



ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΠΛΗΡΟΦΟΡΙΑΚΑ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

**Μοντέλα Πρόβλεψης Θνητότητας με Εφαρμογή Μηχανικής
Μάθησης σε Ασθενείς με Εγκεφαλικό.**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Δημόπουλου Δημήτριου
AM 3252020007

Επιβλέπων : Κωστούλας Θεόδωρος

Μέλη εξεταστικής επιτροπής: Βλάχου Ακριβή, Συμεωνίδης Παναγιώτης

Σάμος, 06/2022

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πρόλογος και ευχαριστίες

Η εκπόνηση της διπλωματικής εργασίας αποτέλεσε ένα ευχάριστο ταξίδι διαπροσωπικών σχέσεων, συζητήσεων, ανταλλαγής απόψεων και γνώσης. Τα μονοπάτια που αποκαλύφθηκαν κατά την υλοποίησης της άνοιξαν τις πόρτες για έρευνα σε πολλαπλά επίπεδα. Έτσι η εργασία όχι μόνο μοιάζει, αλλά είναι όντως τόσο ξεχωριστή. Ο τίτλος που θα ταίριαζε σε αυτή τη διαδρομή, θα ήταν από το ποίημα του Καβάφη “Το Πρώτο Σκαλί”. Όπως ο ποιητής στην αρχή της σταδιοδρομίας του, έτσι ακριβώς και εγώ, εδώ στην αρχή της δικής μου. Ένας στοίχος που θα χαρακτήριζε την δική μου την πορεία μέσα σε αυτή τη διαδρομή θα ήταν:

*“Αλλοίμονον, είν’ υψηλή το βλέπω,
πολύ υψηλή της Ποίσεως η Σκάλα·
κι απ’ το σκαλί το πρώτο εδώ που είμαι,
ποτέ δεν θ’ ανέβω ο δυστυχισμένος.”*

Ευχαριστίες πολλές αξίζουν στον Καθηγητή PhD Κωστούλα Θεόδωρο όπου ανταποκρίθηκε επάξια στον ρόλο του επιβλέποντος καθηγητή. Η συμβολή πέρα από καθοριστική αποτέλεσε ταυτόχρονα και μεγάλη έμπνευση μιας και καθόλη την διάρκεια αυτής της συνεργασίας η προσήλωση του ξεπέρασε κάθε αρχική μου προσδοκία. Τα απανωτά μηνύματα που λάμβανε σε καθημερινή βάση και σε ώρες που πέρα από συνηθισμένες θα μπορούσαν να χαρακτηριστούν, τα μετασχημάτιζε σε συναντήσεις ουσίας. Η Ιατρός Αιματολογίας MD, PhD, IBCLC Δανηλάτου Βασιλική κατεύθυνε με την ιατρική της συμβολή την εργασία σε ουσιαστικά μονοπάτια, παράγοντας αξιόλογα αποτελέσματα. Τεράστιες επίσης ευχαριστίες αξίζουν σε όλα αυτά “τι εννοείς;” που αντηχούσαν στ’ αφτιά μου καθόλη την διάρκεια υλοποίησης της διπλωματικής. Σε όλα αυτά τα “τι εννοείς;” που δημιουργήσαν ερωτήματα και έφτασαν σε απαντήσεις. Στα “τι εννοείς;” που με έκαναν να σκεφτώ πριν απαντήσω, να σχεδιάσω πριν υλοποιήσω, να αναζητήσω πριν ρωτήσω. Στα “τι εννοείς;” που με έκαναν πιο κατανοητό και σίγουρο για όλα αυτά που εννοώ. Πίσω από όλα αυτά τα “τι εννοείς;” στόχος και σκοπός να αποκαλυφθούν στις σελίδες που ακολουθούν όλα αυτά για τα οποία κοπίασα και σαν έργο θέλω να αφήσω πίσω. Η στάση όλων αυτών που συνέβαλαν και πίστεψαν τόσο σε εμένα όσο και σε αυτή μου την προσπάθεια θα ήταν μέσα στο ίδιο ποίημα του Καβάφη η απάντηση του Θεόκριτου:

*“Κι αν είσαι στο σκαλί το πρώτο, πρέπει
να ‘σαι υπερήφανος κ’ ευτυχισμένος.
Εδώ που έφτασες, λίγο δεν είναι·
τόσο που έκαμες, μεγάλη δόξα”.*

Το μόνο που έχω ειλικρινά να πω για αυτή σας την προσφορά είναι ένα τεράστιο ευχαριστώ.

© 2022

του

Δημόπουλου Δημήτριου

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Η σελίδα αυτή είναι σκόπιμα λευκή

Πίνακας περιεχομένων

1	Εισαγωγή	1
1.1	Μηχανική μάθηση στην πρόβλεψη θνησιμότητας	1
1.2	Αντικείμενο διπλωματικής	1
1.3	Ηλεκτρονικά αρχεία υγείας	2
1.4	Δομή της διπλωματικής	3
2	Αρχιτεκτονική Προτεινόμενου Συστήματος	4
2.1	Σχήμα συστήματος	4
2.2	MIMIC – IV ver1.0	5
2.2.1	Ενότητα Πυρήνα - Core	7
2.2.2	Ενότητα Νοσοκομείου - Hosp (Hospital)	7
2.2.3	Ενότητα Μονάδων Εντατικής Θεραπείας – ICU	8
2.2.4	Ενότητα Επειγόντων Περιστατικών – ED	9
2.3	Προεπεξεργασία δεδομένων	10
2.3.1	Ασθενείς με ισχαιμικό εγκεφαλικό στο τμήμα ‘Επειγόντων Περιστατικών’	10
2.3.2	Ασθενείς με ισχαιμικό εγκεφαλικό στις ‘Νοσοκομειακές Μονάδες’	11
2.3.3	Ασθενείς με ισχαιμικό εγκεφαλικό στις ‘Μονάδες Εντατικής Θεραπείας’	12
2.3.4	Κλινικές καταγραφές ισχαιμικού εγκεφαλικού	12
2.3.5	Εργαστηριακές ενδείξεις ισχαιμικού εγκεφαλικού	14
2.3.6	Δεδομένα εκπαίδευσης – ελέγχου (Training – Test Set)	16
2.4	Αλγόριθμοι μηχανικής μάθησης	18
3	Πειραματικά Αποτελέσματα	19
3.1	Αλγόριθμοι Μηχανικής Μάθησης (Ταξινομητές)	19
3.2	Στοιχεία ταξινομητών	20
3.2.1	True Positive Rate (TP Rate)	22
3.2.2	False Positive Rate (FP Rate)	22
3.2.3	Precision	22
3.2.4	Recall	22
3.2.5	F – Measure	23
3.2.6	Receiver Operator Characteristic (ROC Area)	23
3.3	Αποτελέσματα και παραμετροποίηση ταξινομητών	23
3.3.1	Ταξινομητής “Logistic Model Tree (LMT)”	23
3.3.2	Ταξινομητής “Bayesian Network”	23

3.3.3	Ταξινομητής “Naïve Bayes”	24
3.3.4	Ταξινομητής “Logistic Regression”	25
3.3.5	Ταξινομητής “Simple Logistic”	25
3.3.6	Ταξινομητής “Decision Table”	26
3.3.7	Ταξινομητής “Sequential Minimal Optimization - SMO”	26
3.3.8	Ταξινομητής “Random Tree”	27
3.3.9	Ταξινομητής “Attribute Selected”	27
3.3.10	Ταξινομητής “K-Nearest Neighbors”	28
3.4	ROC Curves “Receiver Operating Characteristic”	29
3.4.1	Καμπύλες ROC (Not Survive)	30
4	Συμπεράσματα	34
4.1	Ανάλυση αποτελεσμάτων	34
4.2	Επιλογή αποδοτικότερων ταξινομητών	36
	Βιβλιογραφία	38
	Παράρτημα I [Πιστοποίηση CITI]	40
	Παράρτημα II [Αναλυτική βαθμολογία CITI]	41
	Παράρτημα III [Χρήση δεδομένων MIMIC-IV (v1.0)]	42
	Παράρτημα IV [Χρήση Δεδομένων MIMIC-IV-ED (v1.0)]	43
	Παράρτημα V [Πίνακας Τμήμα Επειγόντων]	44
	Παράρτημα VI [Πίνακας Νοσοκομείου]	46
	Παράρτημα VII [Πίνακας Chartevents]	47
	Παράρτημα VIII [Πίνακας Labevents]	55
	Παράρτημα IX [Ενιαίος Πίνακας]	62
	Παράρτημα X [ROC Διαγράμματα]	64

Λίστα Σχημάτων

2.1	Διαχωρισμός συνόλου δεδομένων σε Training & Test Set.	σελ 17
3.1	Διαχωρισμός του συνόλου δεδομένων σε δύο κλάσεις.	σελ 20
3.2	Διαδικασία προβλέψεις.	σελ 21
3.3	Παραγωγή Ψευδών Θετικών & Ψευδών Αρνητικών προβλέψεων.	σελ 21
3.4	ROC Curve “treeLMT”.	σελ 30
3.5	ROC Curve “BayesNet”.	σελ 30
3.6	ROC Curve “Naïve Bayes”.	σελ 31
3.7	ROC Curve “Logistic Regression”.	σελ 31
3.8	ROC Curve “Simple Logistic”.	σελ 31
3.9	ROC Curve “Decision Table”.	σελ 32
3.10	ROC Curve “SMO”.	σελ 32
3.11	ROC Curve “Random Tree”.	σελ 32

Λίστα Πινάκων

2.1	Αρχεία Ενότητας Core	σελ 7
2.2	Αρχεία Ενότητας Hosp	σελ 7
2.3	Αρχεία Ενότητας ICU	σελ 9
2.4	Αρχεία Ενότητας ED	σελ 9
2.5	Τμήμα Επειγόντων Περιστατικών	σελ 11
2.6	Εγγραφές Ζωτικών Οργάνων στο Τμήμα Επειγόντων Περιστατικών.	σελ 11
2.7	Διαγνώσεις εγκεφαλικού στις νοσοκομειακές μονάδες.	σελ 11
2.8	Ασθενείς με εγκεφαλικό στις ΜΕΘ.	σελ 12
2.9	Είδος καταγραφών στον πίνακα καταγραφών ΜΕΘ.	σελ 13
2.10	Ενοποιημένος πίνακας καταγραφών ΜΕΘ.	σελ 14
2.11	Χαρακτηριστικά πίνακα εργαστηριακών εξετάσεων.	σελ 15
2.12	Χαρακτηριστικά ενιαίου πίνακα ασθενών στις ΜΕΘ.	σελ 16
2.13	Δεδομένα Εκπαίδευσης (Training Set) – Ελέγχου (Test Set).	σελ 17
2.14	Ασθενείς που επιβίωσαν ή όχι στα Δεδομένα Εκπαίδευσης και Ελέγχου	σελ 18
3.1	Περίληπτικά αποτελέσματα ταξινομητή treeLMT.	σελ 24
3.2	Περίληπτικά αποτελέσματα ταξινομητή Bayesian Network.	σελ 24
3.3	Περίληπτικά αποτελέσματα ταξινομητή Naïve Bayes.	σελ 25
3.4	Περίληπτικά αποτελέσματα ταξινομητή Logistic Regression.	σελ 25
3.5	Περίληπτικά αποτελέσματα ταξινομητή Simple Logistic.	σελ 26
3.6	Περίληπτικά αποτελέσματα ταξινομητή Decision Table.	σελ 26
3.7	Περίληπτικά αποτελέσματα ταξινομητή Sequential Minimal Optimization – SMO.	σελ 27

3.8	Περίληπτικά αποτελέσματα ταξινομητή Random Tree.	σελ 27
3.9	Περίληπτικά αποτελέσματα ταξινομητή Attribute Selected.	σελ 28
3.10	Περίληπτικά αποτελέσματα ταξινομητή KNN-1.	σελ 28
3.11	Περίληπτικά αποτελέσματα ταξινομητή KNN-2.	σελ 29
3.12	Τιμές AUC.	σελ 29
3.13	AUC treeLMT.	σελ 30
3.14	AUC Bayesian Network	σελ 30
3.15	AUC Naïve Bayes.	σελ 31
3.16	AUC Logistic Regression.	σελ 31
3.17	AUC Simple Logistic.	σελ 31
3.18	AUC Decision Table.	σελ 32
3.19	AUC SMO.	σελ 32
3.20	AUC Random Tree.	σελ 32
3.21	AUC προτεινόμενων αλγορίθμων.	σελ 37

Ακρωνύμια

AUC	Area Under the Curve
CPT	Current Procedural Terminology
ED	Emergency Department
EHR	Electronic Health Record
ICD	International Classification of Disease
ICU	Intensive Care Unit
MIMIC	Medical Information Mart for Intensive Care
MIT	Massachusetts Institute of Technology
PHI	Protected Health Information
ROC	Receiver Operating Characteristic
WEKA	Waikato Environment for Knowledge Analysis
ΜΕΘ	Μονάδες Εντατικής Θεραπείας

Περίληψη

Η συγκεκριμένη μελέτη στοχεύει στην δημιουργία και ανάλυση μοντέλων μηχανικής μάθησης τα οποία έχουν τροφοδοτηθεί με ιατρικά δεδομένα. Ο στόχος που επιδιώκετε με την υλοποίηση αυτής της έρευνας πέρα από την δημιουργία των μοντέλων, καθώς και της εφαρμογής της πρόβλεψης, είναι η επικαιροποίηση των ποσοστών θνησιμότητας σε ασθενείς που πάσχουν από ισχαιμικό εγκεφαλικό. Ο λόγος ύπαρξης της εν λόγω εργασίας είναι να μπορέσει να καταστεί δυνατό μέσω της έγκαιρης πρόβλεψης η εφαρμογή της αποτελεσματικότερης κατανομής τόσο των έμψυχων όσο και άψυχων ιατρικών πόρων στις Μονάδες Εντατικής Θεραπείας (ΜΕΘ). Η πρόβλεψη θνησιμότητας μέσα στις ΜΕΘ αποτελούσε και εξακολουθεί να αποτελεί πρόκληση για το ιατρικό προσωπικό που στελεχώνει τις ΜΕΘ, ειδικότερα την περίοδο της πανδημίας Covid 19 που διανύουμε και αποτέλεσε εφιαλτήριο αυτής της έρευνας. Για την επίτευξη της δημιουργίας και ανάλυσης των μοντέλων έχει συλλεχθεί και αναλυθεί ένας τεράστιος όγκος πληροφοριών που φιλοξενούνται στα ηλεκτρονικά βιβλιάρια υγείας ασθενών. Τα ιατρικά δεδομένα που αποτελούνται ποικιλότροπος τόσο από κλινικές όσο και εργαστηριακές πληροφορίες ταξινομήθηκαν χρονικά μέσα στις εισαγωγές των ασθενών και ομαδοποιήθηκαν μέσα στα οχτάωρα της νοσηλείας στις ΜΕΘ. Η ομαδοποίηση αυτή αποσκοπούσε στην ισορροπημένη επίτευξη της βέλτιστης πρόβλεψης, στον ελάχιστο χρόνο, αξιοποιώντας τον μεγαλύτερο δυνατό όγκο δεδομένων. Για την πειραματική αυτή διαδικασία αξιοποιήθηκε κάτω από κατάλληλες άδειες χρήσης που παραχωρήθηκαν από θυγατρική του Ινστιτούτου Τεχνολογίας της Μασαχουσέτης (MIT) η βάση δεδομένων MIMIC IV.

Τα αποτελέσματα που προκύπτουν από την χρήση των μοντέλων μηχανικής μάθησης υποδεικνύουν επαρκή ικανότητα πρόβλεψης από το τέλος της πρώτης κιόλας μέρας νοσηλείας του ασθενούς στην ΜΕΘ. Από τα διαγράμματα ROC προκύπτει ότι τις πρώτες τέσσερις ημέρες νοσηλείας στις ΜΕΘ, απόδοση του μοντέλου παράγει τιμές της τάξεως του 74% από το τέλος της πρώτης κιόλας ημέρας και κάτι περισσότερο από 77% στη δεύτερη.

Λέξεις Κλειδιά:

Πρόβλεψη Θνησιμότητας, Εγκεφαλικό, Ισχαιμικό Εγκεφαλικό, ΜΕΘ, MIMICIV

Abstract

This examination aims to create and analyze machine-learning models, which have been fed with medical data. The goal of this research is to update the mortality rates in patients with ischemic stroke, in addition to creating models and implementing the prediction. The reason for the existence of this work is to be able through timely forecasting to implement the most efficient distribution of both animate and inanimate medical resources in the Intensive Care Units (ICU). Predicting mortality in ICUs has been and continues to be a challenge for the medical staff staffing ICUs, especially during the Covid 19 pandemic that we are going through and has been the springboard for this research. To achieve the creation and analysis of the models, a huge amount of information has been collected and analyzed, which are hosted in the electronic patient health booklets. Medical data consisting of a variety of both clinical and laboratory information were classified over time in patients' admissions and grouped within the eight hours of ICU treatment. This grouping was aimed at achieving a balanced optimal forecast, in the shortest time, utilizing the largest possible volume of data. The MIMIC IV database was used for this experimental process under appropriate licenses granted by a subsidiary of the Massachusetts Institute of Technology (MIT).

The results obtained from the use of the machine learning models indicate a sufficient ability to predict from the end of the very first day of hospitalization of the patient in the ICU. ROC charts show that during the first four days of ICU treatment, the model yields values of 74% from the end of the first day and just over 77% after the end of the second.

Keywords:

Mortality Prediction, Stroke, Ischemic Stroke, ICU, MIMICIV

1

Εισαγωγή

1.1 Μηχανική μάθηση στην πρόβλεψη θνησιμότητας

Ο σκοπός της διπλωματικής εργασίας αυτής είναι μέσα από τα προς διάθεση δεδομένα να αντλήσει την γνώση αναφορικά με τους εισηγμένους στις Μονάδες Εντατικής Θεραπείας (Μ.Ε.Θ.) ασθενείς που έχουν διαγνωστεί με ισχαιμικό εγκεφαλικό. Σαν συνέπεια αυτού θα ήταν η δημιουργία πρόσφορου εδάφους στην πρόβλεψη θνησιμότητας, όπου η έγκαιρη ταυτοποίηση της θα μπορούσε να επηρεάσει τις ιατρικές κλινικές πρακτικές και συνάμα να επιτρέψει την αποτελεσματική κατανομή των πόρων υγείας στις ΜΕΘ. Αποτέλεσμα όλων αυτών θα ήταν ποιοτικότερες υπηρεσίες υγείας προς τους ασθενείς μέσω της σωστής διαχείρισης των πόρων.

Το ισχαιμικό εγκεφαλικό είναι η δεύτερη πιο συχνή αιτία θανάτου και αναπηρίας παγκοσμίως και προκαλείται τόσο από την μειωμένη ή καθόλου παροχή αίματος σε μέρος του εγκεφάλου. Η αντιμετώπιση του απαιτεί εξειδικευμένη φροντίδα στις μονάδες εντατικής θεραπείας [1].

1.2 Αντικείμενο διπλωματικής

Η διπλωματική αυτή εργασία έχει σαν σκοπό την δημιουργία , την αξιολόγηση και την επιλογή των ιδανικότερων προς χρήση μοντέλων πρόβλεψης θνησιμότητας ασθενών που έχουν διαγνωστεί με ισχαιμικό εγκεφαλικό. Η υλοποίηση των εν λόγω μοντέλων και κατ' επέκταση της αξιοποίησης τους θα μπορούν να συμβάλουν δραστικά μέσω της σωστής διαχείρισης και κατανομής των πόρων στις ΜΕΘ. Το ιατρικό προσωπικό καλείτε καθημερινά μέσα στις μονάδες να αντιμετωπίσει προκλήσεις αποσκοπώντας στη βελτίωση της υγείας των ασθενών. Η έγκαιρη

πρόβλεψη αναφορικά με την πορεία του ασθενούς στο επίπεδο της θνησιμότητας μπορεί να συμβάλει καθοριστικά στην επίτευξη της καλύτερευσης της υγείας των νοσηλευόμενων προσαρμόζοντας κατά περίπτωση οι κλινικοί γιατροί της διαδικασίες νοσηλείας.

Για την επίτευξη του στόχους της διπλωματικής έπρεπε να έχουμε στην διάθεση μας στοιχεία που θα αναφέρονται στα ιατρικά δεδομένα των ασθενών. Δεδομένα και πληροφορίες τόσο κλινικού όσο και εργαστηριακών εξετάσεων επιπέδου. Όλες αυτές οι πληροφορίες αναζητήθηκαν στη βάση δεδομένων της MIMIC. Για την πρόσβαση σε αυτή χρειάστηκαν να εφαρμοστούν διαδικασίες και να υλοποιηθούν ενέργειες τόσο από την δική μου την πλευρά, μιας και θα αποκτούσα προσωπικά, ατομική πρόσβαση όσο και από την πλευρά του επιβλέποντος καθηγητή όπου έπρεπε να επικαιροποιήσει τους λόγους ενασχόλησης μου με αυτή. Η πρόσβαση στη MIMIC έφερε την έρευνα αντιμέτωπη με ένα τεράστιο όγκο δεδομένων όπου έπρεπε μέσα από πολυσύνθετα στάδια επεξεργασίας να ανασχηματιστούν και να πάρουν μια πιο αυστηρή δομημένη μορφή. Ο τεράστιος όγκος των δεδομένων επεξεργασίας εντάσσει τα δεδομένα της MIMIC στα Μεγάλα Δεδομένα (Big Data). Λόγω του μεγάλου όγκου αυτών των δεδομένων ήταν αδύνατο να εφαρμοστούν οι παραδοσιακές μέθοδοι επεξεργασίας. Γι' αυτό το λόγο επιστρατεύτηκε η γλώσσα προγραμματισμού Python μέσω της πλατφόρμας Anaconda.

Η Python αποτέλεσε τον ακρογωνιαίο λίθο για την υλοποίηση της διπλωματικής. Μέσω του κώδικα της έγινε η συλλογή και επεξεργασία δεδομένων αρχικά διαχωρίζοντας τους ασθενείς που έχουν διαγνωστεί με ισχαιμικό εγκεφαλικό από τις άλλες κλινικές περιπτώσεις. Συνέβαλε στην ενοποίηση των στοιχείων – δεδομένων που ήταν αποθηκευμένα σε διαφορετικά σημεία μέσα στη βάση. Επιπλέον με την χρήση της Python στοιχεία μετασχηματίστηκαν από λεκτικά σε αριθμούς, από ακέραιους σε δεκαδικούς και ούτω καθεξής. Τέλος αξιοποιώντας εργαλεία και βιβλιοθήκες έγινε η αξιολόγηση των μοντέλων που δημιουργήθηκαν, και με την παραγωγή τόσο των αποτελεσμάτων όσο και των διαγραμμάτων φτάσαμε στο στάδιο να μπορούμε να αποτυπώσουμε το συμπέρασμα στο τέλος αυτού του συγγράμματος.

1.3 Ηλεκτρονικά αρχεία υγείας

Η χρήση και υιοθέτηση των Ηλεκτρονικών Αρχείων Υγείας (Electronic Health Record – EHR) σε συνδυασμό με την χρήση πιστοποιημένων προϊόντων πληροφορικής τόσο στα νοσοκομεία όσο και στις μονάδες εντατικής θεραπείας όπου είναι το πεδίο δραστηριότητας αυτής της εργασίας αποτέλεσε το βασικό καθοριστικό βήμα για την συλλογή των καταγεγραμμένων προς ανάλυση δεδομένων. Ένα σύστημα ηλεκτρονικών καταγραφών φιλοξενεί σε ψηφιακή τις μέχρι πρότινος σε φυσική μορφή πληροφορίες των ασθενών, οι οποίες αναφέρονται τόσο στο ιατρικό όσο και θεραπείας ιστορικό. Οι πληροφορίες αυτές ξεπέρασαν μέσω του πληροφοριακού συστήματος τα στενά σύνορα και όρια του εκάστοτε χώρου περίθαλψης είτε αυτό ήταν ένα ιατρείο ή νοσοκομείο και κατέστησε αυτές τις πληροφορίες άμεσα διαθέσιμες στο ευρύ εξουσιοδοτημένο προσωπικό [2].

Η απαρίθμηση των πλεονεκτημάτων από την χρήση των EHR από το ιατρικό προσωπικό δεν αποτελεί κομμάτι αυτής της διπλωματικής, αλλά σίγουρα έγινε εκμεταλλεύσιμη τόσο η ακρίβεια όσο και η πληρότητα των πληροφοριών που εμπεριέχονται σε αυτά. Η πλούσια πηγή δεδομένων αποτελεί βασικό πυλώνα των απαντήσεων στα ερευνητικά ερωτήματα [3]. Στη

συνέχεια αυτού του κεφαλαίου θα απαριθμήσουμε αναλυτικά τα στοιχεία που εμπεριέχονται στη βάση δεδομένων της MIMIC IV version 1.0. Με τον καιρό τα στοιχεία που εμπεριέχονται στα EHR εμπλουτίζουν αντίστοιχες βάσεις δεδομένων όπου ο λόγος ύπαρξής τους είναι καθοριστικός στην ερευνητική κοινότητα. Δυστυχώς τα στοιχεία και οι παρατηρήσεις που συλλέγονται και καταχωρούνται κατά την διάρκεια της νοσηλείας γίνεται σύμφωνα με την κλινική ανάγκη του κάθε ασθενή. Αυτό έχει σαν συνέπεια να απέχουν τόσο από τα εθνικά πρότυπα όσο και από την τυποποίηση που παρέχεται από αυτά. Έτσι διαφορετικά στοιχεία ανάμεσα σε διαφορετικά συστήματα μπορεί να αποτυπώνονται με διαφορετικό τρόπο ή το αντίστροφο, δηλαδή διαφορετικά στοιχεία με τον ίδιο. Η πραγματικότητα αυτή καθιστά χρονικά απαγορευτική την εξάλειψη των πλεονασμάτων και ασαφειών, την κανονικοποίηση του περιεχομένου, την ανάλυση, επεξήγηση και ερμηνεία των στοιχείων. Μεγαλύτερη προσπάθεια απαιτείται σίγουρα κατά την συγκέντρωση στοιχείων από διαφορετικούς φορείς μιας και πέρα από τον καθαρισμό των δεδομένων απαιτείται και η αντιστοίχιση των μεταβλητών μεταξύ τους. Από την άλλη το Ινστιτούτο Τεχνολογίας της Μασαχουσέτης (MIT) εμπλεκόμενο στην παραπάνω διαδικασία έχει καταφέρει να παραδώσει ένα αξιόλογο προς χρήση προϊόν για την εξυπηρέτηση ερευνητικών σκοπών. Κάτω από συγκεκριμένες εκπαιδευτικές και εξεταστικές διαδικασίες που όπου γίνεται λόγος και αναφορά στα προσωπικά δεδομένα παρέχεται η έγκριση πρόσβασης στη βάση της MIMIC και στις διάφορες εκδόσεις της. Σήμερα που γίνεται η συγγραφή αυτής της διπλωματικής η MIMIC έχει φτάσει στην τέταρτη έκδοση της. Κατά την ερευνητική διαδικασία η οποία έχει διαρκέσει περίπου ένα ημερολογιακό έτος υπήρχε στην διάθεση του ερευνητικού και ακαδημαϊκού κοινού η τρίτη έκδοση της. Κατά την ερευνητική διαδικασία έγινε διαθέσιμη σε δύο φάσεις λόγω επιδιορθώσεων (version 0.4 και version 1.0) η τέταρτη έκδοση και έτσι χρειάστηκε να γίνει εκ νέου αναθεώρηση και επαναπροσδιορισμός της ερευνητικής διαδικασίας.

1.4 Δομή της διπλωματικής

Το σύγγραμμα έχει στο σύνολο του χωριστεί σε τέσσερα κεφάλαια ακολουθούμενα από τα αντίστοιχα παραρτήματα. Στο κεφάλαιο 1^ο παραθέσαμε την εισαγωγή, τα βασικά στοιχεία της διπλωματικής. Στο κεφάλαιο 2^ο γίνεται παρουσίαση της αρχιτεκτονικής του προτεινόμενου σχήματος. Παρουσιάζονται τμηματικά τα στάδια υλοποίησης της πειραματικής διαδικασίας ξεκινώντας από την ανάπτυξη των βημάτων υλοποίησης. Συνεχίζουμε με την βάση δεδομένων MIMIC όπου παρουσιάζουμε τα τέσσερα βασικά της μέρη. Ακολουθεί η ανάλυση της προεπεργασίας των δεδομένων έτσι ώστε να επιτευχθεί συγκέντρωση τόσο των ζωτικών όσο και εργαστηριακών παρατηρήσεων ομαδοποιημένα ανά οχτάωρο σε ένα ενιαίο πίνακα και διαχωρισμός αυτού σε δεδομένα εκπαίδευσης – ελέγχου. Στο κεφάλαιο 3^ο παρατίθενται για τον κάθε ένα από τους έντεκα αλγόριθμους μηχανικής μάθησης που χρησιμοποιήθηκαν στην πειραματική διαδικασία η περιγραφή τους, τα αριθμητικά αποτελέσματα τους, τα AUC (Area Under the Curve) αποτελέσματα, καθώς και τα ROC (Receiver Operating Characteristic) διαγράμματα τους. Στο 4^ο και τελευταίο κεφάλαιο παρουσιάζουμε την ανάλυση που εφαρμόσαμε στα αποτελέσματα του 3^{ου} κεφαλαίου και προτείνουμε τους τρεις αποδοτικότερους προς χρήση αλγόριθμους.

2

Αρχιτεκτονική Προτεινόμενου Συστήματος

2.1 Σχήμα συστήματος

Στις σελίδες του κεφαλαίου που ακολουθούν θα αναλυθούν τα λογικά πεπερασμένα βήματα της συγκέντρωσης των δεδομένων που θα τροφοδοτήσουν το περιβάλλον WEKA (Waikato Environment for Knowledge Analysis) για την δημιουργία των μοντέλων μηχανικής μάθησης. Θα παρουσιαστεί μέσα από τις αναφορές στα παραρτήματα ο κώδικας που συγγράφτηκε για την παραγωγή τόσο των διαγραμμάτων όσο και του ενιαία ημερολογιακά στοιχισμένου ως προς τους ασθενείς πίνακα στοιχείων που αναφέρονται στις διαγνώσεις ισχαιμικού εγκεφαλικού. Στη συνέχεια μέρος των παραγόμενων αποτελεσμάτων από το περιβάλλον του WEKA και πιο συγκεκριμένα η πρόβλεψη της επιβίωσης (survival class) που εφαρμόζει το μοντέλο αξιοποιήθηκε κάτω από κατάλληλες διεργασίες για την δημιουργία των διαγραμμάτων ROC. Οι κατάλληλες αυτές διεργασίες αναφέρονται στον λόγο (Αθροιστική Πρόβλεψη Επιβίωσης / Αριθμό των παρατηρήσεων την δεδομένη στιγμή) για κάθε ασθενή. Τα διαγράμματα ROC θα εξυπηρετήσουν τόσο την αξιολόγηση όσο και απόδοση των μοντέλων. Έτσι θα μπορούμε συνδυάζοντας επιπλέον πληροφορίες να υποδείξουμε τα αποδοτικότερα προς χρήση. Η αρχική προσέγγιση ήταν να συγκεντρωθούν και να χρησιμοποιηθούν δεδομένα από τα τρία τμήματα των μονάδων περίθαλψης ασθενών όπου περιλαμβάνονται στην Βάση Δεδομένων της MIMIC IV version 1.0, τα οποία είναι 'Τμήμα Επειγόντων Περιστατικών', 'Νοσοκομείων' και 'Μονάδων Εντατικής Θεραπείας (ΜΕΘ)'. Έτσι από τα αντίστοιχα μέρη της βάσης δεδομένων απομονώθηκαν τα στοιχεία εκείνα που αποτυπώνουν ενδείξεις ζωτικών οργάνων καθώς και των εργαστηριακών εξετάσεων. Πάνω στα στοιχεία έγινε ομαδοποίηση των κοινών ενδείξεων και εφαρμόστηκε ο μέσος όρος τιμών τους ανά οχτάωρο. Τέλος αποτυπώθηκε σε μια ξεχωριστή στήλη αν ο ασθενής επιβίωσε ή όχι της νοσηλείας. Μετά την συγκέντρωση αυτών των στοιχείων

παρατηρήθηκε μεγάλη διασπορά στα δεδομένα των μονάδων εκτός ΜΕΘ. Αυτό είχε σαν συνέπεια να μην μπορεί να εξαχθεί έγκυρο αποτέλεσμα από τα τμήματα αυτά, σε αντίθεση με τις Μονάδες Εντατικής Θεραπείας όπου η πυκνότητα των δεδομένων είναι ικανοποιητική μιας και στις ΜΕΘ υπάρχει ανά ώρα καταγραφή ενδείξεων. Η δομή του παραγόμενου πίνακα είναι τέτοια όπου κάθε στήλη αναφέρεται σε μια ξεχωριστή ένδειξη είτε αυτή είναι ζωτικού οργάνου ή εργαστηριακή παρατήρηση και η κάθε γραμμή αποτυπώνει τις ανά οχτάωρο μετρήσεις του εκάστοτε ασθενούς στην εισαγωγή του. Η διαφοροποίηση της προσέγγισης αυτού του σχήματος σε σχέση με τα μέχρι τώρα που έχουν μελετηθεί είναι ως προς την αντιμετάθεση γραμμών και στηλών. Σε αυτό το σχήμα έχουμε παραθέσει σε στήλες τις κατηγορίες εξετάσεων και κλινικών παρατηρήσεων και σε γραμμές τους ασθενείς χρονικά ταξινομημένους μέσα στις εισαγωγές τους. Αυτή η προσέγγιση μας έδωσε την ευελιξία της προαναφερθείσας ομαδοποίησης, των ενδείξεων δηλαδή σε οχτάωρα. Κατ' επέκταση μέσω αυτής της προσέγγισης υπάρχει η δυνατότητα παραγωγής συμπερασμάτων όχι μόνο κατά την ολοκλήρωση της ημέρας, δηλαδή στο τέλος του εικοσιτετράωρου αλλά και κατά την διάρκεια αυτού μέσα στα οχτάωρα. Καταφέραμε με αυτή την προσέγγιση να έχουμε την ημέρα χωρισμένη σε τρία τμήματα και έτσι να ισορροπήσουμε ανάμεσα στον όγκο των δεδομένων και την παραγωγή του βέλτιστου αποτελέσματος. Επιπλέον η τιμή του οχτάωρου επιλέχτηκε για να υπάρχει ισορροπία μεταξύ του μεγάλου αριθμού τιμών που θα έλειπαν αν η τιμή ήταν μικρή και της χαμένης χρονικής πτυχής αν η τιμή της παρατήρησης ήταν μεγάλη. Ένας επιπλέον λόγος της επιλογής του οχτάωρου ήταν γιατί ταίριαζε με τις νοσηλευτικές βάρδιες και τις πιθανές διαφορετικές καταγραφές από το προσωπικό. Τα στάδια δημιουργίας του ενιαίου πίνακα χωρίζονται σε τρεις φάσεις. Στις δύο πρώτες συλλέγουμε, ομαδοποιούμε και αποτυπώνουμε ανεξάρτητα σε διαφορετικούς πίνακες τις καταχωρήσεις των ΜΕΘ (αρχείο *chartevents*), καθώς και των εργαστηριακών εξετάσεων (αρχείο *labevents*). Μετά το πέρας της προαναφερθείσας διαδικασίας οι δύο αυτοί ανεξάρτητοι πίνακες ενοποιούνται και στοιχίζονται ημερολογιακά ως προς τους ασθενείς. Έχοντας ενημερωθεί ως προς τα δεδομένα που φιλοξενούνται στο *chartevents* γνωρίζαμε ότι μέρος των στοιχείων του *labevents* εμπεριεχόντουσαν εξ' αρχής σε αυτό. Έπρεπε να δοθεί η δέουσα προσοχή αρχικά στα κοινά στοιχεία ανάμεσα σε αυτούς τους δύο πίνακες και επιπλέον να παρατηρηθούν τυχόν διαφοροποιήσεις στις τιμές ανάμεσα τους. Σε μια τέτοια περίπτωση τα στοιχεία του *labevents* θεωρούνται ακριβέστερα. Εν συνεχεία γίνεται έλεγχος για διπλές εγγραφές και γίνεται καθαρισμός από αυτές έτσι ώστε να έχουμε ένα ενιαίο μη επικαλυπτόμενο πίνακα για την τροφοδότηση των αλγορίθμων – ταξινομητών. Για να μπορέσει το σχήμα μας να ολοκληρωθεί και να τροφοδοτήσει τους αλγορίθμους μηχανικής μάθησης θα πρέπει ο ενιαία στοιχισμένος πίνακας να διαχωριστεί σε δύο κομμάτια. Φιλοξενώντας στο ένα εξ αυτών το εβδομήντα της εκατό των εγγραφών και το τριάντα της εκατό στο δεύτερο. Αυτά τα νέα αρχεία θα χρησιμοποιηθούν έως αρχεία εκπαίδευσης και ελέγχου.

2.2 MIMIC – IV ver1.0

Ξεκινώντας την αναφορά μας στην MIMIC παρατηρούμε ότι το ακρωνύμιο της παραμένει ίδιο ανάμεσα στις εκδόσεις παρόλο που ο τίτλος της έχει αλλάξει. Από Multiparameter Intelligent

Monitoring in Intensive Care όπου αναφερόταν μέχρι και την δεύτερη έκδοση της, έχει μετατραπεί σε Medical Information Mart for Intensive Care στην τρίτη και τέταρτη. Ανάμεσα στις εκδόσεις η βάση έχει διαφορετική δομή με την πιο ουσιαστική διαφοροποίηση στην τέταρτη έκδοση όπου στηρίχτηκε και η υλοποίηση της εν λόγω διατριβής. Η υλοποίηση της βάσης διαμορφώθηκε στηριζόμενη στην συλλογή και απόκτηση των δεδομένων από τα τρία τμήματα που απαρτίζουν τις νοσοκομιακές μονάδες. Από το τμήμα επειγόντων περιστατικών, το νοσοκομείο καθώς και από τις μονάδες εντατικής θεραπείας του Beth Israel Deaconess Medical Center. Στη συνέχεια τα δεδομένα αναδιοργανώθηκαν και απλοποιήθηκαν περιορίζοντας σε μεγάλο βαθμό την επανάληψη τους. Αυτό έγινε μέχρι το σημείο εκείνο όπου διασφαλίστηκε ο αντικατοπτρισμός του πραγματικού κλινικού αποτελέσματος.

Η επεξεργασία των προσωπικών δεδομένων που εμπεριέχονται στη βάση έγινε με την χρήση αλγορίθμων κρυπτογράφησης, επαναπροσδιορισμού ελεύθερου κειμένου για την αφαίρεση προστατευόμενων πληροφοριών υγείας “PHI” (Protected Health Information) καθώς και μετατόπισης ημερομηνιών στο μέλλον είτε αναφέρονται σε γέννησης, εισαγωγής, εξαγωγής ή θανάτου. Θα παρουσιάσω στις επόμενες παραγράφους αναλυτικότερα τι στοιχεία περιλαμβάνονται στη βάση. Είναι καθοριστικής σημασίας για την κατανόηση της MIMIC και κατ’ επέκταση της διατριβής ο τρόπος υλοποίησης και ο συσχετισμός της μετατόπισης των ημερομηνιών. Σε κάθε ασθενή εφαρμόζεται ανεξάρτητη μετατόπιση στο μέλλον των ημερομηνιών που εμπεριέχονται στη MIMIC όπως ανέφερα προηγουμένως. Τα χρονικά διαστήματα ανάμεσα στις μετρήσεις παρόλο που οι ημερομηνίες έχουν μετατοπιστεί στο μέλλον παραμένουν σταθερά. Ο περιορισμός όμως που υπάρχει είναι στην ημερολογιακή σύγκριση ανάμεσα σε δυο ασθενείς. Η ημερολογιακή μετατόπιση ανάμεσα στους ασθενείς με δεν γίνεται με τον ίδιο γνώμονα. Παρόλο που μπορεί να φαίνονται οι ίδιες ημερολογιακές εγγραφές μεταξύ ασθενών, δεν είναι απαραίτητο να αναφέρονται στην ίδια πραγματική ημερομηνία [4].

Στα πλαίσια των προσωπικών δεδομένων όπου ο σεβασμός τους αποτελεί βασική προϋπόθεση για την διανομή της MIMIC αφαιρέθηκαν χαρακτηριστικά και στοιχεία τα οποία θα μπορούσαν να ταυτοποιήσουν ασθενείς. Για την πρόσβαση στην εν λόγω βάση έπρεπε να ολοκληρωθεί από την μεριά μου επιτυχώς το πρόγραμμα Collaborative Institutional Training Initiative το οποίο είναι αφιερωμένο στην εξυπηρέτηση των εκπαιδευτικών αναγκών Πανεπιστημίων και Τεχνολογικών Ερευνητικών Οργανισμών, καθώς προάγουν την ακεραιότητα και την επαγγελματική εξέλιξη. Στο Παράρτημα I και II έχω παραθέσει τόσο το πιστοποιητικό επιτυχούς ολοκλήρωσης του προγράμματος καθώς και την αναλυτική βαθμολογία αυτού. Η συμφωνία – ρήτρα χρήσης της MIMIC περιλαμβάνει δέκα σημεία τα οποία αναφέρονται στο παράρτημα III και IV. Από τις βασικές δεσμεύσεις που έλαβα για την πρόσβαση σε αυτή μαζί με την ρήτρα εμπιστευτικότητας είναι τόσο η άμεση αναφορά στην περίπτωση που γίνει αντιληπτή ταυτοποίηση ασθενή καθόλη την διάρκεια της επεξεργασίας των δεδομένων καθώς και η κοινοποίηση του κώδικα που χρησιμοποιήθηκε για την επεξεργασία των δεδομένων. Για την υλοποίηση της διπλωματικής παράχθηκε κώδικας εκ του μηδενός χωρίς να γίνει χρήση κάποιου έτοιμου κομματιού ή χρήση τμήματος από αυτά που υπάρχουν στα διαδικτυακά αποθετήρια. Για τον λόγο αυτό ο κώδικας μου είναι διαθέσιμος όπως ορίζει και η συμφωνία στην ηλεκτρονική διεύθυνση του αποθετηρίου μου στο github:

<https://github.com/DimopDim/MimicIV/tree/main/Thesis>

2.2.1 Ενότητα Πυρήνα - Core

Στην συγκεκριμένη ενότητα βρίσκουμε στοιχεία των ασθενών που ο διαχωρισμός και η κατανομή τους γίνεται μέσα σε τρεις πίνακες όπως αποτυπώνονται στον πίνακα 2.1 που ακολουθεί.

Ονομασία Αρχείου	Περιγραφή
admissions	Παρέχονται πληροφορίες αναφορικά με την εισαγωγή ασθενών στο νοσοκομείο.
patients	Στοιχεία που αναφέρονται στον ασθενή.
transfers	Αποτυπώνονται ενδονοσοκομειακές μετακινήσεις ασθενών.

Πίνακας 2.1 Αρχεία Ενότητας Core

2.2.2 Ενότητα Νοσοκομείου - Hosp (Hospital)

Στη ενότητα του νοσοκομείου φιλοξενούνται εγγραφές που προέρχονται από το EHR. Μετρήσεις οι οποίες μπορεί να έχουν συλλεχθεί από το ίδιο το νοσοκομείο κατά την διάρκεια παραμονής του ασθενούς, είτε από εξωτερικά εργαστήρια. Μέσα σε αυτό περιλαμβάνονται πίνακες χωρισμένοι σε δυο βασικές κατηγορίες. Αρχικά στους πίνακες που περιλαμβάνουν πληροφορίες, επεξηγήσεις και περιγραφές των κωδικοποιημένων πληροφοριών που περιλαμβάνονται στην δεύτερη κατηγορία πινάκων όπου φιλοξενούν τις καταγραφές των ασθενών. Στον πίνακα 2.2 που ακολουθεί παρουσιάζονται τα ονόματα των πινάκων που αποτελούν την ενότητα Hosp και μια περιγραφή για το τι περιέχεται μέσα σε αυτούς.

Ονομασία Αρχείου	Περιγραφή
d_hcps	Περιέχει περιγραφές των CPT κωδικών όπου αναφέρονται στο αρχείο hrcsevents. Οι κωδικοί αυτοί αναφέρονται σε ιατρικές, διαγνωστικές καθώς και χειρουργικές υπηρεσίες. Οι CPT κωδικοί επηρεάζουν άμεσα το ποσό που θα πληρώσει ένα ασθενής για την ιατρική περίθαλψη.
d_icd_diagnoses	Περιέχει περιγραφές των ICD-9 και ICD-10 διαγνώσεων. Οι ICD κωδικοί είναι ένα παγκοσμίως χρησιμοποιούμενο διαγνωστικό εργαλείο.
d_icd_procedures	Αυτός ο πίνακας προσδιορίζει τους κωδικούς που εκχωρούνται στο τέλος της παραμονής ενός ασθενούς στο νοσοκομείο και χρησιμοποιούνται για τον υπολογισμό του κόστους παραμονής.
d_labitems	Περιέχει ορισμούς για τα στοιχεία που σχετίζονται με εργαστηριακές μετρήσεις.

diagnoses_icd	Περιέχει διαγνώσεις με τις οποίες χρεώθηκε κατά την παραμονή του ο ασθενής στο νοσοκομείο.
drgcodes	Οι κωδικοί αυτοί απευθύνονται στον κύριο λόγο για τον οποίο παρέμεινε ο ασθενής στο νοσοκομείο.
emar	Περιλαμβάνει στοιχεία αναφορικά με την καταγραφή των χορηγήσεων ενός φαρμάκου σε ένα ασθενή. Οι καταγραφές αυτές αναφέρονται σαν μετρήσεις κομοδίνου μιας και έχουν γίνει από το νοσηλευτικό προσωπικό δίπλα στο κρεβάτι του ασθενή.
emar_detail	Περιέχει πληροφορίες για τις χορηγήσεις φαρμάκων που γίνονται στον πίνακα emar.
hpcsevents	Περιλαμβάνει τιμολογημένα συμβάντα κατά την διάρκεια νοσηλείας.
labevents	Σε αυτό τον πίνακα εμπεριέχονται όλες οι εργαστηριακές μετρήσεις που έγιναν σε κάθε ασθενή.
microbiologyevents	Μικροβιολογικές εξετάσεις.
pharmacy	Αναλυτικές πληροφορίες συνταγογραφημένων φαρμάκων.
poe	Πληροφορίες παραγγελιών.
poe_detail	Συμπληρωματικές πληροφορίες για παραγγελίες που γίνονται από παρόχους στο νοσοκομείο.
prescriptions	Περιέχει πληροφορίες σχετικά με τα συνταγογραφημένα φάρμακα.
procedures_icd	Περιλαμβάνει τις διαδικασίες με τις οποίες χρεώθηκε ένας ασθενής
services	Οι υπηρεσίες που φρόντισαν τον ασθενή κατά την διάρκεια νοσηλείας.

Πίνακας 2.2 Αρχεία Ενότητας Hosp

2.2.3 Ενότητα Μονάδων Εντατικής Θεραπείας – ICU

Όπως στην ενότητα του νοσοκομείου έτσι και στην μονάδα εντατικής θεραπείας φιλοξενούνται εγγραφές που προέρχονται από το σύστημα EHR. Τα δεδομένα που περιλαμβάνονται σε αυτούς τους πίνακες είναι περιεκτικότερα μιας και στις μονάδες εντατικής θεραπείας οι μετρήσεις είναι εντατικότερες λόγω της σημαντικότητας του τμήματος και τις σοβαρής κατάστασης που βρίσκονται οι ασθενείς σε αυτό. Στον πίνακα 2.3 που ακολουθεί παρουσιάζονται τα ονόματα των πινάκων που αποτελούν την ενότητα ICU και μια περιγραφή για το τι περιέχεται μέσα σε αυτούς.

Ονομασία Αρχείου	Περιγραφή
d_items	Ορίζει έννοιες που καταγράφονται στον πίνακα συμβάντων.
chartevents	Από τους πιο σημαντικούς πίνακες που περιλαμβάνονται στην ενότητα. Περιέχει την πλειοψηφία των πληροφοριών που έχουν συλλεχθεί κατά την παραμονή ασθενών στις ΜΕΘ. Ένα από αυτά τα στοιχεία είναι το ηλεκτρονικό διάγραμμα όπου εμπεριέχει τα ζωτικά σημεία, στοιχεία ρύθμισης αναπνευστήρα, εργαστηριακές τιμές κ.α. Μέσα στο chartevent επαναλαμβάνονται οι τιμές των εργαστηριακών εξετάσεων που έχουμε στην διάθεση μας και από το labevents της ενότητας hosp.
datetime	Περιλαμβάνει όλες τις ημερολογιακές μετρήσεις ασθενών στις ΜΕΘ.
icustays	Χρονικές πληροφορίες αναφορικά με την παραμονή ασθενών στις ΜΕΘ
inputevents	Ο πίνακας αυτός περιλαμβάνει πληροφορίες αναφορικά με την έγχυση φαρμάκων στους ασθενείς. Η ιδιαιτερότητα του είναι ότι διαχωρίζει στις εγγραφές τα συστατικά των φαρμάκων καθώς και την περιεκτικότητά τους. Έτσι ακόμα και μια μικρή ποσοστιαία μεταβολή στη δόση ενός συστατικού θα είχε σαν συνέπεια την εκ νέου εγγραφή και αποτύπωση του στη βάση.
outputevents	Αναφέρεται στα στοιχεία εξόδου υγρών από τον ασθενή, το ακριβώς αντίθετο από το inputevents όπως για παράδειγμα της παροχέτευσης κ.α.
procedureevents	Διαδικασίες που τεκμηριώνονται κατά την παραμονή στη ΜΕΘ.

Πίνακας 2.3 Αρχεία Ενότητας ICU

2.2.4 Ενότητα Επειγόντων Περιστατικών – ED

Όπως και στις προαναφερθείσες ενότητες έτσι και σε αυτή συλλέγονται δεδομένα ασθενών οι οποίοι όμως τώρα έχουν επισκεφτεί το τμήμα επειγόντων περιστατικών. Τα δεδομένα αυτά παρατίθενται στον πίνακα 2.4

Ονομασία Αρχείου	Περιγραφή
diagnosis	Περιέχει τιμολογημένες διαγνώσεις οι οποίες καθορίζονται μετά την έξοδο από το τμήμα επειγόντων περιστατικών
edstays	Αποτελεί τον βασικό πίνακα παρακολούθησης ασθενών κατά την επίσκεψή τους στο τμήμα επειγόντων περιστατικών. Ένα από τα σημαντικά στοιχεία αυτού είναι ο χρόνος παραμονής στα επείγοντα.
medrecon	Αποτυπώνονται σε αυτόν τα φάρμακα που λαμβάνει ο ασθενής

	κατά την είσοδο του στα επείγοντα.
pyxis	Παρέχει πληροφορίες σχετικά με τις χορηγήσεις φαρμάκων.
triage	Ο πίνακας διαλογής όπως προσδίδει και το όνομα του περιέχει τις αρχικές πληροφορίες όπως για παράδειγμα τα ζωτικά όργανα έτσι ώστε να κριθεί η είσοδος του ασθενή στα επείγοντα.
vitalsign	Περιλαμβάνει τις ενδείξεις ζωτικών σημείων όπου η συλλογή αυτών γίνεται σε διάστημα από μία έως τέσσερις ώρες.
vitalsign_hl7	Τα στοιχεία αυτού του πίνακα περιλαμβάνουν ενδείξεις ζωτικών σημείων τα οποία έχουν καταγραφεί με τηλεμετρία.

Πίνακας 2.4 Αρχεία Ενότητας ED

2.3 Προεπεξεργασία δεδομένων

Για την επεξεργασία και συγκέντρωση των δεδομένων που φιλοξενούνται μέσα στην μεγάλη σε έκταση βάση Mimic IV έγινε χρήση της ερμηνευτικής γλώσσας προγραμματισμού Python. Ο κώδικας που έχει συγγραφεί για την εξαγωγή του ενιαίου πίνακα όπου παρουσιάστηκε στην παράγραφο 2.1 παρατίθεται στα παραρτήματα που ακολουθούν στο τέλος του παρόντος συγγράμματος. Τα βασικότερα πλεονεκτήματα από την χρήση της python είναι η ολοκλήρωση μακροχρόνιων εργασιών σε σύντομο χρονικό διάστημα ακόμα και με την χρήση απλών υπολογιστικών συστημάτων, όπως αυτό που χρησιμοποιήθηκε για την ολοκλήρωση αυτής της έρευνας. Ένας επιπλέον λόγος είναι η ύπαρξη πολλαπλών και ελεύθερων προς χρήση βιβλιοθηκών αναφορικά με την ανάλυση και οπτικοποίηση των δεδομένων. Ο κώδικας ο οποίος παρατίθεται στο παράρτημα αυτής της εργασίας συγγράφτηκε και εκτελέστηκε με την χρήση της πλατφόρμας Anaconda σε ένα επιτραπέζιο υπολογιστή με επεξεργαστή Intel Core i5 στα 3.20 GHz και μνήμη RAM 24GB.

2.3.1 Ασθενείς με ισχαιμικό εγκεφαλικό στο τμήμα 'Επείγοντων Περιστατικών'

Ο σκοπός όπως προκύπτει και από τον τίτλο της συγκεκριμένης παραγράφου είναι να ανασυρθούν από την βάση που φιλοξενεί τις εισαγωγές στο τμήμα επείγοντων περιστατικών, τους ασθενείς που διαγνώστηκαν με την πάθηση του ισχαιμικού εγκεφαλικού. Πριν την παραπομπή στο Παράρτημα V και ποιο συγκεκριμένα από το βήμα ένα έως έξι όπου εμπεριέχεται ο κώδικας συνοδευόμενος από σχόλια θα πρέπει να αναφέρουμε το χαρακτηριστικό εκείνο το οποίο λειτούργησε ως φίλτρο για τον διαχωρισμό των ασθενών αναφορικά με την πάθηση και διάγνωση που τους έγινε. Στον πίνακα 2.2 έχει παρατεθεί η περιγραφή του αρχείου d_icd_diagnoses όπου περιέχει την κωδικοποίηση των διαγνώσεων. Έτσι με βάση τους κωδικούς του ισχαιμικού εγκεφαλικού αναζητούμε και απομονώνουμε τους ασθενείς με την συγκεκριμένη διάγνωση. Στον πίνακα 2.5 που ακολουθεί παρατίθενται τα πρώτα στατιστικά στοιχεία που εξάγονται από την προεπεξεργασία των δεδομένων.

Περιγραφή	Αριθμός Ασθενών
Συνολικές εγγραφές στον πίνακα διαγνώσεων:	949.172
Εγγραφές ασθενών που νοσηλεύτηκαν:	448.972
Διαγνώσεις εγκεφαλικού:	2.292
Διαγνώσεις εγκεφαλικού που νοσηλεύτηκαν:	2.233
Μοναδικοί ασθενείς που διαγνώστηκαν με εγκεφαλικό:	2.145
Μοναδικοί ασθενείς που νοσηλεύτηκαν με εγκεφαλικό:	2.094
Εισαγωγές με εγκεφαλικό που δεν νοσηλεύτηκαν:	59
Μοναδικοί ασθενείς με εγκεφαλικό που δεν νοσηλεύτηκαν:	59

Πίνακας 2.5 Τμήμα Επειγόντων Περιστατικών

Έχοντας τώρα στη διάθεση μας τους ασθενείς με ισχαιμικό εγκεφαλικό που έχουν περάσει από το τμήμα επειγόντων περιστατικών ενσωματώνουμε στο επόμενο στάδιο τις ενδείξεις των ζωτικών οργάνων που εμπεριέχονται στο αρχείο Vitalsign. Από το έβδομο έως το εντέκατο βήμα του παραρτήματος V έχει παρατεθεί ο σχετικός κώδικας. Τα στατιστικά αποτελέσματα του κώδικα που εκτελέστηκε αποτυπώνονται στον πίνακα 2.6.

Περιγραφή	Αριθμός Εγγραφών
Σύνολο μετρήσεων:	1.651.119
Σύνολο ομαδοποιημένων μετρήσεων οχταώρου:	752.146
Ομαδοποιημένες μετρήσεις οχταώρου ασθενών με εγκεφαλικό:	2.645

Πίνακας 2.6 Εγγραφές Ζωτικών Οργάνων στο Τμήμα Επειγόντων Περιστατικών.

2.3.2 Ασθενείς με ισχαιμικό εγκεφαλικό στις 'Νοσοκομειακές Μονάδες'

Την τεχνική που εφαρμόσαμε για την συγκέντρωση στοιχείων από το τμήμα επειγόντων περιστατικών ακολουθούμε και στις νοσοκομειακές μονάδες. Στο παράρτημα VI έχει παρατεθεί ο κώδικας και στον πίνακα 2.7 τα στοιχεία των διαγνώσεων εγκεφαλικού στις νοσοκομειακές μονάδες.

Περιγραφή	Αριθμός Εγγραφών
Σύνολο νοσοκομειακών διαγνώσεων:	5.280.351
Εισαγωγές στις νοσοκομειακές μονάδες:	523.740
Διαγνώσεις ασθενών με εγκεφαλικό:	12.015
Μοναδικές διαγνώσεις ασθενών με εγκεφαλικό (Δεν περιλαμβάνονται οι ενδονοσοκομειακές μετακινήσεις):	9.926

Μοναδικοί ασθενείς με εγκεφαλικό (Χωρίς τις επανεισαγωγές):	8.320
Επιβίωσαν στις νοσοκομειακές μονάδες (89,31%)	7.431
Απεβίωσαν στις νοσοκομειακές μονάδες (10,68%)	889

Πίνακας 2.7 Διαγνώσεις εγκεφαλικού στις νοσοκομειακές μονάδες.

2.3.3 Ασθενείς με ισχαιμικό εγκεφαλικό στις 'Μονάδες Εντατικής Θεραπείας'

Ακολουθώντας και στο τμήμα των Μονάδων Εντατικής Θεραπείας θέλουμε να βρούμε ποιοι ασθενείς έχουν εισαχθεί με ισχαιμικό εγκεφαλικό. Στα βήματα από 1 έως 5 του παραρτήματος VII έχει παρατεθεί ο κώδικας και στον πίνακα 2.8 τα στοιχεία των διαγνώσεων εγκεφαλικού στις Μονάδες Εντατικής Θεραπείας.

Περιγραφή	Αριθμός Εγγραφών
Σύνολο εισαγωγών:	5.436
Μοναδικές εισαγωγές ασθενών (Δεν περιλαμβάνονται οι εσωτερικές μετακινήσεις):	3.809
Μοναδικοί ασθενείς (Δεν περιλαμβάνονται οι επανεισαγωγές):	3.646
Επιβίωσαν στις ΜΕΘ (73,71%)	2.887
Απεβίωσαν στις ΜΕΘ (26,29%)	759

Πίνακας 2.8 Ασθενείς με εγκεφαλικό στις ΜΕΘ.

2.3.4 Κλινικές καταγραφές ισχαιμικού εγκεφαλικού

Έχοντας στην διάθεση μας πλέον απομονωμένους τους ασθενείς που έχουν διαγνωστεί με ισχαιμικό εγκεφαλικό θα πρέπει να συλλέξουμε τα στοιχεία που αναφέρονται σε αυτούς από το αρχείο των καταγραφών ΜΕΘ (chartevents). Μέχρι αυτή τη στιγμή στις γραμμές του κώδικα που έχω παραθέσει στα παραρτήματα η επεξεργασία των δεδομένων εφαρμοζόταν στο σύνολο των αρχείων που εμπεριέχονται στη βάση της MIMIC. Τα βήματα που εφαρμοζόντουσαν ήταν η φόρτωση του εκάστοτε αρχείου της βάσης, ακολουθούσε η επεξεργασία αυτού, συνεχίζαμε με την συλλογή των χρήσιμων προς την εξυπηρέτηση του σχήματος στοιχείων από αυτό και δημιουργούσαμε ένα νέο αρχείο όπου αποτελούσε την βάση για τα επόμενα βήματα μας μέχρι να φτάσουμε στον επιθυμητό τελικό ενιαίο πίνακα. Τα στοιχεία των καταγραφών ΜΕΘ (chartevents) εμπεριέχονται σε ένα από τους μεγαλύτερους πίνακες που καλέστηκα να επεξεργαστώ στην βάση δεδομένων της MIMIC IV. Η επεξεργασία του στο σύνολο του αρχείου απαιτούσε αρκετούς υπολογιστικούς πόρους με αποτέλεσμα η διαχείριση του να μην είναι δυνατόν να υλοποιηθεί μονομιάς στο σύνολο του αρχείου στον υπολογιστή που χρησιμοποιήθηκε κατά την διεξαγωγή της έρευνας. Η προσέγγιση για την επίλυση αυτού του προβλήματος ήταν η τμηματοποίηση του αρχείου σε επιμέρους κομμάτια τα οποία ανήλθαν στα επτά. Έπειτα ακολούθησε η συλλογή και αποθήκευση σε επτά ανεξάρτητα αρχεία των στοιχείων που απομονώθηκαν. Τέλος τα αρχεία συγχωνεύτηκαν σε ένα νέο αρχείο. Για την

αρχική τμηματοποίηση του αρχείου των κλινικών στοιχείων χρησιμοποιήθηκε ένα πρόγραμμα τρίτου κατασκευαστή του οποίου ο κώδικας δεν είναι διαθέσιμος προς κοινοποίηση και ο οποίος υλοποιήθηκε με χρήση της γλώσσα Python. Από το έκτο έως το δέκατο τέταρτο βήμα του παραρτήματος VII παρατίθεται ο κώδικας της επεξεργασίας του πίνακα των καταγραφών ΜΕΘ και η παραγωγή αρχείου που περιλαμβάνει μόνο τις καταγραφές των ασθενών με ισχαιμικό εγκεφαλικό.

Οι διαγνωστικοί έλεγχοι στους οποίους επέβαλα τα αποτελέσματα μου μέχρι τώρα είχαν σαν συνέπεια ο κώδικας που έχω παρουσιάσει να εξάγει τα επιθυμητά αποτελέσματα. Ο κώδικας που έχει δημιουργηθεί για τον έλεγχο της ορθότητας των αποτελεσμάτων είναι διαθέσιμος στο ηλεκτρονικό αποθετήριο του github όπου η διεύθυνση του έχει δηλωθεί στην παράγραφο 2.2. Ο κώδικας ελέγχου δεν έχει παρατεθεί στα παραρτήματα αποσκοπώντας σε ένα λιτό και ευανάγνωστο κώδικα. Ο πίνακας των καταγραφών ΜΕΘ μετά από τους απαραίτητους ελέγχους διαπιστώθηκε ότι έπρεπε να υποβληθεί σε μια επιπλέον επεξεργασία πριν οι καταγραφές του περάσουν στο στάδιο της ομαδοποίησης ανά οχτάωρο. Ο σκοπός της επεξεργασίας αυτής είναι η αφαίρεση των καταγραφών που έχουν σαν δεδομένα στις περιοχές τους κείμενα. Ο σκοπός του σταδίου της εργασίας που βρισκόμαστε είναι η περισυλλογή των αριθμητικών στοιχείων που απευθύνονται σε εργαστηριακές και κλινικές παρατηρήσεις. Τα κείμενα δεν μπορούν να ομαδοποιηθούν κατά την τεχνική που εφαρμόζεται στην εργασία, οπότε η παραμονή τους θα έχει σαν συνέπεια την αύξηση του μεγέθους του τελικού πίνακα με στοιχεία τα οποία δεν θα ήταν χρήσιμα, προκαλώντας κατά την ανάλυση μεγαλύτερη αραίωση στα δεδομένα. Θα δούμε και θα εξηγήσουμε περισσότερα κατά την ανάλυση του αποτελέσματος. Στο παράρτημα VII και στα βήματα από το δέκατο πέμπτο έως το δέκατο όγδοο αναπτύσσετε κώδικας που υλοποιεί τις παραπάνω ενέργειες. Εν συνεχεία στο προαναφερθέν παράρτημα και στο στάδιο των βημάτων από το δέκατο ένατο έως το εικοστό τρίτο υλοποιείτε η ομαδοποίηση των εγγραφών ανά οχτάωρο. Στον πίνακα 2.9 αποτυπώνονται τα στοιχεία που αφορούν των πίνακα κλινικών καταγραφών.

Περιγραφή	Αριθμός Εγγραφών
Συνολικές εγγραφές που αφορούν ασθενείς με εγκεφαλικό:	26.600.037
Εγγραφές που περιέχουν κείμενα:	16.725.563
Εγγραφές που περιέχουν τιμές:	9.874.474
Ομαδοποιημένες εγγραφές που περιέχουν τιμές	4.177.458
Αναγνωριστικοί κωδικοί κλινικών μετρήσεων	707

Πίνακας 2.9 Είδος καταγραφών στον πίνακα καταγραφών ΜΕΘ.

Συνοψίζοντας τα βήματα που έχουν υλοποιηθεί κατά την επεξεργασία των καταγραφών ΜΕΘ στην συγκεκριμένη παράγραφο τεκμαίρεται ότι πλέον διαθέτουμε αρχείο πίνακα όπου περιέχει τις μέσες τιμές οχτάωρου των εγγραφών παρατηρήσεων που αναφέρονται στους ασθενείς με ισχαιμικό εγκεφαλικό στις Μονάδες Εντατικής Θεραπείας. Στην συνέχεια του παραρτήματος VII και από το εικοστό τέταρτο βήμα έως το τριακοστό πρώτο παρατίθεται ο κώδικας δημιουργίας ενός κενού πίνακα όπου η κάθε στήλη φιλοξενεί τα είδη των

παρατηρήσεων όπως για παράδειγμα την θερμοκρασία, την πίεση κ.α. που στο σύνολο τους ανέρχονται στις επτακόσιες επτά και τις πλειάδες που αντιστοιχούν στις ομαδοποιημένες ανά οχτάωρο εισαγωγές ασθενών.

Πλέον έχουμε στην διάθεση μας τόσο τον πίνακα που περιέχει τις ομαδοποιημένες ανά οχτάωρο καταγραφές καθώς και ένα κενό με εναλλακτική ως προς τις στήλες πίνακα προς συμπλήρωση. Στην συνέχεια του παραρτήματος VII και πιο συγκεκριμένα από το τριακοστό δεύτερο βήμα έως το τεσσαρακοστό δεύτερο γίνεται η συμπλήρωση μαζί με τις απαραίτητες τροποποιήσεις αυτού. Πλέον έχουμε εμπράκτως στα χέρια μας μία πρώτη έκδοση τόσο του πίνακα και βρισκόμαστε πιο κοντά στον στόχο που έχει δηλωθεί στην αρχιτεκτονική του συστήματος μας. Υπενθυμίζοντας τον ο στόχος είναι ένας ενιαίος πίνακας όπου οι γραμμές θα αποτελούνται από τις παρατηρήσεις οχτάωρου των ασθενών και οι στήλες θα μπορούν ανεξάρτητα να φιλοξενούν ανά κατηγορία τις καταγεγραμμένες ενδείξεις.

Περιγραφή	Αριθμός Εγγραφών
Σύνολο στηλών πίνακα:	711
Στήλες πίνακα που αποτυπώνουν στοιχεία ασθενών	4
Στήλες πίνακα που αποτυπώνουν κατηγορίες ενδείξεων	707
Συνολικές γραμμές πίνακα	65.189
Σύνολο κελιών πίνακα	46.088.623
Συμπληρωμένα κελιά	4.177.458
Κενά κελιά	41.911.165

Πίνακας 2.10 Ενοποιημένος πίνακας καταγραφών ΜΕΘ.

2.3.5 Εργαστηριακές ενδείξεις ισχαιμικού εγκεφαλικού

Με την ολοκλήρωση των βημάτων και έχοντας πλήρως εφαρμοστεί η τεχνική που έχει παρατεθεί μέχρι τώρα για την παραγωγή του ενιαίου πίνακα που φιλοξενεί τις καταγραφές των ΜΕΘ, ακολουθεί ο δεύτερος σε μέγεθος πίνακας καταγραφών και παρατηρήσεων που καλούμαστε να συγκεντρώσουμε στοιχεία. Το αρχείο labevents όπως παρουσιάστηκε και στον πίνακα 2.2 εμπεριέχει όλες τις εργαστηριακές μετρήσεις που έγιναν στους ασθενείς που νοσηλεύτηκαν στα διάφορα τμήματα των μονάδων υγείας. Οπότε ο στόχος είναι η απομόνωση μέσα από αυτό των εξετάσεων που έγιναν σε ασθενείς με διάγνωση ισχαιμικού εγκεφαλικού στις ΜΕΘ. Λόγω του μεγάλου μεγέθους του αρχείου των εργαστηριακών εξετάσεων η εργασία αυτή ήρθε εκ νέου αντιμέτωπη με την έλλειψη υπολογιστικών πόρων. Πρόβλημα που αρχικά εμφανίστηκε κατά την επεξεργασία του chartevents. Ίδιο πρόβλημα σε διαφορετικό αρχείο σημαίνει και κατ' επέκταση ίδια προσέγγιση ως προς την λύση. Εφαρμόστηκε επιμερισμός του αρχείου σε περισσότερα από ένα και πιο συγκεκριμένα σε τρία, στα οποία εφαρμόστηκε ξεχωριστή επεξεργασία ως προς την περισυλλογή των καταγραφών που αναφέρονται στους ασθενείς στόχο και ενοποιήθηκαν στο τελείωμα τους. Έτσι πλέον έχουμε σε ένα ενιαίο αρχείο συγκεντρωμένες όλες τις εγγραφές εργαστηριακών εξετάσεων των αναφερόμενων ασθενών.

Στη συνέχεια οι καταγεγραμμένες εργαστηριακές ενδείξεις ομαδοποιήθηκαν ως προς το οχτάωρο. Στο παράρτημα VIII και πιο συγκεκριμένα από το πρώτο έως το δωδέκατο βήμα έχει παρατεθεί ο κώδικας υλοποίησης της προαναφερθείσας διαδικασίας.

Ακολουθεί η ανάλυση και εξαγωγή των κωδικών που αναφέρονται στις εξετάσεις που έχουμε συμπεριλάβει στο παραγόμενο αρχείο του δωδέκατου βήματος του παραρτήματος VIII. Ο σκοπός αυτής της διενέργειας είναι για να παραθέσω αυτά τα στοιχεία ως κεφαλίδες των στηλών του πίνακα που θέλω να συμπεριλάβω τις εγγραφές των εργαστηριακών εξετάσεων. Στο ίδιο παράρτημα και στα βήματα από το δέκατο τρίτο έως το εικοστό πρώτο έχω αποτυπώσει τον κώδικα που υλοποιεί την συγκεκριμένη σκέψη.

Πλέον υπάρχει στη διάθεση μας ο κενός και έτοιμος προς συμπλήρωση πίνακας που θα φιλοξενεί τις εργαστηριακές εξετάσεις και από το εικοστό δεύτερο βήμα του παραρτήματος VIII έως τριακοστό πρώτο γίνεται αναφορά στην τεχνική που χρησιμοποιήθηκε αποτυπώνοντας τον κώδικα υλοποίησης. Στον πίνακα 2.11 που ακολουθεί εμπεριέχονται τα χαρακτηριστικά του αρχείου που φιλοξενεί τις εργαστηριακές εξετάσεις.

Περιγραφή	Σύνολο
Σύνολο στηλών πίνακα:	392
Στήλες πίνακα που αποτυπώνουν στοιχεία ασθενών:	4
Στήλες πίνακα που αποτυπώνουν κατηγορίες ενδείξεων:	388
Συνολικές γραμμές πίνακα:	68.191
Σύνολο κελιών πίνακα:	26.458.108
Συμπληρωμένα κελιά:	1.276.856
Κενά κελιά:	25.181.252

Πίνακας 2.11 Χαρακτηριστικά πίνακα εργαστηριακών εξετάσεων.

Έχουμε φτάσει στο σημείο εκείνο όπου έχουμε συγκεντρώσει όλα εκείνα τα στοιχεία που αφορούν τους ασθενείς που έχουν εισαχθεί με διάγνωση ισχαιμικού εγκεφαλικού στις μονάδες εντατικής θεραπείας σε δυο ξεχωριστούς πίνακες. Αυτό που απομένει να κάνουμε τώρα είναι να τα συγκεντρώσουμε όλα αυτά μέσω της ενοποίησης στο ίδιο. Στην περίπτωση που υπάρξουν επικαλυπτόμενες στήλες ή γραμμές ο κώδικας μας αποδοτικά θα τις αντιμετωπίσει διατηρώντας τις τιμές. Βέβαια καθόλη την διάρκεια υλοποίησης του έγιναν οι απαραίτητοι έλεγχοι έτσι ώστε να μην αντιμετωπίσουμε ένα τέτοιο πρόβλημα. Επικαλυπτόμενες τιμές θα σήμαινε και πιθανές απώλειες σε δεδομένα. Στο παράρτημα IX από το πρώτο έως το πέμπτο βήμα παρατίθεται ο κώδικας υλοποίησης αυτής της συγχώνευσης. Στον πίνακα 2.12 αποτυπώνεται το σύνολο των χαρακτηριστικών του ενιαίου πίνακα που αναφέρονται στους ασθενείς με εγκεφαλικό στις ΜΕΘ.

Περιγραφή	Σύνολο
Αριθμός στηλών:	1.095
Αριθμός γραμμών:	90.817
Κελιά που απαρτίζουν τον πίνακα:	99.444.615
Εγγράψιμα στοιχεία:	5.454.314
Κενές περιοχές:	93.990.301

Πίνακας 2.12 Χαρακτηριστικά ενιαίου πίνακα ασθενών στις ΜΕΘ.

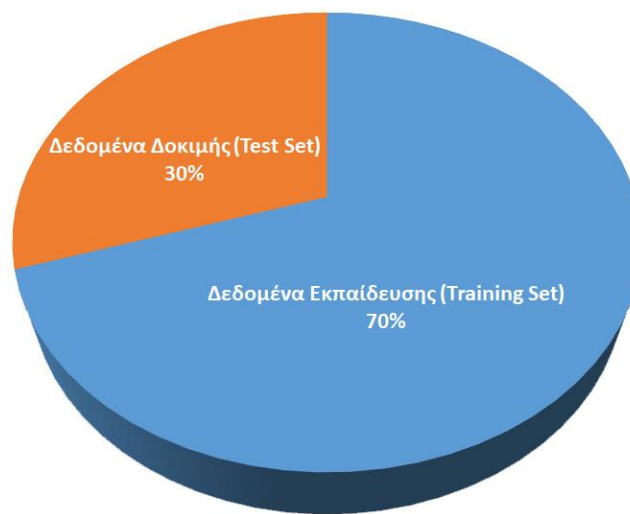
Από τον πίνακα μας υπολείπεται μια τελευταία αλλά ζωτικής σημασίας στήλη. Θα πρέπει να τοποθετήσουμε στο τέλος αυτού το στοιχείο της θνησιμότητας και να το ορίσουμε λεκτικά μιας και μέσα στην Mimic αποτυπώνεται η θνησιμότητα με το νούμερο ένα και η επιβίωση με το νούμερο μηδέν. Ο λόγος που θα πρέπει να γίνει αυτή η μετατροπή από νούμερα σε λέξεις είναι για να μπορέσει ο εκάστοτε αλγόριθμος παραγωγής μοντέλου στο WEKA να βλέπει την εγγραφή σαν ονομαστική οντότητα και όχι σαν νούμερο. Με αυτό τον τρόπο θα μπορέσει να εκπαιδευτεί κατάλληλα βλέποντας το πρόβλημα που καλείτε να επιλύσει ως πρόβλημα ταξινόμησης. Στο παράρτημα ΙΧ και πιο συγκεκριμένα από το έκτο έως το ενδέκατο βήμα γίνεται αναφορά στον κώδικα υλοποίησης.

2.3.6 Δεδομένα εκπαίδευσης – ελέγχου (Training – Test Set)

Τα δεδομένα μας πλέον έχουν συγκεντρωθεί κάτω από ένα ενιαίο πίνακα όπου σαν σκοπό θα έχουν να τροφοδοτήσουν το μοντέλο μας και να συμβάλουν στην πρόγνωση και πρόβλεψη της θνησιμότητας σε ασθενείς που έχουν διαγνωστεί με ισχαιμικό εγκεφαλικό. Στα πλαίσια αυτής της εργασίας γίνεται διαχωρισμός του συνόλου των δεδομένων σε Δεδομένα Εκπαίδευσης (Training Set) και Δεδομένα Ελέγχου (Test Set). Η τεχνική αυτή του διαχωρισμού αποσκοπεί στην δημιουργία ενός αξιόπιστου μοντέλου το οποίο θα παράγει μη προκατειλημμένα αποτελέσματα.

Η ιδέα πίσω από τα δεδομένα εκπαίδευσης – ελέγχου (Training – Test Set) είναι η δημιουργία ενός μοντέλου από τα δεδομένα εκπαίδευσης το οποίο θα προσπαθήσει να προβλέψει τα δεδομένα ελέγχου. Όταν ολοκληρωθεί επιτυχώς αυτή η διαδικασία τότε θα μπορούμε να τον χρησιμοποιήσουμε με άγνωστα ως προς το αποτέλεσμα δεδομένα. Στα χέρια μας έχουμε ένα σύνολο δεδομένων (100 %) όπου σε αυτό εμπεριέχονται πληροφορίες που συλλέξαμε από το κώδικα που δημιουργήσαμε στις προηγούμενες παραγράφους. Επιπλέον μέσα σε αυτά εμπεριέχεται και η βασική προς το αποτέλεσμα στήλη που δεν είναι άλλη από αυτή που δείχνει αν ο ασθενής επιβίωσε ή όχι. Κατά τον διαχωρισμό των δεδομένων έπρεπε να διασφαλιστεί ότι το αποτέλεσμα είναι αντιπροσωπευτικό του συνόλου δεδομένων και δεν έχουμε επικάλυψη σε αυτά. Δηλαδή τα δεδομένα εκπαίδευσης δεν επαναλαμβάνονται στα δεδομένα ελέγχου. Αν δεν ελέγχαμε την επικάλυψη τότε ο αλγόριθμος μας θα έβγαζε λανθασμένα ή για να είμαστε πιο ακριβείς, θα παρατηρούσαμε αυξημένα επίπεδα στην ακρίβεια και προσαρμογή του. Αν μη τι άλλο αυτό είναι και το αναμενόμενο, όταν εκπαιδεύεις έναν αλγόριθμο με συγκεκριμένα δεδομένα και τον τροφοδοτείς με τα ίδια για να το ελέγξεις

τότε η ακρίβεια του αποτελέσματος θα είναι αυξημένη. Η σειρά προτεραιότητας στην διαδικασία παραγωγής του μοντέλου πρόβλεψης ξεκίνησε τροφοδοτώντας τον αλγόριθμο με τα δεδομένα εκπαίδευσης. Από το σύνολο των στηλών συλλέγει τα καταγεγραμμένα δεδομένα και κάνει την αντιστοίχιση με την στήλη της θνησιμότητας. Έτσι εκπαιδεύεται και κρίνει αν με τις αναφερόμενες ενδείξεις της κάθε πλειάδας ο ασθενής κατάφερε να επιβιώσει ή όχι. Αφού ολοκληρωθεί αυτό το στάδιο ακολουθεί ο έλεγχος του μοντέλου. Τροφοδοτήσαμε το παραγόμενο μοντέλο με τα δεδομένα ελέγχου όπου εφάρμοσε συγκρίσεις ανάμεσα τους χωρίς όμως να κάνει χρήση της στήλης που αποτυπώνει την επιβίωση. Ο σκοπός είναι να παραχθεί ένα καινούργιο αποτέλεσμα που αφορά την θνησιμότητα στα δεδομένα εκπαίδευσης. Έτσι θα έχουμε την δυνατότητα της σύγκρισης του αποτελέσματος που παρήγαγε το μοντέλο εκπαίδευσης με το πραγματικό που βρίσκεται αποτυπωμένο στην στήλη της θνησιμότητας. Μετά από αυτόν τον έλεγχο θα μπορεί να διαπιστωθεί αν το εξαγώγιμο αποτέλεσμα είναι σωστό ή όχι. Στην εικόνα 2.1 αποτυπώνεται το ποσοστό των δεδομένων που κατά τον διαχωρισμό δόθηκαν για εκπαίδευση (70%) και δοκιμή (30%). Στο δωδέκατο βήμα του κώδικα του παραρτήματος ΙΧ αποτυπώνεται το σημείο διαχωρισμού των πλειάδων μέσα στο στοιχισμένο αρχείο. Στον πίνακα 2.13 οι πλειάδες που απαρτίζουν τα νέα αρχεία εκπαίδευσης (Training.csv) και ελέγχου (Test.csv).



Εικόνα 2.1 Διαχωρισμός συνόλου δεδομένων σε Training & Test Set

Περιγραφή	Σύνολο
Αριθμός πλειάδων ενιαίου αρχείου:	90.817
Αριθμός πλειάδων αρχείου εκπαίδευσης (Training Set) :	63.568
Αριθμός πλειάδων αρχείου ελέγχου (Test Set) :	27.249

Πίνακας 2.13 Δεδομένα Εκπαίδευσης (Training Set) – Ελέγχου (Test Set).

Από τους ασθενείς που επιλέχθηκαν αρχικά για το σετ δοκιμών (Training Set), δεκάπεντε ασθενείς είτε δεν επέζησαν είτε πήραν εξιτήριο εντός της πρώτης ημέρας. Αυτοί οι ασθενείς αποκλείστηκαν από τα πειράματα μιας και η παρατήρηση μας γίνεται στο τέλος του κάθε εικοσιτετράωρου. Η διαφορά στον αριθμό των ασθενών για διαφορετικές ημέρες οφείλεται στο γεγονός ότι ορισμένοι ασθενείς πήραν εξιτήριο ή δεν επέζησαν στο μεταξύ. Στον πίνακα 2.14 παρατίθενται τα σχετικά αποτελέσματα.

	Δεδομένα Εκπαίδευσης	Κατά την διάρκεια της 1 ^{ης} ημέρας	Δεδομένα Ελέγχου (Test Set)			
			1 ^η Ημέρα	2 ^η Ημέρα	3 ^η Ημέρα	4 ^η Ημέρα
Ασθενείς που επιβίωσαν	2051	1	866	824	719	593
Ασθενείς που δεν επιβίωσαν	518	14	196	155	121	104
Σύνολο	2569	15	1062	979	840	697

Πίνακας 2.14 Ασθενείς που επιβίωσαν ή όχι στα Δεδομένα Εκπαίδευσης και Ελέγχου.

2.4 Αλγόριθμοι μηχανικής μάθησης

Έχοντας πλέον ολοκληρώσει και κατέχοντας τα αρχεία εκπαίδευσης και ελέγχου ήρθε η ώρα για την υποβολή τους στους αλγορίθμους μηχανικής μάθησης. Έντεκα στο σύνολο τους οι αλγόριθμοι οι οποίοι υποβλήθηκαν στη διαδικασία της εκπαίδευσης, μοντελοποίησης και αξιολόγησης, τα αποτελέσματα των οποίων μαζί με περιγραφή αυτών θα παρατεθούν στο κεφάλαιο 3 που ακολουθεί όπου θα γίνει παρουσίαση των πειραματικών αποτελεσμάτων.

3

Πειραματικά Αποτελέσματα

3.1 Αλγόριθμοι Μηχανικής Μάθησης (Ταξινομητές)

Τα δεδομένα τα οποία έχουν παραχθεί κατά την χρήση των αλγορίθμων της προηγούμενης ενότητας θα χρησιμοποιηθούν για την εκπαίδευση μιας πλειάδας μοντέλων μηχανικής μάθησης. Μοντέλα που αποσκοπούν στην πρόβλεψη θνησιμότητας σε ασθενείς με ισχαιμικό εγκεφαλικό. Για την επιλογή του κατάλληλου αλγορίθμου υποβάλαμε έντεκα στο σύνολο με διαφορετικές ρυθμίσεις στην διαδικασία εκπαίδευσης και αξιολόγησης. Μετά την ολοκλήρωση της αξιολόγησης παρήχθησαν αποτελέσματα τα οποία συγκρίθηκαν μεταξύ τους. Μετά από αυτή τη σύγκριση μπορέσαμε και προτείναμε τους τρεις αποδοτικότερους σε σύγκριση ταξινομητές. Η εργαλειοθήκη WEKA (Waikato Environment for Knowledge Analysis) συνέβαλε στην υλοποίηση αυτής της εργασίας στα στάδια που αναφέραμε προηγουμένως.

Το WEKA είναι μια συλλογή αλγορίθμων μηχανικής εκμάθησης και προεπεξεργασίας δεδομένων που περιλαμβάνει πολύ περισσότερους αλγορίθμους σε σχέση με αυτούς που αξιοποιήθηκαν σε αυτή την έρευνα [5]. Παρείχε στην εργασία μας εκτεταμένη υποστήριξη στα δεδομένα τόσο με την στατιστική αξιολόγηση τους όσο και με την οπτικοποίηση τους. Άλλοι λόγοι που το έκαναν ιδανικό και αποτέλεσαν σημαντικό παράγοντα στην επιλογή του είναι ότι ανήκει στην κατηγορία του ανοιχτού κώδικα, τρέχει σε πολλαπλά περιβάλλοντα όπως Windows, Mac OS X και Linux και περιέχει αλγορίθμους τελευταίας τεχνολογίας [6]

3.2 Στοιχεία ταξινομητών

Ο σκοπός των ταξινομητών όπου καλό είναι να τους ορίσουμε ως στατιστικά υπολογιστικά μοντέλα, είναι να διαιρούν ένα σύνολο δεδομένων σε ομάδες, όπου κατά την ολοκλήρωση της ταξινόμησης είναι σημαντικό να τους υποβάλουμε στην διαδικασία της αξιολόγησης. Υλοποίηση της αξιολόγησης όπου γίνεται μέσω της ερμηνείας των παραγόμενων αποτελεσμάτων μας δίνει την δυνατότητα να εξάγουμε συμπεράσματα όπου αφορούν την αποτελεσματικότητα τους και κατ' επέκταση της χρησιμότητας τους σε σύγκριση με άλλους ταξινομητές ή μεθόδους που θα της χαρακτηρίζαμε ανταγωνίστριες. Τα μέτρα απόδοσης των ταξινομητών απεικονίζονται είτε μέσω γραφημάτων είτε με αριθμητικά αποτελέσματα. Στις μελέτες της βιοπληροφορικής οι ταξινομητές έχουν εφαρμοστεί τα τελευταία χρόνια με αρκετά μεγάλη επιτυχία [7]. Στην κατηγορία των ιατρικών προβλημάτων η οποία εμπεριέχει ως επί το πλείστον έντονα μη ισορροπημένα δεδομένα [8] όπως και τα δεδομένα αυτής της εργασίας, γίνεται συνδυαστική χρήση αποτελεσμάτων για την εξαγωγή συμπερασμάτων. Η αναφορά στα μη ισορροπημένα δεδομένα αυτής της εργασίας γίνεται μιας και η ταξινόμηση μας γίνεται με κριτήριο την επιβίωση των ασθενών. Οι ασθενείς που επιβίωσαν (73,71%) είναι κατά πολλοί περισσότεροι σε σχέση με αυτούς που δεν επιβίωσαν (26,29%). Εκεί παρατηρείται και η ανισορροπία των δεδομένων. Τα κοινά χρησιμοποιούμενα μέτρα της απόδοσης του ταξινομητή στη φάση της κατασκευής του μοντέλου είναι η ακρίβεια, το ποσοστό σφάλματος και η περιοχή λειτουργικών χαρακτηριστικών δέκτη (ROC) [9]. Στις παραγράφους που ακολουθούν γίνεται παρουσίαση και επεξήγηση των στοιχείων ακριβείας που παράγονται κατά την αξιολόγηση των αλγορίθμων για να ακολουθήσει η σύγκριση τους και μέσω αυτής η επιλογή του ιδανικού προς χρήση.

Πριν προχωρήσουμε στην εξειδικευμένη επεξήγηση των αριθμητικών απεικονίσεων καλό είναι να δούμε πως ακριβώς λειτουργεί ένας ταξινομητής. Το σύνολο των δεδομένων διαχωρίζονται σε δυο διαφορετικές κλάσεις όπως φαίνονται στην εικόνα 3.1.

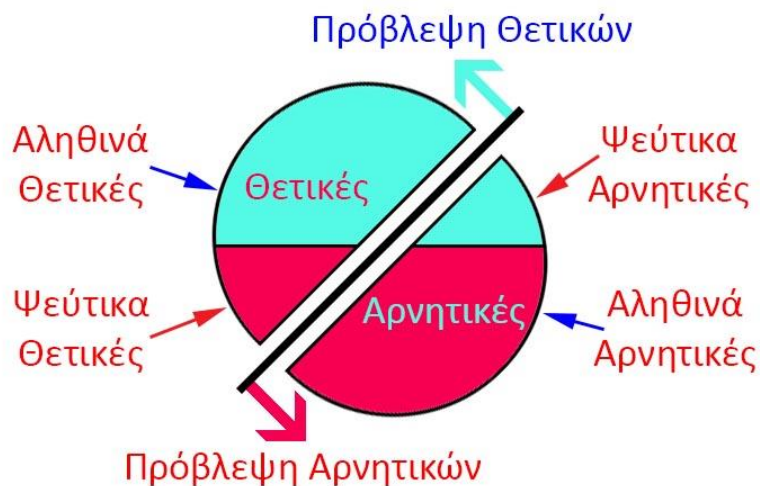


Εικόνα 3.1 Διαχωρισμός του συνόλου δεδομένων σε δύο κλάσεις.

Ο ταξινομητής έπειτα ταξινομεί τα δεδομένα είτε σαν θετικά είτε σαν αρνητικά. Όπως φαίνεται και στην εικόνα 3.2 κατά την διάρκεια που ο ταξινομητής κάνει την ταξινόμηση ο κύκλος μετατοπίζεται αριστερόστροφα με αποτέλεσμα τα στοιχεία των θετικών προβλέψεων να μπαίνουν στη ζώνη των αρνητικών και το αντίστροφο. Έτσι η ταξινόμηση παράγει τέσσερις αντί για δύο κατηγορίες δεδομένων όπως φαίνεται και στην εικόνα 3.3 μιας και στην περιοχή των θετικών προβλέψεων εμπεριέχονται πλέον ψευδείς θετικές προβλέψεις και στην περιοχή των αρνητικών προβλέψεων εμπεριέχονται ψευδείς αρνητικές.



Εικόνα 3.2 Διαδικασία προβλέψεις.



Εικόνα 3.3 Παραγωγή Ψευδών Θετικών & Ψευδών Αρνητικών προβλέψεων.

3.2.1 True Positive Rate (TP Rate)

Ένα αποτέλεσμα δοκιμής που υποδεικνύει σωστά την παρουσία μιας κατάστασης ή ενός χαρακτηριστικού ανήκει στην κατηγορία του True Positive Rate. Επιπλέον ο πραγματικός θετικός ρυθμός TPR ονομάζεται και ευαισθησία και είναι ο λόγος «True Positive / (True Positive + False Negative)». Με λίγα λόγια το True Positive Rate είναι η πιθανότητα ένα αποτέλεσμα να είναι θετικό και να χαρακτηριστεί σαν τέτοιο [10].

3.2.2 False Positive Rate (FP Rate)

Το ποσοστό False Positive Rate (FPR) όπου και αυτό είναι ένα μέτρο ακρίβειας για ένα μοντέλο μηχανικής μάθησης υποδεικνύει λανθασμένα την παρουσία μιας κατάστασης. Σε τεχνικούς όρους, το ψευδώς θετικό ποσοστό ορίζεται ως η πιθανότητα ψευδούς απόρριψης της μηδενικής υπόθεσης. Στην επιστήμη των δεδομένων, το ποσοστό ψευδώς θετικών μετρά το ποσοστό των ψευδώς θετικών έναντι όλων των θετικών προβλέψεων (το άθροισμα των ψευδώς θετικών και των αληθινών θετικών) σε ένα πρόβλημα δυαδικής ταξινόμησης. Το ποσοστό ψευδώς θετικών βασίζεται στο πόσα πραγματικά αρνητικά το μοντέλο προέβλεψε λανθασμένα. Αυτή η μέτρηση είναι συμπληρωματική του αληθινού θετικού ποσοστού ή της ανάκλησης, που δείχνει πόσα πραγματικά θετικά προέβλεψε σωστά το μοντέλο [10].

3.2.3 Precision

Η ακρίβεια (Precision) είναι μια μέτρηση που ποσοτικοποιεί τον αριθμό των σωστών θετικών προβλέψεων που έγιναν. Υπολογίζεται ως η αναλογία των σωστά προβλεπόμενων θετικών παραδειγμάτων διαιρεμένη με τον συνολικό αριθμό των θετικών παραδειγμάτων που είχαν προβλεφθεί « Precision = True Positive / (True Positive + False Positive) ». Η ακρίβεια αξιολογεί το κλάσμα των σωστών ταξινομημένων περιπτώσεων μεταξύ αυτών που ταξινομούνται ως θετικές.

3.2.4 Recall

Η ανάκληση (recall) είναι αντίστοιχο του True Positive Rate (TP Rate), είναι μια μέτρηση που ποσοτικοποιεί τον αριθμό των σωστών θετικών προβλέψεων που έγιναν από όλες τις θετικές προβλέψεις που θα μπορούσαν να είχαν γίνει. Σε αντίθεση με την ακρίβεια (precision) που σχολιάζει μόνο τις σωστές θετικές προβλέψεις από όλες τις θετικές προβλέψεις, η ανάκληση παρέχει μια ένδειξη χαμένων θετικών προβλέψεων. Με αυτόν τον τρόπο, η ανάκληση παρέχει κάποια έννοια της κάλυψης της θετικής τάξης.

3.2.5 F – Measure

Το μέτρο F (F – Measure) υπολογίζεται ως ο αρμονικός μέσος όρος συνοψίζοντας την ακρίβεια και ανάκληση, δίνοντας στο καθένα την ίδια στάθμιση. Επιτρέπει την αξιολόγηση ενός μοντέλου λαμβάνοντας υπόψη τόσο την ακρίβεια όσο και την ανάκληση, χρησιμοποιώντας μια ενιαία βαθμολογία, η οποία είναι χρήσιμη κατά την περιγραφή της απόδοσης του μοντέλου και τη σύγκριση μοντέλων. Αποτελεί μια σημαντική μέτρηση όταν η ακρίβεια (Precision) και η ανάκληση (Recall) είναι σημαντικοί παράμετροι. Αναφορικά με το F-measure υπάρχει και μια παραλλαγή του την οποία δεν βλέπουμε στην αναφορά του weka με την ονομασία Fbeta-measure. Το Fbeta-measure είναι μια γενίκευση που προσθέτει μία επιπλέον παράμετρο όπου από προεπιλογή της έχει δοθεί η τιμή ένα. Η τιμή ίση με ένα παράγει το ίδιο αποτέλεσμα με το f-Measure μιας και δίνεται η ίδια βαρύτητα τόσο στην ακρίβεια όσο και στην ανάκληση. Μία τιμή της τάξεως του 0,5 δίνει μεγαλύτερη βαρύτητα στην ακρίβεια ενώ μια τιμή της τάξεως του δύο μεγαλύτερη στην ανάκληση. Έτσι το Fbeta-measure φαίνεται χρήσιμο όταν χρειάζεται να δοθεί κάπου έμφαση.

3.2.6 Receiver Operator Characteristic (ROC Area)

Ένα γράφημα χαρακτηριστικών λειτουργίας δέκτη (ROC) είναι μια τεχνική για την οπτικοποίηση, την οργάνωση και την επιλογή ταξινομητών με βάση την απόδοσή τους. Τα γραφήματα ROC χρησιμοποιούνται εδώ και πολύ καιρό στη θεωρία ανίχνευσης σήματος για να απεικονίσουν την ανταλλαγή μεταξύ των ρυθμών αληθινών και ψευδών προβλέψεων των ταξινομητών [11]. Τα τελευταία χρόνια παρατηρείται αύξηση στη χρήση γραφημάτων ROC στην κοινότητα μηχανικής μάθησης, εν μέρει λόγω της συνειδητοποίησης ότι η απλή ακρίβεια ταξινόμησης είναι συχνά μια κακή μέτρηση για τη μέτρηση της απόδοσης [12].

Το γράφημα ROC αποτελεί σημαντικό κομμάτι στην αξιολόγηση της αποδοτικότητας των αλγορίθμων ταξινόμησης που έχουν εφαρμοστεί στην εργασία όπου θα παρατεθούν και οι απεικονίσεις τους.

3.3 Αποτελέσματα και παραμετροποίηση ταξινομητών

3.3.1 Ταξινομητής “Logistic Model Tree (LMT)”

Ο ταξινομητής treeLMT (Logistic Model Trees) δημιουργεί δέντρα ταξινόμησης με συναρτήσεις λογιστικής παλινδρόμησης. Η φιλοσοφία πίσω από τον treeLMT είναι με την χρήση του αλγορίθμου LogitBoost να εκτελεί προσθετική λογιστική παλινδρόμηση σε κάθε επανάληψη του, βρίσκοντας έτσι την απλή συνάρτηση με το μικρότερο σφάλμα και προσθέτοντας την στο τελικό μοντέλο. Ο ταξινομητής εκτελέστηκε διατηρώντας της προκαθορισμένες τιμές στις παραμέτρους του.

TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	Correctly Classified Instances	Incorrectly Classified Instances
0,901	0,621	0,883	0,901	0,892	0,292	0,688	0,884	survival	81,6727	18,3273
0,379	0,099	0,422	0,379	0,4	0,292	0,688	0,336	death		
0,817	0,537	0,809	0,817	0,813	0,292	0,688	0,796	Avg.		

Πίνακας 3.1 Περίληπτικά αποτελέσματα ταξινομητή *treeLMT*.

3.3.2 Ταξινομητής “Bayesian Network”

Το δίκτυο Bayesian (Bayesian Network) είναι ένας συνδυασμός πιθανολογικού μοντέλου και μοντέλου γραφήματος. Έχει πολλές εφαρμογές αλλά το συναντάμε συστηματικά τόσο στη μηχανική μάθηση, την εξόρυξη δεδομένων όσο και την διάγνωση επειδή έχει ένα συμπέρασμα βασισμένο σε στοιχεία που είναι οικείο στην ανθρώπινη διαίσθηση [17]. Σε ότι αφορά την παραμετροποίηση του αλγορίθμου εφαρμόσαμε ανάμεσα σε μια πλειάδα από επιλογές μεθόδων αναζήτησης την K2. Ο αλγόριθμος K2 αποτελεί μία αξιόπιστη κατάταξη κόμβου για την καλύτερη ανακάλυψη της δομή του δικτύου. Επιπλέον είναι ο πιο διάσημος αλγόριθμος που βασίζεται σε σκορ στο Bayesian δίκτυο τις τελευταίες δύο δεκαετίες [18].

TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	Correctly Classified Instances	Incorrectly Classified Instances
0,814	0,467	0,901	0,814	0,855	0,297	0,756	0,937	survival	76,8358	23,1642
0,533	0,186	0,354	0,533	0,425	0,297	0,756	0,383	death		
0,768	0,422	0,813	0,768	0,786	0,297	0,756	0,848	Avg.		

Πίνακας 3.2 Περίληπτικά αποτελέσματα ταξινομητή *Bayesian Network*.

3.3.3 Ταξινομητής “Naïve Bayes”

Ο ταξινομητής Naïve Bayes χτίζει μια σταθερή δομή για προβλήματα ταξινόμησης είτε δύο, είτε πολλαπλών κλάσεων. Οι υπολογισμοί που υλοποιούνται βασίζονται στο θεώρημα του Bayes εφαρμόζοντας μία υπόθεση ανεξαρτησίας μεταξύ των προγνωστικών. Σχετικά με την παραμετροποίηση του Naïve Bayes παραμείναμε στις προκαθορισμένες τιμές τόσο στην επιλογή του εκτιμητή πυρήνα όσο και στην επιλογή της εποπτευόμενης διανομής. Στην περίπτωση του εκτιμητή πυρήνα ο αλγόριθμος αποφασίζει τότε θα τον χρησιμοποιήσει ή όχι

και στην περίπτωση της εποπτευόμενης διανομής αποφασίζει την μετατροπή των αριθμητικών χαρακτηριστικών σε ονομαστικών. Όπως αναφέρθηκε και προηγουμένως παραμείναμε στις αρχικές επιλογές μιας και το παραγόμενο αποτέλεσμα παρέμενε το ίδιο σε οποιαδήποτε παραμετροποίηση.

TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	Correctly Classified Instances	Incorrectly Classified Instances
0,617	0,322	0,909	0,617	0,735	0,218	0,696	0,901	survival	62,6372	37,3628
0,678	0,383	0,253	0,678	0,369	0,218	0,696	0,314	death		
0,626	0,332	0,803	0,626	0,676	0,218	0,696	0,807	Avg.		

Πίνακας 3.3 Περιληπτικά αποτελέσματα ταξινομητή Naïve Bayes.

3.3.4 Ταξινομητής “Logistic Regression”

Ο ταξινομητής Logistic Regression (λογιστικής παλινδρόμησης) χρησιμοποιείται κατά την επεξεργασία κατηγορικών, δυαδικών δεδομένων. Θεωρείτε ένας από τους απλούστερους αλγορίθμους μηχανικής μάθησης και ανήκει στην κατηγορία των εποπτευόμενων ταξινομητών. Κατά την εκτέλεση του αλγορίθμου χρησιμοποιήθηκαν οι παράμετροι αρχικοποίησης του.

TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	Correctly Classified Instances	Incorrectly Classified Instances
0,929	0,692	0,875	0,929	0,901	0,279	0,732	0,906	survival	82,8958	17,1015
0,308	0,071	0,454	0,308	0,367	0,279	0,732	0,351	death		
0,829	0,592	0,807	0,829	0,815	0,279	0,732	0,817	Avg.		

Πίνακας 3.4 Περιληπτικά αποτελέσματα ταξινομητή Logistic Regression.

3.3.5 Ταξινομητής “Simple Logistic”

Η απλή λογιστική παλινδρόμηση είναι μια στατιστική δοκιμή που χρησιμοποιείται για την πρόβλεψη μιας μεμονωμένης δυαδικής μεταβλητής χρησιμοποιώντας μια άλλη. Χρησιμοποιείται επίσης για τον προσδιορισμό της αριθμητικής σχέσης μεταξύ δύο τέτοιων μεταβλητών. Κατά την εκτέλεση ο αλγόριθμος παραμετροποιήθηκε ως προς το να αποφασίζει να βρει τον καλύτερο ρυθμό επαναλήψεων χρησιμοποιώντας μια εσωτερική διαδικασία διασταυρούμενης επικύρωσης η οποία επαναλαμβάνεται πέντε φορές.

TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	Correctly Classified Instances	Incorrectly Classified Instances
0,942	0,721	0,872	0,942	0,906	0,28	0,752	0,922	survival	83,5774	16,4226
0,279	0,058	0,482	0,279	0,354	0,28	0,752	0,372	death		
0,836	0,614	0,809	0,836	0,817	0,28	0,752	0,834	Avg.		

Πίνακας 3.5 Περιληπτικά αποτελέσματα ταξινομητή Simple Logistic.

3.3.6 Ταξινομητής “Decision Table”

Οι πίνακες αποφάσεων “Decision Tables”, όπως τα δέντρα αποφάσεων ή τα νευρωνικά δίκτυα, είναι μοντέλα ταξινόμησης που χρησιμοποιούνται για την πρόβλεψη. Προκαλούνται από αλγόριθμους μηχανικής μάθησης. Ένας πίνακας αποφάσεων αποτελείται από έναν ιεραρχικό πίνακα στον οποίο κάθε καταχώρηση σε έναν πίνακα υψηλότερου επιπέδου αναλύεται με τις τιμές ενός ζεύγους πρόσθετων χαρακτηριστικών για να σχηματιστεί ένας άλλος πίνακας. Η δομή είναι παρόμοια με τη διαστασιακή στοίβαξη [14]. Η παραμετροποίηση του ταξινομητή παρέμεινε στις αρχικές επιλογές διατηρώντας ακόμα και την επιλογή του αλγορίθμου αναζήτησης για την εύρεση των καλύτερων συνδυαστικών αποφάσεων στην επιλογή Best First.

TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	Correctly Classified Instances	Incorrectly Classified Instances
0,967	0,71	0,877	0,967	0,92	0,36	0,745	0,922	survival	85,816	14,184
0,29	0,033	0,628	0,29	0,397	0,36	0,745	0,457	death		
0,858	0,601	0,837	0,858	0,836	0,36	0,745	0,847	Avg.		

Πίνακας 3.6 Περιληπτικά αποτελέσματα ταξινομητή Decision Table.

3.3.7 Ταξινομητής “Sequential Minimal Optimization - SMO”

Ο ταξινομητής Sequential Minimal Optimization – SMO χρησιμοποιείται για την επίλυση προβλημάτων σε μεγάλο σετ δεδομένων. Η εκπαίδευση μιας μηχανής διανυσμάτων υποστήριξης απαιτεί τη λύση ενός πολύ μεγάλου προβλήματος βελτιστοποίησης τετραγωνικού προγραμματισμού (QP). Το SMO διασπά αυτό το πρόβλημα μεγάλου QP σε μια σειρά από μικρότερα δυνατά προβλήματα QP. Αυτά τα μικρά προβλήματα QP επιλύονται αναλυτικά, γεγονός που αποφεύγει τη χρήση μιας χρονοβόρας αριθμητικής βελτιστοποίησης QP ως

εσωτερικό βρόχο. Η ποσότητα της μνήμης που απαιτείται για το SMO είναι γραμμική στο μέγεθος του σετ προπόνησης, κάτι που επιτρέπει στο SMO να χειρίζεται εύκολα πολύ μεγάλα σετ [16].

TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	Correctly Classified Instances	Incorrectly Classified Instances
0,962	0,798	0,863	0,962	0,91	0,245	0,582	0,862	survival	83,959	16,041
0,202	0,038	0,504	0,202	0,288	0,245	0,582	0,23	death		
0,84	0,676	0,805	0,84	0,81	0,245	0,582	0,76	Avg.		

Πίνακας 3.7 Περιληπτικά αποτελέσματα ταξινομητή *Sequential Minimal Optimization – SMO*.

3.3.8 Ταξινομητής “Random Tree”

Ο Random Tree αλγόριθμος δημιουργεί ένα δέντρο που θεωρεί τυχαία N δείγματα ως υποψήφια χαρακτηριστικά μέσα σε κάθε κόμβο. Όπως προσδίδει και το όνομα του μιλάμε ακριβώς για ένα τυχαίο δέντρο. Η έκδοση του Random Tree που χρησιμοποιήθηκε δεν εφαρμόζει την τεχνική του κλαδέματος. Σε ότι αφορά την παραμετροποίηση του διατηρήσαμε την προκαθορισμένη τιμή που είναι ίση με το μηδέν τόσο για το μέγιστο βάθος όσο για την προσαρμογή στην ανάπτυξη του δέντρου [13]. Στον πίνακα που ακολουθεί εμπεριέχονται τα αποτελέσματα μετά την επιτυχή ολοκλήρωση της διαδικασίας ταξινόμησης του αλγορίθμου.

TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	Correctly Classified Instances	Incorrectly Classified Instances
1	0,946	0,846	1	0,917	0,212	0,842	0,961	survival	84,7701	15,229
0,054	0	0,992	0,054	0,102	0,212	0,842	0,605	death		
0,848	0,794	0,87	0,848	0,786	0,212	0,842	0,904	Avg.		

Πίνακας 3.8 Περιληπτικά αποτελέσματα ταξινομητή *Random Tree*.

3.3.9 Ταξινομητής “Attribute Selected”

Ο ταξινομητής Attribute Selected στηρίζει την λειτουργία του στην επιλογή ενός μέρους από το σύνολο των χαρακτηριστικών που απαρτίζουν τα δεδομένα. Έτσι κατ’ αυτόν τον τρόπο μειώνει κατά πολύ τον χρόνο επεξεργασίας και αυξάνει την απόδοση του ταξινομητή μιας και κατά την

διαδικασία εξόρυξης δεδομένων εμπλέκονται μόνο αυτά τον επιλεγόμενων χαρακτηριστικών [15].

TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	Correctly Classified Instances	Incorrectly Classified Instances
0,991	0,938	0,846	0,991	0,913	0,151	0,734	0,916	survival	84,1866	15,8134
0,062	0,009	0,58	0,062	0,112	0,151	0,734	0,357	death		
0,842	0,789	0,804	0,842	0,784	0,151	0,734	0,826	Avg.		

Πίνακας 3.9 Περιληπτικά αποτελέσματα ταξινομητή *Attribute Selected*.

3.3.10 Ταξινομητής “K-Nearest Neighbors”

Ο ταξινομητής K-Nearest Neighbors γνωστός και ως KNN είναι ένας εποπτευόμενος ταξινομητής ο οποίος κάνει υποθέσεις με βάση τα παρόμοια σημεία που μπορούν να βρεθούν κοντά του. με αυτόν τον τρόπο χτίζει τις υποθέσεις του. Από τις βασικές παραμετροποιήσεις του είναι η αριθμητική επιλογή των γειτόνων και στην περίπτωση κάναμε τις επιλογές (1, 3, 7, 11, 21). Πριν την εκτέλεση του ορίσαμε την παράμετρο *crossValidate* σε *True*. Η παράμετρος αυτή αξιολογεί την αριθμητική μας επιλογή αναφορικά με τους γειτόνους που έχουμε ορίσει και κρίνει στο πεδίο τιμών που έχει δημιουργηθεί ανάμεσα στην τιμή ένα και την τιμή που έχουμε ορίσει ποια θεωρείτε ιδανικότερη. Οι υπόλοιπες επιλογές παρέμειναν στις προεπιλογές. Μετά την εκτέλεση του ταξινομητή στις προαναφερθείσες τιμές (1, 3, 7, 11, 21) λάβαμε αποτελέσματα για κοντινούς γειτόνους στην τιμή ένα και δύο.

TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	Correctly Classified Instances	Incorrectly Classified Instances
0,965	0,921	0,845	0,965	0,901	0,08	0,522	0,845	survival	82,2232	17,7768
0,079	0,035	0,3	0,079	0,125	0,08	0,522	0,172	death		
0,822	0,779	0,758	0,822	0,776	0,08	0,522	0,737	Avg.		

Πίνακας 3.10 Περιληπτικά αποτελέσματα ταξινομητή *KNN-1*.

TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	Correctly Classified Instances	Incorrectly Classified Instances
0,99	0,97	0,842	0,99	0,91	0,065	0,565	0,857	survival	83,5664	16,4336
0,03	0,01	0,369	0,03	0,056	0,065	0,565	0,192	death		
0,836	0,815	0,766	0,836	0,773	0,065	0,565	0,75	Avg.		

Πίνακας 3.11 Περιληπτικά αποτελέσματα ταξινόμητη KNN-2.

3.4 ROC Curves “Receiver Operating Characteristic”

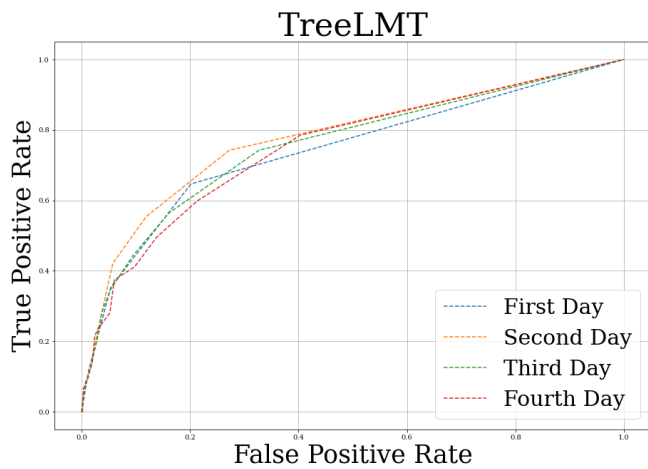
Ο όρος ROC Curve – “Receiver Operating Characteristic” (Καμπύλη Λειτουργίας Δέκτη) είναι μία γραφική παράσταση που απεικονίζει τη διαγνωστική ικανότητα ενός δυαδικού συστήματος ταξινόμητη καθώς το όριο διάκρισης του ποικίλλει. Οι καμπύλες ROC χρησιμοποιούνται και στα κλινικά δεδομένα για την επιλογή του καταλληλότερου διαχωρισμού στις παρατηρήσεις. Η καλύτερη αποκοπή έχει το υψηλότερο πραγματικό θετικό ποσοστό μαζί με το χαμηλότερο ποσοστό ψευδώς θετικών. Η περιοχή που δημιουργείτε κάτω από την καμπύλη ROC η οποία φέρει την ονομασία AUC (Area Under the Curve) δίνει μια ιδέα για το όφελος ή μη από τη χρήση της εν λόγω προσέγγισης που έχουμε εφαρμόσει. Όσο μεγαλύτερη είναι η τιμή της η οποία κυμαίνεται από το μηδέν έως το ένα τόσο πιο θετική χαρακτηρίζεται η προσέγγιση μας. Βέβαια υπάρχει και το ενδεχόμενο της υπερβολικά μεγάλης απόδοσης που μπορεί να παρατηρηθεί με τιμές αποτελέσματος που αγγίζουν την τιμή της μονάδας. Αυτό δημιουργεί υποψίες αναφορικά με την υπερπροσαρμογή του αλγορίθμου ή της εσφαλμένης εφαρμογής του. Θα μπορούσαμε από τον πίνακα 3.12 που ακολουθεί να χαρακτηρίσουμε το αποτέλεσμα της AUC θεωρώντας ένα μοντέλο με τιμές AUC μικρότερες του 0.7 ως φτωχό και για μεγαλύτερες ως αποδεκτό ή καλύτερο.

Τιμές AUC “Area Under the Curve”				
0.5	0.5 – 0.7	0.7 – 0.8	0.8 – 0.9	>0.9
Καμία διάκριση	Κακή διάκριση	Αποδεκτές Διακρίσεις	Άριστη διάκριση	Εξαιρετική διάκριση

Πίνακας 3.12 Τιμές AUC.

Τα διαγράμματα που ακολουθούν έχουν παραχθεί με την χρήση του κώδικα του παραρτήματος Χ και είναι κοινός για όλα. Η διαφοροποίηση βρίσκεται στα δεδομένα όπου τροφοδοτούνται όπου για το κάθε ένα είναι και διαφορετικά. Αναφορά σε αυτό έχει γίνει στο κεφάλαιο 2 στην παράγραφο 2.1 όπου έγινε η παρουσίαση του σχήματος του συστήματος. Τα δεδομένα που τροφοδοτούν τα διαγράμματα είναι ο λόγος (Πρόβλεψη Επιβίωσης / Επιβίωση).

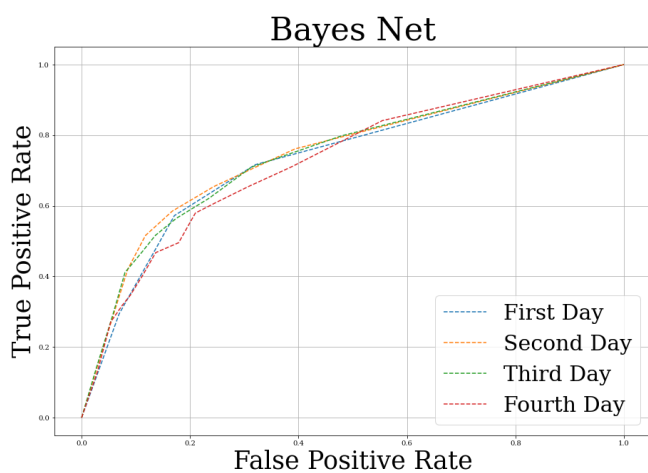
3.4.1 Καμπύλες ROC (Not Survive)



Ημέρα	Area Under the Curve (AUC)
1 ^η	0.7416
2 ^η	0.7759
3 ^η	0.7548
4 ^η	0.7503

Εικόνα 3.4 ROC Curve "treeLMT".

Πίνακας 3.13 AUC treeLMT.

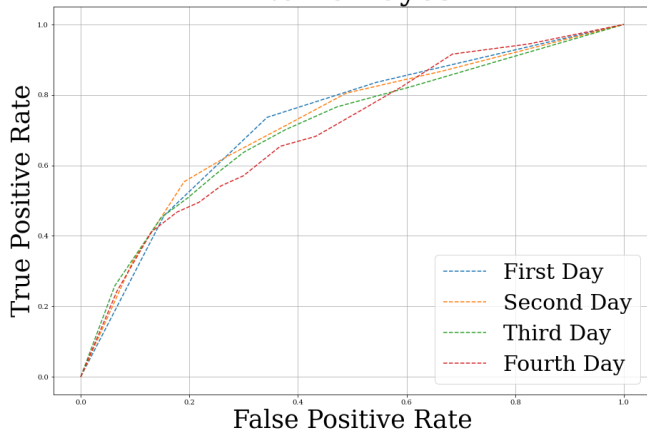


Ημέρα	Area Under the Curve (AUC)
1 ^η	0.7337
2 ^η	0.7479
3 ^η	0.7443
4 ^η	0.7285

Εικόνα 3.5 ROC Curve "BayesNet".

Πίνακας 3.14 AUC BayesNet.

Naive Bayes

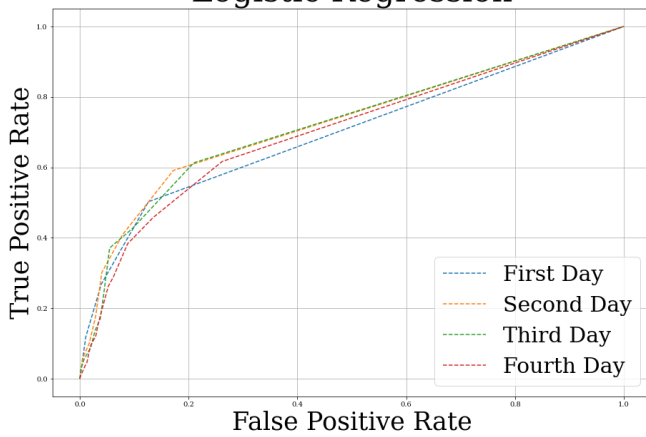


Εικόνα 3.6 ROC Curve “Naïve Bayes”.

Ημέρα	Area Under the Curve (AUC)
1 ^η	0.7246
2 ^η	0.7208
3 ^η	0.7076
4 ^η	0.6994

Πίνακας 3.15 AUC Naïve Bayes.

Logistic Regression

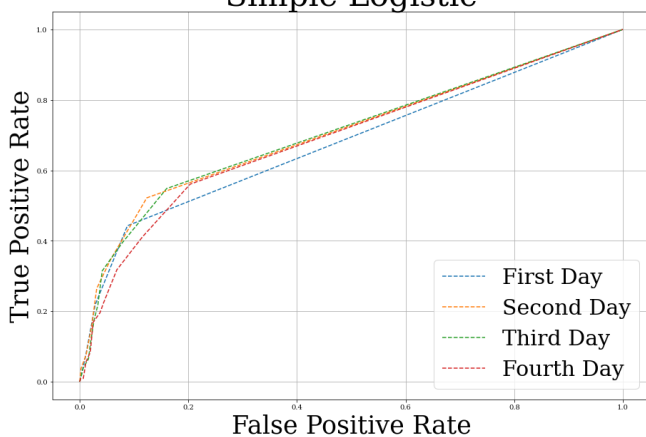


Εικόνα 3.7 ROC Curve “Logistic Regression”.

Ημέρα	Area Under the Curve (AUC)
1 ^η	0.6961
2 ^η	0.7246
3 ^η	0.7210
4 ^η	0.7025

Πίνακας 3.16 AUC Logistic Regression.

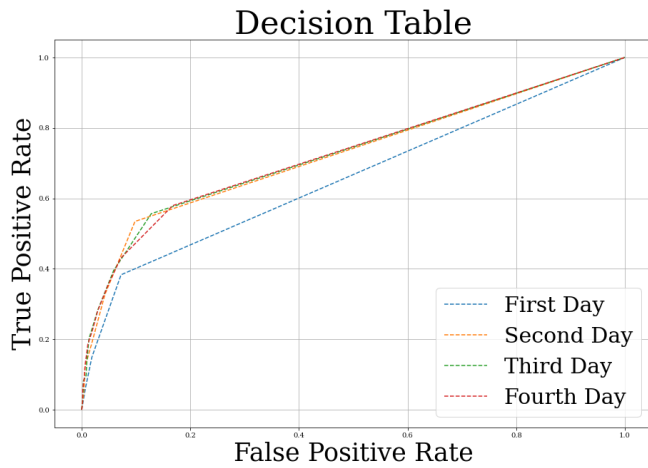
Simple Logistic



Εικόνα 3.8 ROC Curve “Simple Logistic”.

Ημέρα	Area Under the Curve (AUC)
1 ^η	0.6806
2 ^η	0.7075
3 ^η	0.7073
4 ^η	0.6933

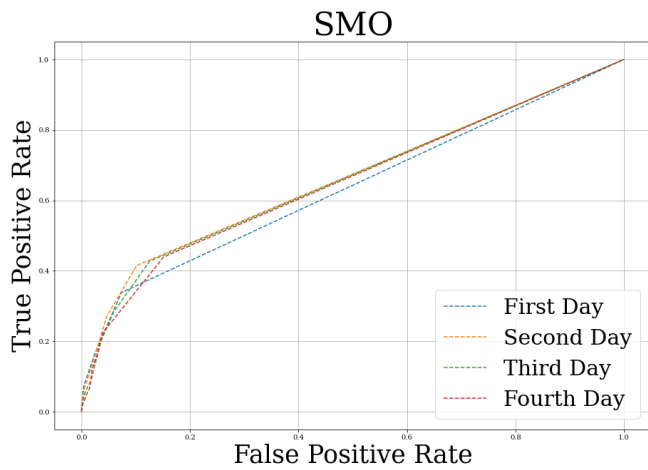
Πίνακας 3.17 AUC Simple Logistic.



Εικόνα 3.9 ROC Curve “Decision Table”.

Ημέρα	Area Under the Curve (AUC)
1 ^η	0.6572
2 ^η	0.7243
3 ^η	0.7268
4 ^η	0.7263

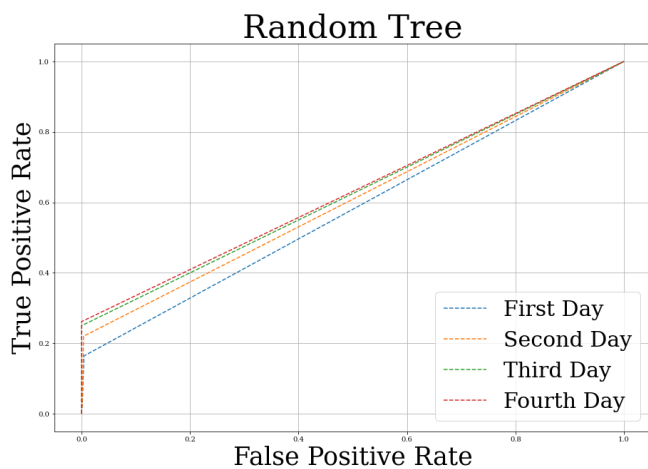
Πίνακας 3.18 AUC Decision Table.



Εικόνα 3.10 ROC Curve “SMO”.

Ημέρα	Area Under the Curve (AUC)
1 ^η	0.6349
2 ^η	0.6617
3 ^η	0.6571
4 ^η	0.6523

Πίνακας 3.19 AUC SMO.



Εικόνα 3.11 ROC Curve “Random Tree”.

Ημέρα	Area Under the Curve (AUC)
1 ^η	0.5798
2 ^η	0.6082
3 ^η	0.6250
4 ^η	0.6308

Πίνακας 3.20 AUC Random Tree.

Έχοντας παραθέσει τα αποτελέσματα των οχτώ με αύξουσα, συγκριτικά με την απόδοση σειρά, από τους έντεκα αλγόριθμους ταξινόμησης θα συνεχίσουμε στο επόμενο κεφάλαιο με την ανάλυση τους και την επιλογή των τριών πιο αποδοτικών για την επίλυση της πρόβλεψης θνησιμότητας σε ασθενείς με διάγνωση ισχαιμικού εγκεφαλικού. Οι τρεις αλγόριθμοι που δεν παραθέσαμε τα διαγράμματα τους είχαν απόδοση στις τιμές της Area Under the Roc (AUR) που κυμαινόντουσαν από 0.49 έως 0.57. Η προγνωστική ικανότητα τους ανήκει στην κατηγορία της τυχειότητας και με βάση των πίνακα 3.12 οι τιμές αυτές ταξινομούνται στο 'Καμία Διάκριση'.

4

Συμπεράσματα

4.1 Ανάλυση αποτελεσμάτων

Έχοντας στην διάθεση μας τα στοιχεία που έχουμε παραθέσει στο κεφάλαιο των πειραματικών αποτελεσμάτων μπορούμε εφαρμόζοντας συνδυαστική αξιολόγηση να βαθμολογήσουμε την απόδοση των αλγορίθμων. Μέσα από την αξιολόγηση τους θα μπορούμε να υποδείξουμε τους πιο αποδοτικούς από αυτούς. Συνδυαστική αξιολόγηση εφαρμόζουμε αναλύοντας τα επί του συνόλου των δεδομένων στοιχεία από τα περιληπτικά αποτελέσματα της παραγράφου 3.3 που έχω παραθέσει στους πίνακες από τον 3.1 έως 3.11 σε σχέση με τα διαγράμματα ROC των πρώτων τεσσάρων ημερών της παραγράφου 3.4 έτσι ώστε να έχουμε ένα συνδυαστικά πιο αξιόπιστο αποτέλεσμα. Τους αλγορίθμους θα τους αναλύσω με διαφορετική σειρά από την παρουσίαση τους στα προηγούμενα κεφάλαια για να υποδείξω και τις σχέσης ανάμεσα στις τιμές που εμπεριέχονται στις αξιολογήσεις τους.

Ξεκινώντας με τον ταξινομητή **Random Tree** και πηγαίνοντας στον πίνακα 3.8 βλέπουμε ότι η απόδοση του ως προς τις σωστά ταξινομημένες περιπτώσεις (στήλη Correctly Classified Instances) είναι της τάξεως του 84,77%. Θα μπορούσε άνετα να χαρακτηριστεί ένας αρκετά καλός αλγόριθμος αν η ανάλυση μας σταματούσε εκεί. Στην στήλη Recall όμως βλέπουμε σε ότι αφορά την τιμή της κλάσης death (στήλη class) η αντίστοιχη τιμή της ισούται με 0,054 με άριστα την τιμή της μονάδας. Έχοντας αναφέρει στην παράγραφο 3.2.4 το τι αποτυπώνει η Recall και κατ' επέκταση την σημαντικότητα της, προκύπτει ότι ο ταξινομητής Random Tree δεν αποδεικνύεται αξιόλογος για την επίλυση της πρόβλεψης θνησιμότητας στους ασθενείς με ισχαιμικό εγκεφαλικό. Αναλύοντας και το διάγραμμα ROC παρατηρούμε ότι η απόδοση των διαγραμμίσεων που αναφέρονται στις μέρες κινούνται περίπου στο μέσω του διαγράμματος. Η περιοχή που δημιουργείτε κάτω από την ROC (Area Under the Curve – AUC) σε επίπεδο τιμών

κυμαίνεται ανάμεσα στο 0.57 έως 0.63 πίνακας 3.19. Με βάση την απόδοση του ταξινομητή και στηριζόμενοι στον πίνακα 3.12 οι τιμές αυτές για την AUC χαρακτηρίζονται ως κακής διάκρισης.

Συνεχίζοντας με τον **Naïve Bayes** και διαβάζοντας την σύνοψη από τον πίνακα 3.3 παρατηρούμε ότι η Recall και στις δύο κλάσεις (survival – death) έχει τιμές 0,617 και 0,678 αντίστοιχα. Αυτό δηλώνει μια κατάσταση όπου και στις δύο κλάσεις των δεδομένων υπάρχει μία ικανοποιητική ισορροπία. Συνδυάζοντας και το διάγραμμα ROC εικόνα 3.6 και τις τιμές του πίνακα 3.15 των πρώτων τεσσάρων ημέρων με τιμές AUC ξεκινούν από το 0.72 τις δύο πρώτες μέρες και κυμαίνονται έως το 0.7 από την τρίτη έως τέταρτη μέρα, συμπεραίνουμε αυτό που η σύνοψη αρχικά μας είχε δηλώσει. Ο αλγόριθμος μπορεί να εφαρμόσει μία ικανοποιητική διάκριση ανάμεσα στις κλάσεις και η εφαρμογή του θα παρήγαγε αξιοσημείωτα αποτελέσματα.

Ακολουθεί ο **treeLMT** τα στοιχεία του οποίου αποτυπώνονται στον πίνακα 3.1 και παρατηρούμε απόδοση ως προς τις σωστά ταξινομημένες περιπτώσεις (στήλη Correctly Classified Instances) της τάξεως του 81,67%. Ελάχιστα πιο μειωμένη σε σχέση με του Random Tree αλλά όχι σε πλαίσια που θα θεωρούνταν μη αποδοτικά. Μεταφέροντας την ανάλυση μας στην Recall παρατηρούμε μια ουσιαστική συνδυαστική διαφοροποίηση. Η τιμή της κλάσης survival διατηρείτε σε υψηλά επίπεδα της τάξεως του 0,901 με άριστα την μονάδα, συνδυαζόμενη ταυτόχρονα με αξιοσημείωτη απόδοση της τιμής στην κλάση death της τάξεως του 0,379. Μπορούμε να παρατηρήσουμε ταυτόχρονα και το Precision όπου αποδίδει την ποσοτικοποίηση των σωστών θετικών προβλέψεων όπου για την κλάση survival ισούται με 0,883 ενώ για την death είναι ίση με 0,422. Για την ώρα έχουμε μια εμφανώς καλύτερη απόδοση σε σχέση με τους ταξινομητές που έχουμε αναλύσει. Περνώντας στην καμπύλη ROC και την περιοχή που βρίσκεται κάτω από αυτή (AUC) παρατηρούμε οι τιμές τις να κυμαίνονται από 0.74 έως 0.77, πίνακας 3.13. Το πεδίο τιμών μέσα στο τετράημερο αναφορικά με τον πίνακα 3.12 έχουν σαν συνέπεια να μπορούμε να χαρακτηρίσουμε τις διακρίσεις αυτές κάτι παραπάνω από αποδεκτές.

Εν συνεχεία ο **Decision Table** έχοντας λίγο καλύτερη απόδοση συγκριτικά με τον treeLMT ως προς τις σωστά ταξινομημένες περιπτώσεις (στήλη Correctly Classified Instances) με τιμή ίση με 85,81%, διαφορά της τάξεως του 4,14% με τον treeLMT δεν τον καθιστά αυτόματα και ιδανικότερο. Περνώντας στις επόμενες καταγεγραμμένες παρατηρήσεις του πίνακα 3.6 όπου αποτυπώνονται οι τιμές της Recall και του Precision έχουν στο σύνολο τους καλύτερη απόδοση του Decision Table σε σχέση πάντα με του treeLMT. Μεταφέροντας την ανάλυση στην καμπύλη ROC και τις τιμές της AUC παρατηρούμε και εδώ καθημερινή αύξηση σε αυτές ξεκινώντας από την πρώτη μέρα με τον χαρακτηρισμό της απόδοσης ως “Κακής Διάκρισης” αλλά μετά το πέρας του πρώτου εικοσιτετράωρου μεταβαίνει σε αποδεκτές τιμές.

Με τον **Sequential Minimal Optimization – SMO** θα συνεχίσουμε μεταφέροντας την παρατήρηση μας στον πίνακα 3.7 και την απόδοση ως προς τις σωστά ταξινομημένες περιπτώσεις (στήλη Correctly Classified Instances) να ακουμπά τιμές της τάξεως του 83,95%. Οι τιμές του Precision είναι ικανές να κατατάξουν σε υψηλά βαθμολογικά επίπεδα τον ταξινομητή καθώς και η τιμή της κλάσης survival στη Recall. Εμφανώς μειωμένη είναι η τιμή της Recall στην κλάση death όπου σε σύγκριση με το διάγραμμα ROC εικόνα 3.10 και της περιοχή κάτω από την ROC να βαδίζει σε τιμές από 0.63 έως 0.66 δεν μπορούν να το κατατάξουν στους αξιόλογους προς την επίλυση αυτής της πρόβλεψης ταξινομητές.

Ο **Logistic Regression** που ακολουθεί αποδίδει αρκετά ικανοποιητική απόδοση ως προς τις σωστά ταξινομημένες περιπτώσεις (στήλη Correctly Classified Instances) με τιμή ίση με 82,89% και τιμές σε Precision και Recall αρκετά αποδοτικές. Σε όλες μέχρι τώρα τις παρατηρήσεις εκεί που έγκειται η διαφοροποίηση στον χαρακτηρισμό ενός ταξινομητή είναι στην απόδοση της κλάσης death όπου αν συνδυαστεί με μια υψηλή απόδοση στην κλάση survival με τιμές άνω του 0,9 τότε ο ταξινομητής φαίνεται να αποδίδει τα μέγιστα των επιδόσεων του. Στον συγκεκριμένο αλγόριθμο παρατηρείτε απόδοση στην κλάση death τόσο της Recall όσο και του Precision της τάξεως του 0,308 και 0,454 αντίστοιχα. Αναλύοντας και το διάγραμμα ROC της εικόνας 3.7 και της περιοχής που δημιουργείτε κάτω από τις καμπύλες μετά το πέρας της πρώτης μέρα η απόδοση χαρακτηρίζεται “Αποδεκτή” μιας και διατηρεί τιμές που κυμαίνονται από το 0,69 έως 0,72 έτσι όπως αποτυπώνονται στον πίνακα 3.16.

Μετέπειτα έχουμε εφαρμόσει τον **Simple Logistic** του οποίου μέσα από την καμπύλη ROC παρατηρούμε μια γενικά μειωμένη απόδοση σε σχέση με άλλους ταξινομητές που έχουμε παρουσιάσει. Τα στοιχεία που φιλοξενούνται στον πίνακα 3.17 φαίνεται να ακροβατούν ανάμεσα στον χαρακτηρισμό των αποδοτικών και μη διακρίσεων. Κάτι επιπλέον που έχει παρατηρηθεί ότι τα διαγράμματα ROC καθώς και οι περιοχές κάτω από τις καμπύλες παράγουν καλύτερα αποτελέσματα όταν η Recall στην κλάση death φιλοξενεί τιμές μεγαλύτερες από το 0,3. Έτσι και στην περίπτωση του Simple Logistic όπου η Recall βρίσκεται στην τιμή 0.279 ο ταξινομητής μετά την δεύτερη ημέρα περνά στο στάδιο των αποδεκτών διακρίσεων.

Τελειώνοντας ο αλγόριθμος **Bayesian Network** όπου εξ αρχής φαίνεται να έχει την δεύτερη χαμηλότερη απόδοση όπως αποτυπώνεται στον πίνακα 3.2 με τιμή ίση με 76,83% στις σωστά ταξινομημένες περιπτώσεις (στήλη Correctly Classified Instances). Οπότε έχουμε το πρώτο αρνητικό δείγμα για αξιολόγηση του. Συνεχίζοντας με την Recall βλέπουμε εμφανώς ανεβασμένη την τιμή της στην κλάση death με τιμή ίση με 0,533 και μια αντίστοιχη αποδοτικά στην κλάση του survival που ισοδυναμεί με 0,814. Όχι βέβαια με αυτές που έχουμε παρατηρήσει σχεδόν σε όλους τους ταξινομητές που κυμαίνονται πάνω από το 0.9 αλλά αρκετά αξιόλογης απόδοσης. Όπως αναφέραμε και σε προηγούμενο ταξινομητή η Recall όταν αποτυπώνεται με υψηλές τιμές άνω του 0,9 για την survival και άνω του 0,3 για την death τότε παράγονται επιθυμητά αποτελέσματα. Σε αυτή την περίπτωση η Recall του Survival είναι σαφώς πιο μειωμένη αλλά είναι ένα ικανοποιητικό δείγμα. Ακολουθώντας το διάγραμμα ROC στην εικόνα 3.5 και τις τιμές της AUC στον πίνακα 3.14 όπου κυμαίνονται από 0.73 έως σχεδόν 0.75 βλέπουμε ότι υπάρχει μία εμφανή αποδοτικότητα όπου χαρακτηρίζεται με βάση τον πίνακα 3.12 αποδεκτή.

4.2 Επιλογή αποδοτικότερων ταξινομητών

Με βάση την ανάλυση που παρατέθηκε στην προηγούμενη παράγραφο μπορούμε να ξεχωρίσουμε οριακά τρεις αλγόριθμους οι οποίοι μπορούν αποδοτικά να εφαρμοστούν ως μοντέλα πρόβλεψης στους ασθενείς που έχουν διαγνωστεί με ισχαιμικό εγκεφαλικό. Αυτοί και με σειρά απόδοσης είναι ο treeLMT όπου ανεβάζει την απόδοση του από την πρώτη στην δεύτερη ημέρα και παγιώνεται από την τρίτη ημέρα και μετά. Ακολουθεί ο Bayesian Network όπου ακολουθεί τα βήματα του treeLMT με πιο σταθερή απόδοση ανάμεσα στην δεύτερη και

τρίτη ημέρα και τέλος ο Naïve Bayes όπου παρατηρείται σταθερή απόδοση την πρώτη και δεύτερη μέρα και μικρής κλίμακας πτώση από την τρίτη και μετά. Κοινό στοιχείο σε σχέση με τον πίνακα 3.12 και των τριών προτεινόμενων αλγορίθμων είναι ότι οι τιμές της Area Under the Rock (AUC) κατηγοριοποιούνται στις αποδεκτές διακρίσεις.

Αλγόριθμοι	AUC(%)	1 ^η Ημέρα	2 ^η Ημέρα	3 ^η Ημέρα	4 ^η Ημέρα
Logistic Model Tree (tree LMT)		74,16%	77,59%	75,48%	75,03%
Bayesian Network		73,37%	74,79%	74,43%	72,85%
Naïve Bayes		72,46%	72,08%	70,76%	69,94%

Πίνακας 3.21 AUC προτεινόμενων αλγορίθμων.

Βιβλιογραφία

- [1] Dimopoulos Dimitrios, Vasiliki Danilatu, Theodoros Kostoulas (2022), Workshop of SETN 2022: “Mortality Prediction in ICU Patients Suffering from Stroke”, abstract.
- [2] Official Website of The Office of the National Coordinator for Health Information Technology (ONC), <https://www.healthit.gov/faq/what-electronic-health-record-ehr> .
- [3] Swapna Abhyankar, Dina Demner-Fushman, Clement J. McDonald, Standardizing clinical laboratory data for secondary use, page 1.
- [4] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, Roger Mark (16 March 2021) “Database of patients admitted to the Beth Israel Deaconess Medical Center”.
- [5] Eibe Frank, Mark A. Hall and Ian H. Witten (2016), The WEKA Workbench. “Data Mining: Practical Machine Learning Tools and Techniques”, 4th Edition, page 7.
- [6] Jason Brownlee (2016), Machine Learning Mastery With Weka. "Analyze Data, Develop Models and Work Through Projects", Edition: v1.1, page 12.
- [7] Tarca AL, Carey VJ, Chen XW, Romero R, Draghici S. (2007), "Machine learning and its applications to biology", e116.
- [8] Takaya Saito and Marc Rehmsmeier (2015), Machine Learning Mastery With Weka. "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets", page 1.
- [9] Hanley JA, McNeil BJ (1982), “The meaning and use of the area under a receiver operating characteristic (ROC) curve”.
- [10] Ron Kohavi, Foster Provost (1998), “Special Issue on Applications of Machine Learning and the Knowledge Discovery Process”.
- [11] Egan, J.P., (1975.) Signal detection theory and ROC analysis, Series in Cognition and Perception. Academic Press, New York.
- [12] Provost, F., Fawcett, T., (1998). Robust classification systems for imprecise environments. In: Proc. AAAI-98. AAAI Press, Menlo Park, CA.
- [13] Alex G. C. de Sa, Cristiano G. Pimenta, Gisele L. Pappa, Alex A. Freitas (31 Jul 2020). A Robust Experimental Evaluation of Automated Multi-Label Classification Methods.
- [14] B.G. Becker.(Silicon Graphics, Inc., Mountain View, CA, USA), (19 October 1998), Visualizing decision table classifiers.

- [15] Dr Gnanambal S, Dr Thangaraj M, Dr Meenatchi V.T, Dr Gayathri V, (4 April 2018), Classification Algorithms with Attribute Selection: an evaluation study using WEKA.
 - [16] John C. Platt, (April 21, 1998) Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines.
 - [17] Loc Nguyen (Lambert Academic Publishing), (February 2022), Overview of Bayesian Network.
 - [18] G. Cooper and E. Herskovitz, A Bayesian method for the induction of probabilistic networks from data, Machine Learning 9 (1992), 330–347.
- Wes McKinney and the Pandas Development Team (Feb 12, 2022), Pandas: powerful Python data analysis toolkit, Release 1.4.1
- Joel Grus (2019), Data Science from Scratch Second Edition, Απόδοση – Επιστημονική Επιμέλεια Αριστείδης Θ. Παγουρτζής Αθήνα 2021
- Jake Lever, Martin Krzywinski & Naomi Altman (July 28, 2016), Classification evaluation, Nat. Methods 13, 603-604

Παράρτημα Ι [Πιστοποίηση CITI]



Completion Date 17-Oct-2021
Expiration Date 16-Oct-2024
Record ID 45635380

This is to certify that:

Dimitrios Dimopoulos

Has completed the following CITI Program course:

Not valid for renewal of certification through CME.

Human Research
(Curriculum Group)

Data or Specimens Only Research
(Course Learner Group)

2 - Refresher Course
(Stage)

Under requirements set by:

Massachusetts Institute of Technology Affiliates

CITI
Collaborative Institutional Training Initiative

Verify at www.citiprogram.org/verify/?wbabd6ade-f96b-47f4-8c38-2f8cec253067-45635380

Παράρτημα II [Αναλυτική βαθμολογία CITI]

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 1 OF 2 COURSEWORK REQUIREMENTS*

* NOTE: Scores on this Requirements Report reflect quiz completions at the time all requirements for the course were met. See list below for details. See separate Transcript Report for more recent quiz scores, including those on optional (supplemental) course elements.

- **Name:** Dimitrios Dimopoulos (ID: 10611828)
- **Institution Affiliation:** Massachusetts Institute of Technology Affiliates (ID: 1912)
- **Institution Email:** icd520007@icad.aegean.gr
- **Institution Unit:** Information and communication system engineering
- **Curriculum Group:** Human Research
- **Course Learner Group:** Data or Specimens Only Research
- **Stage:** Stage 2 - Refresher Course
- **Record ID:** 45635380
- **Completion Date:** 17-Oct-2021
- **Expiration Date:** 16-Oct-2024
- **Minimum Passing:** 90
- **Reported Score*:** 100

REQUIRED AND ELECTIVE MODULES ONLY	DATE COMPLETED	SCORE
SBE Refresher 1 – Defining Research with Human Subjects (ID: 15029)	13-Oct-2021	2/2 (100%)
SBE Refresher 1 – Privacy and Confidentiality (ID: 15035)	13-Oct-2021	4/4 (100%)
SBE Refresher 1 – Assessing Risk (ID: 15034)	15-Oct-2021	2/2 (100%)
SBE Refresher 1 – Research with Children (ID: 15036)	15-Oct-2021	2/2 (100%)
SBE Refresher 1 – International Research (ID: 15028)	15-Oct-2021	2/2 (100%)
Instructions (ID: 764)	16-Oct-2021	No Quiz
Biomed Refresher 2 – History and Ethical Principles (ID: 511)	16-Oct-2021	3/3 (100%)
Biomed Refresher 2 – Regulations and Process (ID: 512)	16-Oct-2021	2/2 (100%)
Biomed Refresher 2 – SBR Methodologies in Biomedical Research (ID: 515)	16-Oct-2021	4/4 (100%)
Biomed Refresher 2 – Genetics Research (ID: 518)	17-Oct-2021	2/2 (100%)
Biomed Refresher 2 – Records-Based Research (ID: 516)	17-Oct-2021	3/3 (100%)
Biomed Refresher 2 – Populations in Research Requiring Additional Considerations and/or Protections (ID: 519)	17-Oct-2021	1/1 (100%)
Biomed Refresher 2 – HIPAA and Human Subjects Research (ID: 526)	17-Oct-2021	5/5 (100%)
Biomed Refresher 2 – Conflicts of Interest in Research Involving Human Subjects (ID: 17545)	17-Oct-2021	5/5 (100%)
Biomed Refresher 2 - Conclusion (ID: 922)	17-Oct-2021	No Quiz

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?x0501cccf-c846-426e-a8e6-d725f5ee85c5-45635380

Collaborative Institutional Training Initiative (CITI Program)
Email: support@citiprogram.org
Phone: 888-529-5929
Web: <https://www.citiprogram.org>

Παράρτημα III [Χρήση δεδομένων MIMIC-IV (v1.0)]

PhysioNet Credentialed Health Data Use Agreement 1.5.0 Data Use Agreement for the MIMIC-IV (v1.0)

If I am granted access to the database:

1. I will not attempt to identify any individual or institution referenced in PhysioNet restricted data.
2. I will exercise all reasonable and prudent care to avoid disclosure of the identity of any individual or institution referenced in PhysioNet restricted data in any publication or other communication.
3. I will not share access to PhysioNet restricted data with anyone else.
4. I will exercise all reasonable and prudent care to maintain the physical and electronic security of PhysioNet restricted data.
5. If I find information within PhysioNet restricted data that I believe might permit identification of any individual or institution, I will report the location of this information promptly by email to PHI-report@physionet.org, citing the location of the specific information in question.
6. I have requested access to PhysioNet restricted data for the sole purpose of lawful use in scientific research, and I will use my privilege of access, if it is granted, for this purpose and no other.
7. I have completed a training program in human research subject protections and HIPAA regulations, and I am submitting proof of having done so.
8. I will indicate the general purpose for which I intend to use the database in my application.
9. If I openly disseminate my results, I will also contribute the code used to produce those results to a repository that is open to the research community.
10. This agreement may be terminated by either party at any time, but my obligations with respect to PhysioNet data shall continue after termination.

SIGNED: Dimitrios Dimopoulos

DATED: Dec. 10, 2021

Παράρτημα IV [Χρήση Δεδομένων MIMIC-IV-ED (v1.0)]

PhysioNet Credentialed Health Data Use Agreement 1.5.0 Data Use Agreement for the MIMIC-IV-ED (v1.0)

If I am granted access to the database:

1. I will not attempt to identify any individual or institution referenced in PhysioNet restricted data.
2. I will exercise all reasonable and prudent care to avoid disclosure of the identity of any individual or institution referenced in PhysioNet restricted data in any publication or other communication.
3. I will not share access to PhysioNet restricted data with anyone else.
4. I will exercise all reasonable and prudent care to maintain the physical and electronic security of PhysioNet restricted data.
5. If I find information within PhysioNet restricted data that I believe might permit identification of any individual or institution, I will report the location of this information promptly by email to PHI-report@physionet.org, citing the location of the specific information in question.
6. I have requested access to PhysioNet restricted data for the sole purpose of lawful use in scientific research, and I will use my privilege of access, if it is granted, for this purpose and no other.
7. I have completed a training program in human research subject protections and HIPAA regulations, and I am submitting proof of having done so.
8. I will indicate the general purpose for which I intend to use the database in my application.
9. If I openly disseminate my results, I will also contribute the code used to produce those results to a repository that is open to the research community.
10. This agreement may be terminated by either party at any time, but my obligations with respect to PhysioNet data shall continue after termination.

SIGNED: Dimitrios Dimopoulos

DATED: Nov. 19, 2021

Παράρτημα V [Πίνακας Τμήμα Επειγόντων]

1. Εισαγωγή της βιβλιοθήκης Panda.

```
import pandas as pd
```

2. Ανάθεση σε μεταβλητές των στοιχείων εισαγωγής και διαγνώσεων.

```
df_patients=pd.read_csv('D:\Thesis\mimic-iv-1.0\ed\diagnosis.csv.gz')
```

```
df_edstays=pd.read_csv('D:\Thesis\mimic-iv-1.0\ed\edstays.csv.gz')
```

3. Με την χρήση των ICD9 απομονώνω τους ασθενείς με ισχαιμικό εγκεφαλικό.

```
ed_cerebral_infarction_diagnoses = df_patients[  
    (df_patients['icd_code'] == "43300") | (df_patients['icd_code'] == "43301") |  
    (df_patients['icd_code'] == "43310") | (df_patients['icd_code'] == "43311") |  
    (df_patients['icd_code'] == "43320") | (df_patients['icd_code'] == "43321") |  
    (df_patients['icd_code'] == "43330") | (df_patients['icd_code'] == "43331") |  
    (df_patients['icd_code'] == "43380") | (df_patients['icd_code'] == "43381") |  
    (df_patients['icd_code'] == "43390") | (df_patients['icd_code'] == "43391") |  
    (df_patients['icd_code'] == "43400") | (df_patients['icd_code'] == "43401") |  
    (df_patients['icd_code'] == "43410") | (df_patients['icd_code'] == "43411") |  
    (df_patients['icd_code'] == "43491") | (df_patients['icd_code'] == "43490") |  
    (df_patients['icd_code'] == "436") | (df_patients['icd_code'] == "431") |  
    (df_patients['icd_code'] == "430") | (df_patients['icd_code'] == "4350") |  
    (df_patients['icd_code'] == "4351") | (df_patients['icd_code'] == "4352") |  
    (df_patients['icd_code'] == "4353") | (df_patients['icd_code'] == "4358") |  
    (df_patients['icd_code'] == "4359")]
```

4. Ενοποίηση των στοιχείων έτσι ώστε να κρατήσουμε τους ασθενείς που έχουν hadm_id πράγμα που σημαίνει ότι νοσηλεύτηκαν στο νοσοκομείο.

```
rslt_df=pd.merge(df_edstays,ed_cerebral_infarction_diagnoses,on=['subject_id',
```

'stay_id']])

5. Κρατώ μόνο τις γραμμές που έχουν τιμές στο *hadm_id* γιατί αυτό σημαίνει νοσηλεία.
rslt_df = rslt_df[rslt_df['hadm_id'].notna()]

6. Εξάγω τους ασθενείς που νοσηλεύτηκαν με ισχαιμικό σε csv αρχείο.
rslt_df.to_csv('D:\Thesis\exports\o1_ed_patients_to_hosp.csv', index=False)

7. Εισάγω τα αρχεία τόσο των ασθενών όσο και το ζωτικών οργάνων.
df_vitalsign=pd.read_csv('D:\Thesis\mimic-iv-1.0\ed\Vitalsign.csv.gz')
df_ed_patients=pd.read_csv('D:\Thesis\exports\o1_ed_patients_to_hosp.csv',
usecols=['subject_id','hadm_id','stay_id'])

8. Μετατρέπω την στήλη που εμπεριέχει ημερολογιακά στοιχεία από *object* σε *datetime*.
df_vitalsign['charttime']=pd.to_datetime(df_vitalsign['charttime'])

9. Ομαδοποιώ τις καταγραφές των στηλών ανά οχτάωρο.
grouped_multiple = (df_vitalsign.groupby(['subject_id', 'stay_id',
pd.Grouper(key='charttime',
freq='8H'])).agg({'temperature': ['mean'],
'heartrate': ['mean'],
'resprate': ['mean'],
'o2sat': ['mean'],
'sbp': ['mean'],
'dbp': ['mean']})
grouped_multiple.columns = (['Mean_temperature',
'Mean_heartrate',
'Mean_resprate',
'Mean_o2sat',
'Mean_sbp',
'Mean_dbp'])
grouped_multiple=grouped_multiple.reset_index()

10. Ομαδοποιώ τους ασθενείς με τα ζωτικά σημεία
rslt_df=pd.merge(df_ed_patients,grouped_multiple,on=['stay_id'])

11. Εξαγωγή ασθενών σε αρχείο.
rslt_df.to_csv('D:\Thesis\exports\o2_ed_mean_vitals.csv',index=False)

Παράρτημα VI [Πίνακας Νοσοκομείου]

1. Εισαγωγή της βιβλιοθήκης *Panda*.

```
import pandas as pd
```

2. Ανάθεση σε μεταβλητή των διαγνώσεων.

```
df_patients = pd.read_csv('D:\Thesis\mimic-iv-1.0\hosp\diagnoses_icd.csv.gz', usecols =  
['subject_id', 'hadm_id', 'icd_code'])
```

3. Κωδικοί Ασθενών με ισχαιμικό εγκεφαλικό.

```
hosp_patients = df_patients[  
    (df_patients['icd_code'] == "43300") | (df_patients['icd_code'] == "43301") |  
    (df_patients['icd_code'] == "43310") | (df_patients['icd_code'] == "43311") |  
    (df_patients['icd_code'] == "43320") | (df_patients['icd_code'] == "43321") |  
    (df_patients['icd_code'] == "43330") | (df_patients['icd_code'] == "43331") |  
    (df_patients['icd_code'] == "43380") | (df_patients['icd_code'] == "43381") |  
    (df_patients['icd_code'] == "43390") | (df_patients['icd_code'] == "43391") |  
    (df_patients['icd_code'] == "43400") | (df_patients['icd_code'] == "43401") |  
    (df_patients['icd_code'] == "43410") | (df_patients['icd_code'] == "43411") |  
    (df_patients['icd_code'] == "43491") | (df_patients['icd_code'] == "43490") |  
    (df_patients['icd_code'] == "436") | (df_patients['icd_code'] == "431") |  
    (df_patients['icd_code'] == "430") | (df_patients['icd_code'] == "4350") |  
    (df_patients['icd_code'] == "4351") | (df_patients['icd_code'] == "4352") |  
    (df_patients['icd_code'] == "4353") | (df_patients['icd_code'] == "4358") |  
    (df_patients['icd_code'] == "4359")]
```

4. Εξαγωγή ασθενών σε αρχείο.

```
hosp_patients.to_csv('D:\Thesis\exports\o3_hosp_patients.csv', index=False)
```

Παράρτημα VII [Πίνακας Chartevents]

1. Εισαγωγή της βιβλιοθήκης Panda.

```
import pandas as pd
```

2. Ανάθεση μεταβλητών.

```
df_hosp_patients=pd.read_csv('D:\Thesis\exports\o3_hosp_patients.csv')  
df_icd=pd.read_csv('D:\Thesis\mimic-iv-1.0\hosp\icd_diagnoses.csv.gz')  
df_icu_stay = pd.read_csv('D:\Thesis\mimic-iv-1.0\icu\icustays.csv.gz', usecols =  
['subject_id', 'hadm_id'])
```

3. Μοναδικές εισαγωγές στις νοσοκομειακές μονάδες χωρίς περιλαμβάνονται οι ενδονοσοκομειακές μετακινήσεις.

```
df_hosp_patients.drop_duplicates(subset=['hadm_id']).info()
```

4. Συνολικές εισαγωγές στις ΜΕΘ.

```
rslt_df=pd.merge(df_hosp_patients, df_icu_stay, on=['subject_id','hadm_id'])  
rslt_df.to_csv('D:\Thesis\exports\o4_icu_total_admissions.csv', index = False)
```

5. Μοναδικές εισαγωγές στις ΜΕΘ.

```
rslt_df.drop_duplicates(subset=['hadm_id']).to_csv('D:\Thesis\exports\o5_icu_unique_admissions.csv', index = False)
```

6. Ανάθεση μεταβλητών.

```
df_patients=pd.read_csv('D:\Thesis\exports\o5_icu_unique_admissions.csv',usecols=['subject_id','hadm_id'])
```

7. Επεξεργασία πρώτου μέρους του αρχείου chartevent.

```
df_chartevents=pd.read_csv('D:\Thesis\Split_Merge_tables\Splited_Chartevent\chartevents_0.csv')  
rslt_df=pd.merge(df_patients, df_chartevents, on=['subject_id','hadm_id'])  
rslt_df.to_csv('D:\Thesis\Split_Merge_tables\Merged_Chartevent\cerebral_infarction_chartevent_0.csv',index=False)
```

del (df_chartevents, rslt_df) # Διαγραφή των dataframe για απελευθέρωση μνήμης.

8. Επεξεργασία δεύτερου μέρους του αρχείου chartevent.

```
df_chartevents=pd.read_csv('D:\Thesis\Split_Merge_tables\Split_Chartevent\chartevents_1.csv',names=['subject_id','hadm_id','seq_num','icd_code','icd_version','stay_id','charttime','storetime','itemid','value','valuenum','valueuom','warning'])
rslt_df=pd.merge(df_patients, df_chartevents, on=['subject_id','hadm_id'])
rslt_df.to_csv('D:\Thesis\Split_Merge_tables\Merged_Chartevent\cerebral_infarction_chartevent_1.csv',index=False)
del (df_chartevents, rslt_df) # Διαγραφή των dataframe για απελευθέρωση μνήμης.
```

9. Επεξεργασία τρίτου μέρους του αρχείου chartevent.

```
df_chartevents=pd.read_csv('D:\Thesis\Split_Merge_tables\Split_Chartevent\chartevents_2.csv',names=['subject_id','hadm_id','seq_num','icd_code','icd_version','stay_id','charttime','storetime','itemid','value','valuenum','valueuom','warning'])
rslt_df=pd.merge(df_patients, df_chartevents, on=['subject_id','hadm_id'])
rslt_df.to_csv('D:\Thesis\Split_Merge_tables\Merged_Chartevent\cerebral_infarction_chartevent_2.csv',index=False)
del (df_chartevents, rslt_df) # Διαγραφή των dataframe για απελευθέρωση μνήμης.
```

10. Επεξεργασία τέταρτου μέρους του αρχείου chartevent.

```
df_chartevents=pd.read_csv('D:\Thesis\Split_Merge_tables\Split_Chartevent\chartevents_3.csv',names=['subject_id','hadm_id','seq_num','icd_code','icd_version','stay_id','charttime','storetime','itemid','value','valuenum','valueuom','warning'])
rslt_df=pd.merge(df_patients, df_chartevents, on=['subject_id','hadm_id'])
rslt_df.to_csv('D:\Thesis\Split_Merge_tables\Merged_Chartevent\cerebral_infarction_chartevent_3.csv',index=False)
del (df_chartevents, rslt_df) # Διαγραφή των dataframe για απελευθέρωση μνήμης.
```

11. Επεξεργασία πέμπτου μέρους του αρχείου chartevent.

```
df_chartevents=pd.read_csv('D:\Thesis\Split_Merge_tables\Split_Chartevent\chartevents_4.csv',names=['subject_id','hadm_id','seq_num','icd_code','icd_version','stay_id','charttime','storetime','itemid','value','valuenum','valueuom','warning'])
rslt_df=pd.merge(df_patients, df_chartevents, on=['subject_id','hadm_id'])
rslt_df.to_csv('D:\Thesis\Split_Merge_tables\Merged_Chartevent\cerebral_infarction_chartevent_4.csv',index=False)
del (df_chartevents, rslt_df) # Διαγραφή των dataframe για απελευθέρωση μνήμης.
```

Επεξεργασία έκτου μέρους του αρχείου chartevent.

```
df_chartevents=pd.read_csv('D:\Thesis\Split_Merge_tables\Split_Chartevent\chartevents_5.csv',names=['subject_id','hadm_id','seq_num','icd_code','icd_version','stay_id','charttime','storetime','itemid','value','valuenum','valueuom','warning'])
rslt_df=pd.merge(df_patients, df_chartevents, on=['subject_id','hadm_id'])
```

```
rslt_df.to_csv('D:\Thesis\Split_Merge_tables\Merged_Chartevent\cerebral_infarction_ch
artevent_5.csv',index=False)
del (df_chartevents, rslt_df) # Διαγραφή των dataframe για απελευθέρωση μνήμης.
```

12. Επεξεργασία έβδομου μέρους του αρχείου chartevent.

```
df_chartevents=pd.read_csv('D:\Thesis\Split_Merge_tables\Split_Chartevent\charteven
ts_6.csv',names=['subject_id','hadm_id','seq_num','icd_code','icd_version','stay_id','chartt
ime','storetime','itemid','value','valuenum','valueuom','warning'])
rslt_df=pd.merge(df_patients, df_chartevents, on=['subject_id','hadm_id'])
rslt_df.to_csv('D:\Thesis\Split_Merge_tables\Merged_Chartevent\cerebral_infarction_ch
artevent_6.csv',index=False)
del (df_chartevents, rslt_df) # Διαγραφή των dataframe για απελευθέρωση μνήμης.
```

13. Φορτώνουμε τα επτά παραγόμενα από τα προηγούμενα βήματα (7 έως 12) αρχεία

```
df_patient_chartevent_0=pd.read_csv('D:\Thesis\Split_Merge_tables\Merged_Charteven
t\cerebral_infarction_chartevent_0.csv',usecols=['subject_id','hadm_id','stay_id','charttim
e','storetime','itemid','value','valuenum','valueuom','warning'])
```

```
df_patient_chartevent_1=pd.read_csv('D:\Thesis\Split_Merge_tables\Merged_Charteven
t\cerebral_infarction_chartevent_1.csv',usecols=['subject_id','hadm_id','stay_id','charttim
e','storetime','itemid','value','valuenum','valueuom','warning'])
```

```
df_patient_chartevent_2=pd.read_csv('D:\Thesis\Split_Merge_tables\Merged_Charteven
t\cerebral_infarction_chartevent_2.csv',usecols=['subject_id','hadm_id','stay_id','charttim
e','storetime','itemid','value','valuenum','valueuom','warning'])
```

```
df_patient_chartevent_3=pd.read_csv('D:\Thesis\Split_Merge_tables\Merged_Charteven
t\cerebral_infarction_chartevent_3.csv',usecols=['subject_id','hadm_id','stay_id','charttim
e','storetime','itemid','value','valuenum','valueuom','warning'])
```

```
df_patient_chartevent_4=pd.read_csv('D:\Thesis\Split_Merge_tables\Merged_Charteven
t\cerebral_infarction_chartevent_4.csv',usecols=['subject_id','hadm_id','stay_id','charttim
e','storetime','itemid','value','valuenum','valueuom','warning'])
```

```
df_patient_chartevent_5=pd.read_csv('D:\Thesis\Split_Merge_tables\Merged_Charteven
t\cerebral_infarction_chartevent_5.csv',usecols=['subject_id','hadm_id','stay_id','charttim
e','storetime','itemid','value','valuenum','valueuom','warning'])
```

```
df_patient_chartevent_6=pd.read_csv('D:\Thesis\Split_Merge_tables\Merged_Charteven
t\cerebral_infarction_chartevent_6.csv',usecols=['subject_id','hadm_id','stay_id','charttim
e','storetime','itemid','value','valuenum','valueuom','warning'])
```

14. Ενσωματώνουμε τα μέρη του *chartevents* σε ένα αρχείο.

```
pd.concat(  
    [df_patient_chartevent_0,  
    df_patient_chartevent_1,  
    df_patient_chartevent_2,  
    df_patient_chartevent_3,  
    df_patient_chartevent_4,  
    df_patient_chartevent_5,  
    df_patient_chartevent_6]).to_csv('D:\Thesis\exports\o6_patient_chartevent.csv',  
    index=False)
```

15. Εισαγωγή των dataframe

```
df_patient_chart=pd.read_csv('D:\Thesis\exports\o6_patient_chartevent.csv')
```

16. Κρατώ μόνο τις γραμμές που περιέχουν τιμές στην *valuenum*

```
without_nan=df_patient_chart[df_patient_chart['valuenum'].notna()]
```

17. Δημιουργία αρχείου του προηγούμενου βήματος.

```
without_nan.to_csv('D:\Thesis\exports\o7_patient_chartevent_without_nan.csv',  
    index=False)
```

18. Καταμέτρηση των εξετάσεων που κάνουν οι ασθενείς με εγκεφαλικό και μετατροπή από *series* σε *dataframe*.

```
rslt_df = pd.DataFrame(without_nan['itemid'].value_counts(dropna=True))
```

Το παραγόμενο αρχείο έχει σαν *index* την στήλη που περιέχει τον κωδικό των εξετάσεων, οπότε την κάνουμε *reset*

```
rslt_df = rslt_df.reset_index()
```

Ορίζουμε νέες ονομασίες στις στήλες για να μπορούν να είναι ευανάγνωστες.

```
rslt_df = rslt_df.rename(columns={'index': 'itemid', 'itemid': 'times'})
```

Εξαγωγή σε αρχείο

```
rslt_df.to_csv('D:\Thesis\exports\o8_chartevents_counts.csv', index=False)
```

19. Εισαγωγή dataframe

```
df_patients_chart = pd.read_csv  
('D:\Thesis\exports\o7_patient_chartevent_without_nan.csv')
```

20. Μετατροπή της στήλης charttime από object σε datetime

```
df_patients_chart['charttime'] = pd.to_datetime(df_patients_chart['charttime'])
```

21. Ομαδοποίηση στοιχείων ανά οχτάωρο

```
grouped_multiple = (df_patients_chart.groupby(['subject_id', 'hadm_id', 'itemid',
                                                pd.Grouper(key='charttime',
                                                            freq='8H')]).agg({'valuenum': ['mean']}))
grouped_multiple.columns = ['Mean_Chart']
grouped_multiple = grouped_multiple.reset_index()
```

22. Διαχωρισμός σε δυο καινούργιες στήλες της ημερομηνίας και ώρας

```
grouped_multiple['date'] = pd.to_datetime(grouped_multiple['charttime']).dt.date
grouped_multiple['time'] = pd.to_datetime(grouped_multiple['charttime']).dt.time
```

23. grouped_multiple.to_csv('D:\Thesis\exports\o9_chartevent_8_hour_mean.csv', index=False)

24. Εισαγωγή dataframe

```
grouped_multiple.to_csv('D:\Thesis\exports\o9_chartevent_8_hour_mean.csv',
index=False)
```

25. Δημιουργώ τις στήλες από το αρχείο που περιλαμβάνει τις μοναδικές εξετάσεις

```
df_columns = pd.read_csv('D:\Thesis\exports\o8_chartevents_counts.csv',
                        usecols = ['itemid'], index_col=False)
```

Ορίζω το header ως index έτσι ώστε όταν αλλάξω την κατεύθυνση του csv να μπουν τα στοιχεία στην κεφαλίδα.

```
df_columns = df_columns.set_index('itemid')
```

Μεταφέρω τις γραμμές σε στήλες

```
df_columns = df_columns.transpose()
```

26. Δημιουργώ ένα μοναδικό αναγνωριστικό με την ένωση των στοιχείων από στήλες subject, hadm_id, date και time. Ανάμεσα στα στοιχεία προβάλλω ένα κενό το οποίο θα μου φανεί χρήσιμο αργότερα για τον διαχωρισμό τους ξανά σε διαφορετικές στήλες.

```
df_rows["concat"] = (df_rows["subject_id"].astype(str)
                    + " "
                    +df_rows["hadm_id"].astype(str)
                    + " ")
```

```
+df_rows["date"].astype(str)
+" "+
df_rows["time"].astype(str))
```

27. Αλλάζω την σειρά των στηλών του dataframe.

```
df_rows = df_rows[['concat', 'subject_id', 'hadm_id', 'charttime', 'date', 'time', 'itemid',
'Mean_Chart']]
```

28. Από το `df_rows` κρατάω τις μοναδικές εγγραφές.

```
df_copy_rows = pd.DataFrame(df_rows.concat.unique())
```

Δημιουργώ καινούργιο πίνακα ενσωματώνοντας στον πίνακα `df` columns τις στήλες με τις κενές για την ώρα εξετάσεις.

```
chart_tbl = pd.concat([df_copy_rows, df_columns], axis=1)
```

Μετονομάζω την στήλη 0 σε `concat` για να είναι ίδιες στους δυο πίνακες

```
chart_tbl = chart_tbl.rename({0:'concat'}, axis=1)
```

29. Διαχωρισμός του `concat` σε τέσσερις στήλες.

```
chart_tbl[['subject_id', 'hadm_id', 'date', 'time']] =
chart_tbl['concat'].str.split(expand=True)
```

30. Διαγράφω την στήλη `concat` γιατί θα πρέπει να την ξαναδημιουργήσω χωρίς κενά ανάμεσα στα ενοποιημένα στοιχεία, οπότε θα την έχω διπλή

```
chart_tbl = chart_tbl.drop('concat', 1)
```

31. Εξαγωγή του άδειου πίνακα.

```
chart_tbl.to_csv('D:\Thesis\exports\o10_empty_chart_tbl.csv', index=False)
```

32. Εισαγωγή dataframe

```
df_chart_tbl = pd.read_csv('D:\Thesis\exports\o10_empty_chart_tbl.csv',
index_col=False)
```

```
df_mean_chart = pd.read_csv('D:\Thesis\exports\o9_chartevent_8_hour_mean.csv',
index_col=False)
```

```
df_data_chart = pd.read_csv('D:\Thesis\mimic-iv-1.0\icu\d_items.csv.gz')
```

33. Φέρνω τις τελευταίες στήλες, πρώτες

```
for i in range (1, 5):
```

```
cols = list(df_chart_tbl.columns)
cols = [cols[-1]] + cols[:-1]
df_chart_tbl = df_chart_tbl[cols]
```

34. Δημιουργώ εκ νέου την στήλη *concat* χωρίς κενά ανάμεσα στα στοιχεία. Θα χρησιμοποιήσω αυτή τη στήλη σαν *index* γιατί τα κενά μπορεί να δημιουργήσουν πρόβλημα.

```
df_chart_tbl["concat"] = (df_chart_tbl["subject_id"].astype(str)
                          + df_chart_tbl["hadm_id"].astype(str)
                          + df_chart_tbl["date"].astype(str)
                          + df_chart_tbl["time"].astype(str))
```

Αφαιρώ τους χαρακτήρες - και : από της τιμές της *concat* για να αποφύγω πιθανά προβλήματα

```
df_chart_tbl['concat'] = df_chart_tbl['concat'].str.replace('-', '')
df_chart_tbl['concat'] = df_chart_tbl['concat'].str.replace(':', '')
```

35. Ορίζω την στήλη *concat* ως *index* των γραμμών

```
df_chart_tbl = df_chart_tbl.set_index('concat')
```

36. Δημιουργώ την στήλη *concat* και στον πίνακα που περιλαμβάνει τις μετρήσεις για να είναι το κοινό σημείο ανάμεσα στους δυο πίνακες

```
df_mean_chart["concat"] = (df_mean_chart["subject_id"].astype(str)
                          + df_mean_chart["hadm_id"].astype(str)
                          + df_mean_chart["date"].astype(str)
                          + df_mean_chart["time"].astype(str))
```

Αφαιρώ τους χαρακτήρες - και : από της τιμές της *concat* για να αποφύγω πιθανά προβλήματα

```
df_mean_chart['concat'] = df_mean_chart['concat'].str.replace('-', '')
df_mean_chart['concat'] = df_mean_chart['concat'].str.replace(':', '')
```

37. Γεμίζω τον πίνακα

```
for row in df_mean_chart.iteruples():
    con = row.concat
    item = row.itemid
    value = row.Mean_Chart
```

```
df_chart_tbl.at['{}'.format(con), '{}'.format(item)] = value
```

38. Αποθήκευση του τελικού αρχείου με επικεφαλίδες τους κωδικούς των παρατηρήσεων

```
df_chart_tbl.to_csv('D:\Thesis\exports\o11_fill_chart_itemid_header_tbl.csv',  
index=False)
```

39. `df_data_chart = df_data_chart.set_index('itemid')`

40. Ενοποίηση τιμών σε κελιά

```
df_data_chart["header"] = (df_data_chart["label"].astype(str)  
+ "  
+ df_data_chart["unitname"].astype(str)  
+ ")")
```

41. Ελέγχω τις κεφαλίδες και αν δεν ανήκουν στις τέσσερις πρώτες τιμές κάνω τις αντικατάσταση.

```
for col in df_chart_tbl.columns:  
    if ((col != 'subject_id') and  
        (col != 'hadm_id') and  
        (col != 'date') and  
        (col != 'time')):  
        temp = (df_data_chart.loc[int(col), 'header'])  
        df_chart_tbl.rename(columns={'{}'.format(col): '{}'.format(temp)}, inplace=True)
```

42. Αποθήκευση του τελικού αρχείου με επικεφαλίδες τις ονομασίες των παρατηρήσεων

```
df_chart_tbl.to_csv('D:\Thesis\exports\o12_fill_chart_label_header_tbl.csv', index=False)
```

Παράρτημα VIII [Πίνακας Labevents]

1. Εισαγωγή της βιβλιοθήκης Panda.

```
import pandas as pd
```

2. Εισαγωγή dataframe

```
df_patients = pd.read_csv('D:\Thesis\exports\o5_icu_unique_admissions.csv')
```

3. Επεξεργασία πρώτου μέρους του αρχείου labevents

```
df_labevents =  
pd.read_csv('D:\Thesis\Split_Merge_tables\Splited_Labevents\labevents_0.csv')  
df_labevents.columns = ["labevent_id", "subject_id", "hadm_id", "specimen_id", "itemid",  
                        "charttime", "storetime", "value", "valuenum", "valueuom",  
                        "ref_range_lower", "ref_range_upper", "flag", "priority", "comments"]
```

```
rslt_df=pd.merge(df_patients, df_labevents, on=['subject_id','hadm_id'])
```

```
rslt_df.to_csv('D:\Thesis\Split_Merge_tables\Merged_Labevents\cerebral_infarction_labevents_0.csv', index=False)
```

```
del (df_labevents, rslt_df) # Διαγραφή των dataframe για απελευθέρωση μνήμης.
```

4. Επεξεργασία δεύτερου μέρους του αρχείου labevents

```
df_labevents =  
pd.read_csv('D:\Thesis\Split_Merge_tables\Splited_Labevents\labevents_1.csv')  
df_labevents.columns = ["labevent_id", "subject_id", "hadm_id", "specimen_id", "itemid",  
                        "charttime", "storetime", "value", "valuenum", "valueuom",  
                        "ref_range_lower", "ref_range_upper", "flag", "priority", "comments"]
```

```
rslt_df=pd.merge(df_patients, df_labevents, on=['subject_id','hadm_id'])
```

```
rslt_df.to_csv('D:\Thesis\Split_Merge_tables\Merged_Labevents\cerebral_infarction_labevents_1.csv', index=False)
```

`del (df_labevents, rslt_df) # Διαγραφή των dataframe για απελευθέρωση μνήμης.`

5. Επεξεργασία τρίτου μέρους του αρχείου *labevents*

```
df_labevents =
pd.read_csv('D:\Thesis\Split_Merge_tables\Splited_Labevents\labevents_2.csv')
df_labevents.columns = ["labevent_id", "subject_id", "hadm_id", "specimen_id", "itemid",
                        "charttime", "storetime", "value", "valuenum", "valueuom",
                        "ref_range_lower", "ref_range_upper", "flag", "priority", "comments"]

rslt_df=pd.merge(df_patients, df_labevents, on=['subject_id','hadm_id'])

rslt_df.to_csv('D:\Thesis\Split_Merge_tables\Merged_Labevents\cerebral_infarction_labevents_2.csv', index=False)
```

`del (df_labevents, rslt_df) # Διαγραφή των dataframe για απελευθέρωση μνήμης.`

6. Φορτώνουμε τα τρία παραγόμενα από τα προηγούμενα βήματα (7 έως 12) αρχεία

```
df_labevents_0 =
pd.read_csv('D:\Thesis\Split_Merge_tables\Merged_Labevents\cerebral_infarction_labevents_0.csv',
            usecols = ["labevent_id", "subject_id", "hadm_id", "specimen_id", "itemid",
                      "charttime", "storetime", "value", "valuenum", "valueuom",
                      "ref_range_lower", "ref_range_upper", "flag", "priority", "comments"])

df_labevents_1 =
pd.read_csv('D:\Thesis\Split_Merge_tables\Merged_Labevents\cerebral_infarction_labevents_1.csv',
            usecols = ["labevent_id", "subject_id", "hadm_id", "specimen_id", "itemid",
                      "charttime", "storetime", "value", "valuenum", "valueuom",
                      "ref_range_lower", "ref_range_upper", "flag", "priority", "comments"])
```

```
df_labevents_2 =
pd.read_csv('D:\Thesis\Split_Merge_tables\Merged_Labevents\cerebral_infarction_labevents_2.csv',
            usecols = ["labevent_id", "subject_id", "hadm_id", "specimen_id", "itemid",
                      "charttime", "storetime", "value", "valuenum", "valueuom",
                      "ref_range_lower", "ref_range_upper", "flag", "priority", "comments"])
```

7. Ενσωματώνουμε τα μέρη του *labevent* σε ένα αρχείο.

```
pd.concat(
    [df_labevents_0,
     df_labevents_1,
     df_labevents_2]).to_csv('D:\Thesis\exports\o13_icu_labevents.csv', index=False)
```

- # Διαγραφή των dataframe για απελευθέρωση μνήμης.
`del (df_labevents_0,
df_labevents_1,
df_labevents_2)`
8. Εξαγωγή του παραγόμενου αρχείου.
`df_labevents = pd.read_csv('D:\Thesis\exports\o13_icu_labevents.csv')`
9. Μετατροπή της στήλης charttime από object σε datetime
`df_labevents['charttime'] = pd.to_datetime(df_labevents['charttime'])`
10. Απομονώνω τις γραμμές που δεν περιλαμβάνουν τιμές στην valuenum
`df_labevents = df_labevents[df_labevents['valuenum'].notna()]`
11. Ομαδοποίηση των τιμών του labevents ανά οχτάωρο.
`grouped_multiple = df_labevents.groupby(['subject_id', 'hadm_id', 'itemid',
pd.Grouper(key='charttime',
freq='8H')]).agg({'valuenum': ['mean']})
grouped_multiple.columns = ['Mean_Lab']
grouped_multiple = grouped_multiple.reset_index()
grouped_multiple`
12. Εξαγωγή των ομαδοποιημένων ανά οχτάωρο εργαστηριακών εξετάσεων.
`grouped_multiple.to_csv('D:\Thesis\exports\o14_labevents_grouped_8hours.csv',
index=False)`
13. Δήλωση dataframe
`grouped_multiple.to_csv('D:\Thesis\exports\o14_labevents_grouped_8hours.csv',
index=False)`
- Καταμέτρηση των εξετάσεων που κάνουν οι ασθενείς και μετατροπή από series σε dataframe.
`rslt_df = pd.DataFrame(df_labevents['itemid'].value_counts(dropna=True))`
- Το παραγόμενο αρχείο έχει σαν index την στήλη που περιέχει τον κωδικό των εξετάσεων, οπότε την κάνουμε reset
`rslt_df = rslt_df.reset_index()`
- Και ορίζουμε νέες ονομασίες στις στήλες για να μπορούν να είναι ευανάγνωστες.
`rslt_df.rename (columns = {'index': 'itemid', 'itemid':
'times'}).to_csv('D:\Thesis\exports\o15_labevents_counts.csv', index=False)`

14. Δημιουργώ τον πίνακα που θα περιλαμβάνει τις τιμές των εργαστηριακών μετρήσεων ασθενών με εγκεφαλικό που έχουν εισαχθεί στις ΜΕΘ

```
df_rows = pd.read_csv('D:\Thesis\exports\o14_labevents_grouped_8hours.csv')
```

```
# Μετατροπή της στήλης charttime από object σε datetime
```

```
df_rows['charttime'] = pd.to_datetime(df_rows['charttime'])
```

```
# Διαχωρισμός σε δυο καινούργιες στήλες της ημερομηνίας και ώρας
```

```
df_rows['date'] = pd.to_datetime(df_rows['charttime']).dt.date
```

```
df_rows['time'] = pd.to_datetime(df_rows['charttime']).dt.time
```

15. Αποθηκεύω το αρχείο που πλέον περιλαμβάνει τις δυο καινούργιες στήλες, date και time.

```
df_rows.to_csv('D:\Thesis\exports\o16_labevents_grouped_8hours_date_time.csv',
              index=False)
```

16. Δημιουργώ τις στήλες από το αρχείο που περιλαμβάνει τις μοναδικές εξετάσεις

```
df_columns = pd.read_csv('D:\Thesis\exports\o15_labevents_counts.csv',
                        usecols = ['itemid'], index_col=False)
```

Ορίζω το header ως index

```
df_columns = df_columns.set_index('itemid')
```

Μεταφέρω τις γραμμές σε στήλες

```
df_columns = df_columns.transpose()
```

17. Δημιουργώ ένα μοναδικό αναγνωριστικό με την ένωση των στοιχείων από στήλες subject, hadm_id, date και time. Ανάμεσα στα στοιχεία προβάλλω ένα κενό το οποίο θα μου φανεί χρήσιμο αργότερα για τον διαχωρισμό τους ξανά σε διαφορετικές στήλες.

```
df_rows["concat"] = (df_rows["subject_id"].astype(str)
                    + " "
                    + df_rows["hadm_id"].astype(str)
                    + " "
                    + df_rows["date"].astype(str)
                    + " "+
                    df_rows["time"].astype(str))
```

18. Από το df_rows κρατάω τις μοναδικές εγγραφές

```
df_copy_rows = pd.DataFrame(df_rows.concat.unique())
```

Δημιουργώ καινούργιο πίνακα ενσωματώνοντας στον πίνακα `df` columns τις στήλες με τις κενές για την ώρα εξετάσεις

```
lab_tbl = df_columns.append(df_copy_rows)
```

Μετονομάζω την στήλη 0 σε `concat` για να είναι ίδιες στους δυο πίνακες

```
lab_tbl = lab_tbl.rename({0:'concat'}, axis=1)
```

19. Διαχωρισμός του `concat` σε τέσσερις στήλες

```
lab_tbl[['subject_id', 'hadm_id', 'date', 'time']] = lab_tbl['concat'].str.split(expand=True)
```

20. Διαγράφω την στήλη `concat` γιατί θα πρέπει να την ξαναδημιουργήσω χωρίς κενά ανάμεσα στα ενοποιημένα στοιχεία, οπότε θα την έχω διπλή.

```
lab_tbl = lab_tbl.drop('concat', 1)
```

21. Αποθήκευση κενού πίνακα.

```
lab_tbl.to_csv('D:\Thesis\exports\o17_empty_lab_tbl.csv', index=False)
```

22. `df_lab_tbl = pd.read_csv('D:\Thesis\exports\o17_empty_lab_tbl.csv', index_col = False)`

```
df_data_labevents = pd.read_csv('D:\Thesis\mimic-iv-1.0\hosp\d_labitems.csv.gz')
```

```
df_mean_lab =
```

```
pd.read_csv('D:\Thesis\exports\o16_labevents_grouped_8hours_date_time.csv')
```

23. Φέρνω τις τελευταίες στήλες, πρώτες

```
for i in range (1, 5):
```

```
cols = list(df_lab_tbl.columns)
```

```
cols = [cols[-1]] + cols[:-1]
```

```
df_lab_tbl = df_lab_tbl[cols]
```

24. Δημιουργώ εκ νέου την στήλη `concat` χωρίς κενά ανάμεσα στα στοιχεία. Θα χρησιμοποιήσω αυτή τη στήλη σαν `index` και τα κενά μπορεί να δημιουργήσουν πρόβλημα.

```
df_lab_tbl["concat"] = (df_lab_tbl["subject_id"].astype(str)
```

```
    + df_lab_tbl["hadm_id"].astype(str)
```

```
    + df_lab_tbl["date"].astype(str)
```

```
    + df_lab_tbl["time"].astype(str))
```

Αφαιρώ τους χαρακτήρες - και : από της τιμές της concat
`df_lab_tbl['concat'] = df_lab_tbl['concat'].str.replace('-', '')`
`df_lab_tbl['concat'] = df_lab_tbl['concat'].str.replace(':', '')`

Ορίζω την στήλη concat ως index των γραμμών
`df_lab_tbl = df_lab_tbl.set_index('concat')`

25. Δημιουργώ την στήλη concat και στον πίνακα που περιλαμβάνει τις μετρήσεις για να είναι το κοινό σημείο ανάμεσα στους δυο πίνακες.

```
df_mean_lab["concat"] = (df_mean_lab["subject_id"].astype(str)
                        + df_mean_lab["hadm_id"].astype(str)
                        + df_mean_lab["date"].astype(str)
                        + df_mean_lab["time"].astype(str))
```

Αφαιρώ τους χαρακτήρες - και : από της τιμές της concat
`df_mean_lab['concat'] = df_mean_lab['concat'].str.replace('-', '')`
`df_mean_lab['concat'] = df_mean_lab['concat'].str.replace(':', '')`

26. Γεμίζω τον πίνακα

```
for row in df_mean_lab.itertuples():
    con = row.concat
    item = row.itemid
    value = row.Mean_Lab
    df_lab_tbl.at['{}'.format(con), '{}'.format(item)] = value
```

27. Αποθήκευση του τελικού αρχείου με επικεφαλίδες τους κωδικούς των παρατηρήσεων

```
df_lab_tbl.to_csv('D:\Thesis\exports\o18_fill_lab_itemid_header_tbl.csv', index=False)
```

28. `df_data_labevents = df_data_labevents.set_index('itemid')`

29. Ενοποίηση τιμών σε κελιά

```
df_data_labevents["header"] = (df_data_labevents["label"].astype(str)
                                + " ("
                                + df_data_labevents["fluid"].astype(str)
                                + ")")
```

30. Ελέγχω τις κεφαλίδες και αν δεν ανήκουν στις τέσσερις πρώτες τιμές κάνω τις αντικαταστάσεις.

```
for col in df_lab_tbl.columns:
    if ((col != 'subject_id')
        and (col != 'hadm_id')
        and (col != 'date')
        and (col != 'time')):
        temp = (df_data_labevents.loc[int(col), 'header'])
        df_lab_tbl.rename(columns={'{}'.format(col): '{}'.format(temp)}, inplace=True)
```

31. Αποθήκευση του τελικού αρχείου με επικεφαλίδες τα label των παρατηρήσεων
`df_lab_tbl.to_csv('D:\Thesis\exports\o19_fill_lab_label_header_tbl.csv', index=False)`

Παράρτημα ΙΧ [Ενιαίος Πίνακας]

1. Εισαγωγή της βιβλιοθήκης Panda.

```
import pandas as pd
```

2. Εισαγωγή dataframe

```
df_icu_chart = pd.read_csv('D:\Thesis\exports\o12_fill_chart_label_header_tbl.csv')
```

```
df_lab_chart = pd.read_csv('D:\Thesis\exports\o19_fill_lab_label_header_tbl.csv')
```

3. Ελέγχω αν μετά την συγχώνευση υπάρχουν διπλές ίδιες στήλες

```
rslt=pd.merge(df_icu_chart, df_lab_chart, on=['subject_id','hadm_id', 'date', 'time'],  
             how="outer", suffixes=("", '_delme'))
```

Και αν υπάρχουν τις αφαιρώ

```
rslt = rslt[[c for c in rslt.columns if not c.endswith('_delme')]]
```

4. Εισάγω στήλη με την τιμές την ονομασία του τμήματος

```
rslt.insert(0, "Department", "ICU")
```

5. Εξαγωγή σε csv

```
rslt.to_csv('D:\Thesis\exports\o20_ICU_merged_chart_lab.csv', index=False)
```

6. `df_icu = pd.read_csv('D:\Thesis\exports\o20_ICU_merged_chart_lab.csv')`

```
df_death= pd.read_csv('D:\Thesis\mimic-iv-1.0\core\Admissions.csv.gz', usecols =  
['subject_id', 'hadm_id', 'hospital_expire_flag'])
```

```
df_death = df_death.loc[df_death['hospital_expire_flag'] == 1]
```

7. Ενοποιώ τα *dataframes* του συνολικού πίνακα με της θνησιμότητας.

```
df_rslt=pd.merge(df_icu, df_death, on=['subject_id','hadm_id'],  
                 how="left", suffixes=(", '_delme'))
```

8. Εισάγω την τιμή μηδέν στα κενά κελιά της *hospital_expire_flag*

```
df_rslt[['hospital_expire_flag']] = df_rslt[['hospital_expire_flag']].fillna(0)
```

Μετατρέπω την στήλη *hospital_expire_flag* σε ακέραιο.

```
df_rslt['hospital_expire_flag'] = df_rslt['hospital_expire_flag'].apply(int)
```

9. Αντικατάσταση τιμών από μηδέν σε *survival* και από ένα σε *death* στην στήλη *hospital_expire_flag*

```
df_rslt["hospital_expire_flag"] = np.where(df_rslt["hospital_expire_flag"] == 0, "survival",  
                                           "death")
```

10. Μετατροπή του *date* από *object* σε *date* για να μπορέσει να εκτελεστεί η στοίχιση που ακολουθεί.

```
df_rslt['date'] = pd.to_datetime(df_rslt.date)
```

```
df_rslt = df_rslt.sort_values(by=["subject_id", "date", "time"], ascending=[True, True,  
True])
```

11. Εξαγωγή σε *csv*

```
df_rslt.to_csv('D:\Thesis\exports\o31_ICU_final.csv', index=False)
```

12. *df_icu* = *pd.read_csv('D:\Thesis\exports\o31_ICU_final.csv')*

```
df_1 = df_icu.iloc[:63568,:]
```

```
df_2 = df_icu.iloc[63568:,:]
```

```
df_1.to_csv('D:\Thesis\Weka\Train_Test_Sets\Sorted Set\Train_70.csv', index=False)
```

```
df_2.to_csv('D:\Thesis\Weka\Train_Test_Sets\Sorted Set\Test_30.csv', index=False)
```

Παράρτημα Χ [ROC Διαγράμματα]

1. Εισαγωγή βιβλιοθηκών για την δημιουργία ROC καμπυλών των πρώτων τεσσάρων ημερών.

```
import pandas as pd
import numpy as np
from sklearn.datasets import make_classification
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score
from matplotlib import pyplot
import matplotlib.pyplot as plt
```

2. Διάβασμα από csv

```
df_classifier = pd.read_csv('D:\Thesis\Weka\ROC\treeLMT.csv', encoding=
'unicode_escape')
```

Ανάθεση σε μεταβλητές των στοιχείων του πίνακα των πρώτων τεσσάρων ημερών.

```
day1 = df_classifier.loc[df_classifier['Denominator'] == 3]
day2 = df_classifier.loc[df_classifier['Denominator'] == 6]
day3 = df_classifier.loc[df_classifier['Denominator'] == 9]
day4 = df_classifier.loc[df_classifier['Denominator'] == 12]
```

Ανάθεση σε μεταβλητές της στήλης Ground Truth Alive της κάθε μέρας και μετατροπή της από dataframe σε list.

```
ground_truth_alive1 = day1["Weka (Ground Truth Alive)"].tolist()
ground_truth_alive2 = day2["Weka (Ground Truth Alive)"].tolist()
ground_truth_alive3 = day3["Weka (Ground Truth Alive)"].tolist()
ground_truth_alive4 = day4["Weka (Ground Truth Alive)"].tolist()
```

Ανάθεση σε μεταβλητές της στήλης *Prob - Alive* της κάθε μέρας και μετατροπή της από *dataframe* σε *list*.

```
prob_alive1 = day1["Prob - Alive"].tolist()
prob_alive2 = day2["Prob - Alive"].tolist()
prob_alive3 = day3["Prob - Alive"].tolist()
prob_alive4 = day4["Prob - Alive"].tolist()
```

Ανάθεση σε μεταβλητές της στήλης *Prob - Dead* της κάθε μέρας και μετατροπή της από *dataframe* σε *list*.

```
prob_dead1 = day1["Prob - Dead"].tolist()
prob_dead2 = day2["Prob - Dead"].tolist()
prob_dead3 = day3["Prob - Dead"].tolist()
prob_dead4 = day4["Prob - Dead"].tolist()
```

3. *Ground Truth Dead* ανά ημέρα

Για την δημιουργία του *Ground Truth Dead* ανά ημέρα, αφαιρώ από την *ground_truth_alive* της κάθε μέρας το 1 και με την χρήση της *abs* παίρνω την απόλυτη τιμή της. Έτσι έχω τις τιμές χωρίς πρόσημο. Με αυτή τη διεργασία οι τιμές της *Ground Truth Alive* και *Ground Truth Dead* είναι αντεστραμμένες.

```
#Day 1
ground_truth_dead1=abs(np.subtract(ground_truth_alive1, 1))
```

```
#Day 2
ground_truth_dead2=abs(np.subtract(ground_truth_alive2, 1))
```

```
#Day 3
ground_truth_dead3=abs(np.subtract(ground_truth_alive3, 1))
```

```
#Day 4
ground_truth_dead4=abs(np.subtract(ground_truth_alive4, 1))
```

4. *Roc* τεσσάρων πρώτων ημερών

```
print(roc_auc_score(ground_truth_alive1, prob_alive1))
print(roc_auc_score(ground_truth_alive2, prob_alive2))
print(roc_auc_score(ground_truth_alive3, prob_alive3))
```

```
print(roc_auc_score(ground_truth_alive4, prob_alive4))

pyplot.figure(figsize=(15, 10))
pyplot.plot(fpr1, tpr1, linestyle='--', label='First Day')
pyplot.plot(fpr2, tpr2, linestyle='--', label='Second Day')
pyplot.plot(fpr3, tpr3, linestyle='--', label='Third Day')
pyplot.plot(fpr4, tpr4, linestyle='--', label='Fourth Day')

pyplot.xlabel('False Positive Rate') # Εκτυπώνει τον τίτλο του X
pyplot.ylabel('True Positive Rate') # Εκτυπώνει τον τίτλο του Y
pyplot.legend() # Εκτυπώνει το label
pyplot.grid(True) # Εκτυπώνει τα τετράγωνα
pyplot.title("Logistic")
pyplot.show()
```