεδομένων των παραμέτρων του μοντέλου.

2. Decision Tree

Το δένδρο απόφασης (decision tree) είναι ένας ταξινομητής πολύ απλός στη σύλληψη του και μοιάζει με ένα διάγραμμα ροής. Από το ένα επίπεδο στο άλλο περνάει μέσω λογικών συνθηκών. Ο κάθε κόμβος αποτελείται από ένα χαρακτηριστικό του συνόλου δεδομένων. Το κριτήριο με το οποίο κάθε χαρακτηριστικό γίνεται κόμβος είναι το είτε το Gini Index, είτε το Information Gain. Το μεν Gini Index ορίζεται ως:

$$Gini = 1 - \sum[p(j|t)]^2$$

Όπου p(j|t) η σχετική συχνότητα της κλάσης t στον κόμβο j. Το Gini είναι μέτρο του impurity κάθε διαχωρισμού, δηλαδή του πόσο αναμεμιγμένος είναι. Ένας κόμβος ο οποίος διαχωρίζεται σε 2 φύλλα ονομάζεται pure όταν κάθε φύλλο έχει 100% πιθανότητα.

Το δε το Information Gain ορίζεται ως:

$$\text{Gain(S,A)} = \text{E(S)} - \text{I(S,A)}$$

Όπου το μέγεθος $\mathrm{E(S)} = \sum p_i * \log(\frac{1}{p_i})$ καλείται συνολική εντροπία, ενώ το p_i συμβολίζει την πιθανότητα εμφάνισης της κλάσης i στο j και $\mathrm{I(S,A)} = \sum \frac{|S_j|}{|S|}$*E(S) είναι η εντροπία του διαχωρισμού, ενώ το $|S_j|$ συμβολίζει τον πληθάριθμο του διαχωρισμού J.

Το decision tree καθώς έχει την τάση να αναπτύσσεται έως ότου καλύψει όλα τα σημεία του συνόλου εκπαίδευσης έχει το μειονέκτημα ότι κάνει εύκολα overfit, το οποίο θα οδηγήσει σε μεγάλη διακύμανση στις προβλέψεις στο σύνολο ελέγχου. Για να αποφευχθεί αυτό συνηθίζεται το κλάδεμα (pruning) του δέντρου.

Το decision tree έχει ως υπερπαράμετρο το max_depth που είναι το μέγιστο επίπεδο ως το οποίο θα αναπτυχθεί το δένδρο. Σκοπός αυτής της παραμέτρου είναι η αποφυγή του overfitting.

Μία άλλη υπερπαράμετρος του είναι το max_leaf_nodes το οποίο καθορίζει το μέγιστο πλήθος των τελικών φύλλων που θα έχει τελικά το δένδρο.

3. Random Forest

Ο Random Forest θεωρείται ο χαρακτηριστικότερος εκπρόσωπος του bagging. Χρησιμοποιεί bootstrapping δηλαδή δειγματοληψία με επανάθεση πάνω στις εγγραφές του συνόλου δεδομένων φτιάχνοντας έτσι ένα υποσύνολο του αρχικού συνόλου. Στη συνέχεια με βάση αυτό το υποσύνολο δεδομένων δημιουργεί πολλαπλά Decision Trees χρησιμοποιώντας σε κάθε ένα από αυτά, ένα τυχαίο, αλλά ίδιο στη πλήθος, υποσύνολο των χαρακτηριστικών. Τελικά κάθε εγγραφή περνάει μέσα από όλα τα Decision Trees που έχουν φτιαχτεί και η τελική απόφαση είναι η απόφαση της πλειοψηφίας. Αν για παράδειγμα έχουν φτιαχτεί 100 δέντρα και τα 70 κατατάσσουν μία εγγραφή στην κλάση 0 και τα υπόλοιπα 30 την κατατάσσουν στην κλάση 1 τότε η εγγραφή θα καταταχτεί στην κλάση 0.

Ο Random Forest έχει ως υπερπαράμετρο το class_weight και ενημερώνει τον αλγόριθμο για τα βάρη που θα πρέπει να έχουν οι κλάσεις σε σχέση με την κατανομή τους ως προς τη μεταβλητή στόχος. Αν οι 2 κλάσεις είναι ισοκατανεμημένες τότε class_weight=1 που είναι και η default τιμή, διαφορετικά έχουμε τις τιμές balanced όπου το μοντέλο υπολογίζει μόνο του τα βάρη που θα πρέπει να έχει κάθε label και τέλος την τιμή balanced_subsample όπου το μοντέλο κάνει το ίδιο όπως στην περίπτωση balanced, όμως το εφαρμόζει στο υποσύνολο από το οποίο αντλεί τυχαία τις εγγραφές με επανάθεση.

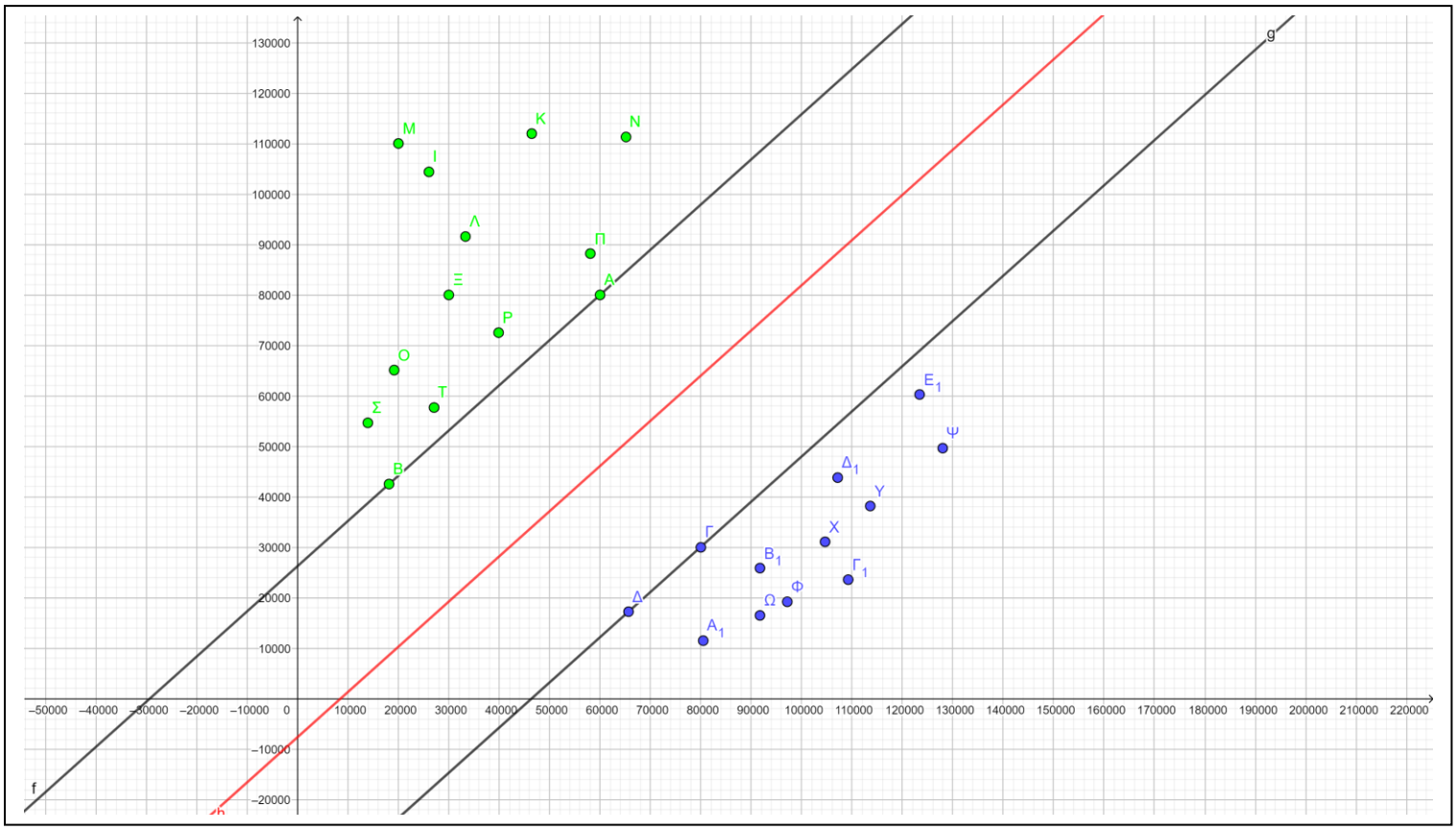
Άλλη υπερπαράμετρος του είναι το max_depth που αφορά το κάθε δένδρο που δημιουργείται και η λειτουργία της είναι ίδια με αυτή που έχει για το decision tree.

4. Support Vector Machine

Πρόκειται για έναν από τους σημαντικότερους ταξινομητές. Το SVM χρησιμοποιείται σε γραμμικά διαχωρίσιμα προβλήματα αλλά και μη γραμμικά διαχωρίσιμα προβλήματα. Πρόκειται για ένα ταξινομητή μέγιστου περιθωρίου ο οποίος όμως επιτρέπει ως ένα βαθμό λάθος ταξινομήσεις, τα περιθώρια δηλαδή είναι ελαστικά, προκειμένου εν τέλει να βελτιστοποιήσει την προγνωστική του δυνατότητα. Οι ταξινομητές μέγιστου περιθωρίου προσπαθούν να βρουν ποιο είναι το μέγιστο δυνατό περιθώριο ανάμεσα σε σημεία δύο εκ των προτέρων γνωστές κλάσεις. Διαισθητικά αυτό που προσπαθεί να κάνει το SVM είναι να δημιουργήσει την καλύτερη γραμμή -κριτήριο απόφασης- η οποία να διαχωρίζει το επίπεδο σε δύο ημιεπίπεδα τα οποία θα αντιστοιχούν στις δύο κλάσεις.

Σημαντικό ρόλο στα SVMs παίζουν ορισμένοι μετασχηματισμοί οι οποίοι ονομάζονται kernels. Οι μετασχηματισμοί αυτοί είναι ικανοί να μετατρέψουν ένα μη γραμμικά διαχωρίσιμο στο επίπεδο πρόβλημα σε γραμμικά διαχωρίσιμο μέσω ενός υπερεπιπέδου ανεβάζοντας τις γεωμετρικές διαστάσεις σε ανώτερο επίπεδο.

Το SVM έχει κι αυτό την υπερπαράμετρο C το οποίο αφορά στη δημιουργία του συνόρου απόφασης και αν θα είναι σχετικά απλό και θα γενικεύει καλά ή αν θα είναι πιο περίπλοκο με τον κίνδυνο σε αυτή την περίπτωση να δημιουργήσει overfitting.



Γεωμετρική αναπαράσταση ενός SVM στο επίπεδο

Στο παραπάνω σχήμα οι ευθείες f και g ορίζουν τα όρια μεταξύ των 2 κλάσεων και η κόκκινη γραμμή το διαχωριστή μέγιστου περιθωρίου

5. Adaboost

Ο Adaboost σε αντίθεση με τον random Forest, ο οποίος αφήνει το κάθε δέντρο να αναπτυχθεί πλήρως, δημιουργεί δέντρα με 1 κόμβο και μόνο 2 φύλλα. Αυτά τα δέντρα καλούνται stumps και ο Adaboost φτιάχνει ένα Forest of stumps [4]. Το

κάθε stump αφού αποτελείται από ένα μόνο κόμβο θα βασίζει την πρόβλεψη του σε ένα και μόνο χαρακτηριστικό του συνόλου δεδομένων. Και πάλι σε αντίθεση με το random Forest κάθε stump δεν έχει την ίδια βαρύτητα στην τελική απόφαση του αλγορίθμου και επίσης η σειρά με την οποία φτιάχνονται τα stumps έχει σημασία καθώς τα σφάλματα του προηγούμενου stump επηρεάζουν το επόμενο. Το σφάλμα του t-οστού stump υπολογίζεται από τον εξής τύπο

$$error = \varepsilon(t) = \frac{\sum w_i \, I(y_i \neq h_j(x_i))}{\sum w_i}$$

Το οποίο σημαίνει ότι το σφάλμα κάθε stump ισούται με το άθροισμα των βαρών των εγγραφών που προβλέφθηκαν λανθασμένα δια το άθροισμα όλων των βαρών. Η συνάρτηση που καθορίζει το πόσο επηρεάζει το κάθε stump την απόφαση του αλγορίθμου δίδεται από τον εξής τύπο

$$a(t) = 0{,}5 * \log\left(\frac{1-\varepsilon(t)}{\varepsilon(t)}\right)$$

Ενώ τα βάρη από το ένα stump στο άλλο αναπροσαρμόζονται μέσω της συνάρτησης:

$$a(t+1) = \begin{cases} w_i * e^{a(t)}, \textit{για τις λανθασμένα ταξινομημένες εγγραφές} \\ w_i * e^{-a(t)}, \textit{για τις σωστά ταξινομημένες εγγραφές} \end{cases}$$

Από την παραπάνω σχέση προκύπτει ότι οι λανθασμένα ταξινομημένες εγγραφές λαμβάνουν μεγαλύτερο βάρος από τις ορθά ταξινομημένες και αυτό συμβαίνει ούτως ώστε στο επόμενο βήμα να επικεντρωθεί σε αυτές ο αλγόριθμος.

Ο Adaboost έχει ως υπερπαράμετρο το n_estimators η οποία καθορίζει το πλήθος των δένδρων (stumps) που θα φτιαχτούν. Η τιμή της συγκεκριμένης παραμέτρου πρέπει να επιλέγεται με προσοχή ούτως ώστε να μην καταλήξει το μοντέλο σε overfitting.

6. Gradient Boosting

Το Gradient Boosting ξεκινά από μία κοινή πρόβλεψη πιθανοτικού τύπου για κάθε εγγραφή του συνόλου δεδομένων. Στη συνέχεια προσπαθεί συνεχώς να βελτιώσει τα υπόλοιπα (residuals) της πρόβλεψης κάθε εγγραφής σε σχέση με την πραγματική τιμή η οποία είναι 0 για την κλάση 0 και 1 για την κλάση 1. Ως weak learner χρησιμοποιεί το απλό decision tree και για κάθε φύλλο του υπολογίζεται η ποσότητα:

$$\frac{\sum residuals}{\sum p(1-p)}$$

Όπου p η πιθανότητα που είχε υπολογιστεί για τη συγκεκριμένη εγγραφή στο προηγούμενο βήμα-επίπεδο του δένδρου.

Η ποσότητα αυτή στη συνέχεια περνάει μέσα από τη λογιστική συνάρτηση (logistic function) για να λάβει την πιθανοτικού τύπου πρόβλεψη. Η τελική πρόβλεψη διαμορφώνεται ως το άθροισμα όλων των προβλέψεων των weak learners πολλαπλασιασμένο επί το ρυθμό μάθησης (learning rate)

Το learning rate είναι η υπερπαράμετρος του Gradient Boosting η οποία καθορίζει το πόσο γρήγορα μαθαίνει το κάθε δένδρο. Καθορίζει δηλαδή το πόση επίδραση θα έχει κάθε δένδρο που προστίθεται στο μοντέλο στην τελική απόφαση του αλγορίθμου. Οι τιμές αυτής της υπερπαραμέτρου συνηθίζεται να είναι μεταξύ 0,1 και 0,3. Όσο μικρότερος βέβαια είναι ο ρυθμός μάθησης τόσο περισσότερο χρόνο θα κάνει το μοντέλο να εκπαιδευτεί. Ο ρυθμός μάθησης συνδυάζεται με μία άλλη υπερπαράμετρο του Gradient Boosting, το n_estimators, η οποία όπως στο Adaboost καθορίζει το πλήθος των δένδρων που θα φτιαχτούν [8]. Έτσι αν επιλεγεί χαμηλός ρυθμός μάθησης τότε θα πρέπει να επιλεγεί υψηλό n_estimators και αντίστροφα.

7. XGBoost

Ο XGBoost ανήκει και αυτός στις ensemble μεθόδους. Αποτελεί μια βελτίωση του gradient boosting και είναι ιδανικός για χρήση σε μεγάλα σύνολα δεδομένων καθώς είναι εξαιρετικά γρήγορος. Σε σχέση με το gradient boosting με τον οποίο μοιράζονται πολλά κοινά διαφοροποιείται στη χρήση μιας παραμέτρου κανονικοποίησης λ, η οποία έχει ρόλο να μειώσει τη διακύμανση και να αποτρέψει το overfitting μειώνοντας την πολυπλοκότητα του μοντέλου.

Ως base learner χρησιμοποιεί ένα δικό του δένδρο απόφασης, το XGBoost tree, το οποίο διαφέρει από ένα κλασικό δένδρο απόφασης στο κριτήριο με το οποίο διαχωρίζει τα δεδομένα από το ένα επίπεδο στο άλλο. Ο XGBoost χρησιμοποιεί ως κριτήριο το μέγεθος gain, το οποίο βασίζεται σε ένα άλλο μέγεθος την ομοιότητα (similarity) το οποίο με τη σειρά του βασίζεται στα υπόλοιπα (residuals) που αφήνει η πρόβλεψη σε σχέση με την πραγματική τιμή. Η πρόβλεψη που κάνει κάθε φορά ο XGBoost αφορά στην πιθανότητα κάθε εγγραφή να ανήκει στη μία ή την άλλη κλάση. Έτσι τελικά αν η υπολογισμένη πιθανότητα είναι <0,5 κατηγοριοποιεί την εγγραφή στην κλάση 0 διαφορετικά στην κλάση 1. Τα μεγέθη similarity και gain υπολογίζονται πάνω στα φύλλα του κάθε XGBoost tree.

Η ομοιότητα ορίζεται ως εξής:

$$Sim = \frac{(\sum residuals)^2}{\sum p(1-p) + \lambda}$$

Όπου p η πιθανότητα που είχε υπολογιστεί για τη συγκεκριμένη εγγραφή στο προηγούμενο βήμα- επίπεδο του δένδρου.

Το μέγεθος gain ορίζεται ως εξής:

$$GAIN=RIGHT_{Sim} + LEFT_{Sim} - ROOT_{Sim}$$

Όπου $RIGHT_{Sim}$ η ομοιότητα του δεξιού φύλλου, $LEFT_{Sim}$ του αριστερού και $ROOT_{Sim}$ του κόμβου στον οποίο ανήκουν. Το χαρακτηριστικό με το μεγαλύτερο GAIN επιλέγεται ως κόμβος και ακολουθεί με την ίδια λογική η παραγωγή του υπόλοιπου δέντρου. Το κάθε φύλλο αποδίδει την ομοιότητα ως τιμή εξόδου, η οποία στη συνέχεια περνάει μέσα από τη λογιστική συνάρτηση (logistic function) για να λάβει ο αλγόριθμος την πιθανοτικού τύπου πρόβλεψη όπως και στο Gradient Boosting.

Ο XGBoost έχει την υπερπαράμετρο scale_pos_weight η οποία αναθέτει βάρη στις κλάσεις ανάλογα με την κατανομή τους όπως το class_weight στο Random Forest [9].

8. Voting Classifier

Ο Voting Classifier αποτελεί μια περίπτωση ensemble μεθόδου. Σε αντίθεση με τις υπόλοιπες ensemble μεθόδους που χρησιμοποιούν ως base learner το decision tree, ο Voting Classifier χρησιμοποιεί οποιοδήποτε μοντέλο ως base learner και η απόφαση λαμβάνεται είτε δια της πλειοψηφίας (hard voting), όπως στην περίπτωση του random Forest, ενώ στην περίπτωση του soft voting η πρόβλεψη της κλάσης βασίζεται στο argmax των αθροισμάτων των προβλεφθεισών πιθανοτήτων. Υπάρχει δε η δυνατότητα να δοθεί διαφορετικό βάρος ως προς την επίδραση της ψήφου καθενός από τους base learners στην τελική απόφαση μέσω της υπερπαραμέτρου weights.

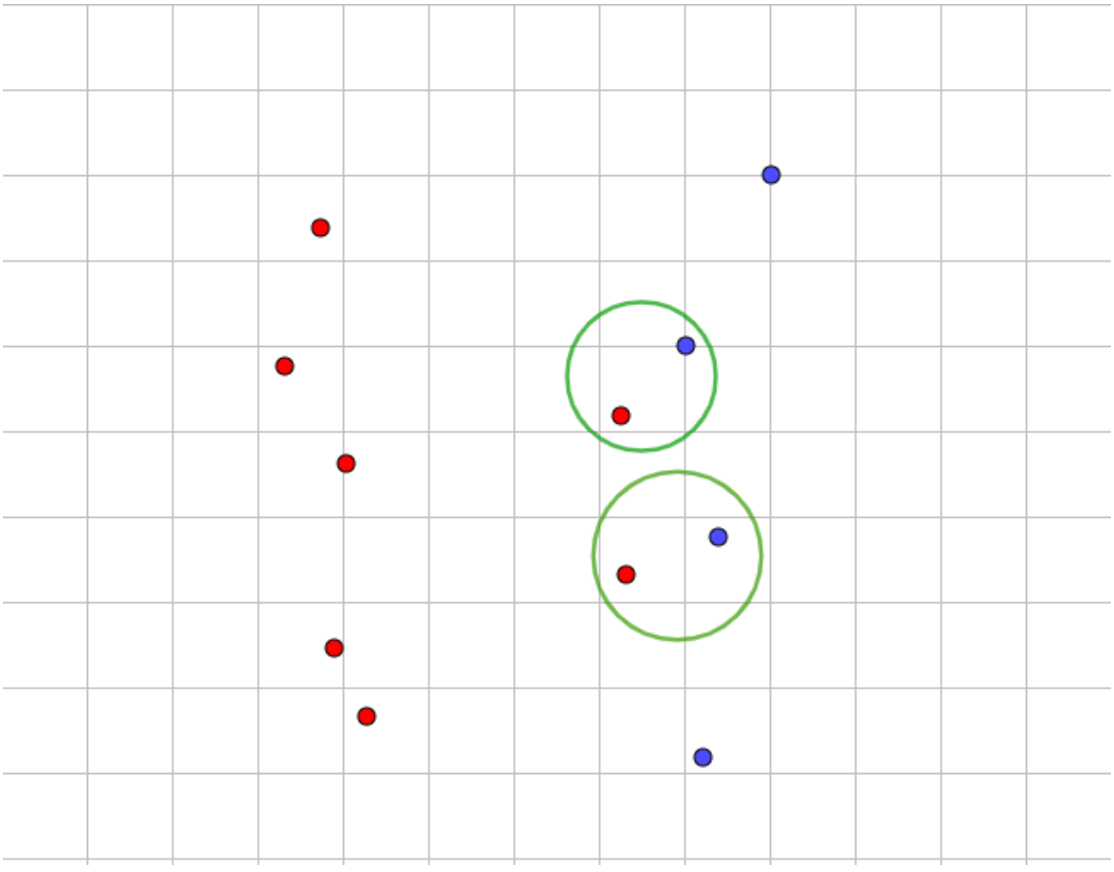
Στα πλαίσια της εργασίας ο Voting Classifier αξιολογήθηκε μόνο με cross validation χρησιμοποιώντας ως base learners τα 3 μοντέλα που απέδιδαν καλύτερα κάθε φορά.

4.3 Μεθοδολογίες εκπαίδευσης σε ανισοκατανεμημένες κλάσεις

4.3.1 Tomek links

Εκτός από την τεχνική του υπερδειγματισμού υπάρχουν αντίστοιχα τεχνικές υποδειγματισμού (undersampling). Στην περίπτωση αυτή αφαιρούνται παραδείγματα εκπαίδευσης από την πλειονοτική κλάση προκειμένου οι δύο κλάσεις να ισορροπήσουν αριθμητικά. Οι τεχνικές αυτές έχουν ως μειονέκτημα την απώλεια πληροφορίας από την πλειονοτική κλάση, μπορούν ωστόσο να φανούν χρήσιμες. Η πιο διαδεδομένη undersampling τεχνική είναι η Tomek Links.

Ως Tomek links καλούνται εκείνα τα ζεύγη παραδειγμάτων εκπαίδευσης που βρίσκονται σε κοντινή απόσταση μεταξύ τους αλλά ανήκουν σε διαφορετικές κλάσεις. Η τεχνική είναι απλή στη σύλληψη της και αφαιρεί από το training set τα παραδείγματα εκπαίδευσης της πλειονοτικής κλάσης που ανήκουν σε Tomek links. Η λογική πίσω από την τεχνική είναι ότι τα παραδείγματα αυτά είναι δύσκολο να προσφέρουν στον αλγόριθμο τις κατάλληλες πληροφορίες καθώς βρίσκονται κοντά στο σύνορο απόφασης και έτσι απορρίπτονται. [11]



Οπτική αναπαράσταση Tomek Links

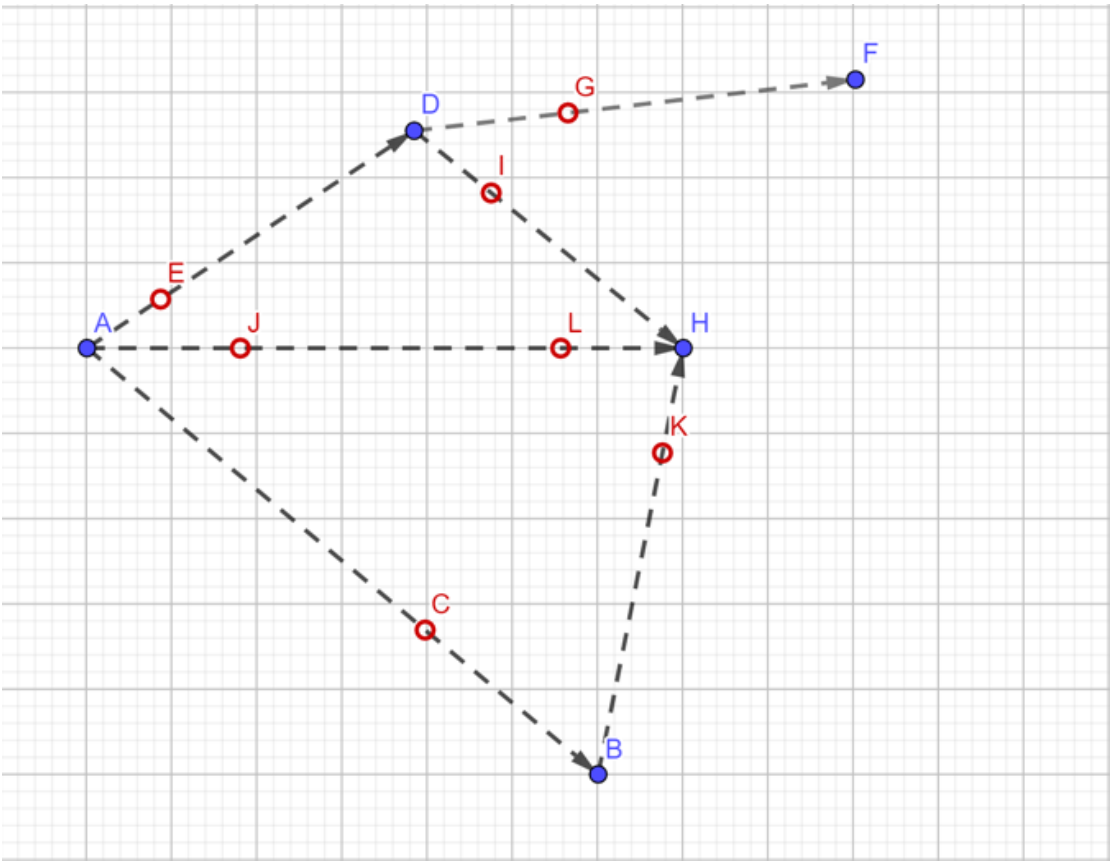
Στο παραπάνω σχήμα τα σημεία εντός των 2 πράσινων κύκλων αποτελούν Tomek links. Στην περίπτωση αυτή τα σημεία κόκκινου χρώματος θα αφαιρούντο από το σύνολο εκπαίδευσης

4.3.2 SMOTE

Το πρόβλημα που αντιμετωπίζει ένας αλγόριθμος μηχανικής μάθησης όταν υπάρχουν ανισοκατανεμημένες κλάσεις είναι ότι δεν υπάρχουν επαρκή παραδείγματα από την κλάση της μειονότητας προκειμένου να βρει το κατάλληλο σύνορο απόφασης. Μια τεχνική για να ξεπεραστεί αυτό το πρόβλημα είναι ο υπερδειγματισμός (oversampling) της μειονοτικής κλάσης. Ο υπερδειγματισμός αυτός γίνεται μέσω της δημιουργίας συνθετικών παραδειγμάτων εκπαίδευσης της μειονοτικής κλάσης. Τα παραδείγματα αυτά όμως δε θα παρέχουν καινούριες πληροφορίες στον αλγόριθμο μηχανικής μάθησης.

Μια από τις πιο ευρέως διαδεδομένες προσεγγίσεις για τη δημιουργία συνθετικών παραδειγμάτων είναι το SMOTE (Synthetic Minority Over-sampling Technique). Το SMOTE λειτουργεί ως εξής: Αρχικά διαλέγει ένα τυχαίο σημείο της μειονοτικής κλάσης έστω Α και βρίσκει τους k κοντινότερους γείτονες του, που επίσης ανήκουν στη μειονοτική κλάση, χρησιμοποιώντας την ευκλείδεια απόσταση. Στη συνέχεια διαλέγει τυχαία ένα από αυτά τα k κοντινότερα σημεία έστω Β και το ενώνει με μία γραμμή με το αρχικό σημείο Α. Το συνθετικό σημείο δημιουργείται ως κυρτός συνδυασμός των Α και Β. Κατ’ αυτό τον τρόπο μπορούν να δημιουργηθούν όσα σημεία της μειονοτικής κλάσης χρειάζεται ούτως ώστε να οι 2 κλάσεις να εξισορροπηθούν αριθμητικά.

Η αποτελεσματικότητα της μεθόδου έγκειται στο ότι τα συνθετικά παραδείγματα βρίσκονται κοντά στο χώρο χαρακτηριστικών της μειονοτικής κλάσης. Σε περιπτώσεις βέβαια που οι δύο κλάσεις υπερκαλύπτονται η αποτελεσματικότητα της μεθόδου μειώνεται καθώς τα συνθετικά παραδείγματα θα βρίσκονται κοντά τόσο στη μειονοτική όσο και στην πλειονοτική κλάση.



Γράφημα δημιουργίας συνθετικών σημείων με τη μέθοδο SMOTE

Στο παραπάνω γράφημα τα σημεία με το μωβ χρώμα αναπαριστούν παραδείγματα της μειονοτικής κλάσης και τα σημεία με το κόκκινο χρώμα είναι συνθετικά σημεία που κείνται επί των ευθυγράμμων τμημάτων που ορίζουν τα σημεία της μειονοτικής κλάσης.

4.3.3 Random oversampling

Η τεχνική του τυχαίου υπερδειγματισμού (random oversampling) περιλαμβάνει την αντιγραφή παραδειγμάτων της μειονοτικής κλάσης με τυχαίο τρόπο έως ότου ο πληθάριθμος των δύο κλάσεων ισορροπήσει. Τα ίδια παραδείγματα μπορούν να επιλεχθούν πολλαπλές φορές άρα έχουμε δειγματοληψία με επανάθεση.

Το random oversampling βοηθάει τους αλγόριθμους οι οποίοι επηρεάζονται από την ανισοκατανομή μεταξύ των δύο κλάσεων όπως νευρωνικά δίκτυα, SVMs και decision trees. Η αντιγραφή παραδειγμάτων της μειονοτικής κλάσης πολλαπλές φορές όμως έχει το μειονέκτημα ότι μπορεί να οδηγήσει τον αλγόριθμο σε overfitting, καθώς υπάρχει συνεχής αναπαραγωγής της ίδια πληροφορίας, και κατά συνέπεια δυνητικά σε σφάλμα γενίκευσης (generalization error) όταν κληθεί να προβλέψει πάνω νέα παραδείγματα.

5 Προβλέψεις

Η μεθοδολογία αξιολόγησης των προβλέψεων σε κάθε σενάριο είναι κοινή. Αρχικά αξιολογούνται όλοι οι ταξινομητές με ένα απλό cross validation και τα αποτελέσματα αυτά αποτελούν τα αποτελέσματα βάσης για τις ακόλουθες αξιολογήσεις των ταξινομητών μέσω των διαφορετικών μεθοδολογιών εκπαίδευσης.

Μέσω της εξαντλητικής μεθόδου του gridsearch δοκιμάζονται οι συνδυασμοί μεταξύ διαφορετικών υπερπαραμέτρων κάθε αλγορίθμου για ένα εύλογο κάθε φορά εύρος τιμών. Τα αποτελέσματα που παρατίθενται αφορούν κάθε φορά τον αλγόριθμο που απέδωσε την υψηλότερη ακρίβεια με τις ανάλογες τιμές των υπερπαραμέτρων του.

5.1 Χρησιμοποιώντας όλα τα διαθέσιμα δεδομένα

5.1.1 Αξιολόγηση μέσω cross validation

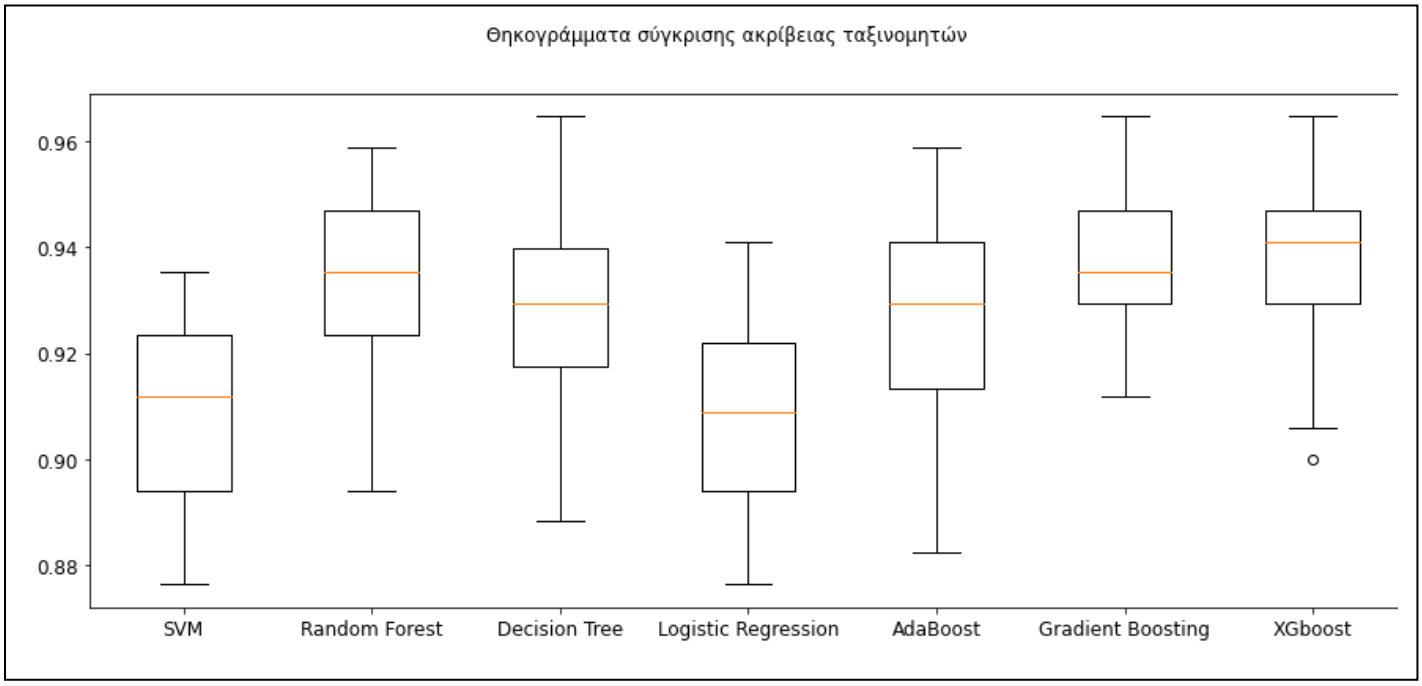
	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	90%	93%	90%	91%	93%	93%	94%
Precision	75%	98%	68%	76%	82%	90%	91%
Recall	56%	58%	67%	60%	69%	66%	67%
F1	0,63	0,72	0,67	0,66	0,75	0,76	0,77

Η πρώτη αξιολόγηση μέσω του cross validation δίνει αρκετά ενθαρρυντικά αποτελέσματα. Η ακρίβεια όλων των μοντέλων είναι άνω του 90% και στην περίπτωση του XGBoost φτάνει ως το 94%. Πέραν της ακρίβειας υπάρχει μια ξεκάθαρη διαφορά των Adaboost, Gradient Boosting και XGBoost στο f1 score σε σχέση με τους υπόλοιπους ταξινομητές όπου οι μεν πρώτοι πετυχαίνουν f1 score από 0,75 ως 0,77 ενώ αντίθετα οι υπόλοιποι είναι μεταξύ 0,63 και 0,67. Στο ενδιάμεσο βρίσκεται ο Random Forest με f1 score 0,72 και ακρίβεια 93%.

5.1.2 Αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost	Voting Classifier
Accuracy	91%	93%	93%	91%	93%	94%	94%	94%
Precision	81%	94%	97%	80%	83%	89%	89%	94%
Recall	56%	62%	56%	57%	69%	69%	69%	66%
F1	0,66	0,74	0,71	0,66	0,75	0,77	0,77	0,77

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost	Voting Classifier
Υπερ-παράμετροι	**C**=*0,16*	**class_weight**=*balanced_subsample* **max_depth**=*7*	**max_depth**=*4* **max_leaf_nodes**=*4*	**C**=*0,35* **Penalty**=*l1*	**n_estimators**=*100*	**n_estimators**=*150* **learning_rate**=*0,14*	**scale_pos_weight**=*2*	**Voting=***soft* **Weights=***(1, 1, 2)*



Βρίσκοντας τις βέλτιστες υπερπαραμέτρους υπάρχει μια βελτίωση σε αυτούς τους ταξινομητές οι οποίοι δεν πέτυχαν τόσο καλή ακρίβεια αρχικά. Πιο αξιοσημείωτη βελτίωση στην ακρίβεια είχε το decision tree όπου από το 90% ανήλθε στο 93%. Adaboost και Gradient Boosting ανέβασαν την ακρίβεια τους και το f1 score τους σε 94% και 0,77 αντίστοιχα επιτυγχάνοντας ίδια αποτελέσματα με τον XGBoost.

5.1.3 Χρήση της μεθόδου SMOTE και αξιολόγηση μέσω cross validation με ρύθμιση υπερπαραμέτρων

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	87%	93%	91%	88%	93%	94%	94%
Precision	58%	98%	80%	60%	81%	89%	91%
Recall	79%	61%	63%	79%	73%	72%	69%
F1	0,67	0,74	0,7	0,68	0,76	0,79	0,78

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Υπερ-παράμετροι	**C**=*0,38*	**class_weight**=*balanced_subsample* **max_depth**=*9*	**max_depth**=*8* **max_leaf_nodes**=10	**C**=*0,25* **Penalty**=*l1*	**n_estimators**=*110*	**n_estimators**=*190* **learning_rate**=*0,1*	**scale_pos_weight**=*1*

Η μέθοδος SMOTE αν και δε βελτιώνει την ακρίβεια των ταξινομητών, εντούτοις ανεβάζει το f1 score τους, ειδικά δε το f1 score του Gradient Boosting φτάνει στο 0,79 με τον XGBoost να ακολουθεί με 0,78.

5.1.4 Αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων με χρήση της μεθόδου Tomek links

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	91%	93%	93%	91%	93%	94%	94%
Precision	81%	93%	97%	82%	82%	98%	90%
Recall	55%	62%	56%	55%	69%	61%	67%
F1	0,65	0,74	0,71	0,66	0,74	0,75	0,76

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Υπερ-παράμετροι	**C**=*0,09*	**class_weight**=*balanced_subsample* **max_depth**=*7*	**max_depth**=4 **max_leaf_nodes**=4	**C**=*0,2* **Penalty**=*l1*	**n_estimators**=*50*	**n_estimators**=*50* **learning_rate**=*0,06*	**scale_pos_weight**=*1*

Με τη μέθοδο Tomek links παρατηρείται η αύξηση της ακρίβειας του decision tree από το 90% στο 93% και του f1 score στο 0,74 από 0,67. Οι υπόλοιποι ταξινομητές συνεχίζουν να αποδίδουν εξίσου καλά.

5.1.5 Αξιολόγηση των μοντέλων με τις βέλτιστες υπερπαραμέτρους μέσω cross validation και στάθμιση πιθανοτήτων

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	Xgboost
Accuracy	90%	93%	93%	90%	93%	94%	94%
Precision	86%	93%	92%	80%	90%	96%	93%
Recall	46%	64%	58%	54%	63%	63%	67%
F1	0,6	0,75	0,71	0,64	0,73	0,76	0,77

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	Xgboost
Μέθοδος στάθμισης	**Sigmoid cv=4**	**Isotonic cv=5**	**Sigmoid cv=5**	**Isotonic cv=4**	**Isotonic cv=5**	**Sigmoid cv=3**	**Isotonic cv=5**

Λαμβάνοντας τις σταθμισμένες πιθανότητες παρατηρείται άνοδος του f1 score του random Forest στο 0,75, αλλά και πτώση του f1 score του SVM στο 0,6. Στους υπόλοιπους ταξινομητές δεν παρατηρείται κάποια ιδιαίτερη μεταβολή.

5.1.6 Αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων με χρήση της μεθόδου random oversampling

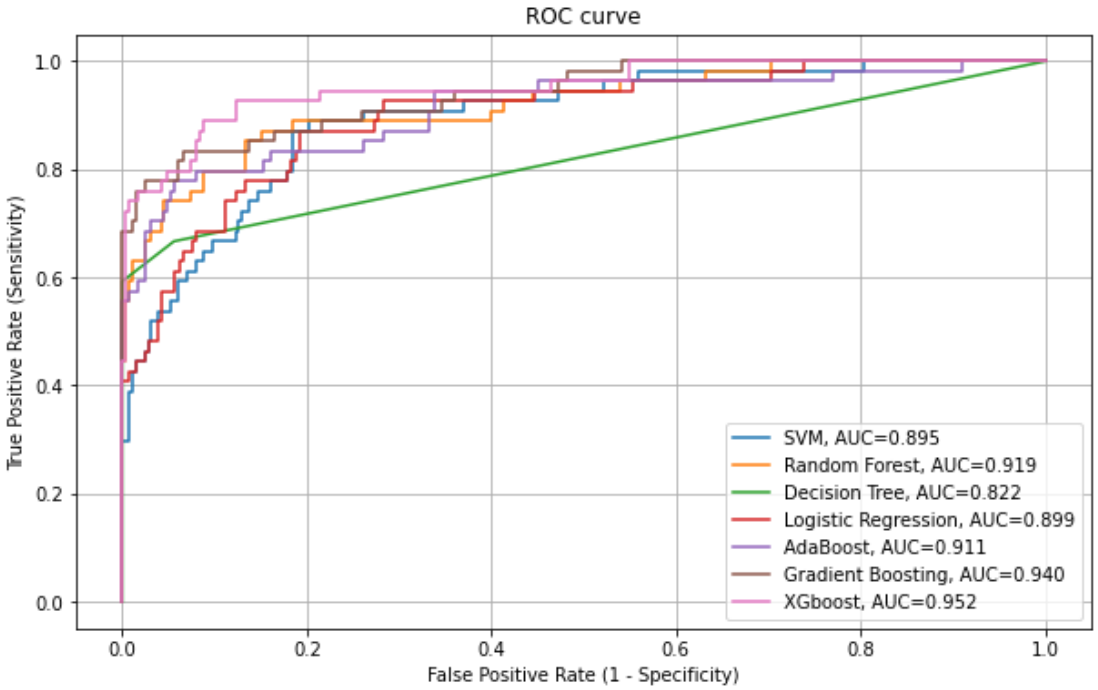
	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	86%	94%	90%	87%	91%	93%	93%
Precision	55%	92%	69%	58%	71%	81%	85%
Recall	80%	66%	66%	80%	75%	77%	71%
F1	0,65	0,76	0,67	0,67	0,73	0,78	0,77

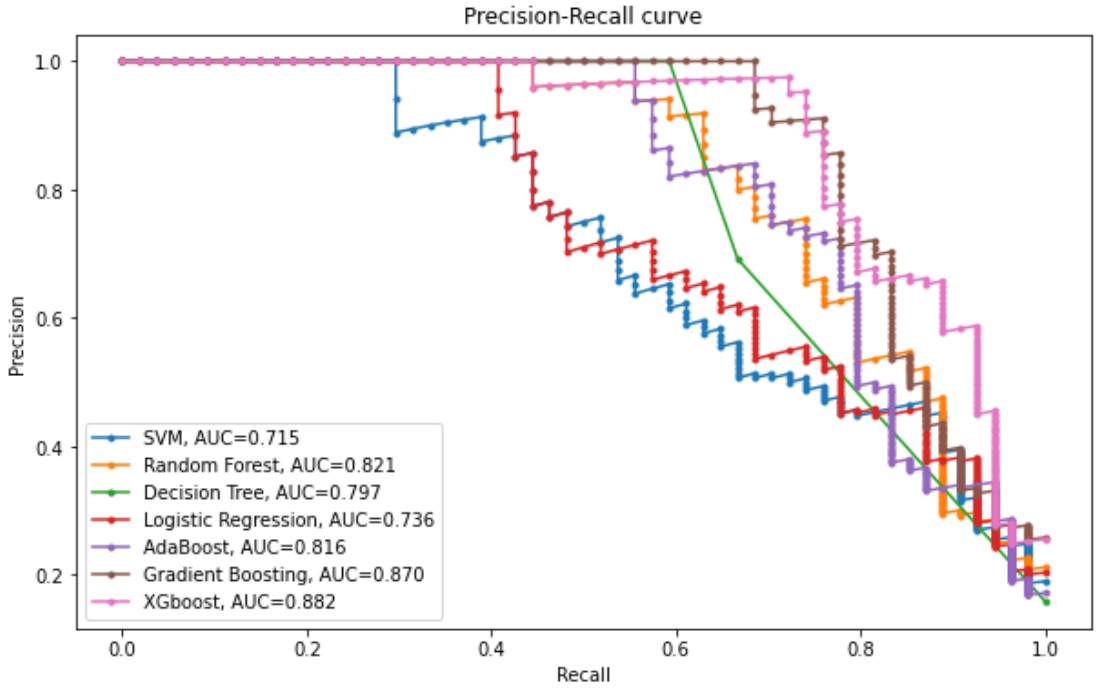
	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Υπερ-παράμετροι	**C**=*0,13*	**class_weight**=*balanced_subsample* **max_depth**=*8*	**max_depth**=4 **max_leaf_nodes**=3	**C**=*0,7* **Penalty**=*l1*	**n_estimators**=*500*	**n_estimators**=*350* **learning_rate**=*0,08*	**scale_pos_weight**=*1*

Μέσω του random oversampling ο random Forest επιτυγχάνει την καλύτερη ακρίβεια με 94%, η οποία είναι και η καλύτερη επίδοση αυτού του ταξινομητή συνολικά σε αυτό το σενάριο. Το καλύτερο f1 score έρχεται από το Gradient Boosting με 0,78

5.1.7 Αξιολόγηση με απλό διαχωρισμό holdout

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	89%	93%	94%	89%	93%	95%	95%
Precision	76%	89%	100%	70%	81%	91%	97%
Recall	52%	63%	59%	48%	70%	74%	70%
F1	0,62	0,74	0,74	0,57	0,75	0,82	0,82





Καταλήγοντας στην τελευταία και απλούστερη μέθοδο αξιολόγησης, τον απλό διαχωρισμό holdout, η ακρίβεια των Gradient Boosting, και XGBoost ανεβαίνει στο 95% και το f1 score τους στο 0,82. Οι επιδόσεις των μοντέλων αποτυπώνονται και στις καμπύλες ROC και precision-recall όπου και στις δύο επιβεβαιώνεται η ανωτερότητα των Gradient boosting και XGBoost έναντι των υπολοίπων ταξινομητών. Πιο συγκεκριμένα στην καμπύλη ROC ο XGBoost δημιουργεί το μεγαλύτερο εμβαδό κάτω από την καμπύλη (AUC) με 0,952, ενώ ο Gradient Boosting είναι αυτός που δημιουργεί το μεγαλύτερο εμβαδό στην καμπύλη precision-recall με 0,87

5.1.8 Επισκόπηση σεναρίου 1 και επιλογή μοντέλου

Τα μοντέλα δοκιμάζονται αρχικά σε ένα απλό 10-fold cross validation. Τα αποτελέσματα που προκύπτουν αποτελούν μία βάση σύγκρισης για την απόδοση κάθε αλγορίθμου πάνω στα δεδομένα. Η ακρίβεια όλων των ταξινομητών είναι τουλάχιστον 90%, ενώ σε μια πρώτη ματιά ξεχωρίζουν οι ταξινομητές που βασίζονται στο boosting -Adaboost, Gradient Boosting, XGBoost- καθώς και το random Forest, οι οποίοι αποδίδουν 93-94% ακρίβεια.

Καθώς η μεταβλητή στόχος είναι ανισοκατανεμημένη πρέπει να λαμβάνεται πάντα υπόψιν η επίδοση κάθε μοντέλου στην κλάση της μειονότητας. Πράγματι οι 4 προαναφερθέντες ταξινομητές πετυχαίνουν καλά f1 scores από 0,72 ο random Forest ως 0,77 ο XGBoost.

Περνώντας σε δεύτερο στάδιο στην εύρεση των βέλτιστων υπερπαραμέτρων υπάρχει μια μικρή βελτίωση στους υπόλοιπους ταξινομητές - SVM, logistic regression, decision tree - και έτσι έχουμε το μικρότερο ποσοστό ακρίβειας να ανέρχεται πλέον στο 91% καθώς επίσης και μια μικρή βελτίωση στα f1 scores τους.

Εξετάζοντας τις ειδικές τεχνικές εκπαίδευσης σε ανισοκατανεμημένες κλάσεις η τεχνική SMOTE ανέβασε τα f1 scores των Adaboost, Gradient Boosting, XGBoost με το f1 score του Gradient Boosting να φτάνει στο 0,79 διατηρώντας παράλληλα την ακρίβεια τους στο 94%. Αντίθετα στους SVM, decision tree και logistic regression υπήρξε πτώση της ακρίβειας.

Η τεχνικές Tomek links και random oversampling δεν προσέφεραν κάποια ουσιαστική βελτίωση ούτε στην ακρίβεια ούτε στο f1 score με εξαίρεση τον random Forest όπου με random oversampling απέδωσε ακρίβεια 94% και 0,76 f1 score που είναι και η καλύτερη επίδοση του συνολικά.

Τέλος η αξιολόγηση μέσω της απλής μεθόδου του holdout set αποδίδει ελαφρώς καλύτερα αποτελέσματα καθώς τα αποτελέσματα αυτής της μεθόδου εξαρτώνται από το πόσο “ευνοϊκός” θα είναι ο διαχωρισμός (split) για τα μοντέλα. Έτσι Gradient boosting και XGBoost αποδίδουν 95% ακρίβεια και 0,82 f1 score.

Αν έπρεπε να γίνει επιλογή μοντέλου και μεθοδολογίας εκπαίδευσης για το σενάριο αυτό τότε αυτό θα ήταν το Gradient Boosting το οποίο με SMOTE έδωσε 94% ακρίβεια και 0,79 f1 score.

5.2 Χρησιμοποιώντας τα δεδομένα ως την 3η ημέρα νοσηλείας

5.2.1 Αξιολόγηση μέσω cross validation

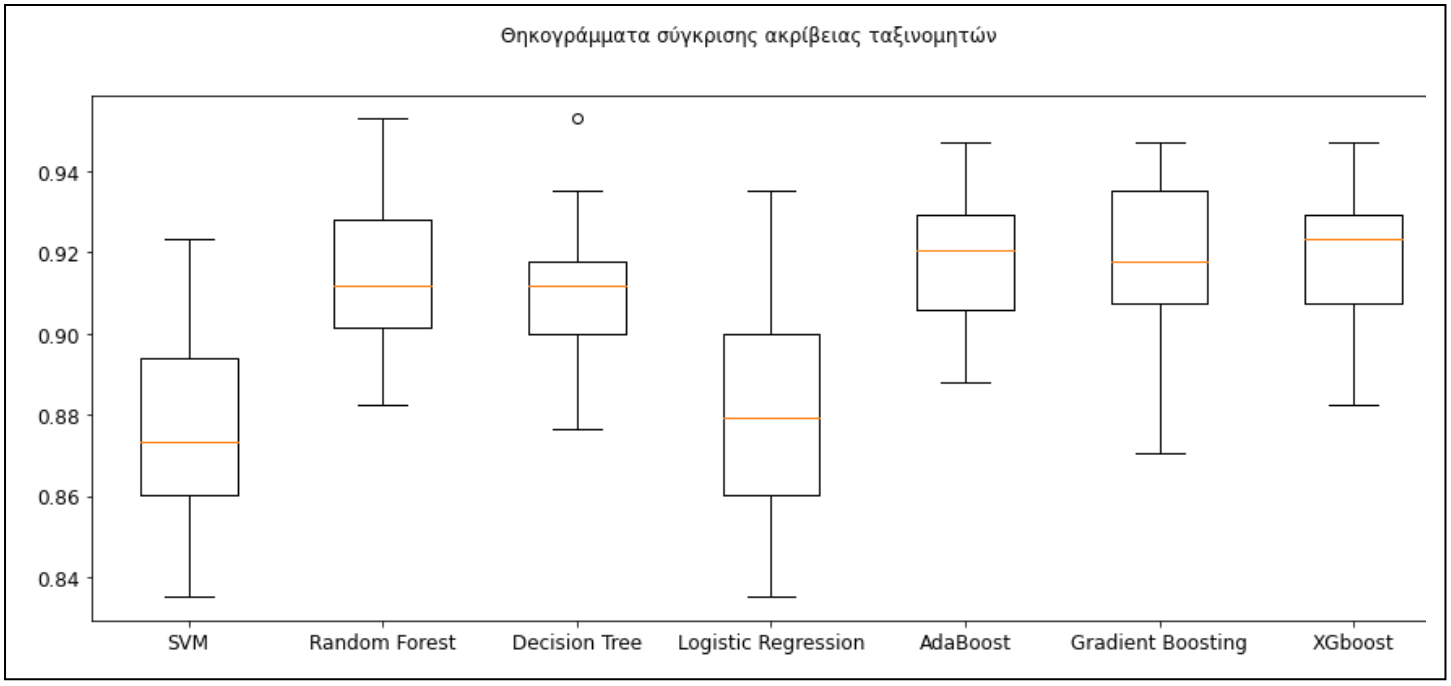
	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	88%	91%	86%	88%	91%	92%	92%
Precision	68%	98%	56%	66%	80%	86%	87%
Recall	42%	46%	60%	44%	60%	58%	59%
F1	0,51	0,62	0,57	0,52	0,69	0,69	0,7

Στο δεύτερο σενάριο της πειραματικής διαδικασίας κατά το οποίο οι αλγόριθμοι τροφοδοτούνται με 11 λιγότερα χαρακτηριστικά σε σχέση με το πρώτο οι επιδόσεις των μοντέλων παραμένουν σε υψηλά επίπεδα και πιο συγκεκριμένα Gradient boosting και XGBoost όπως και στο προηγούμενο σενάριο ξεκινούν με 92% ακρίβεια και fi score στο 0,69 και 0,7 αντίστοιχα. Χαμηλότερη απόδοση παρουσιάζει το decision tree, ενώ adaboost και random Forest έχουν ακρίβεια στο 91%.

5.2.2 Αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost	Voting Classifier
Accuracy	88%	91%	91%	88%	92%	92%	92%	92%
Precision	73%	93%	97%	72%	83%	85%	87%	90%
Recall	38%	50%	45%	41%	62%	60%	59%	58%
F1	0,5	0,65	0,61	0,52	0,7	0,7	0,7	0,7

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost	Voting Classifier
Υπερ-παράμετροι	**C**=*0,17*	**class_weight**=*balanced_subsample* **max_depth**=8	**max_depth**=*2* **max_leaf_nodes**=*3*	**C**=*0,35* **Penalty**=*l1*	**n_estimators**=*40*	**n_estimators**=*180* **learning_rate**=*0,09*	**scale_pos_weight**=*1*	**Voting**=hard **Weights**=(1, 2, 1)



Ρυθμίζοντας τις υπερπαραμέτρους των αλγορίθμων παρατηρείται σε όλους μια μικρή βελτίωση. Η ακρίβεια του Adaboost ανεβαίνει το 92%, ενώ του decision tree σημειώνει σημαντική άνοδο από το 86% στο 91%. Καθώς και στο προηγούμενο σενάριο παρατηρήθηκε στο decision tree η ίδια μεγάλη αύξηση της ακρίβειας μετά τη ρύθμιση των υπερπαραμέτρων συμπεραίνεται ότι η αρχική κακή του επίδοση οφείλεται σε overfitting το οποίο ξεπερνιέται μέσω του

ορισμού των τιμών των υπερπαραμέτρων max depth και max leaf nodes σε 2 και 3 αντίστοιχα.

5.2.3 Χρήση της μεθόδου SMOTE και αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	80%	92%	90%	82%	91%	92%	92%
Precision	43%	95%	72%	46%	73%	86%	85%
Recall	73%	50%	55%	75%	66%	63%	61%
F1	0,54	0,65	0,62	0,57	0,69	0,72	0,70

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Υπερ-παράμετροι	**C**=*0,25*	**class_weight**=*balanced_subsample* **max_depth**=*11*	**max_depth**=4 **max_leaf_nodes**=10	**C**=*0,3* **Penalty**=*l1*	**n_estimators**=*60*	**n_estimators**=*110* **learning_rate**=*0,17*	**scale_pos_weight**=*1*

Το SMOTE οδηγεί τους SVM και Logistic Regression σε πολύ χαμηλά επίπεδα ακρίβειας - 80% και 82% αντίστοιχα - σε σχέση με τα έως τώρα αποτελέσματα και αυτό είναι ένδειξη ότι οι 2 αυτοί αλγόριθμοι έκαναν underfit στο συνθετικό σύνολο δεδομένων που φτιάχνει το SMOTE. Από την άλλη μεριά ο Gradient Boosting μέσω του SMOTE πετυχαίνει f1 score 0,72 και ακρίβεια 92% που είναι ο καλύτερος συνδυασμός για αυτή την τεχνική.

5.2.4 Αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων με χρήση της μεθόδου Tomek links

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	88%	91%	91%	88%	91%	92%	92%
Precision	73%	90%	97%	71%	80%	85%	85%
Recall	40%	50%	45%	43%	61%	59%	61%
F1	0,51	0,66	0,61	0,53	0,69	0,69	0,7

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Υπερ-παράμετροι	**C**=*0,09*	**class_weight**=*balanced_subsample* **max_depth**=*7*	**max_depth**=*2* **max_leaf_nodes**=*3*	**C**=*0,4* **Penalty**=*l1*	**n_estimators**=*40*	**n_estimators**=*110* **learning_rate**=*0,1*	**scale_pos_weight**=*2*

Η μέθοδος Tomek links δε φαίνεται να προσθέτει κάποια αξιοσημείωτη βελτίωση σε σχέση με το cross validation. Η ακρίβεια των καλύτερων ταξινομητών παραμένει στα ίδια υψηλά επίπεδα.

5.2.5 Αξιολόγηση των μοντέλων με τις βέλτιστες υπερπαραμέτρους μέσω cross validation και στάθμιση πιθανοτήτων

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	87%	91%	91%	88%	92%	92%	92%
Precision	77%	92%	95%	75%	88%	91%	92%
Recall	30%	50%	44%	36%	55%	54%	55%
F1	0,42	0,64	0,6	0,48	0,67	0,68	0,68

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Μέθοδος στάθμισης	**Sigmoid cv=5**	**Isotonic cv=4**	**Sigmoid cv=3**	**Isotonic cv=4**	**Isotonic cv=5**	**Sigmoid cv=4**	**Sigmoid cv=4**

Τα σταθμισμένα μοντέλα αποδίδουν εν πολλοίς ίδια αποτελέσματα με το cross validation

5.2.6 Αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων με χρήση της μεθόδου random oversampling

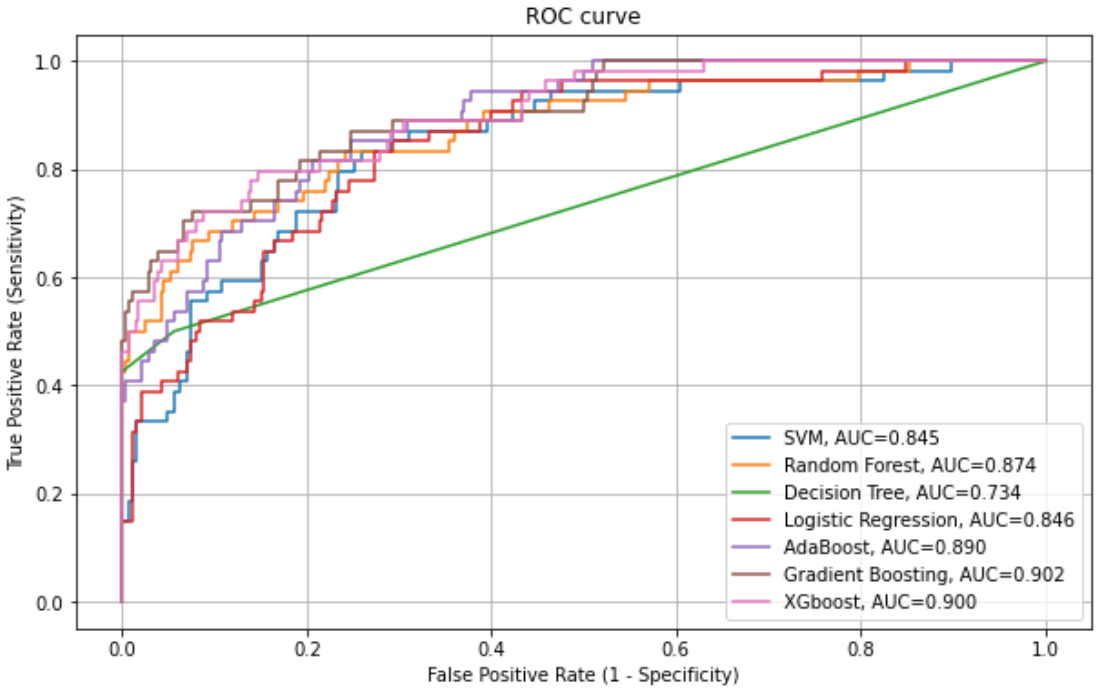
	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	81%	92%	86%	82%	88%	92%	92%
Precision	44%	95%	58%	46%	61%	79%	78%
Recall	74%	51%	62%	73%	70%	67%	66%
F1	0,55	0,66	0,6	0,56	0,66	0,72	0,71

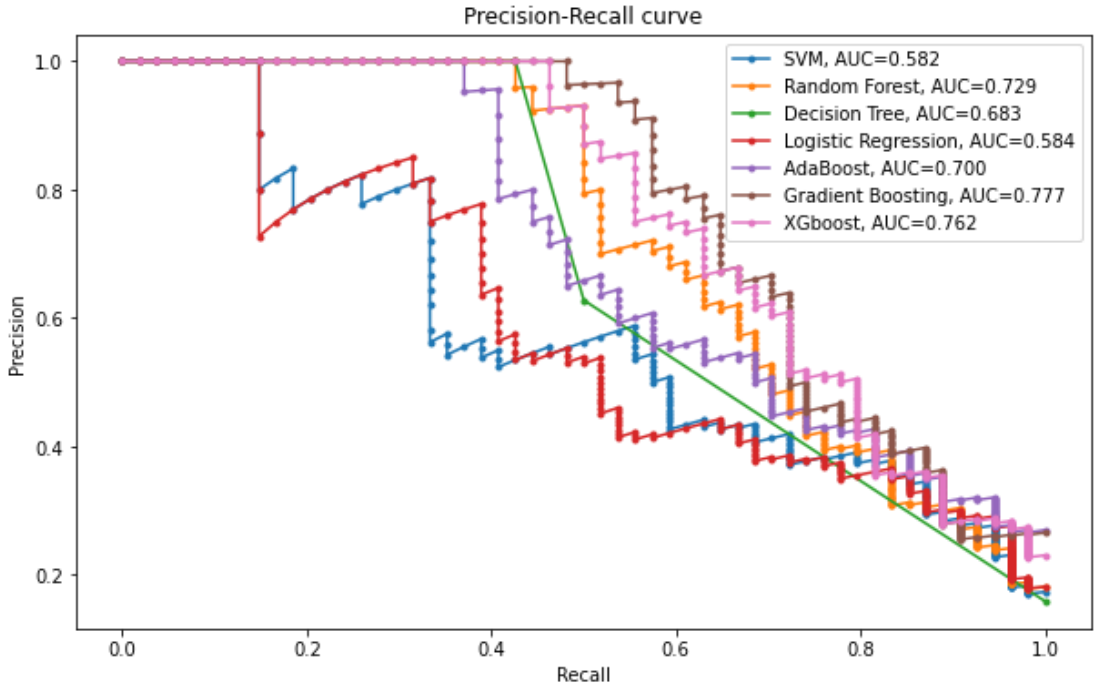
	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Υπερ-παράμετροι	**C**=*0,11*	**class_weight**=*balanced* **max_depth**=*12*	**max_depth**=10 **max_leaf_nodes**=3	**C**=*0,55* **Penalty**=*l2*	**n_estimators**=*400*	**n_estimators**=*500* **learning_rate**=*0,17*	**scale_pos_weight**=4

Η μέθοδος random oversampling δεν οδηγεί σε κάποια αξιοσημείωτη άνοδο της ακρίβειας, επιτυγχάνεται όμως βελτίωση του f1 score των Gradient Boosting και XGBoost στο 0,72 και 0,71 αντίστοιχα.

5.2.7 Αξιολόγηση με απλό διαχωρισμό holdout

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	86%	91%	91%	87%	88%	92%	91%
Precision	60%	93%	100%	68%	65%	91%	79%
Recall	33%	46%	43%	39%	48%	56%	56%
F1	0,43	0,62	0,6	0,49	0,55	0,69	0,65





Στην αξιολόγηση μέσω του holdout set οι επιδόσεις των μοντέλων εν πολλοίς δε διαφοροποιούνται. Έτσι έχουμε το Gradient Boosting να δίνει ακρίβεια 92% και decision tree, random Forest και XGBoost να έπονται με 91%. Οι καμπύλες ROC και precision-recall επιβεβαιώνουν τα ευρήματα με το Gradient Boosting να δημιουργεί εμβαδό 0,902 στην πρώτη και 0,777 στη δεύτερη.

5.2.8 Επισκόπηση σεναρίου 2 και επιλογή μοντέλου

Στο δεύτερο κατά σειρά πρόβλημα μηχανικής μάθησης που εξετάζεται οι αλγόριθμοι καλούνται να εκπαιδευτούν σε ένα μικρότερο σύνολο χαρακτηριστικών σε σχέση με το πρώτο. Οι διαθέσιμες πληροφορίες σε αυτό το σενάριο σταματούν στην τρίτη ημέρα νοσηλείας. Εύλογο αποτέλεσμα αυτού του γεγονότος είναι η μικρή πτώση της ακρίβειας. Παρ' όλα αυτά οι ταξινομητές επιτυγχάνουν ποσοστά ακρίβειας που κινούνται σταθερά πάνω από το ποσοστό της πλειοψηφικής κλάσης (84%) αποδεικνύοντας έτσι ότι διαθέτουν ικανότητα πρόβλεψης και δεν είναι απλά αφελείς ταξινομητές (dummy classifiers) που επιτυγχάνουν την ακρίβεια τους μοναχά κατηγοριοποιώντας όλες τις εγγραφές ως εγγραφές της πλειονοτικής κλάσης.

Το 10 fold cross validation δίνει και πάλι τη βάση πάνω στην οποία θα μετρηθούν οι επιδόσεις των διαφορετικών τεχνικών και παραμετροποιήσεων. Σε αυτό ξεχωρίζουν ξανά οι XGBoost και gradient boosting με 92% ακρίβεια και f1 score 0,7 και 0,69 αντίστοιχα. Προχωρώντας στην εύρεση των καλύτερων υπερπαραμέτρων για τα μοντέλα παρατηρείται αξιοσημείωτη βελτίωση μόνο στο decision tree του οποίου η ακρίβεια ανεβαίνει από το 86% στο 91%. SMOTE, Tomek links και random oversampling δε φαίνεται να βοηθάνε τα μοντέλα ιδιαίτερα να ανεβάσουν τις επιδόσεις τους με εξαίρεση το gradient boosting όπου και με SMOTE και με random oversampling πετυχαίνει 92% ακρίβεια και 0,72 f1 score. Οι καμπύλες ROC και Precession-Recall επιβεβαιώνουν τα αρχικά συμπεράσματα σε σχέση με την αξιολόγηση των μοντέλων, ότι δηλαδή Gradient Boosting και XGBoost πραγματοποιούν σταθερά καλύτερες προβλέψεις έναντι των υπόλοιπων.

Στην επιλογή του καλύτερου μοντέλου και τεχνικής εκπαίδευσης για αυτό το σενάριο θα μπορούσαν να γίνουν δύο επιλογές οι οποίες έχουν τις ίδιες επιδόσεις. Η μια είναι το SMOTE και η δεύτερη το random oversampling οι οποίες συνδυαζόμενες αμφότερες με το gradient boosting δίνουν 92% ακρίβεια και 0,72 f1 score.

5.3 Χρησιμοποιώντας τα διαθέσιμα δεδομένα ως τη 2η μέρα νοσηλείας

5.3.1 Αξιολόγηση μέσω cross validation

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	88%	90%	84%	87%	90%	90%	90%
Precision	69%	94%	50%	67%	75%	82%	81%
Recall	42%	39%	53%	43%	56%	49%	52%
F1	0,52	0,55	0,51	0,52	0,63	0,61	0,63

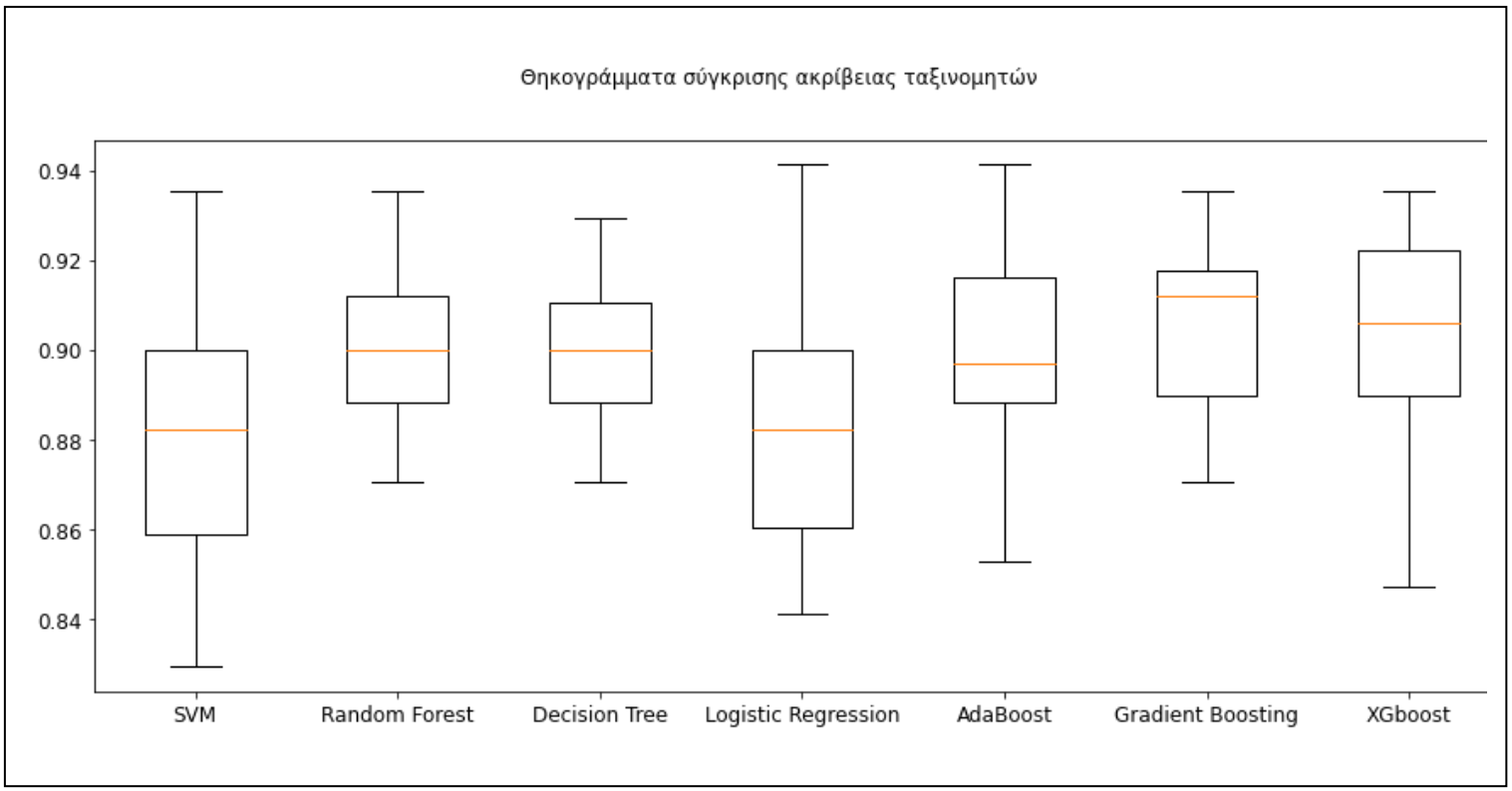
Από το cross validation ξεχωρίζουν οι AdaBoost, Gradient Boosting και XGboost με 90% ακρίβεια, ενώ στα ίδια επίπεδα ακρίβειας κυμαίνεται και ο Random Forest ο οποίος όμως αποδίδει κακό precision στο 39% με αποτέλεσμα και το f1 score του να είναι στο 0,55, ενώ τα 3 άλλα μοντέλα έχουν f1 score μεταξύ 0,61 και 0,63. Κακή αρχική επίδοση δείχνει να έχει το decision tree με ακρίβεια μόλις στο 84%.

5.3.2 Αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost	Voting Classifier
Accuracy	88%	90%	90%	88%	90%	91%	90%	91%
Precision	75%	90%	95%	74%	76%	89%	78%	86%
Recall	37%	42%	38%	39%	56%	46%	57%	52%
F1	0,49	0,57	0,53	0,5	0,64	0,61	0,65	0,64

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost	Voting Classifier
Υπερ-παράμετροι	**C**=*0,12*	**class_weight**=*balanced* **max_depth**=*9*	**max_depth**=*2* **max_leaf_nodes**=3	**C**=*0,2* **Penalty**=*l1*	**n_estimators**=*90*	**n_estimators**=*60* **learning_rate**=*0,1*	**scale_pos_weight**=*4*	**Voting**=*hard* **Weights**=*(1, 1, 1)*

Η ρύθμιση των υπερπαραμέτρων βελτιώνει την ακρίβεια του Gradient Boosting στο 91% ,ενώ ίδια ακρίβεια επιτυγχάνει και ο Voting Classifier. Το καλύτερο f1 score αποδίδει ο XGboost με 0,65. Αξιοσημείωτη είναι η άνοδος της ακρίβειας του decision tree από το 84% στο 90% το οποίο όμως αποδίδει χαμηλό recall στο 38%. Σε χαμηλά επίπεδα κινούνται τα f1 scores και των SVM, Random Forest, Logistic Regression.



5.3.3 Χρήση της μεθόδου SMOTE και αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	80%	92%	90%	81%	90%	91%	91%
Precision	44%	87%	92%	46%	71%	82%	81%
Recall	74%	47%	39%	73%	60%	54%	55%
F1	0,55	0,60	0,54	0,56	0,64	0,65	0,65

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Υπερ-παράμετροι	**C**=*0,28*	**class_weight**=*balanced_subsample* **max_depth**=*7*	**max_depth**=9 **max_leaf_nodes**=5	**C**=*0,4* **Penalty**=*l1*	**n_estimators**=*150*	**n_estimators**=*110* **learning_rate**=*0,9*	**scale_pos_weight**=*1*

Η μέθοδος SMOTE ανεβάζει την ακρίβεια του XGBoost στο 91% και το f1 score του στο 0,65 και μαζί με το Gradient Boosting αποδίδουν τις καλύτερες επιδόσεις. SVM και Logistic Regression έχουν χαμηλές επιδόσεις και φαίνεται πως δεν ανταποκρίνονται στη συγκεκριμένη μεθοδολογία εκπαίδευσης.

5.3.4 Αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων με χρήση της μεθόδου Tomek links

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	88%	90%	90%	88%	90%	91%	91%
Precision	75%	84%	95%	73%	75%	78%	82%
Recall	39%	45%	38%	40%	57%	56%	53%
F1	0,51	0,58	0,54	0,51	0,64	0,65	0,64

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Υπερ-παράμετροι	**C**=*0,1*	**class_weight**=*balanced_subsample* **max_depth**=*8*	**max_depth**=2 **max_leaf_nodes**=3	**C**=*0,2* **Penalty**=*l1*	**n_estimators**=*70*	**n_estimators**=*190* **learning_rate**=*0,14*	**scale_pos_weight**=1

Με τη μέθοδο Tomek links και πάλι ο Gradient Boosting επιτυγχάνει τις καλύτερες επιδόσεις με 91% ακρίβεια και 0,65 f1 score με τον XGBoost να ακολουθεί με ίδια ακρίβεια και f1 score 0,64.

5.3.5 Αξιολόγηση των μοντέλων με τις βέλτιστες υπερπαραμέτρους μέσω cross validation και στάθμιση πιθανοτήτων

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	88%	90%	89%	88%	90%	91%	91%
Precision	78%	86%	84%	77%	87%	90%	89%
Recall	31%	47%	31%	36%	47%	47%	48%
F1	0,44	0,60	0,45	0,48	0,60	0,61	0,62

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Μέθοδος στάθμισης	**Sigmoid cv=5**	**Isotonic cv=5**	**Isotonic cv=5**	**Sigmoid cv=5**	**Isotonic cv=4**	**Sigmoid cv=4**	**Sigmoid cv=4**

Η ακρίβεια των μοντέλων με τις σταθμισμένες πιθανότητες είναι στο 91% για Gradient Boosting και XGBoost, ενώ υψηλή ακρίβεια παρουσιάζουν και τα υπόλοιπα μοντέλα από 88% ως 90%. Το καλύτερο f1 score αποδίδει ο XGBoost με 0,62.

5.3.6 Αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων με χρήση της μεθόδου random oversampling

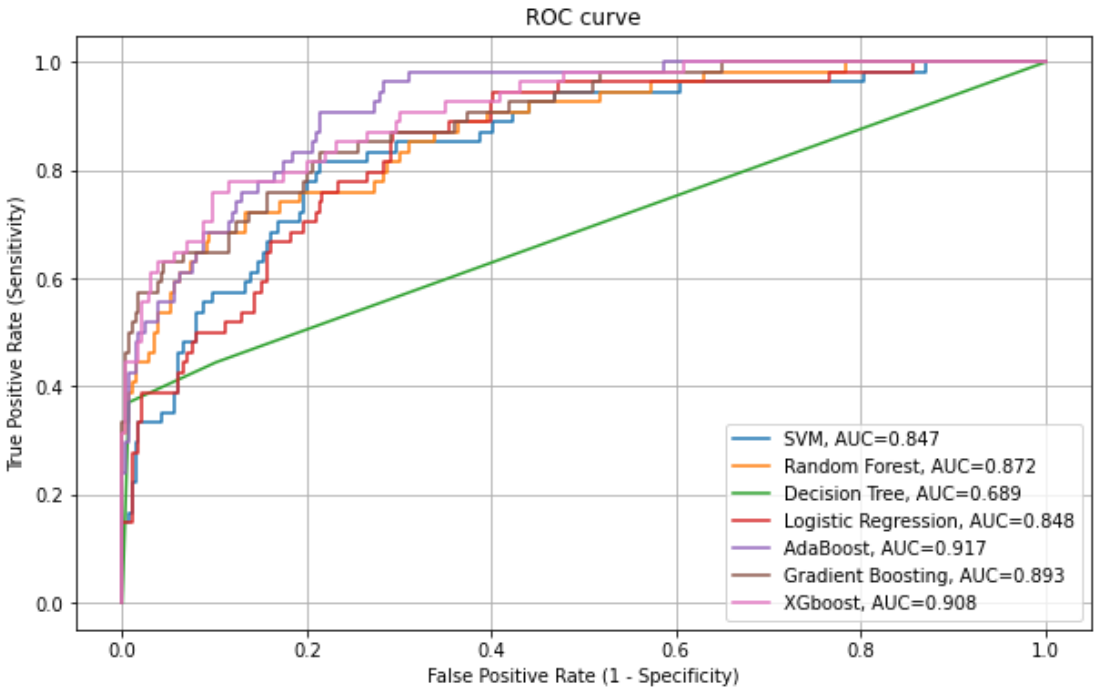
	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	80%	90%	78%	82%	87%	90%	90%
Precision	44%	86%	45%	46%	59%	72%	76%
Recall	74%	46%	55%	74%	68%	65%	57%
F1	0,55	0,59	0,46	0,56	0,63	0,65	0,65

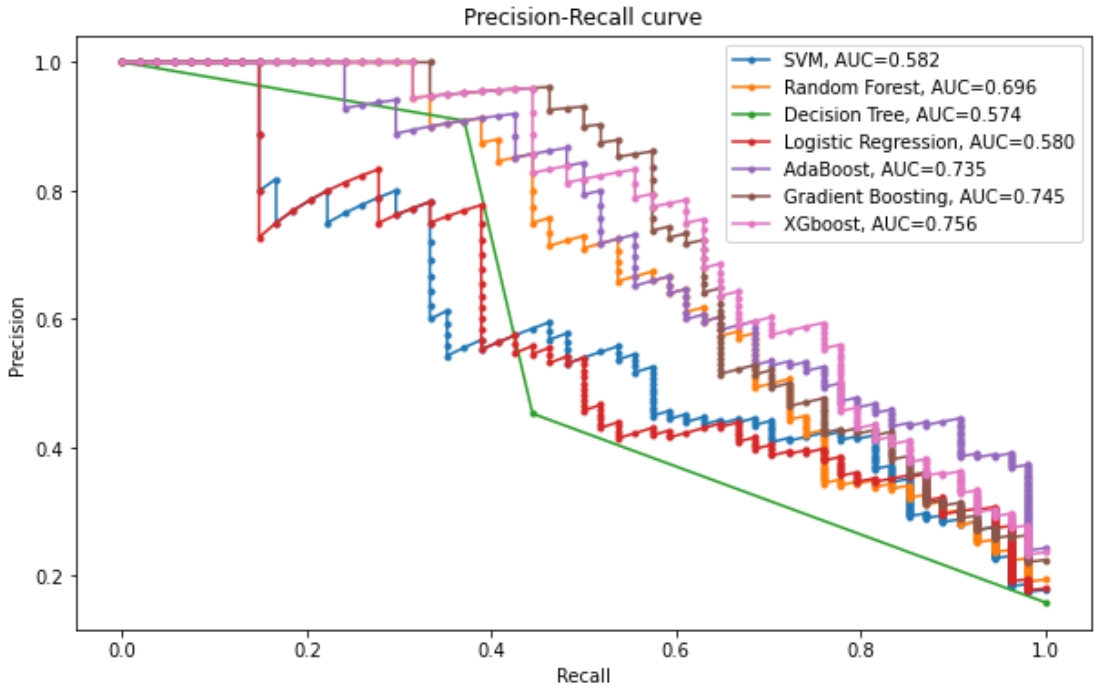
	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Υπερ-παράμετροι	**C**=*0,11*	**class_weight**=*balanced_subsample* **max_depth**=*12*	**max_depth**=8 **max_leaf_nodes**=3	**C**=*0,5* **Penalty**=*l1*	**n_estimators**=*500*	**n_estimators**=*500* **learning_rate**=*0,11*	**scale_pos_weight**=1

Το random oversampling βοηθάει τους Gradient Boosting και XGBoost να επιτύχουν ακρίβεια 90% και f1 score 0,65. Ιδιαίτερα κακή επίδοση εμφανίζει το decision tree με 78% ακρίβεια.

5.3.7 Αξιολόγηση με απλό διαχωρισμό holdout

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	86%	89%	89%	88%	89%	91%	91%
Precision	64%	80%	91%	72%	71%	88%	79%
Recall	33%	44%	37%	39%	56%	54%	56%
F1	0,44	0,57	0,53	0,51	0,63	0,67	0,65





Μέσω του απλού διαχωρισμού με holdout set εξακολουθεί ο Gradient Boosting να αποδίδει τα καλύτερα αποτελέσματα με ακρίβεια 91% και f1 score 0,67 και ο XGBoost ακολουθεί με 91% ακρίβεια και πάλι και 0,65 f1 score. Στην καμπύλη ROC μεγαλύτερο εμβαδό δημιουργεί ο Adaboost 0,917 και ακολουθεί ο XGBoost με 0,908. Στη δε καμπύλη precision-recall ο XGBoost είναι αυτός που δημιουργεί το μεγαλύτερο εμβαδό με 0,756.

5.3.8 Επισκόπηση σεναρίου 3 και επιλογή μοντέλου

Καθώς στο τρίτο σενάριο μειώνονται ακόμα περισσότερο τα διαθέσιμα δεδομένα επέρχεται και η αδυναμία των μοντέλων να κάνουν σωστές προβλέψεις πάνω στη μειονοτική κλάση. Παρά το ότι η ακρίβεια παραμένει υψηλή από 88% ως 91% τα f1 scores κυμαίνονται κάτω από 0,6 για τα περισσότερα μοντέλα με εξαίρεση τους 3 αλγορίθμους που βασίζονται στο boosting οι οποίοι αποδίδουν σταθερά f1 scores 0,6 ως 0,65.

Ως επιλογή μοντέλου για αυτό το σενάριο θα μπορούσε να είναι είτε ο Gradient Boosting εκπαιδευόμενος με SMOTE ή Tomek links, είτε ο XGboost εκπαιδευόμενος με SMOTE. Και στις 3 περιπτώσεις η ακρίβεια είναι 91% και το f1 score 0,65.

5.4 Χρησιμοποιώντας τα διαθέσιμα δεδομένα ως τη 1η μέρα νοσηλείας

5.4.1 Αξιολόγηση μέσω cross validation

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	88%	87%	81%	88%	88%	89%	89%
Precision	70%	85%	42%	67%	68%	76%	76%
Recall	43%	23%	43%	44%	50%	42%	48%
F1	0,52	0,36	0,42	0,52	0,57	0,54	0,58

Στο τέταρτο κατά σειρά εξεταζόμενο σενάριο η ακρίβεια των περισσότερων ταξινομητών κυμαίνεται από 87% ως 89%. Το decision tree παρουσιάζει όπως και σε προηγούμενο σενάριο ιδιαίτερα χαμηλή ακρίβεια λόγω overfitting. Τα f1 scores όλων των μοντέλων είναι χαμηλά.

5.4.2 Αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost	Voting Classifier
Accuracy	88%	88%	86%	88%	89%	89%	89%	89%
Precision	75%	74%	87%	74%	72%	75%	76%	75%
Recall	38%	42%	15%	40%	49%	46%	48%	49%
F1	0,5	0,53	0,27	0,51	0,58	0,56	0,58	0,59

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost	Voting Classifier
Υπερ-παράμετροι	**C**=*0,15*	**class_weight**=*balanced* **max_depth**=*8*	**max_depth**=*2* **max_leaf_nodes**=2	**C**=*0,3* **Penalty**=*l1*	**n_estimators**=*30*	**n_estimators**=*110* **learning_rate**=*0,13*	**scale_pos_weight**=*1*	**Voting=***soft* **Weights=***(1, 1, 2)*

Σε αυτό το σημείο γίνεται αντιληπτό ότι μόνο το μοντέλα που βασίζονται στο boosting αποδίδουν επαρκώς στο εξεταζόμενο σενάριο και έτσι η από εδώ και πέρα πειραματική διαδικασία θα συνεχιστεί μόνο με αυτούς τους τρεις ταξινομητές.

5.4.3 Χρήση της μεθόδου SMOTE και αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων

	AdaBoost	Gradient Boosting	XGboost
Accuracy	88%	89%	89%
Precision	63%	72%	70%
Recall	55%	53%	54%
F1	0,59	0,61	0,60

	AdaBoost	Gradient Boosting	XGboost
Υπερ-παράμετροι	**n_estimators**=*180*	**n_estimators**=*140* **learning_rate**=*0,16*	**scale_pos_weight**=3

Αν και η ακρίβεια δε βελτιώνεται, η μέθοδος SMOTE ανεβάζει ελαφρώς το f1 score και στα 3 μοντέλα σε σχέση με τα αποτελέσματα που απέδωσε το cross validation και ιδιαιτέρως στο gradient boosting υπάρχει η μεγαλύτερη βελτίωση όπου το f1 score από 0,54 ανέρχεται στο 0,61.

5.4.4 Αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων με χρήση της μεθόδου Tomek links

	AdaBoost	Gradient Boosting	XGboost
Accuracy	88%	89%	89%
Precision	67%	73%	72%
Recall	52%	48%	53%
F1	0,58	0,58	0,61

	AdaBoost	Gradient Boosting	XGboost
Υπερ-παράμετροι	**n_estimators**=*150*	**n_estimators**=*170* **learning_rate**=*0,1*	**scale_pos_weight**=*2*

Όπως και στην περίπτωση της μεθόδου SMOTE δεν παρατηρείται κάποια διαφοροποίηση στην ακρίβεια σε κανένα από τα μοντέλα. Η μέθοδος Tomek links βοηθά όμως τον αλγόριθμο XGboost να αυξήσει το f1 score του στο 0,61.

5.4.5 Αξιολόγηση των μοντέλων με τις βέλτιστες υπερπαραμέτρους μέσω cross validation και στάθμιση πιθανοτήτων

	AdaBoost	Gradient Boosting	XGboost
Accuracy	89%	89%	89%
Precision	81%	73%	77%
Recall	39%	48%	46%
F1	0,51	0,54	0,57

	AdaBoost	Gradient Boosting	XGboost
Μέθοδος στάθμισης	**Isotonic cv=4**	**Sigmoid cv=5**	**Isotonic cv=5**

Οι σταθμισμένες πιθανότητες δεν αλλάζουν τίποτα σε σχέση με την αρχική αξιολόγηση μέσω cross validation

5.4.6 Αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων με χρήση της μεθόδου random oversampling

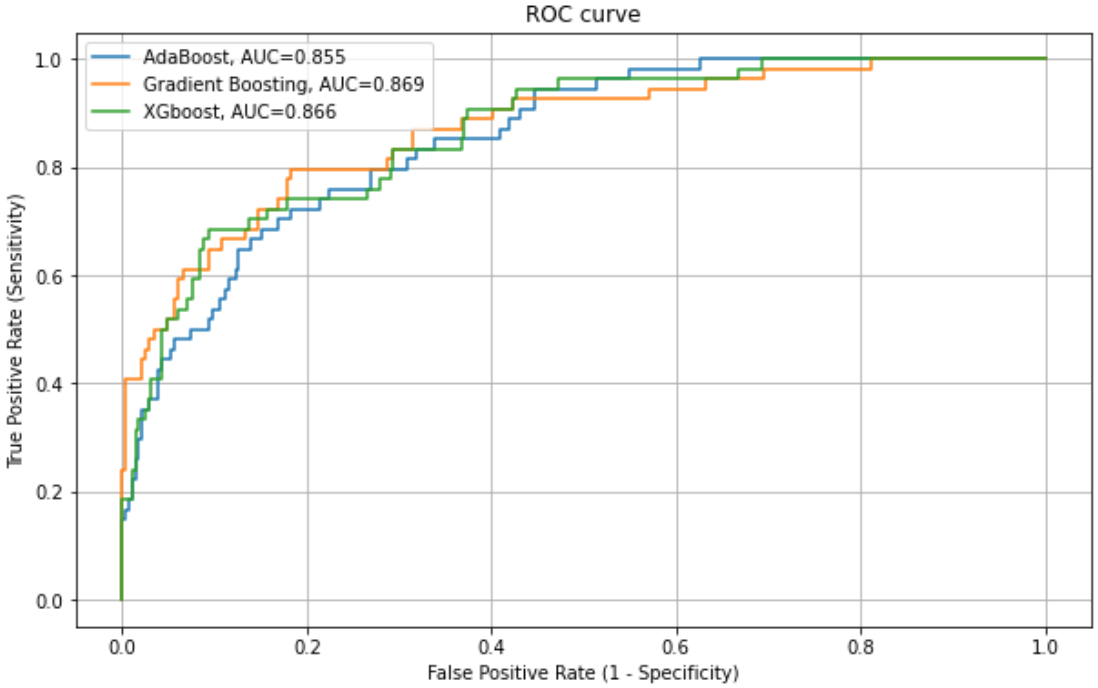
	AdaBoost	Gradient Boosting	XGboost
Accuracy	85%	88%	89%
Precision	53%	65%	59%
Recall	66%	59%	54%
F1	0,58	0,62	0,60

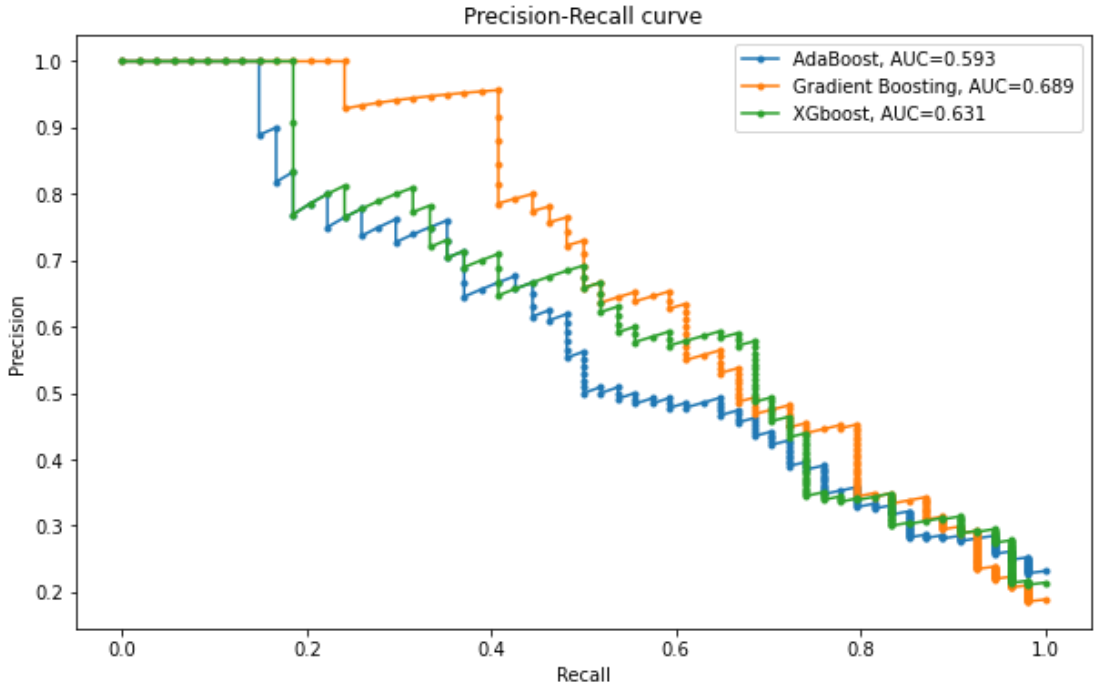
	AdaBoost	Gradient Boosting	XGboost
Υπερ-παράμετροι	**n_estimators**=*400*	**n_estimators**=*400* **learning_rate**=*0,19*	**scale_pos_weight**=*2*

Μέσω του random oversampling πετυχαίνει ο gradient boosting f1 score 0,62 το οποίο είναι και το καλύτερο ως τώρα για αυτό το σενάριο.

5.4.7 Αξιολόγηση με απλό διαχωρισμό holdout

	AdaBoost	Gradient Boosting	XGboost
Accuracy	88%	89%	87%
Precision	67%	76%	65%
Recall	44%	46%	41%
F1	0,53	0,57	0,5





Η αξιολόγηση μέσω του holdout set μας δίνει ακρίβεια 89% για το Gradient Boosting ο οποίος δημιουργεί ταυτόχρονα και το μεγαλύτερο εμβαδό τόσο στην καμπύλη ROC όσο και στην precision-recall με 0,869 και 0,689 αντίστοιχα.

5.4.8 Επισκόπηση σεναρίου 4 και επιλογή μοντέλου

Στο σενάριο 4 αποτυπώνεται μέσω του f1 score ότι η επίδοση των ταξινομητών στην πρόβλεψη της μειονοτικής κλάσης πέφτει σε σχέση με το αμέσως προηγούμενο σενάριο όπου το καλύτερο f1 score συνολικά ήταν στο 0,65. Κατά τα άλλα η ακρίβεια παραμένει στα επίπεδα του 89% για τους 3 εξεταζόμενους ταξινομητές. Ο καλύτερος συνδυασμός μοντέλου και τεχνικής εκπαίδευσης είναι ο Gradient Boosting που αποδίδει 89% ακρίβεια και 0,61 f1 score εκπαιδευόμενος με SMOTE, αλλά και ο XGBoost ο οποίος πετυχαίνει τις ίδιες επιδόσεις εκπαιδευόμενος με Tomek links.

5.5 Χρησιμοποιώντας τα διαθέσιμα δεδομένα κατά την εισαγωγή στην εντατική

5.5.1 Αξιολόγηση μέσω cross validation

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost
Accuracy	87%	87%	81%	88%	89%	88%	88%
Precision	68%	86%	42%	68%	70%	74%	73%
Recall	39%	24%	50%	43%	48%	41%	46%
F1	0,49	0,36	0,43	0,52	0,56	0,52	0,56

Random Forest, SVM και Logistic Regression έχουν πετυχαίνουν ακρίβεια 87%-88% με αρκετά χαμηλό f1 score εντούτοις. Το decision tree στη συγκεκριμένη περίπτωση έχει ακρίβεια 81% που είναι και το χαμηλότερο. Οι AdaBoost, Gradient Boosting και XGboost από την άλλη αποδίδουν 88%-89% ακρίβεια με το f1 score των AdaBoost και XGboost να φτάνει στο 0,56.

5.5.2 Αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost	Voting Classifier
Accuracy	87%	88%	86%	88%	89%	89%	89%	89%
Precision	79%	71%	87%	70%	72%	74%	73%	76%
Recall	29%	45%	15%	41%	48%	46%	50%	49%
F1	0,41	0,54	0,26	0,51	0,57	0,56	0,59	0,59

	SVM	Random Forest	Decision Tree	Logistic Regression	AdaBoost	Gradient Boosting	XGboost	Voting Classifier
Υπερ-παράμετροι	**C**=*0,08*	**class_weight**=*balanced* **max_depth**=*8*	**max_depth**=*2* **max_leaf_nodes**=2	**C**=*0,4* **Penalty**=*l2*	**n_estimators**=*40*	**n_estimators**=*130* **learning_rate**=*0,12*	**scale_pos_weight**=*2*	**Voting=***soft* **Weights=***(1, 1, 1)*

Όπως στο προηγούμενο σενάριο έτσι και τώρα γίνεται αντιληπτό πως τα μόνα μοντέλα που χρήζουν περεταίρω διερεύνησης είναι αυτά που χρησιμοποιούν το boosting καθώς οι επιδόσεις των υπολοίπων στη μειονοτική κλάση δεν είναι ικανοποιητικές. Η εύρεση των βέλτιστων υπερπαραμέτρων των Gradient Boosting και XGboost ανέβασε την ακρίβειά τους στο 89% ενώ ο τελευταίος απέδωσε f1 score 0,59 που είναι το καλύτερο μαζί με αυτό του Voting Classifier.

5.5.3 Χρήση της μεθόδου SMOTE και αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων

	AdaBoost	Gradient Boosting	XGboost
Accuracy	88%	89%	89%
Precision	63%	72%	74%
Recall	53%	50%	49%
F1	0,58	0,58	0,59

	AdaBoost	Gradient Boosting	XGboost
Υπερ-παράμετροι	**n_estimators**=*160*	**n_estimators**=*90* **learning_rate**=*0,14*	**scale_pos_weight**=*1*

Η μέθοδος SMOTE δεν οδηγεί σε κάποια αξιοσημείωτη μεταβολή με την ακρίβεια να παραμένει στο 88%-89% για όλους τους ταξινομητές και το f1 score μεταξύ 0.58 και 0,59.

5.5.4 Αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων με χρήση της μεθόδου Tomek links

	AdaBoost	Gradient Boosting	XGboost
Accuracy	88%	89%	89%
Precision	68%	73%	73%
Recall	50%	47%	47%
F1	0,57	0,57	0,57

	AdaBoost	Gradient Boosting	XGboost
Υπερ-παράμετροι	**n_estimators**=*50*	**n_estimators**=*150* **learning_rate**=*0,11*	**scale_pos_weight**=1

Όπως με τη μέθοδο SMOTE έτσι και με την Tomek links δεν παρουσιάζεται κάποια ιδιαίτερη βελτίωση στην εικόνα των αποτελεσμάτων.

5.5.5 Αξιολόγηση των μοντέλων με τις βέλτιστες υπερπαραμέτρους μέσω cross validation και στάθμιση πιθανοτήτων

	AdaBoost	Gradient Boosting	XGboost
Accuracy	89%	89%	89%
Precision	77%	76%	77%
Recall	41%	45%	45%
F1	0,53	0,56	0,57

	AdaBoost	Gradient Boosting	XGboost
Μέθοδος στάθμισης	**Isotonic cv=5**	**Isotonic cv=5**	**Isotonic cv=5**

Η στάθμιση πιθανοτήτων βγάζει και για τα 3 μοντέλα 89% ακρίβεια με το καλύτερο f1 score να είναι αυτό του XGBoost στο 0,57.

5.5.6 Αξιολόγηση μέσω cross validation και ρύθμιση υπερπαραμέτρων με χρήση της μεθόδου random oversampling

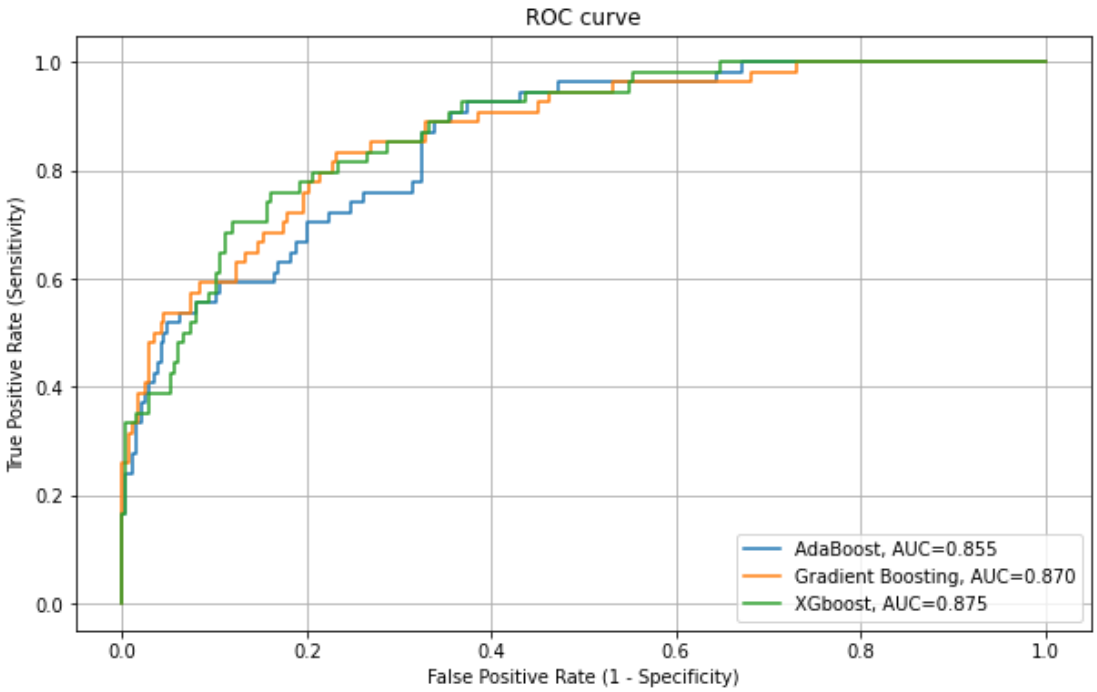
	AdaBoost	Gradient Boosting	XGboost
Accuracy	85%	88%	89%
Precision	53%	64%	69%
Recall	65%	59%	54%
F1	0,58	0,61	0,6

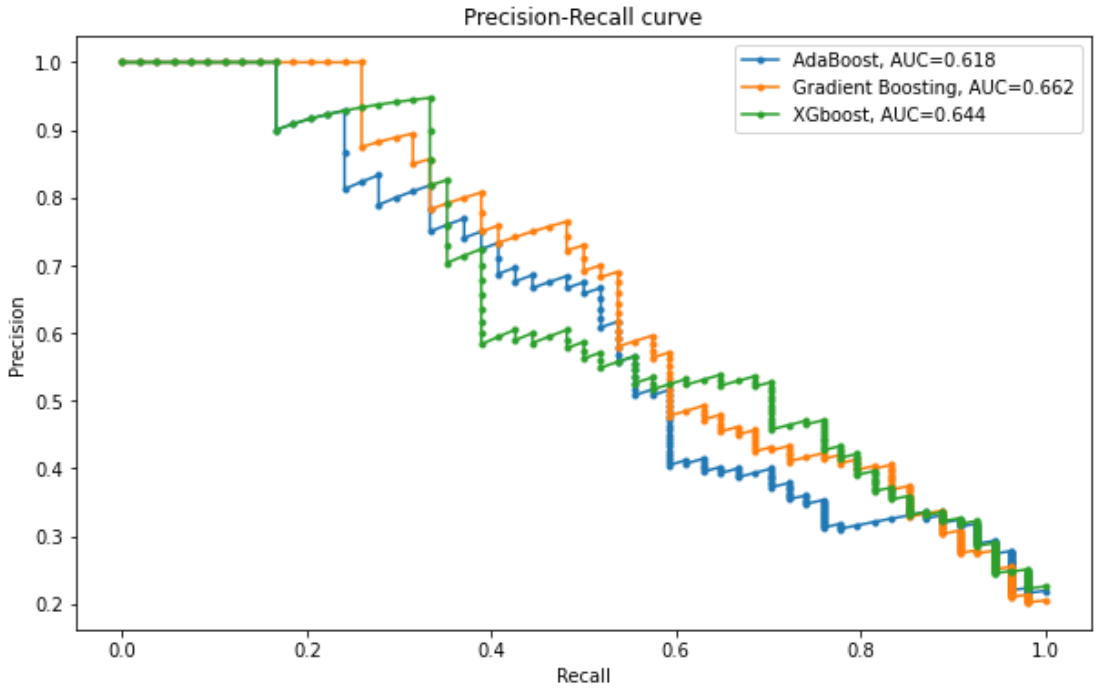
	AdaBoost	Gradient Boosting	XGboost
Υπερ-παράμετροι	**n_estimators**=*350*	**n_estimators**=*400* **learning_rate**=*0,16*	**scale_pos_weight**=*1*

Μέσω της μεθόδου random oversampling οι Gradient Boosting και XGboost αποδίδουν το καλύτερο f1 score σε αυτό το σενάριο 0,6 και 0,61 αντίστοιχα με την ακρίβεια τους στο 88% και 89%. Ο AdaBoost αντίθετα αποδίδει αισθητά χαμηλότερη ακρίβεια και χαμηλότερο f1 score.

5.5.7 Αξιολόγηση με απλό διαχωρισμό holdout

	AdaBoost	Gradient Boosting	XGboost
Accuracy	88%	89%	87%
Precision	73%	74%	68%
Recall	41%	43%	39%
F1	0,52	0,54	0,49





Στην αξιολόγηση μέσω του holdout set ο Gradient Boosting δίνει ακρίβεια 89% και f1 score 0,54 ενώ στη μεν καμπύλη ROC το μεγαλύτερο εμβαδό παράγει ο XGBoost με 0,875, στη δε καμπύλη precision-recall το μεγαλύτερο εμβαδό παράγει ο Gradient Boosting με 0,662.

5.5.8 Επισκόπηση σεναρίου 5 και επιλογή μοντέλου

Το πέμπτο και τελευταίο σενάριο είναι και το δυσκολότερο για τους αλγορίθμους υπό την έννοια ότι έχει τις λιγότερες διαθέσιμες πληροφορίες. Εν τούτοις δεν υπήρξε πτώση στην ακρίβεια των τριών ταξινομητών που εξετάστηκαν, σε σχέση με το αμέσως προηγούμενο σενάριο. Η ακρίβεια κυμάνθηκε από 88% ως 89% για όλους τους ταξινομητές και τις τεχνικές εκπαίδευσης. Ωστόσο το f1 score ήταν ως επί το πλείστον χαμηλό καταδεικνύοντας τη δυσχέρεια των μοντέλων να εκπαιδευτούν πάνω στην κλάση της μειονότητας στο σύνολο δεδομένων του σεναρίου 5.

Καλύτερη συνολικά επίδοση σε αυτό το σενάριο είχε ο XGBoost εκπαιδευόμενος μέσω της τεχνικής του random oversampling με ακρίβεια 89% και f1 score 0,6 και αν έπρεπε να επιλεγεί ένα μοντέλο και τεχνική εκπαίδευσης αυτός θα ήταν ο καλύτερος ο συνδυασμός.

6. Μελλοντικές προεκτάσεις

Η εργασία βασίστηκε πάνω στην πρόβλεψη μέσω μηχανικής μάθησης μίας από τις πιθανές επιπτώσεις του εμφράγματος μυοκαρδίου αυτή της επιβίωσης ή μη του ασθενούς χρησιμοποιώντας σε κάθε σενάριο ολοένα και λιγότερα δεδομένα. Από εκεί και πέρα υπάρχουν και άλλες δυνατές μεταβλητές στόχοι που θα μπορούσαν να χρησιμοποιηθούν ως αντικείμενο προβλημάτων μηχανικής μάθησης όπως ενδεικτικά το πνευμονικό οίδημα, η ρήξη μυοκαρδίου και άλλα και η μεθοδολογία της παρούσα εργασίας θα μπορούσε να χρησιμοποιηθεί αυτούσια ως βάση αντιμετώπισης αυτών των προβλημάτων.

Επίσης τα μοντέλα που αναπτύχθηκαν μέσα από την εργασία θα μπορούσαν να χρησιμοποιηθούν συνεπικουρικά στο έργο των γιατρών ούτως ώστε να έχουν μία έγκαιρη πρόβλεψη της έκβασης του ασθενούς. Ειδικά στην περίπτωση των λοιπών επιπλοκών πλην του θανάτου η δυνατότητα τέτοιου είδους πρόβλεψης θα μπορούσε να αποδειχτεί καίρια καθώς θα ήταν δυνατή η κατάλληλη προσαρμογή της θεραπευτικής αγωγής πριν ακόμα εγκατασταθούν οι επιπλοκές στον ασθενή.

Τέλος θα μπορούσε πάνω σε ένα παρόμοιο σύνολο δεδομένων που θα αποτελείτο από καρδιοπαθείς μεν, αλλά χωρίς όλοι οι ασθενείς να έχουν υποστεί έμφραγμα του μυοκαρδίου, να εξεταστεί ένα διαφορετικό πρόβλημα μηχανικής μάθησης αυτό του αν κάποιος από τους ασθενείς θα υποστεί έμφραγμα του μυοκαρδίου ή όχι.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Complications of myocardial infarction: a database for testing recognition and prediction systems, S.E. Golovenkin, A.N. Gorban, E.M.Mirkes, V.A. Shulman, D.A. Rossiev, P.A. Shesternya, S.Yu. Nikulina, Yu.V. Orlova, and M.G. Dorrer

[2] https://www.hygeia.gr/emfragma-toy-myokardioy/

[3] Imbalanced Classification with Python, Jason Brownlee, σελ. 4-7, 26-30, 39-46, 49-51, 57-65, 97-101,113-115, 122-129, 150-152, 263-267

[4] Master Machine Learning Algorithms, Jason Brownlee, σελ. 11, 52-54

[5] Classifier calibration The why, when and how of model calibration for classification tasks, Dimitris Poulopoulos

https://towardsdatascience.com/classifier-calibration-7d0be1e05452

[6] https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[7] Ensemble Learning Algorithms With Python, Jason Brownlee

[8] https://www.machinelearningplus.com/machine-learning/an-introduction-to-gradient-boosting-decision-trees/#:~:text=Using%20a%20low%20learning%20rate,0.3%20gives%20the%20best%20results.

[9] Hands-on-gradient-boosting-with-xgboost-and-scikit-learn-perform-accessible-python-machine-learning-and-extreme-gradient-boosting-with-python-9781839218354 σελ. 254, 270-271

[10] Role of machine learning in medical research: A survey, 2021 Computer Science Review, Arunim Garg, Vijay Mago

[11] https://imbalanced-learn.org/stable/auto_examples/under-sampling/plot_illustration_tomek_links.html

[12] On hyperparameter optimization of machine learning algorithms*:* Theory and practice, Li Yang και Abdallah Shami, 2020. σελ. 295-316,

https://doi.org/10.48550/arXiv.2007.15745

[13] XGBoost: A Scalable Tree Boosting System. ,Tianqi Chen and Carlos Guestrin, σελ. 785–794, 2016. https://doi.org/10.48550/arXiv.1603.02754

[14] Tunability: Importance of Hyperparameters of Machine Learning Algorithms, Probst, Philipp, Anne-Laure Boulesteix and B. Bischl. σελ. 3-5

https://arxiv.org/pdf/1603.02754.pdf

[15] Imbalanced Learning: Foundations, Algorithms and Applications, Haibo He, Yunqian Ma

[16] SMOTE: Synthetic Minority Over-sampling Technique, Chawla, Nitesh & Bowyer, Kevin & Hall, Lawrence & Kegelmeyer

[17] Improving Identification of Difficult Small Classes by Balancing Class Distribution, Laurikkala, Jorma

[18] Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation, D. Powers

[19] Speech and Language Processing. Daniel Jurafsky & James H. Martin. Κεφ. 5