ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΗΣ ΣΧΕΔΙΑΣΗΣ ΚΑΙ ΠΑΡΑΓΩΓΗΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΦΑΡΜΟΓΕΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΕ ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ

MACHINE LEARNING APPLICATIONS IN MEDICAL DATA

ΖΑΝΙ ΠΥΡΡΟ

ΕΠΙΒΛΕΨΗ:

ΝΙΚΟΛΑΟΥ ΓΡΗΓΟΡΗΣ

ΑΘΗΝΑ, ΜΑΡΤΙΟΣ 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΗΣ ΣΧΕΔΙΑΣΗΣ ΚΑΙ ΠΑΡΑΓΩΓΗΣ

ΕΦΑΡΜΟΓΕΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΕ ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ

Μέλη Εξεταστικής Επιτροπής συμπεριλαμβανομένου και του Εισηγητή

Η διπλωματική εργασία εξετάστηκε επιτυχώς από την κάτωθι Εξεταστική Επιτροπή:

Γ. ΝΙΚΟΛΑΟΥ	**Σ. ΒΑΣΙΛΕΙΑΔΟΥ**	**Χ. ΔΡΟΣΟΣ**
ΛΕΚΤΟΡΑΣ ΕΦΑΡΜΟΓΩΝ	ΕΠΙΚΟΥΡΗ ΚΑΘΗΓΗΤΡΙΑ	Ε.ΔΙ.Π
GRIGORIOS NIKOLAOU Digitally signed by GRIGORIOS NIKOLAOU Date: 2023.03.23 12:17:12 +02'00'	Soultana Vasileiadou Digitally signed by Soultana Vasileiadou Date: 2023.03.23 15:28:35 +02'00'	Christos Drosos Digitally signed by Christos Drosos Date: 2023.03.27 22:00:40 +03'00'

ΑΘΗΝΑ, ΜΑΡΤΙΟΣ 2023

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο κάτωθι υπογεγραμμένος Ζάνι Πύρρο του Βίκτωρ, με αριθμό μητρώου 71446021 φοιτητής του Πανεπιστημίου Δυτικής Αττικής της Σχολής Μηχανικών του Τμήματος Μηχανικών Βιομηχανικής Σχεδίασης και Παραγωγής, **δηλώνω υπεύθυνα** ότι:

«*Είμαι συγγραφέας αυτής της πτυχιακής/διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος. Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου*».

Ο Δηλών:

ΖΑΝΙ ΠΥΡΡΟ

Ευχαριστίες

Πρώτα απ' όλα θα ήθελα να εκφράσω θερμά τις ευχαριστίες μου στον κ. Νικολάου Γρηγόρη όχι μόνο για την καθοδήγηση, την βοήθεια και τις απόψεις του για την παρούσα εργασία αλλά και για την καθοδήγηση και εκμάθηση σε όλα τα υπόλοιπα μαθήματα που μας δίδαξε. Θα ήθελα επίσης να εκφράσω ένα τεράστιο ευχαριστώ σε όλους τους φίλους μου που με στήριξαν στην εκπόνηση της εργασίας αυτής.

Τέλος, αφιερώνω την εργασία αυτή στην οικογένεια μου, τον Βίκτωρ και την Ντρίτα και παράλληλα τους ευχαριστώ διότι χωρίς την πολύτιμη πνευματική στήριξη, την απόλυτη εμπιστοσύνη και την αγάπη που μου έδειξαν καθ' όλη την διάρκεια των σπουδών μου, δεν θα κατάφερνα να τις ολοκληρώσω.

Περίληψη

Ο σκοπός της διπλωματικής αυτής εργασίας είναι να γίνει αντιληπτό από τον αναγνώστη το πως μπορεί να εφαρμοστεί η Μηχανική Μάθηση και ποιοι μέθοδοι χρησιμοποιούνται στα τωρινά προβλήματα που αφορούν την επιστήμη της ιατρικής.

Αρχικά, παρουσιάζονται η έννοια της Μηχανικής Μάθησης και θεωρητικά η διαδικασία εφαρμογής αυτής ανά βήμα.

Έπειτα, παρουσιάζονται διαφορετικά παραδείγματα εφαρμογής της Μηχανικής Μάθησης σε διάφορους τομείς της ιατρικής, με την μέθοδο που αντιμετωπίστηκαν και την επίδοση που παρουσίασαν.

Τέλος, με την βοήθεια της γλώσσας προγραμματισμού Python, πρακτικά έγινε η εφαρμογή της Μηχανικής Μάθησης για δύο διαφορετικά σύνολα πραγματικών δεδομένων όπου το πρώτο αφορά ασθενείς εγκεφαλικών επεισοδίων και το δεύτερο σύνολο αφορά δεδομένα καρδιοτοκογραφίας εγκύων γυναικών. Στο πρώτο σύνολο εξετάστηκε αν η Μηχανική Μάθηση είναι ικανή να προβλέψει καταστάσεις ασθενών με εγκεφαλικό επεισόδιο, ενώ στο δεύτερο σύνολο αν αυτή είναι ικανή να προβλέψει την κατάσταση της καρδιακής λειτουργίας του εμβρύου.

Λέξεις Κλειδιά: Μηχανική Μάθηση, Κατηγοριοποίηση, Ανομοιογενή Δεδομένα, Ιατρικά Δεδομένα

Abstract

This thesis aims to provide the reader with an understanding of how Machine Learning can be applied and which methods are used in today's problems related to the science of medicine.

Firstly, the concept of Machine Learning and how it is applied step by step are presented theoretically.

Then, various examples of application of Machine Learning in different areas of medicine are introduced, including the method they were dealt with and how they performed.

Finally, by using the Python programming language, the application of Machine Learning was practically done for two different real-world datasets where the first one is for stroke patients and the second dataset is for cardiotocography data of pregnant women. In the first set it was investigated whether Machine Learning is capable of predicting the situations of stroke patients, while in the second set whether it is capable of predicting the Fetal Heart Rate condition.

Keywords: Machine Learning, Classification, Imbalanced Dataset, Medical Dataset

Κατάλογος Περιεχομένων

ΕΥΧΑΡΙΣΤΙΕΣ **1**
ΠΕΡΙΛΗΨΗ **2**
ABSTRACT **3**
ΚΑΤΑΛΟΓΟΣ ΠΕΡΙΕΧΟΜΕΝΩΝ **4**
ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ **6**
ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ **8**
ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ **9**
1.1 Τεχνητή Νοημοσύνη 9
1.2 Κατηγορίες Μηχανικής Μάθησης 12
1.3 Αλγόριθμοι Μηχανικής Μάθησης 13
1.3.1 Decision Tree – Δέντρο απόφασης 14
1.3.2 Random Forest – Random 'Decision' Forest 15
1.3.3 Gradient Boosting 16
1.4 Εφαρμογή της Μηχανικής Μάθησης ανά βήμα 17
1.5 Έλεγχος επίδοσης Αλγορίθμου – Metrics 19
1.5.1 Accuracy 19
1.5.2 Confusion Matrix 19
1.5.3 Precision, Recall και F-measure 20
1.5.4 ROC curve 21
ΚΕΦΑΛΑΙΟ 2: ΕΦΑΡΜΟΓΕΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ **22**
2.1 Ανίχνευση καρδιακών παθήσεων 22
2.2 Αξιολόγηση δεξιοτήτων στη χειρουργική με την ρομποτική υποβοήθηση 23
2.3 Πρόβλεψη των εισαγωγών στο τμήμα επειγόντων περιστατικών για τη βελτίωση της ροής των ασθενών 24
2.4 Ανίχνευση κατάθλιψης από τα μέσα κοινωνικής δικτύωσης 25
2.5 Πρόβλεψη σοβαρού/κρίσιμου συμπτώματος μολυσμένων ασθενών με κορονοϊό 25
ΚΕΦΑΛΑΙΟ 3: ΠΡΟΒΛΕΨΗ ΕΓΚΕΦΑΛΙΚΩΝ ΕΠΕΙΣΟΔΙΩΝ ΑΣΘΕΝΩΝ **27**
3.1 Σοβαρές αιτίες πρόκλησης εγκεφαλικού επεισοδίου 27
3.1.1 Η Υπέρταση 28
3.1.2 Ο Διαβήτης 29

3.1.3 Η Παχυσαρκία 29
3.2 Τύποι εγκεφαλικών επεισοδίων 31
3.2.1 Ισχαιμικό εγκεφαλικό επεισόδιο 31
3.2.2 Αιμορραγικό εγκεφαλικό επεισόδιο 32
3.3 Εφαρμογή της Μηχανικής Μάθησης 33
3.3.1 Ανάλυση και επεξεργασία συνόλου δεδομένων 33
3.3.2 Προετοιμασία δεδομένων για την εκπαίδευση 41
3.3.3 Εκπαίδευση και έλεγχος επίδοσης αλγορίθμου 43
ΚΕΦΑΛΑΙΟ 4: ΠΡΟΒΛΕΨΗ ΚΑΤΑΣΤΑΣΗΣ ΕΜΒΡΥΪΚΗΣ ΚΑΡΔΙΑΚΗΣ ΛΕΙΤΟΥΡΓΙΑΣ 49
4.1 Η καρδιοτοκογραφία 49
4.2 Χαρακτηριστικά καρδιοτοκογραφήματος 50
4.3 Εφαρμογή της Μηχανικής Μάθησης 56
4.3.1 Ανάλυση και επεξεργασία συνόλου δεδομένων 56
4.3.2 Προετοιμασία δεδομένων για την εκπαίδευση 65
4.3.3 Εκπαίδευση και έλεγχος επίδοσης αλγορίθμου 68
ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ 75

Κατάλογος Εικόνων

Εικόνα 1: Απεικόνιση TN, MM και BM [**3**] 10
Εικόνα 2: Απεικόνιση κατηγοριών MM 13
Εικόνα 3: Απεικόνιση μιας δομής Δέντρου Απόφασης [**10**] 14
Εικόνα 4: Απεικόνιση δομής ενός Random Forest [**12**] 15
Εικόνα 5: Απεικόνιση λειτουργίας Gradient Boosting [**13**] 16
Εικόνα 6: Η Μηχανική Μάθηση σαν σύστημα 17
Εικόνα 7: Επτά βήματα στην Μηχανική Μάθηση [**14**] 18
Εικόνα 8: Πίνακας σύγχυσης 20
Εικόνα 9: Παράδειγμα ενός ROC Curve [**18**] 21
Εικόνα 10: Σχηματική απεικόνιση του RMIS με τη συσκευή αφής MRP [**23**] 23
Εικόνα 11: Κυκλοφορία αίματος μέσα στις φλέβες [**45**] 28
Εικόνα 12: Σχήμα BMI σε συνάρτηση βάρους και ύψους [**51**] 30
Εικόνα 13: Παράδειγμα Ισχαιμικού εγκεφαλικού επεισοδίου [**57**] 32
Εικόνα 14: Παράδειγμα Αιμορραγικού εγκεφαλικού επεισοδίου [**57**] 32
Εικόνα 15: Γραφική αναπαράσταση όλων των χαρακτηριστικών 34
Εικόνα 16: Αναπαράσταση χαρακτηριστικών για ελλιπείς τιμές 35
Εικόνα 17: Αναπαράσταση χαρακτηριστικών μετά την αφαίρεση ελλιπών τιμών 36
Εικόνα 18: Πίνακας συσχέτισης Pearson 37
Εικόνα 19: Ραβδόγραμμα αριθμητικών χαρακτηριστικών συσχέτισης με stroke 37
Εικόνα 20: Γραφική αναπαράσταση ηλικίας ανάλογα με το εγκεφαλικό 38
Εικόνα 21: Ραβδόγραμμα πυκνότητας υπέρτασης ανάλογα με το stroke 39
Εικόνα 22: Αναλογία όσων έχουν εγκεφαλικό ή όχι στα δεδομένα 40
Εικόνα 23: Διάγραμμα διασποράς Age και BMI πριν και μετά την υπερδειγματοληψία .. 42
Εικόνα 24: Ραβδόγραμμα δειγμάτων stroke πριν και μετά την υπερδειγματοληψία 43
Εικόνα 25: Πίνακες σύγχυσης ανά μοντέλο πριν το Oversampling 44
Εικόνα 26: Πίνακες σύγχυσης ανά μοντέλο μετά το Oversampling 46
Εικόνα 27: Πίνακες σύγχυσης ανά μοντέλο πριν και μετά το Oversampling 48
Εικόνα 28: Το μηχάνημα της καρδιοτοκογραφίας [**65**] 50
Εικόνα 29: Παράδειγμα γραφικής απεικόνισης της καρδιοτοκογραφίας [**66**] 50

Εικόνα 30: Συσπάσεις της μήτρας [**68**] 51
Εικόνα 31: Βασικός εμβρυϊκός καρδιακός ρυθμός [**68**] 52
Εικόνα 32: Μεταβλητότητα [**68**] 52
Εικόνα 33: Επιταχύνσεις [**68**] 53
Εικόνα 34: Πρώιμες επιβραδύνσεις [**68**] 54
Εικόνα 35: Όψιμες επιβραδύνσεις [**68**] 54
Εικόνα 36: Μεταβαλλόμενες επιβραδύνσεις [**68**] 55
Εικόνα 37: Παρατεταμένες επιβραδύνσεις [**68**] 55
Εικόνα 38: Γραφική αναπαράσταση για κάθε χαρακτηριστικό 58
Εικόνα 39: Ραβδόγραμμα ελλιπών τιμών για κάθε χαρακτηριστικό 59
Εικόνα 40: Πίνακας συσχέτισης Pearson 60
Εικόνα 41: Ραβδόγραμμα χαρακτηριστικών συσχέτισης με NSP 61
Εικόνα 42: Πλήθος τιμών DP ανά NSP 62
Εικόνα 43: Γραφική αναπαράσταση πυκνότητας ASTV ανά NSP 63
Εικόνα 44: Γραφική αναπαράσταση πυκνότητας ALTV ανά NSP 64
Εικόνα 45: Ποσοστό ανά NSP σε γράφημα πίτας 65
Εικόνα 46: Θηκόγραμμα για κάθε χαρακτηριστικό πριν και μετα την κλιμάκωση 66
Εικόνα 47: Διάγραμμα διασποράς ASTV και ALTV πριν και μετά το SMOTE 67
Εικόνα 48: Ραβδόγραμμα δειγμάτων NSP πριν και μετά το SMOTE 67
Εικόνα 49: Διαχωρισμός δεδομένων σε πέντε K-Folds 69
Εικόνα 50: Πίνακας σύγχυσης για Decision Tree 71
Εικόνα 51: Πίνακας σύγχυσης για Random Forest 72
Εικόνα 52: Πίνακας σύγχυσης για Gradient Boosting 73
Εικόνα 53: Πίνακες σύγχυσης για κάθε κλάση και για κάθε μοντέλο 74

Κατάλογος Πινάκων

Πίνακας 1: Παράδειγμα συνόλου δεδομένων καιρού ανά ώρα 11
Πίνακας 2: Συνδυασμοί χαρακτηριστικών με την καλύτερη επίδοση 26
Πίνακας 3: Πληροφορίες για τα χαρακτηριστικά του συνόλου δεδομένων 33
Πίνακας 4: Εύρος τιμών ηλικίας από 0,08 - 1,88 35
Πίνακας 5: Ποσοστά συσχέτισης για κάθε αριθμητικό χαρακτηριστικό με το stroke..... 38
Πίνακας 6: Dataframe πριν την κωδικοποίηση..... 40
Πίνακας 7: Dataframe μετά την κωδικοποίηση 40
Πίνακας 8: Πλήθος δειγμάτων σε σύνολο εκπαίδευσης και δοκιμής 41
Πίνακας 9: Πλήθος δειγμάτων πριν και μετα το Oversampling 42
Πίνακας 10: Ποσοστό ακρίβειας για κάθε μοντέλο πριν το Oversampling..... 43
Πίνακας 11: Ποσοστά Recall για όλα τα μοντέλα ανά κλάση πριν το Oversampling..... 44
Πίνακας 12: Ποσοστό ακρίβειας για κάθε μοντέλο μετά το Oversampling 45
Πίνακας 13: Ποσοστά Recall για όλα τα μοντέλα ανά κλάση μετά το Oversampling 45
Πίνακας 14: Χαρακτηριστική καμπύλη λειτουργίας για όλα τα μοντέλα 47
Πίνακας 15: Μνημονικός κανόνας DR C BRAVADO..... 51
Πίνακας 16: Χαρακτηριστικά συνόλου δεδομένων καρδιοτοκογραφίας 57
Πίνακας 17: Ποσοστά συσχέτισης για κάθε χαρακτηριστικό με το NSP 61
Πίνακας 18: Πλήθος δειγμάτων σε σύνολο εκπαίδευσης και δοκιμής 66
Πίνακας 19: Πλήθος δειγμάτων πριν και μετα το SMOTE 68
Πίνακας 20: Σύγκριση ακρίβειας ανά αλγόριθμο 69
Πίνακας 21: Ποσοστά Recall για κάθε κλάση στο Decision Tree..... 70
Πίνακας 22: Ποσοστά Recall για κάθε κλάση στον Random Forest 72
Πίνακας 23: Ποσοστά Recall για κάθε κλάση στον Gradient Boosting 73

Κεφάλαιο 1: Εισαγωγή στην Μηχανική Μάθηση

Η Μηχανική Μάθηση - ΜΜ ή (Machine Learning - ML) είναι ένα υποσύνολο του κλάδου της Τεχνητής Νοημοσύνης, όπου τα προγράμματα υπολογιστών ή αλγόριθμοι μαθαίνουν συσχετίσεις προγνωστικής ισχύος από παραδείγματα σε δεδομένα. Η Μηχανική Μάθηση είναι πιο απλά η εφαρμογή στατιστικών μοντέλων σε δεδομένα με τη χρήση υπολογιστών, χρησιμοποιεί ένα ευρύτερο σύνολο στατιστικών τεχνικών από αυτές που χρησιμοποιούνται συνήθως στην ιατρική. Οι νεότερες τεχνικές, όπως η βαθιά μάθηση, βασίζονται σε μοντέλα με λιγότερες υποθέσεις σχετικά με τα υποκείμενα δεδομένα και είναι συνεπώς σε θέση να χειριστούν πιο σύνθετα δεδομένα.

Ένας τυπικός αλγόριθμος μηχανικής μάθησης με επίβλεψη αποτελείται από τρία μέρη [1]:

- **Διαδικασία λήψης αποφάσεων**: Γενικά, οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούνται για να κάνουν μια πρόβλεψη ή ταξινόμηση. Με βάση κάποια δεδομένα εισόδου, τα οποία μπορεί να είναι επισημασμένα ή μη επισημασμένα, ο αλγόριθμος θα παράγει μια εκτίμηση σχετικά με ένα μοτίβο στα δεδομένα.

- **Συνάρτηση σφάλματος**: Μια συνάρτηση σφάλματος αξιολογεί την πρόβλεψη του μοντέλου. Εάν υπάρχουν γνωστά παραδείγματα, μια συνάρτηση σφάλματος μπορεί να κάνει μια σύγκριση για την αξιολόγηση της ακρίβειας του μοντέλου.

- **Διαδικασία βελτιστοποίησης μοντέλου**: Μια μέθοδος κατά την οποία ο αλγόριθμος εξετάζει την αστοχία και στη συνέχεια ενημερώνει τον τρόπο με τον οποίο η διαδικασία λήψης αποφάσεων καταλήγει στην τελική απόφαση, ώστε την επόμενη φορά η να υπάρξει μικρότερη αστοχία.

1.1 Τεχνητή Νοημοσύνη

Η Τεχνητή Νοημοσύνη - ΤΝ ή (Artificial Intelligence - AI) είναι ένα αντικείμενο που ανήκει στην Επιστήμη των Υπολογιστών και στην πειθαρχία της Μηχανικής. Αυτό περιορίζεται κυρίως στην επεξεργασία δεδομένων μέσω υπολογιστών. Οι ορισμοί που χρησιμοποιούνται για τον ορισμό της ΤΝ έχουν επίσης τροποποιηθεί κατά καιρούς σύμφωνα με τις ανάγκες της ικανότητας είτε της χρήσης. Οι ορισμοί αυτοί έχουν περιορίσει το πεδίο της ΤΝ σε πολύ περιορισμένο βαθμό και το πραγματικό δεν λαμβάνεται υπόψη. Τα τελευταία χρόνια υπήρξαν σημαντικές εξελίξεις σε εφαρμογές

της ΤΝ όπως η Ρομποτική, η Μηχανική Όραση, η Μηχανική Μάθηση και ο σχεδιασμός ενεργειών. Ένας γενικός ορισμός θα μπορούσε να είναι ο εξής:

«Τεχνητή νοημοσύνη είναι ο τομέας της Επιστήμης των Υπολογιστών που ασχολείται με τη σχεδίαση και την υλοποίηση προγραμμάτων τα οποία είναι ικανά να μιμηθούν τις ανθρώπινες γνωστικές ικανότητες, εμφανίζοντας έτσι χαρακτηριστικά που αποδίδουμε συνήθως σε ανθρώπινη συμπεριφορά, όπως για παράδειγμα η επίλυση προβλημάτων, η αντίληψη μέσω της όρασης, η μάθηση, η εξαγωγή συμπερασμάτων, η κατανόηση φυσικής γλώσσας, κτλ.» [2]

Εικόνα 1: Απεικόνιση ΤΝ, ΜΜ και ΒΜ [3]

Επιστήμη των Δεδομένων

Η Επιστήμη των Δεδομένων είναι ένας διεπιστημονικός ακαδημαϊκός τομέας που χρησιμοποιεί την στατιστική, την υπολογιστική επιστήμη, τις επιστημονικές μεθόδους, διαδικασίες, τους αλγόριθμους και τα συστήματα για την εξαγωγή της γνώσης και της κατανόησης από θορυβώδη, δομημένα και αδόμητα δεδομένα [4].

Ο στόχος της είναι η βελτίωση της λήψης αποφάσεων με βάση τις γνώσεις που εξάγονται από μεγάλα σύνολα δεδομένων. Ως τομέας δραστηριότητας, η Επιστήμη των Δεδομένων περιλαμβάνει ένα σύνολο από βασικές αρχές, ορισμούς προβλημάτων, αλγορίθμους και διαδικασίες για την εξαγωγή μη προφανών και χρήσιμων μοτίβων από μεγάλα σύνολα δεδομένων. Είναι στενά συνδεδεμένη με τους τομείς της εξόρυξης δεδομένων και της Μηχανικής Μάθησης, αλλά έχει ευρύτερο πεδίο εφαρμογής. Σήμερα, η Επιστήμη των Δεδομένων επηρεάζει τη λήψη αποφάσεων σε όλα σχεδόν τα πεδία των σύγχρονων κοινωνιών και ειδικότερα επηρεάζει την καθημερινή μας ζωή, για παράδειγμα καθορίζει ποιες διαφημίσεις σας παρουσιάζονται στο διαδίκτυο, ποια μηνύματα ηλεκτρονικού ταχυδρομείου φιλτράρονται στον φάκελο ανεπιθύμητων μηνυμάτων σας, ποια άτομα σας συστήνονται στα μέσα κοινωνικής δικτύωσης και πολλά άλλα [5].

Σύνολο Δεδομένων – Data Set

Η συλλογή διαφόρων δεδομένων από εφαρμογές, ιστοσελίδες, αισθητήρες στην βιομηχανία στοιβαγμένα σε γραμμές και στήλες όπως έναν πίνακα ή μια Βάση Δεδομένων μπορεί να ονομαστεί ως σύνολο δεδομένων (data-set). Κάθε στήλη ενός πίνακα αντιπροσωπεύει μια συγκεκριμένη μεταβλητή και κάθε γραμμή αντιστοιχεί σε μια δεδομένη εγγραφή του εν λόγω συνόλου δεδομένων. Το σύνολο δεδομένων παραθέτει τιμές για κάθε μία από τις μεταβλητές οι οποίες είναι συνεχείς, κατηγορηματικές ή δυαδικές. Τα σύνολα δεδομένων μπορούν επίσης να αποτελούνται από μια συλλογή εγγράφων ή αρχείων [6].

	Πιθανότητα Βροχής %	**Άνεμος km/h**	**Υγρασία %**	**Αισθητή Θερμοκρασία ℃**	**Θερμοκρασία ℃**
8:00	40	4.3	80	14	16
9:00	30	3.6	65	15	16
10:00	10	3.5	50	18	18
11:00	10	2.7	55	19	20
12:00	15	2.4	63	20	21

Πίνακας 1: Παράδειγμα συνόλου δεδομένων καιρού ανά ώρα

Στατιστική

Για να μπορέσει κανείς να κατανοήσει, να αναλύσει και να επεξεργαστεί καλύτερα ένα σύνολο δεδομένων πολύ σημαντική προϋπόθεση είναι να γνωρίζει βασικές αρχές Στατιστικής.

Η Στατιστική είναι ένας κλάδος των εφαρμοσμένων μαθηματικών που προσφέρει δύο σπουδαίες δυνατότητες αφενός την περιγραφή αριθμητικών συνόλων δεδομένων έρευνας και στη συνέχεια την ανάλυση αυτών. Συνέπεια αυτών των δυνατοτήτων είναι και η βασική διάκρισή της σε περιγραφική και αναλυτική στατιστική [7].

- Στην Περιγραφική στατιστική περιγράφονται τα διάφορα στατιστικά στοιχεία μετά από συλλογή και ταξινόμηση κατά ομάδες των στατιστικών δεδομένων τα οποία ακολούθως παρουσιάζονται υπό μορφή ανάλυσης σε πίνακες, διαγράμματα με χαρακτηριστικές τιμές, ή ιδιότητες.
- Στην Αναλυτική στατιστική, που είναι περισσότερο περίπλοκη, αναζητείται με διάφορες μεθόδους ο προσδιορισμός βαθμού εμπιστοσύνης στην εξαγωγή ασφαλών συμπερασμάτων μέσα όμως από κάποιο περιορισμένο δείγμα στοιχείων ενός γενικότερου συνόλου.

1.2 Κατηγορίες Μηχανικής Μάθησης

Τα μοντέλα μηχανικής μάθησης χωρίζονται σε βασικές κατηγορίες:

- **Μάθηση με Επίβλεψη** (Supervised Learning), γνωστή και ως μάθηση με παραδείγματα, ορίζεται από τη χρήση συνόλων δεδομένων για την εκπαίδευση αλγορίθμων για την ταξινόμηση δεδομένων ή την ακριβή πρόβλεψη αποτελεσμάτων. Καθώς τα δεδομένα εισόδου τροφοδοτούνται στο μοντέλο, το μοντέλο προσαρμόζει τα βάρη του έως ότου προσαρμοστεί κατάλληλα. Αυτό συμβαίνει στην διαδικασία διασταυρούμενης επικύρωσης (cross validation) για να διασφαλιστεί ότι το μοντέλο αποφεύγει την υπερμοντελοποίηση (overfitting) ή την υπομοντελοποίηση (underfitting).

- **Μάθηση χωρίς Επίβλεψη** (Unsupervised Learning), αλλιώς μάθηση από παρατήρηση, χρησιμοποιεί αλγορίθμους μηχανικής μάθησης για την ανάλυση και ομαδοποίηση μη επισημασμένων συνόλων δεδομένων, δηλαδή που δεν προσδιορίζουν χαρακτηριστικά, ιδιότητες ή ταξινομήσεις. Αυτοί οι αλγόριθμοι ανακαλύπτουν κρυμμένα μοτίβα ή ομαδοποιήσεις δεδομένων χωρίς την ανάγκη να παρέμβει ο άνθρωπος. Η ικανότητα αυτής της μεθόδου να ανακαλύπτει ομοιότητες και διαφορές στις πληροφορίες την καθιστούν ιδανική για διερευνητική ανάλυση δεδομένων, κατάτμηση πελατών και αναγνώριση εικόνων και μοτίβων.

- **Μάθηση με Ενίσχυση** (Reinforcement Machine Learning), είναι ένα μοντέλο μηχανικής μάθησης που είναι παρόμοιο με την επιβλεπόμενη μάθηση, αλλά ο αλγόριθμος δεν εκπαιδεύεται με τη χρήση δεδομένων. Αυτό το μοντέλο μαθαίνει στην πορεία χρησιμοποιώντας τη δοκιμή και το σφάλμα. Μια ακολουθία επιτυχημένων αποτελεσμάτων θα ενισχυθεί για να αναπτυχθεί η καλύτερη σύσταση για ένα δεδομένο πρόβλημα.

Εικόνα 2: Απεικόνιση κατηγοριών ΜΜ

1.3 Αλγόριθμοι Μηχανικής Μάθησης

Για να μπορεί κανείς να κατανοήσει καλύτερα τους αλγόριθμους της Μηχανικής Μάθησης πρέπει πρώτα να κατανοήσει την γενική έννοια του αλγορίθμου.

Στα μαθηματικά και την Επιστήμη των Υπολογιστών, ένας αλγόριθμος είναι μια πεπερασμένη ακολουθία αυστηρών οδηγιών, που συνήθως χρησιμοποιούνται για την επίλυση μιας κατηγορίας συγκεκριμένων προβλημάτων ή για την εκτέλεση ενός υπολογισμού. Οι πιο προηγμένοι αλγόριθμοι μπορούν να εκτελούν συμπεράσματα αυτόματα (αυτοματοποιημένη συλλογιστική) και να χρησιμοποιούν μαθηματικές και λογικές δοκιμές για να εκτρέπουν την εκτέλεση του κώδικα μέσω διαφόρων διαδρομών (αυτοματοποιημένη λήψη αποφάσεων) [8].

Σκοπός της Μηχανικής μάθησης είναι η χρήση αλγορίθμων ΜΜ για την ανάλυση δεδομένων. Αξιοποιώντας τη ΜΜ, ένας προγραμματιστής μπορεί να βελτιώσει την αποτελεσματικότητα μιας εργασίας που περιλαμβάνει ένα μεγάλο όγκο δεδομένων χωρίς να χρειαστεί η παρέμβασή του. Σε όλο τον κόσμο, οι ισχυροί αλγόριθμοι Μηχανικής Μάθησης μπορούν να χρησιμοποιηθούν για τη βελτίωση της παραγωγικότητας των επαγγελματιών που εργάζονται στην επιστήμη των δεδομένων, στην Επιστήμη των Υπολογιστών και σε πολλούς άλλους τομείς όπως στον τομέα της υγείας.

Υπάρχει ένας αριθμός αλγορίθμων μηχανικής μάθησης που χρησιμοποιούνται συνήθως από τις σύγχρονες εταιρείες τεχνολογίας. Καθένας από αυτούς τους αλγορίθμους μηχανικής μάθησης μπορεί να έχει πολυάριθμες εφαρμογές σε διάφορα εκπαιδευτικά και επιχειρηματικά περιβάλλοντα.

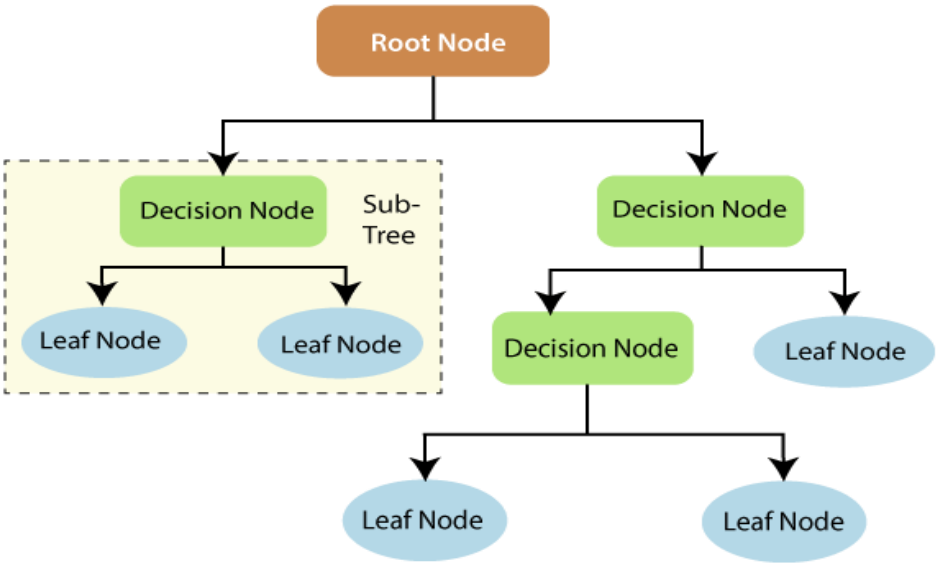
Αλγόριθμοι Μηχανικής Μάθησης για Ταξινόμηση

Οι αλγόριθμοι Μηχανικής Μάθησης με Επίβλεψη αποτελούνται από δύο κατηγορίες, την ταξινόμηση ή κατηγοριοποίηση (classification) και την παλινδρόμηση (regression). Η παρούσα διπλωματική εργασία επικεντρώνεται σε προβλήματα ταξινόμησης.

Στην ταξινόμηση, ένα πρόγραμμα υπολογιστή εκπαιδεύεται σε ένα σύνολο δεδομένων εκπαίδευσης και με βάση την εκπαίδευση κατηγοριοποιεί τα δεδομένα σε διαφορετικές κατηγορίες, με στόχο βέβαια την υψηλότερη ακρίβεια. Αυτός ο αλγόριθμος χρησιμοποιείται για την πρόβλεψη των διακριτών τιμών που στο λεξικό της ΜΜ ονομάζονται κλάσεις.

1.3.1 Decision Tree – Δέντρο απόφασης

Ένα δέντρο απόφασης μοιάζει με δέντρο. Η βάση του δέντρου ονομάζεται ριζικός κόμβος (Root Node). Από τον ριζικό κόμβο απορρέει μια σειρά από κόμβους απόφασης (Decision Node) που απεικονίζουν τις αποφάσεις που πρέπει να ληφθούν. Στον επόμενο κλάδο, από τους κόμβους απόφασης ξεκινούν οι κόμβοι φύλλων (Leaf Nodes) που αναπαριστούν τα αποτελέσματα αυτών των αποφάσεων. Κάθε κόμβος απόφασης αντιπροσωπεύει μια ερώτηση ή ένα σημείο διαχωρισμού και οι κόμβοι φύλλων που απορρέουν από έναν κόμβο απόφασης αντιπροσωπεύουν τις πιθανές απαντήσεις. Οι κόμβοι φύλλων φυτρώνουν από τους κόμβους απόφασης με παρόμοιο τρόπο όπως φυτρώνει ένα φύλλο σε ένα κλαδί δέντρου. Αυτός είναι ο λόγος για τον οποίο αποκαλούμε κάθε υποτμήμα ενός δέντρου αποφάσεων "κλάδο". Όπως φαίνεται στην παρακάτω εικόνα [9]:



Εικόνα 3: Απεικόνιση μιας δομής Δέντρου Απόφασης [10]

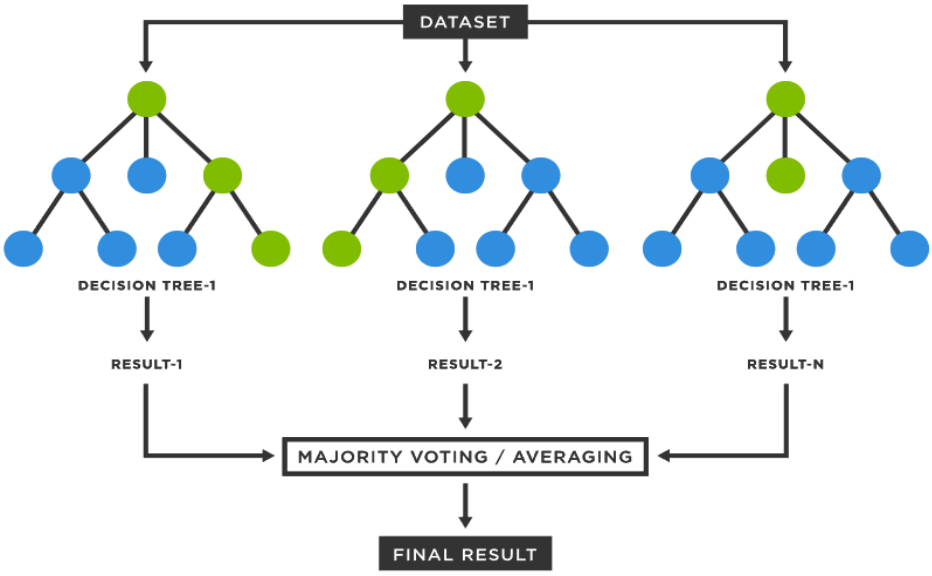
1.3.2 Random Forest – Random 'Decision' Forest

Έχοντας έναν συνδυασμό πολλών δέντρων απόφασης μπορεί να κατασκευαστεί ένα Random Forest ή ακόμη καλύτερα ένα Τυχαίο Δάσος και αποκαλείται "Δάσος" επειδή παράγεται ένα δάσος από δέντρα αποφάσεων.

Ο όρος Random Decision Forest διατυπώθηκε για πρώτη φορά από την επιστήμονα Πληροφορικής, Tin Kam Ho το 1995 [11], η οποία ανέπτυξε έναν τύπο για τη χρήση τυχαίων δεδομένων για τη δημιουργία προβλέψεων, αναφέροντας:

«Η ουσία της μεθόδου είναι η δημιουργία πολλαπλών δέντρων σε τυχαία επιλεγμένα υποδιαστήματα του πεδίου χαρακτηριστικών. Τα δέντρα σε διαφορετικούς υποχώρους γενικεύουν την ταξινόμησή τους με συμπληρωματικούς τρόπους και η συνδυασμένη ταξινόμησή τους μπορεί να βελτιωθεί μονοτονικά.»

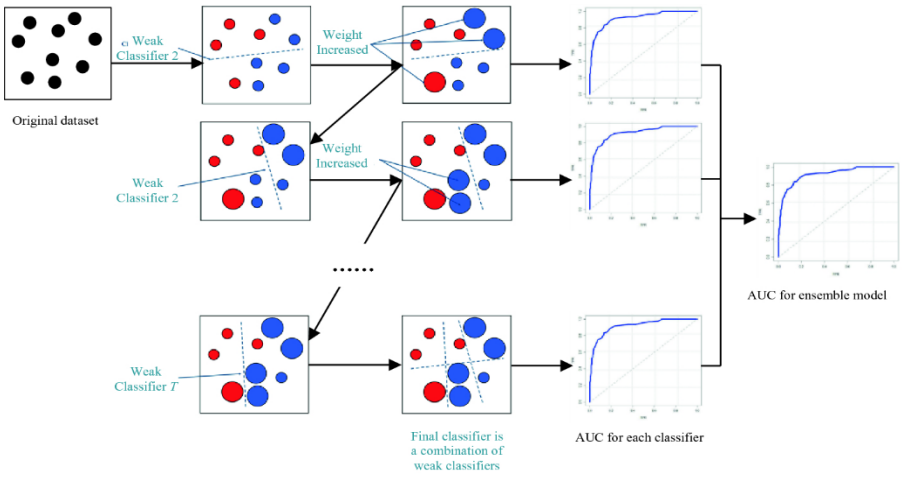
Με απλά λόγια, τα δεδομένα από αυτά τα δέντρα συνδυάζονται μεταξύ τους για να εξασφαλίσουν τις πιο ακριβείς προβλέψεις. Ενώ ένα μεμονωμένο δέντρο αποφάσεων έχει ένα αποτέλεσμα και ένα στενό φάσμα ομάδων, το δάσος εξασφαλίζει ένα πιο ακριβές αποτέλεσμα με μεγαλύτερο αριθμό ομάδων και αποφάσεων. Το πρόσθετο πλεονέκτημα είναι ότι προσθέτει τυχαιότητα στο μοντέλο, βρίσκοντας το καλύτερο χαρακτηριστικό ανάμεσα σε ένα τυχαίο υποσύνολο χαρακτηριστικών.



Εικόνα 4: Απεικόνιση δομής ενός Random Forest [12]

1.3.3 Gradient Boosting

Το Gradient Boosting αποτελείται από τρία κύρια στοιχεία, πρώτον, τη συνάρτηση απώλειας (Loss function), η οποία έχει ως ρόλο να εκτιμά πόσο καλό είναι το μοντέλο στο να κάνει προβλέψεις με τα δεδομένα που του δίνονται. Δεύτερον, ένας αδύναμος μαθητής (weak learner), που είναι συνήθως δέντρα απόφασης, είναι ένας μαθητής που ταξινομεί τα δεδομένα αλλά το κάνει κακώς και ίσως όχι καλύτερα από την τυχαία πρόβλεψη. Και τρίτον, το προσθετικό μοντέλο (Additive model) είναι η επαναληπτική και διαδοχική προσέγγιση της προσθήκης των δέντρων ένα βήμα κάθε φορά, έτσι ώστε κάθε επανάληψη να μειώνει την τιμή της συνάρτησης απωλειών. Ουσιαστικά είναι πολλά δέντρα απόφασης το ένα κάτω από το άλλο, προσπαθώντας το κάθε δέντρο την φορά να ελαχιστοποιεί το εκάστοτε σφάλμα μέχρις ότου η τιμή του σφάλματος να τείνει στο μηδέν.



Εικόνα 5: Απεικόνιση λειτουργίας Gradient Boosting [13]

1.4 Εφαρμογή της Μηχανικής Μάθησης ανά βήμα

Καθημερινά ακούγεται πως ο υπολογιστής με την βοήθεια της Μηχανικής Μάθησης μπορεί να μαθαίνει μόνος του, αφενός αυτό εν μέρει είναι σωστό, αφετέρου χωρίς την παρέμβαση του ανθρώπου αυτό δεν μπορεί να επιτευχθεί . Η ΜΜ εφαρμόζεται για την επίλυση ενός πραγματικού προβλήματος.

Εικόνα 6: Η Μηχανική Μάθηση σαν σύστημα

Για να μπορέσει να γίνει καλύτερα αντιληπτό θα πρέπει να ακολουθηθεί μια συγκεκριμένη διαδικασία για την επίλυση ενός οποιουδήποτε προβλήματος η οποία αποτελείται από διαφορετικά βήματα. Ως είσοδος είναι η κατανόηση του προβλήματος που πρέπει να επιλυθεί, ποια είναι τα ζητήματα και τι είδος προβλήματος είναι π.χ. πρόβλεψη, ταξινόμηση κλπ.

Βήμα 1: Η συλλογή δεδομένων η οποία μπορεί να επιτευχθεί με πολλούς τρόπους αλλά πάντα εξαρτάται από το πρόβλημα της εφαρμογής που πρέπει να επιλυθεί. Για παράδειγμα, στην πρόβλεψη ασθενών για σακχαρώδη διαβήτη, γιατροί πρέπει να συλλέγουν απαραίτητες πληροφορίες από τις εξετάσεις για κάθε νέο ασθενή και έπειτα από πολλούς ασθενείς να δίνονται σε επιστήμονες δεδομένων (data scientists) για να μπορέσουν να προβλέψουν αν κάποιος καινούργιος ασθενείς πάσχει από την νόσο αυτή στο μέλλον.

Βήμα 2: Αφού συλλεχθούν τα δεδομένα, επόμενο βήμα είναι η προετοιμασία των δεδομένων, αυτό θα μπορούσε να είναι καθαρισμός δεδομένων από ασήμαντες πληροφορίες, έλεγχος ελλείπων τιμών, αν υπάρχει πρόβλημα πρόβλεψης τότε θα πρέπει να μετατραπούν τα κατηγορηματικά σε αριθμητικά και πολλά άλλα ανάλογα με το είδος του προβλήματος. Σε κάθε

πρόβλημα πρέπει να γίνεται διαχωρισμός δεδομένων σε δεδομένα εκπαίδευσης και δεδομένα δοκιμής. Να τονιστεί ότι αυτό το βήμα είναι πάρα πολύ σημαντικό διότι όσο καλύτερα γίνει η προετοιμασία και επεξεργασία των δεδομένων τόσο μεγαλύτερο θα είναι το ποσοστό επιτυχίας για την επίλυση του προβλήματος.

Βήμα 3: Στη συνέχεια αφού τα δεδομένα είναι επεξεργασμένα πρέπει να γίνει η επιλογή του κατάλληλου μοντέλου.

Βήμα 4: Αφού επιλεχθεί ο κατάλληλος αλγόριθμος, επόμενο βήμα είναι η εκπαίδευση των δεδομένων εκπαίδευσης.

Βήμα 5: Έπειτα, με τον κατάλληλο αλγόριθμο δημιουργήθηκε το μοντέλο που εκπαιδεύτηκε με τα δεδομένα εκπαίδευσης. Το μοντέλο αυτό εφαρμόζεται στα δεδομένα δοκιμής τα οποία δεν πρέπει να έχουν την οποιαδήποτε επαφή με το μοντέλο, ώστε να μπορέσει να γίνει η αξιολόγηση αν το μοντέλο είναι ικανό να επιλύσει το εκάστοτε πρόβλημα.

Βήμα 6: Το βήμα αυτό αφορά την ρύθμιση παραμέτρων του μοντέλου για την βελτιστοποίηση αυτού από το προηγούμενο βήμα σε περίπτωση που το μοντέλο δεν ανταποκρίθηκε σωστά στις αρχικές παραμέτρους που ορίστηκαν.

Βήμα 7: Τέλος, το μοντέλο μπορεί να χρησιμοποιηθεί σε άλλα καινούργια δεδομένα για την πρόβλεψη και επίλυση του προβλήματος.

Εικόνα 7: Επτά βήματα στην Μηχανική Μάθηση [14]

1.5 Έλεγχος επίδοσης Αλγορίθμου – Metrics

Ανάλογα με το πρόβλημα που πρέπει να επιλυθεί χρησιμοποιούνται και διαφορετικές Μετρήσεις (Metrics) [15] για τον έλεγχο επίδοσης ενός αλγορίθμου. Στην περίπτωση της ταξινόμησης 2 ή περισσοτέρων κλάσεων, όπως εξετάζεται στα δύο τελευταία κεφάλαια, όπου η κλάση έχει την δυαδική τιμή '0' ή '1' αλλιώς True ή False για προβλήματα δύο κλάσεων, η οποία υποδηλώνει ανάλογα με τα χαρακτηριστικά, εάν ο στόχος είναι σωστός ή όχι.

1.5.1 Accuracy

Η πρώτη και βασικότερη από τις μετρικές είναι η ακρίβεια (Accuracy) λαμβάνει τιμές $0 \leq Accuracy \leq 1$, για $Accuracy = 1$ πρέπει να ισχύει $(FP = FN = 0)$, η οποία είναι ο αριθμός των δειγμάτων που ταξινομήθηκαν σωστά από όλα τα δείγματα που υπάρχουν στο σύνολο δεδομένων δοκιμής, όπως φαίνεται παρακάτω:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{Σύνολο\ σωστών\ προβλέψεων}{Σύνολο\ όλων\ των\ προβλέψεων}$$

TP True Positive ή **Πραγματικά Θετικό** αφορά το δείγμα που αντιστοιχεί στη θετική κλάση που ταξινομήθηκε σωστά.

TN True Negative ή **Πραγματικά Αρνητικό** αναφέρεται σε ένα δείγμα που ανήκει στη αρνητική κλάση που ταξινομείται σωστά.

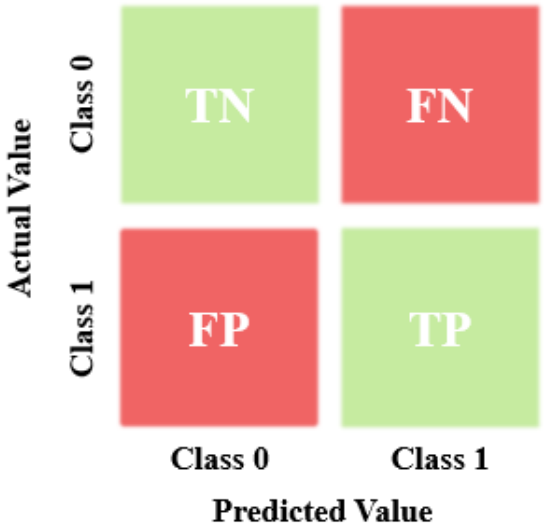
FP False Positive ή **Εσφαλμένα Θετικό** πρόκειται για ένα δείγμα που ανήκει στην αρνητική κλάση αλλά ταξινομείται λανθασμένα ότι ανήκει στη θετική κλάση.

FN False Negative ή **Εσφαλμένα Αρνητικό** αναφέρεται σε ένα δείγμα που ανήκει στη θετική κλάση αλλά ταξινομείται λανθασμένα ως ένα δείγμα που ανήκει στην αρνητική κλάση.

1.5.2 Confusion Matrix

Ένας πιο περιεκτικός τρόπος για την αξιολόγηση των επιδόσεων του μοντέλου και αναπαράστασης των TP, TN, FP, FN είναι ο Πίνακας Σύγχυσης (Confusion Matrix). Ο πίνακας σύγχυσης είναι ένας πιο περιεκτικός τρόπος για την αξιολόγηση των επιδόσεων του μοντέλου. Ένας πίνακας σύγχυσης, όπως υποδηλώνει το όνομα, είναι ένας πίνακας

αριθμών που υποδεικνύει σε ποιο σημείο ένα μοντέλο μπερδεύεται. Πρόκειται για μια κατανομή κατά κλάση της προγνωστικής απόδοσης ενός μοντέλου ταξινόμησης, με άλλα λόγια, ο πίνακας σύγχυσης είναι ένας οργανωμένος τρόπος αντιστοίχισης των προβλέψεων στις αρχικές κλάσεις στις οποίες ανήκουν τα δεδομένα.



Εικόνα 8: Πίνακας σύγχυσης

Σε περίπτωση που το πρόβλημα απαιτεί ταξινόμηση πολλαπλών κατηγοριών ή κλάσεων, οι διαστάσεις του πίνακα αυξάνονται, σε πρόβλημα ταξινόμησης δύο κλάσεων ο πίνακας σύγχυσης θα είναι $(2 * 2)$, ενώ σε πρόβλημα ταξινόμησης N κλάσεων τότε οι διαστάσεις του πίνακα σύγχυσης θα είναι $(N * N)$.

1.5.3 Precision, Recall και F-measure

Precision είναι ο αριθμός των δειγμάτων που πράγματι ανήκουν στη θετική κλάση από όλα τα δείγματα που προέβλεψε το μοντέλο ότι ανήκουν σε αυτήν.

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

Recall – Ανάκληση αποτελεί ένα μέγεθος του πόσα από τα θετικά δείγματα προέβλεψε σωστά ο ταξινομητής, σε σχέση με το σύνολο των θετικών δειγμάτων στα δεδομένα.

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

F-measure ή **F1-score** του συστήματος ορίζεται ο αρμονικός μέσος όρος της ακρίβειας και της ανάκλησής του [16].

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

1.5.4 ROC curve

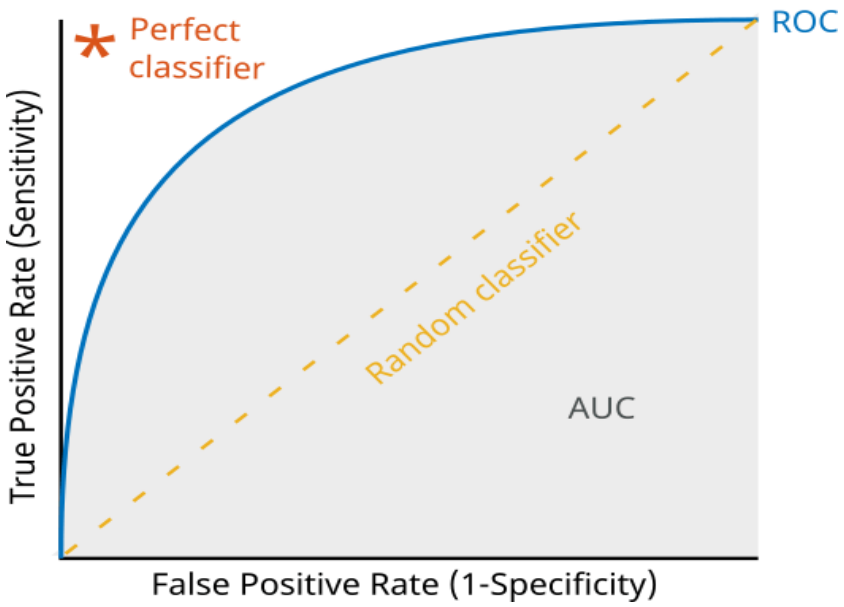
Το ROC (Receiver Operating Characteristic) ή χαρακτηριστική καμπύλη λειτουργίας είναι ένα γράφημα του TPR (True Positive Rate) σε συνάρτηση με το FPR (False Positive Rare) σε διαφορετικές ρυθμίσεις κατωφλίου (threshold) [17].

Το TPR που αποκαλείται και ως ευαισθησία (Sensitivity) είναι η ίδια με την ανάκληση, άρα $Sensitivity = Recall = TPR$, ενώ το $FPR = 1 - Specificity$.

Η ειδικότητα (Specificity) δείχνει σε τι ποσοστό το μοντέλο ταξινόμησε τα χαρακτηριστικά της κλάσης 0, είναι γνωστό ως TNR (True Negative Rate), εστιάζοντας στην κλάση 0 και ισχύει:

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

Στην παρακάτω εικόνα απεικονίζεται η καμπύλη TPR συναρτήσει του FPR και πως πρέπει να είναι η καμπύλη για το καλύτερο αποτέλεσμα.



Εικόνα 9: Παράδειγμα ενός ROC Curve [18]

Κεφάλαιο 2: Εφαρμογές Μηχανικής Μάθησης στην Ιατρική

Η μηχανική μάθηση είναι ένα εργαλείο που χρησιμοποιείται στον τομέα της Ιατρικής για να βοηθήσει τους επαγγελματίες υγείας ιατρούς, μηχανικούς βιοϊατρικής κλπ. ώστε να παρέχουν φροντίδα στους ασθενείς και να διαχειρίζονται τα ιατρικά δεδομένα. Μπορεί να εφαρμοστεί για τη συλλογή και διαχείριση δεδομένων ασθενών, τον εντοπισμό εξελίξεων στην υγειονομική περίθαλψη, τη σύσταση θεραπειών και πολλά άλλα. Τα νοσοκομεία και οι επιχειρήσεις στον τομέα της υγειονομικής περίθαλψης έχουν αρχίσει να αντιλαμβάνονται την ικανότητα της Μηχανικής Μάθησης να συμβάλλει στη βελτίωση της λήψης αποφάσεων και στη μείωση του κινδύνου στον χώρο αυτό.

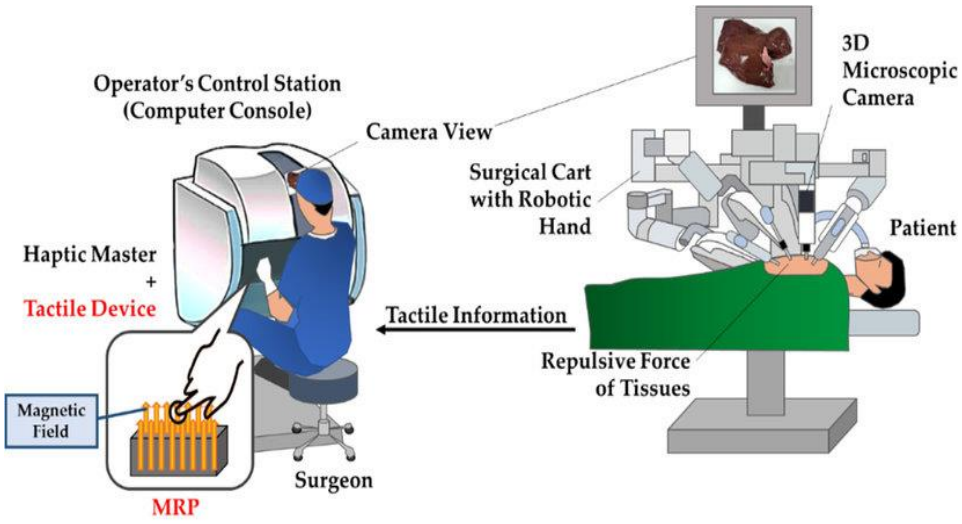
2.1 Ανίχνευση καρδιακών παθήσεων

Η Μηχανική Μάθηση χρησιμοποιείται για την εύρεση μοτίβων από πηγές ιατρικών δεδομένων και παρέχει εξαιρετικές δυνατότητες διάγνωσης και ανίχνευσης ασθενειών. Οι δυνατότητες αυτές είναι ιδιαίτερα κατάλληλες για την ανάπτυξη κλινικών εφαρμογών, ειδικότερα εκείνων που βασίζονται σε προηγμένες μετρήσεις γονιδιωματικής και πρωτεωμικής. Η γονιδιωματική είναι ένας τομέας της βιολογίας που αφορά την δομή, τη λειτουργία, την εξέλιξη, τη χαρτογράφηση και την επεξεργασία των γονιδιωμάτων [19], ενώ η πρωτεωμική ανήκει στον κλάδο της Γενετικής και της Μοριακής Βιολογίας και αφορά την μελέτη της ιδιότητας των πρωτεϊνών, το επίπεδο έκφρασης, μετα-μεταφραστικές τροποποιήσεις, αλληλεπιδράσεις κ.λπ., σε ευρεία κλίμακα για την απόκτηση μιας σφαιρικής, ολοκληρωμένης αντίληψης των διαδικασιών της νόσου, τις κυτταρικές λειτουργίες και τα δίκτυα σε πρωτεϊνικό επίπεδο [20]. Στις ιατρικές εφαρμογές, οι αλγόριθμοι μηχανικής μάθησης θα λαμβάνουν ορθότερες αποφάσεις σχετικά με τα σχέδια θεραπείας των ασθενών [21].

Στην έρευνα που διεξήχθη από τους G. Parthiban και S.K.Srivatsa [22], οι συγγραφείς αναφέρουν ότι οι καρδιακές παθήσεις είναι μία από τις σημαντικότερες αιτίες θανάτου στον κόσμο τα τελευταία κι ο διαβήτης είναι μια χρόνια ασθένεια που προκαλεί σοβαρές επιπλοκές όπως η καρδιακή πάθηση, η νεφρική ανεπάρκεια και η τύφλωση. Το σύνολο δεδομένων που χρησιμοποιήθηκε αφορά διαβητικούς ασθενείς, αρχικά χρησιμοποιήθηκε η εξόρυξη δεδομένων (Data mining) και τέλος χρησιμοποιήθηκαν οι αλγόριθμοι Naïve Bayes και Support Vector Machines (SVM) για την πρόβλεψη των πιθανότερων καρδιακών παθήσεων παρουσιάζοντας πολύ καλή ακρίβεια, με τον αλγόριθμο SVM να αποδίδει 94.60% επιτυχία έναντι του Naïve Bayes που απέδωσε χειρότερα φτάνοντας το 74% της ακρίβειας.

2.2 Αξιολόγηση δεξιοτήτων στη χειρουργική με την ρομποτική υποβοήθηση

Η ρομποτική χειρουργική, που αναφέρεται επίσης ως ρομποτικά υποβοηθούμενη χειρουργική (Robot-Assisted Surgery), επιτρέπει στους γιατρούς να εκτελούν πολλές πολύπλοκες επεμβάσεις με μεγαλύτερη ακρίβεια, ευελιξία και έλεγχο από ό,τι είναι δυνατό με τις συμβατικές τεχνικές. Η ρομποτική χειρουργική συνδέεται συνήθως με την ελάχιστα επεμβατική χειρουργική (Minimally Invasive Surgery) και αναφέρεται και ως Robotic-Assisted Minimally Invasive Surgery (RMIS), διαδικασίες που εκτελούνται μέσω μικροσκοπικών τομών. Το πιο ευρέως χρησιμοποιούμενο κλινικό ρομποτικό χειρουργικό σύστημα περιλαμβάνει ένα βραχίονα κάμερας και μηχανικούς βραχίονες με προσαρτημένα χειρουργικά εργαλεία. Ο χειρούργος ελέγχει τους βραχίονες ενώ κάθεται σε μια κονσόλα υπολογιστή κοντά στο τραπέζι του χειρουργείου. Η κονσόλα παρέχει στον χειρούργο μια υψηλής ευκρίνειας, μεγεθυμένη και τρισδιάστατη εικόνα της χειρουργικής περιοχής, όπως φαίνεται παρακάτω:



Εικόνα 10: Σχηματική απεικόνιση του RMIS με τη συσκευή αφής MRP [23]

Μια κατεξοχήν υποκειμενική υπόθεση είναι η αξιολόγηση των δεξιοτήτων ενός χειρουργού. Η ανάπτυξη αντικειμενικών μεθόδων για την αξιολόγηση των χειρουργικών δεξιοτήτων παρουσιάζει αυξημένο ενδιαφέρον. Μια μελέτη [24], των Mahtab J. Fard, Sattar Ameri, Ratna B. Chinnam, Abhilash K. Pandya, Michael D. Klein και R. Darin Ellis, παρουσίασε την ικανότητα της μηχανικής μάθησης να ταξινομεί αυτόματα αρχάριους και έμπειρους χειρουργούς χρησιμοποιώντας χαρακτηριστικά κίνησης για διάφορες λειτουργίες RMIS. Χρησιμοποιήθηκαν οι αλγόριθμοι SVM και Λογιστική Παλινδρόμηση (Logistic regression) και ως μέθοδο την διασταυρωμένης επιτήρησης

(Cross Validation), το Leave-one-super-trial-out (LOSO) και το Leave-one-user-out (LOUO).

Η **LOSO** είναι μια ειδική μέθοδος διασταυρούμενης επικύρωσης στην οποία παρέχονται δείκτες εκπαίδευσης/δοκιμής για τον διαχωρισμό των δεδομένων σε σύνολα εκπαίδευσης/δοκιμής. Κάθε δείγμα χρησιμοποιείται μία φορά ως σύνολο δοκιμής, ενώ τα υπόλοιπα δείγματα αποτελούν το σύνολο εκπαίδευσης [25], ενώ στο **LOUO**, όλες οι δοκιμές που εκτελούνται από ένα μόνο άτομο παραλείπονται ως σύνολο δοκιμής και οι υπόλοιπες δοκιμές χρησιμοποιούνται για την εκπαίδευση του μοντέλου [26]. Τέλος, η καλύτερη συνολική ακρίβεια που επιτεύχθηκε, ήταν για την εκτέλεση διαδικασίας των ραμμάτων, με 85,7% χρησιμοποιώντας την μέθοδο LOSO και 71,9% χρησιμοποιώντας το LOUO.

2.3 Πρόβλεψη των εισαγωγών στο τμήμα επειγόντων περιστατικών για τη βελτίωση της ροής των ασθενών

Το τμήμα επειγόντων περιστατικών, γνωστό και ως τμήμα ατυχημάτων και επειγόντων περιστατικών, αίθουσα επειγόντων περιστατικών, θάλαμος επειγόντων περιστατικών ή τμήμα ατυχημάτων, είναι μια υγειονομική μονάδα που επικεντρώνεται στην αντιμετώπιση επειγόντων περιστατικών, την οξεία περίθαλψη ασθενών που προσέρχονται χωρίς προηγούμενο ραντεβού, είτε με δικά τους μέσα είτε μέσω ασθενοφόρου. Το τμήμα επειγόντων περιστατικών βρίσκεται συνήθως σε νοσοκομείο ή άλλο κέντρο πρωτοβάθμιας ιατρικής περίθαλψης [27]. Ο συνωστισμός στο τμήμα επειγόντων περιστατικών αποτελεί ένα ευρέως σημαντικό ζήτημα που μπορεί να επηρεάσει την ποιότητα και την πρόσβαση στη υγειονομική περίθαλψη.

Διεξήχθη μια έρευνα [28], που αφορά την αξιολόγηση τριών μοντέλων με σκοπό τη χρήση πληροφοριών που συλλέγονται κατά τη διαλογή για την πρόβλεψη, σε πραγματικό χρόνο, του όγκου των ασθενών του τμήματος επειγόντων περιστατικών που στη συνέχεια θα εισαχθούν σε κλινική μονάδα νοσηλείας και η ανάπτυξη μιας νέας μεθοδολογίας για την εφαρμογή αυτών των προβλέψεων στο νοσοκομείο. Χρησιμοποιήθηκαν οι 3 εξής μέθοδοι, πρώτον, η γνώμη ειδικών πάνω στον τομέα, υπό όρους πιθανότητα Naive Bayes και ένα μοντέλο γενικευμένης γραμμικής παλινδρόμησης (Linear Regression) με χρήση λογαριθμο-γραμμικής παλινδρόμησης (logit-linear). Οι συγγραφείς συμπέραναν ότι, το μοντέλο logit regression είχε την καλύτερη επίδοση, με AUC curve $= 88{,}7\%$, με συντελεστή προσδιορισμού (Coefficient of determination) $R^2 = 58\%$ και ημερήσιο μέσο σφάλμα εκτίμησης 0,19 για το συγκεντρωτικό μοντέλο όσον αφορά τα κρεβάτια των νοσοκομειακών μονάδων. Η συγκεκριμένη μέθοδος στηρίχθηκε σε τέσσερις εισόδους, την ηλικία του ασθενούς, το κύριο παράπονο,

προσδιορισμός τύπου κρεβατιού και τον τρόπο προσέλευσης. Τέλος, αναφέρουν ότι θα μπορούσε να βελτιωθεί το συγκεκριμένο πρόβλημα μελλοντικά.

2.4 Ανίχνευση κατάθλιψης από τα μέσα κοινωνικής δικτύωσης

Τα μέσα κοινωνικής δικτύωσης, που χρησιμοποιούνται καθημερινά και σχεδόν από όλους, αποτελούν μια διαδραστική τεχνολογία που διευκολύνει τη δημιουργία και την ανταλλαγή πληροφοριών, ιδεών, ενδιαφερόντων και άλλων μορφών έκφρασης μέσω διαδικτυακών ομάδων και κοινοτήτων [29]. Η κατάθλιψη είναι μια συχνή αλλά σοβαρή διαταραχή της ψυχικής κατάστασης που προκαλεί σοβαρά συμπτώματα που επηρεάζουν τον τρόπο με τον οποίο κάποιος αισθάνεται, σκέφτεται και αντιμετωπίζει καθημερινές δραστηριότητες, όπως ο ύπνος, το φαγητό, η δουλειά [30]. Μεταξύ άλλων συσχετιζόμενων προβλημάτων, η κατάθλιψη μπορεί να οδηγήσει μέχρι την αυτοκτονία, και έτσι ο μεγάλος αριθμός των ατόμων με κατάθλιψη θεωρείται σοβαρό πρόβλημα. Για αυτό, μια μελέτη διεξήχθη [31], η οποία αφορά την αξιολόγηση αποτελεσματικότητας της χρήσης των δραστηριοτήτων ενός χρήστη στα μέσα κοινωνικής δικτύωσης για την εκτίμηση του μεγέθους της κατάθλιψης με την χρήση της Μηχανικής Μάθησης. Χρησιμοποιώντας ένα ερωτηματολόγιο το οποίο αναρτήθηκε στο διαδίκτυο κατάφεραν να συλλέξουν πληροφορίες σχετικά με τα επίπεδα κατάθλιψης των χρηστών από μια πλατφόρμα κοινωνικής δικτύωσης ακόμη και το ιστορικό των δραστηριοτήτων τους. Οι συγγραφείς σύμφωνα με το αποτέλεσμα της ακρίβειας 69% για την ανίχνευση της κατάθλιψης, έβγαλαν το εξής συμπέρασμα, ότι η χρήση της δραστηριότητας των μέσων κοινωνικής δικτύωσης είναι μια έγκυρη προσέγγιση για την αναγνώριση της κατάθλιψης διότι τα αποτελέσματα προέκυψαν μεταξύ ομάδων χρηστών που μιλούν διαφορετικές γλώσσες [32].

2.5 Πρόβλεψη σοβαρού/κρίσιμου συμπτώματος μολυσμένων ασθενών με κορονοϊό

Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (WHO), αυτή τη στιγμή μία από τις πιο ισχυρότερες μεταδοτικές πανδημίες στην ιστορία είναι αυτή του κορονοϊού (COVID-19) [33], καθώς αρκετοί άνθρωποι αρρωσταίνουν σοβαρά, επομένως υπάρχει ανάγκη για ιατρική φροντίδα. Ωστόσο, οι ηλικιωμένοι άνθρωποι είναι πιο ευάλωτοι και κυρίως όσοι έχουν υποκείμενες ιατρικές παθήσεις όπως καρδιαγγειακές παθήσεις, διαβήτη, χρόνιες αναπνευστικές παθήσεις ή καρκίνο είναι πιο πιθανό να αναπτύξουν σοβαρές ασθένειες με μεγαλύτερες πιθανότητες να αποβιώσουν. Πολλοί επιστήμονες, κλινικοί γιατροί και

ειδικοί που ασχολούνται με την ιατρική φροντίδα σε όλο τον κόσμο συνεχίζουν να αναζητούν μια νέα τεχνολογία που θα βοηθήσει στην αντιμετώπιση της πανδημίας Covid-19. Οι εφαρμογές της Μηχανικής Μάθησης (ML) που χρησιμοποιήθηκαν σε προηγούμενες επιδημίες ενθαρρύνουν τους ερευνητές να αναπτύξουν μια νέα προσέγγιση για την αντιμετώπιση της πανδημίας του κορονοϊού [34].

Σε μια έρευνα που διεξήχθη το 2020 στην Κίνα [34], η οποία αφορά ασθενείς που προσβλήθηκαν από τον ιό στην Σανγκάη, αφού διαγνώστηκαν θετικοί με την βοήθεια των PCR τεστ, έπειτα μεταφέρθηκαν σε συγκεκριμένη κλινική όπου εξετάστηκαν και συλλέχθηκαν στοιχεία δημογραφικού χαρακτήρα όπως (ηλικία, φύλλο, κλπ.), εργαστηριακά δεδομένα και κλινικές πληροφορίες. Στην συνέχεια, οι ερευνητές χρησιμοποίησαν τα δεδομένα αυτά , λαμβάνοντας βέβαια πρώτα την συγκατάθεσή τους, να ερευνήσουν ποιοι είναι οι τέσσερις κυριότεροι λόγοι για το αν κάποιος ασθενής πάσχει από σοβαρό/κρίσιμο σύμπτωμα ή όχι εφαρμόζοντας την Μηχανική Μάθηση. Στην έρευνα το μοντέλο που χρησιμοποιήθηκε ήταν το Support Vector Machine (SVM), σε διαφορετικούς συνδυασμούς με τα τέσσερα χαρακτηριστικά, όπως παρουσιάζει ο **Πίνακας 2**, με καλύτερη επίδοση την μετρική Area Under Curve $AUC = 97{,}57\%$. Τέλος, οι ερευνητές συμπέραναν ότι το μοντέλο είναι αξιόπιστο και αποτελεσματικό στην πρόβλεψη των σοβαρών/κρίσιμων περιστατικών του κορονοϊού λόγω της εξαιρετικής επίδοσης του μοντέλου.

Combinations	Training AUC	Testing AUC
Age, GSH, CD3 ratio, total protein	0.999616858	0.975711
Neutrophil percentage, albumin, GSH, CD4 ratio	0.997318008	0.975711
HCRP, Serum myoglobin, CL, CD4 ratio	0.998357964	0.969466
Age, Cl, Calcium, LDH	0.997318008	0.951748
Age, Serum myoglobin, Retinol binding protein, Acid glycoprotein	0.990960452	0.951423
Neutrophil percentage, Procalcitonin, Serum myoglobin, total protein	0.977024482	0.958362

Πίνακας 2: Συνδυασμοί χαρακτηριστικών με την καλύτερη επίδοση [34]

Κεφάλαιο 3: Πρόβλεψη εγκεφαλικών επεισοδίων ασθενών

Κλινικά το εγκεφαλικό επεισόδιο ορίζεται ως ένα σύνδρομο που εμφανίζει ταχύτατα εξελισσόμενα συμπτώματα ή σημάδια συγκεκριμένης απώλειας της λειτουργίας του εγκεφάλου χωρίς προφανή αιτία εκτός από εκείνη της αγγειακής προέλευσης, ενώ η σοβαρότητα του συνδρόμου κυμαίνεται από την ανάρρωση μιας ημέρας, έως μη πλήρης ανάρρωση, μέχρι τη σοβαρή αναπηρία ή ακόμη και το θάνατο [35]. Σε περίπτωση που κάποιος άνθρωπος έχει βιώσει εγκεφαλικό επεισόδιο στο παρελθόν τότε πολύ πιθανό είναι δεχτεί και άλλα επεισόδια στο μέλλον [36]. Όταν διακόπτεται απότομα η κυκλοφορία του αίματος προς ένα τμήμα του εγκεφάλου τότε προκαλείται το εγκεφαλικό επεισόδιο. Τα εγκεφαλικά κύτταρα σταδιακά νεκρώνονται εάν δεν τους παρέχεται αίμα με αποτέλεσμα να εμφανίζεται αναπηρία ανάλογα με την περιοχή του εγκεφάλου που προσβάλλεται [37].

Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (WHO), παγκοσμίως, το εγκεφαλικό επεισόδιο θεωρείται ότι είναι η δεύτερη κύρια αιτία θανάτου και η τρίτη κύρια αιτία αναπηρίας. Ένας στους τέσσερις ανθρώπους κινδυνεύει με εγκεφαλικό επεισόδιο στη ζωή του. Κάποιοι σοβαροί παράγοντες που αυξάνουν τις πιθανότητες για την πρόκληση του εγκεφαλικού επεισοδίου περιλαμβάνουν το αν οι ασθενείς είναι υπέρβαροι ή παχύσαρκοι, η έλλειψη σωματικής άσκησης, το κάπνισμα και η κατάχρηση αλκοόλ. Άλλοι ιατρικοί παράγοντες περιλαμβάνουν την υψηλή αρτηριακή πίεση, την υψηλή χοληστερόλη, τον διαβήτη μέχρι και το προσωπικό ιστορικό του ασθενούς ή αν κάποιο οικογενειακό μέλος έχει υποστεί εγκεφαλικό επεισόδιο στο παρελθόν [38].

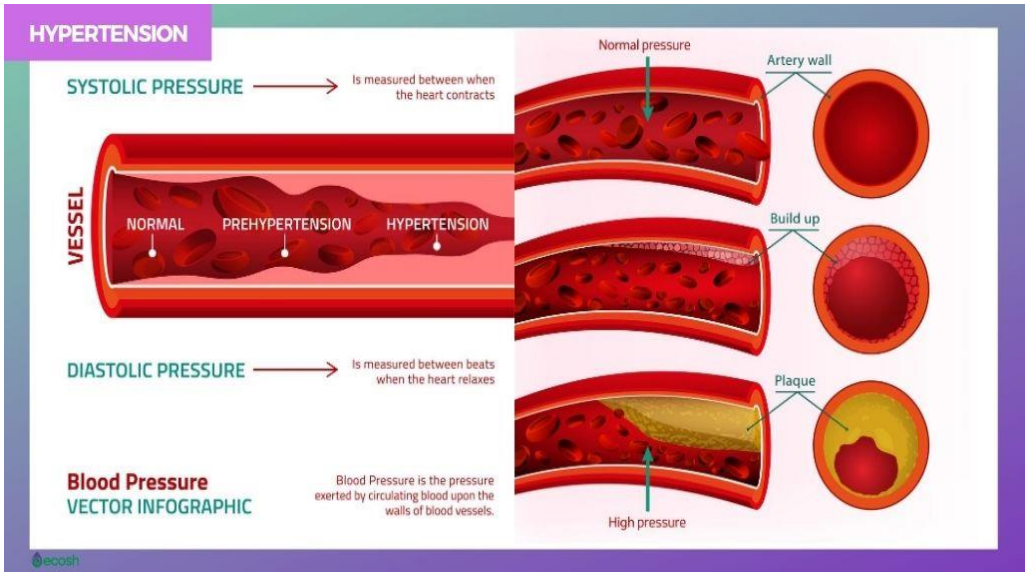
3.1 Σοβαρές αιτίες πρόκλησης εγκεφαλικού επεισοδίου

Πολλές είναι οι αιτίες που μπορούν να προκαλέσουν εγκεφαλικό επεισόδιο σε ασθενείς, κάποιες αιτίες όμως είναι πιο σημαντικές από κάποιες άλλες. Ένας από τους κυριότερους παράγοντες κινδύνου που οδηγούν σε εγκεφαλικό επεισόδιο είναι η υπέρταση. Οι Jehangir Khan, Attique-ur-Rehman, Ashfaq Ali Shah, Asif Jielani πραγματοποίησαν μια έρευνα [39] που αφορά το αν η υπέρταση είναι μια από τις σοβαρότερες αιτίες κινδύνου για πρόκληση εγκεφαλικού στον Πακιστανικό πληθυσμό. Οι ερευνητές συμπέραναν ότι η υπέρταση είναι ένας από τους σημαντικότερους παράγοντες και σαν δεύτερος κύριος παράγοντας να ακολουθεί ο διαβήτης, σύμφωνα με άλλη έρευνα [40] που διεξήχθη από τους Juan Shou, Li Zhou, Shanzhu Zhu, Xiangjie Zhang να επιβεβαιώνει το ίδιο. Ενώ το γυναικείο φύλλο με διαβήτη έχει μεγαλύτερες πιθανότητες να δεχτεί εγκεφαλικό

επεισόδιο σε αντίθεση με το αρσενικό φύλλο, σύμφωνα με την μελέτη των Sanne A E Peters, Rachel R Huxley, Mark Woodward [41].

3.1.1 Η Υπέρταση

Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (WHO), ως υπέρταση, που είναι μια συνηθισμένη, χρόνια, διαταραχή που οφείλεται στην ηλικία [42], χαρακτηρίζεται η υπερβολικά υψηλή αρτηριακή πίεση, η οποία είναι η δύναμη που ασκείται από την κυκλοφορία του αίματος στα τοιχώματα των αρτηριών του σώματος, των κύριων αιμοφόρων αγγείων του σώματος. Η αρτηριακή πίεση καταγράφεται ως δύο αριθμούς. Ο πρώτος αριθμός αντιπροσωπεύει την συστολική πίεση στα αιμοφόρα αγγεία όταν η καρδιά συστέλλεται ή χτυπάει ενώ ο δεύτερος αντιπροσωπεύει την διαστολική πίεση στα αγγεία όταν η καρδιά ηρεμεί μεταξύ των καρδιακών παλμών [43]. Ένα φυσιολογικό επίπεδο αρτηριακής πίεσης είναι χαμηλότερο από <120/80 mmHg, όπου 120 mmHg χαρακτηρίζεται η συστολική, ενώ 80 mmHg η διαστολική [44].



Εικόνα 11: Κυκλοφορία αίματος μέσα στις φλέβες [45]

Σε πολλές ανεπτυγμένες χώρες, η υπέρταση επηρεάζει το 25-35% των ενηλίκων και έως και το 60-70% των ατόμων που έχουν περάσει το εβδομηκοστό έτος της ηλικίας τους [42].

Σημαντικό είναι επίσης το γεγονός ότι, η υπέρταση ενδέχεται να προκαλέσει σοβαρή βλάβη στην καρδιά. Η υπερβολική πίεση είναι δυνατόν να σκληρύνει τις αρτηρίες, μειώνοντας τη ροή του αίματος και του οξυγόνου προς την καρδιά. Αυτή η αυξημένη πίεση και η μειωμένη ροή του αίματος μπορεί να προκαλέσει, πόνους στον θώρακα, καρδιακή προσβολή, καρδιακή ανεπάρκεια, ανώμαλους καρδιακούς παλμούς που μπορεί να οδηγήσουν σε αιφνίδιο θάνατο μέχρι και να προκαλέσει έκρηξη ή φραγμό των

αρτηριών που αιματώνουν τον εγκέφαλο και τον τροφοδοτούν με οξυγόνο, προκαλώντας το εγκεφαλικό επεισόδιο [43].

3.1.2 Ο Διαβήτης

Ο διαβήτης ή σακχαρώδης διαβήτης είναι μια ασθένεια που ζει ανάμεσα μας εδώ και πολλά χρόνια, τονίζοντας, ότι μπορεί να οδηγήσει τον ασθενή ακόμη μέχρι και τον θάνατο. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (WHO) [46], το 2019, ο διαβήτης ήταν η κύρια αιτία για 1,5 εκατομμύριο θανάτους και το 48% όλων των θανάτων λόγω του σακχαρώδη διαβήτη εμφανίστηκαν σε ηλικίες μικρότερες των 70. Άλλοι 460.000 θάνατοι από νεφροπάθεια οφείλονται στον διαβήτη και η αυξημένη γλυκόζη του αίματος προκαλεί περίπου το 20% των καρδιαγγειακών θανάτων.

Ο διαβήτης οφείλεται στο ότι η γλυκόζη του αίματος, που αποκαλείται επίσης και ζάχαρο, είναι πολύ υψηλή. Η γλυκόζη του αίματος είναι η κύρια πηγή ενέργειας και προέρχεται από τα τρόφιμα που καταναλώνει ο άνθρωπος. Η ινσουλίνη, είναι μια ορμόνη που παράγεται από το πάγκρεας, βοηθά τη γλυκόζη από την τροφή να φτάσει στα κύτταρα για να χρησιμοποιηθεί για ενέργεια. Μερικές φορές το σώμα δεν παράγει αρκετή ή καθόλου ινσουλίνη ή δεν την χρησιμοποιεί σωστά, με αποτέλεσμα η γλυκόζη να παραμένει στο αίμα και να μην φτάνει στα κύτταρα του οργανισμού [47]. Οι δύο κύριοι τύποι διαβήτη είναι ο διαβήτης τύπου 1 και ο διαβήτης τύπου 2, καθώς ο δεύτερος να αποτελεί το πιο συνηθισμένο τύπο [48].

Διαβήτης τύπου 1 θεωρείται ότι παρουσιάζει έλλειψη στην παραγωγή της ινσουλίνης και απαιτεί καθημερινή λήψη ινσουλίνης, ενώ συνήθως παρουσιάζεται σε παιδιά και νεαρούς ενήλικες, αν και μπορεί να εμφανιστεί σε οποιαδήποτε ηλικία.

Διαβήτης τύπου 2 προκύπτει από την αναποτελεσματική χρήση της ινσουλίνης από τον οργανισμό, ενώ μπορεί να εκδηλωθεί σε οποιαδήποτε ηλικία αλλά συχνότερα σε άτομα μέσης και τρίτης ηλικίας.[46]

3.1.3 Η Παχυσαρκία

Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (WHO) [49], το 2016, περισσότεροι από 1,9 δισεκατομμύριο ενήλικες, ηλικίας 18 ετών και άνω, ήταν υπέρβαροι. Εκ των οποίων πάνω από 650 εκατομμύρια ήταν παχύσαρκοι. Η παχυσαρκία είναι μια περίπλοκη και ελλιπώς κατανοητή ασθένεια, η οποία είναι ιδιαίτερα σοβαρή επηρεάζοντας όλες τις ηλικιακές ομάδες. Συχνά ορίζεται απλά ως μια κατάσταση ασυνήθιστης ή υπερβολικής συσσώρευσης λίπους στο λιπώδη ιστό, σε τέτοιο βαθμό που μπορεί να βλάψει την υγεία. Η βασική αιτία πρόκλησης της παχυσαρκίας είναι η ενεργειακή ανισορροπία μεταξύ των θερμίδων που λαμβάνονται και των θερμίδων που καταναλώνονται. Αυτό το γεγονός οφείλεται όταν υπάρχει αυξημένη πρόσληψη τροφίμων υψηλής ενεργειακής αξίας με υψηλή περιεκτικότητα σε λιπαρά και σάκχαρα και η απουσία σωματικής άσκησης [50].

Ο δείκτης μάζας σώματος, που είναι γνωστός επίσης και ως Body Mass Index (BMI), είναι ένας απλός δείκτης αναλογίας του βάρους προς το ύψος και χρησιμοποιείται συνήθως για την κατηγοριοποίηση των υπέρβαρων και παχύσαρκων ενηλίκων. Ορίζεται ως το βάρος ενός ατόμου σε κιλά διαιρούμενο με το τετράγωνο του ύψους του σε μέτρα (kg/m2).

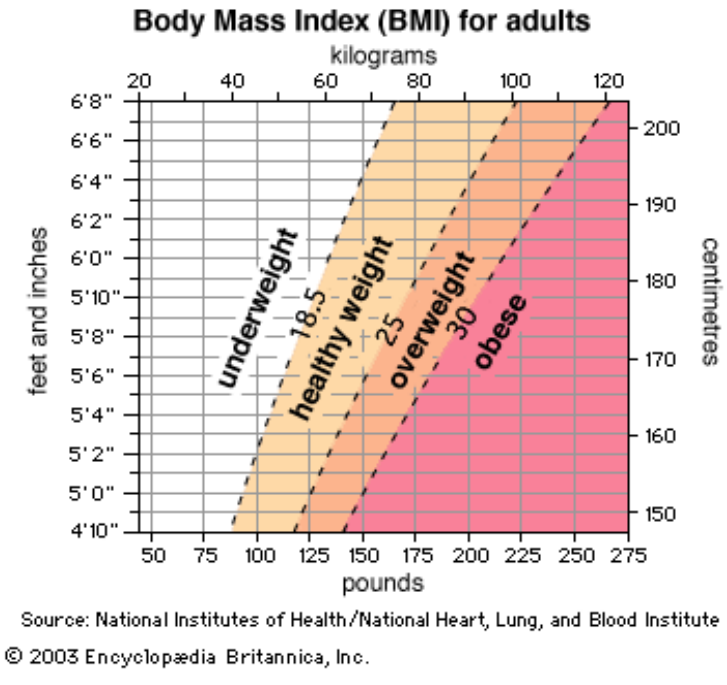
Η κατηγοριοποίηση των ενηλίκων με βάση το BMI είναι [50]:

- Για $BMI < 18{,}50$ ο ενήλικας θεωρείται κάτω από το κανονικό όριο
- Για $BMI < 18.50 - 24.99$ είναι το φυσιολογικό επίπεδο

Για $BMI \geq 25{,}00$ ο ενήλικας θεωρείται ότι είναι υπέρβαρος

- Για $BMI \geq 25{,}00 - 29.99$ ορίζεται η προπαχυσαρκία
- Για $BMI \geq 30.00 - 34.99$ ορίζεται η παχυσαρκία κατηγορίας 1
- Για $BMI \geq 35.00 - 39.99$ ορίζεται η παχυσαρκία κατηγορίας 2
- Για $BMI \geq 40.00$ ορίζεται η παχυσαρκία κατηγορίας 3

Όσο μεγαλύτερη είναι η τιμή του BMI τόσο αυξημένες είναι και πιθανότητες κινδύνου για την πρόκληση εγκεφαλικού επεισοδίου στους ενήλικες.



Εικόνα 12: Σχήμα BMI σε συνάρτηση βάρους και ύψους [51]

Η αντιμετώπιση της παχυσαρκίας μπορεί να αποφευχθεί καθιστώντας σημαντικότερο την επιλογή των πιο υγιεινών τροφίμων, όπως τον περιορισμό της πρόσληψης ενέργειας από λιπαρά και σάκχαρα και την αύξηση της κατανάλωσης φρούτων και λαχανικών και

της τακτικής σωματικής άσκησης τουλάχιστον 150 λεπτών κατά την διάρκεια μιας εβδομάδας για τους ενήλικες.

Κάπνισμα

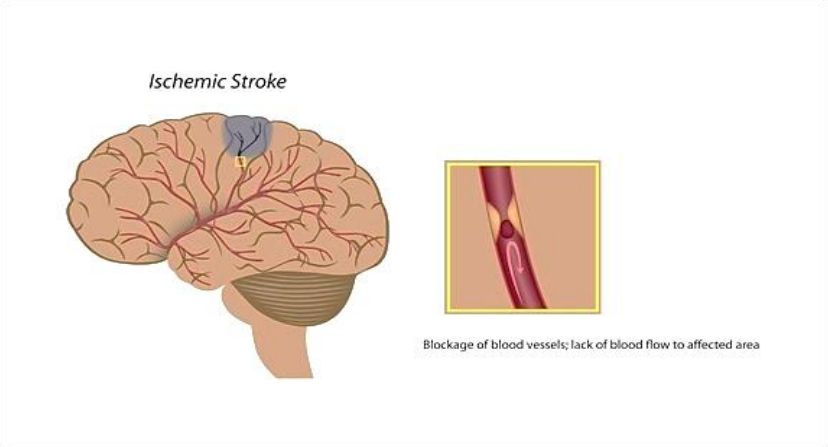
Το κάπνισμα αποτελεί μια από τις μεγαλύτερες απειλές για τη δημόσια υγεία που αντιμετώπισε ποτέ ο άνθρωπος, δεδομένου ότι σκοτώνει πάνω από 8 εκατομμύρια ανθρώπους ετησίως, συμπεριλαμβανομένων περίπου 1,2 εκατομμυρίων θανάτων από την έκθεση στο παθητικό κάπνισμα [52]. Το κάπνισμα είναι η πιο αποτρέψιμη αιτία θανάτου στην κοινωνία σήμερα, η οποία επίσης ευθύνεται για 140.000 εγκεφαλικά επεισόδια και πάνω από 5 εκατομμύρια άνθρωποι χάνουν τη ζωή τους στις ΗΠΑ ετησίως. Επιπλέον το κάπνισμα αυξάνει τον κίνδυνο εγκεφαλικού επεισοδίου κατά τρεις έως τέσσερις φορές και η έκθεση του παθητικού καπνίσματος στο σπίτι αυξάνει τον κίνδυνο εγκεφαλικού επεισοδίου κατά μιάμιση έως δύο φορές [53].

3.2 Τύποι εγκεφαλικών επεισοδίων

Τα εγκεφαλικά επεισόδια χαρακτηρίζονται από το σημείου που προκαλούνται και τον τύπο της διαταραχής. Οι δύο βασικοί τύποι είναι το ισχαιμικό εγκεφαλικό επεισόδιο και το αιμορραγικό εγκεφαλικό επεισόδιο.

3.2.1 Ισχαιμικό εγκεφαλικό επεισόδιο

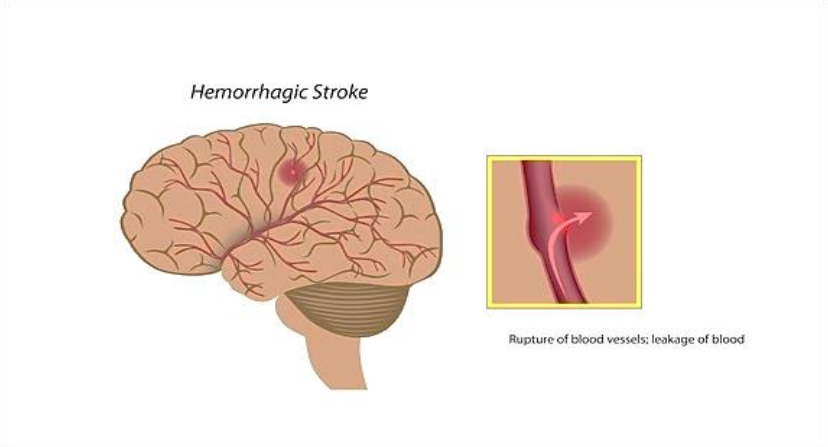
Ο συχνότερος τύπος εγκεφαλικού επεισοδίου είναι το γνωστό ισχαιμικό εγκεφαλικό επεισόδιο ή εγκεφαλικό έμφρακτο. Μια έρευνα που δημοσιεύθηκε σε ένα περιοδικό από τους Klaus Kaae Andersen, Tom Skyhøj Olsen, Christian Dehlendorff και Lars Peter Kammersgaard [54], μεταξύ 39.484 ασθενών, οι 35.491 (89,9%) παρουσίασαν ισχαιμικό εγκεφαλικό επεισόδιο. Το έμφρακτο εμφανίζεται ως αποτέλεσμα ανεπαρκής ή μη ομαλής ροής του αίματος σε μια περιοχή του εγκεφάλου, που συνήθως προκαλείται από την απόφραξη μιας αρτηρίας [55]. Εν ολίγοις, ο εγκεφαλικός ιστός πρέπει να τροφοδοτείται με οξυγόνο, γλυκόζη και άλλα ζωτικά υλικά μέσω της συνεχούς εισροής αίματος με ρυθμό περίπου 50-54 ml αίματος ανά 100 g εγκεφαλικού ιστού ανά λεπτό [56]. Ένας άλλος παράγοντας που συμβάλλει στον κίνδυνο για ισχαιμικό εγκεφαλικό επεισόδιο είναι η αθηροσκλήρυνση, που συμβαίνει όταν οι λιπαρές εναποθέσεις που αποκαλούνται πλάκες μπορούν να προκαλέσουν απόφραξη, καθώς συσσωρεύονται στα αιμοφόρα αγγεία. Με αποτέλεσμα να μπορέσει να αναπτυχθεί ένα πήγμα αίματος που ονομάζεται και θρόμβος και να παρεμποδίσει τη ροή του αίματος. Ο τύπος αυτός του ισχαιμικού εγκεφαλικού επεισοδίου ονομάζεται εγκεφαλική θρόμβωση.



Εικόνα 13: Παράδειγμα Ισχαιμικού εγκεφαλικού επεισοδίου [57]

3.2.2 Αιμορραγικό εγκεφαλικό επεισόδιο

Το αιμορραγικό εγκεφαλικό επεισόδιο συμβαίνει, όταν μία αρτηρία στον εγκέφαλο αιμορραγήσει. Η αιμορραγία από μία αρτηρία μέσα στον εγκέφαλο, που προκαλεί συμπίεση, εκτόπιση και νέκρωση του νευρικού ιστού όπου καλείται ενδοεγκεφαλική αιμορραγία. Απλούστερα, η βασική παθογένεια είναι μια περιοχή αιμορραγίας που προκαλεί άμεση βλάβη στον εγκεφαλικό ιστό. Στο συγκεκριμένο τύπο εγκεφαλικού επεισοδίου ανήκουν το 10-15 % όλων των εγκεφαλικών επεισοδίων και έχουν σημαντικά υψηλότερη νοσηρότητα και θνησιμότητα από ό,τι τα ισχαιμικά εγκεφαλικά επεισόδια [55]. Στην έρευνα που διεξήχθη για τους 39.484 ασθενείς όπως αναφέρθηκε παραπάνω, από τους 3.993 (10,1%) ασθενείς με αιμορραγικό εγκεφαλικό επεισόδιο, οι 1.966 (49,2%) ασθενείς πέθαναν κατά τη διάρκεια παρακολούθησης. Αντιθέτως, από τους 35.491 (89.9%) ασθενείς με ισχαιμικό εγκεφαλικό επεισόδιο, 9.220 (25,9%) πέθαναν κατά τη διάρκεια της παρακολούθησης [54].



Εικόνα 14: Παράδειγμα Αιμορραγικού εγκεφαλικού επεισοδίου [57]

3.3 Εφαρμογή της Μηχανικής Μάθησης

Όπως φαίνεται και από τα προηγούμενα κεφάλαια, το εγκεφαλικό επεισόδιο είναι ένα πάρα πολύ σοβαρό αλλά σίγουρα αποτρέψιμο πρόβλημα. Εδώ είναι που έρχεται η Μηχανική Μάθηση για να βοηθήσει την κατάσταση. Παρακάτω παρουσιάζεται η μέθοδος με την οποία μπορεί να επιλυθεί ένα τέτοιου είδους προβλήματος.

Ποιο είναι το πρόβλημα;

Χρησιμοποιώντας συγκεκριμένο σύνολο δεδομένων με διάφορες, αλλά, σημαντικές πληροφορίες για τον κάθε ασθενή για το αν πρόκειται να αντιμετωπίσει εγκεφαλικό επεισόδιο ή όχι. Σε αυτό το υποκεφάλαιο η διαδικασία που θα ακολουθηθεί θα είναι ίδια όπως προαναφέρθηκε στην Ενότητα **1.4**.

3.3.1 Ανάλυση και επεξεργασία συνόλου δεδομένων

Αρχικά, τα δεδομένα που χρησιμοποιήθηκαν για το συγκεκριμένο πρόβλημα λήφθηκαν από [58]. Στον παρακάτω πίνακα απεικονίζονται όλα τα χαρακτηριστικά του συνόλου δεδομένων που θα αναλυθούν και θα επεξεργαστούν. Το σύνολο δεδομένων αφορά περιπτώσεις 5.110 ασθενών με 12 κλινικά χαρακτηριστικά, όπου το ένα από αυτά τα χαρακτηριστικά αποτελεί τον στόχο, δηλαδή το χαρακτηριστικό που πρέπει να γίνει η πρόβλεψη.

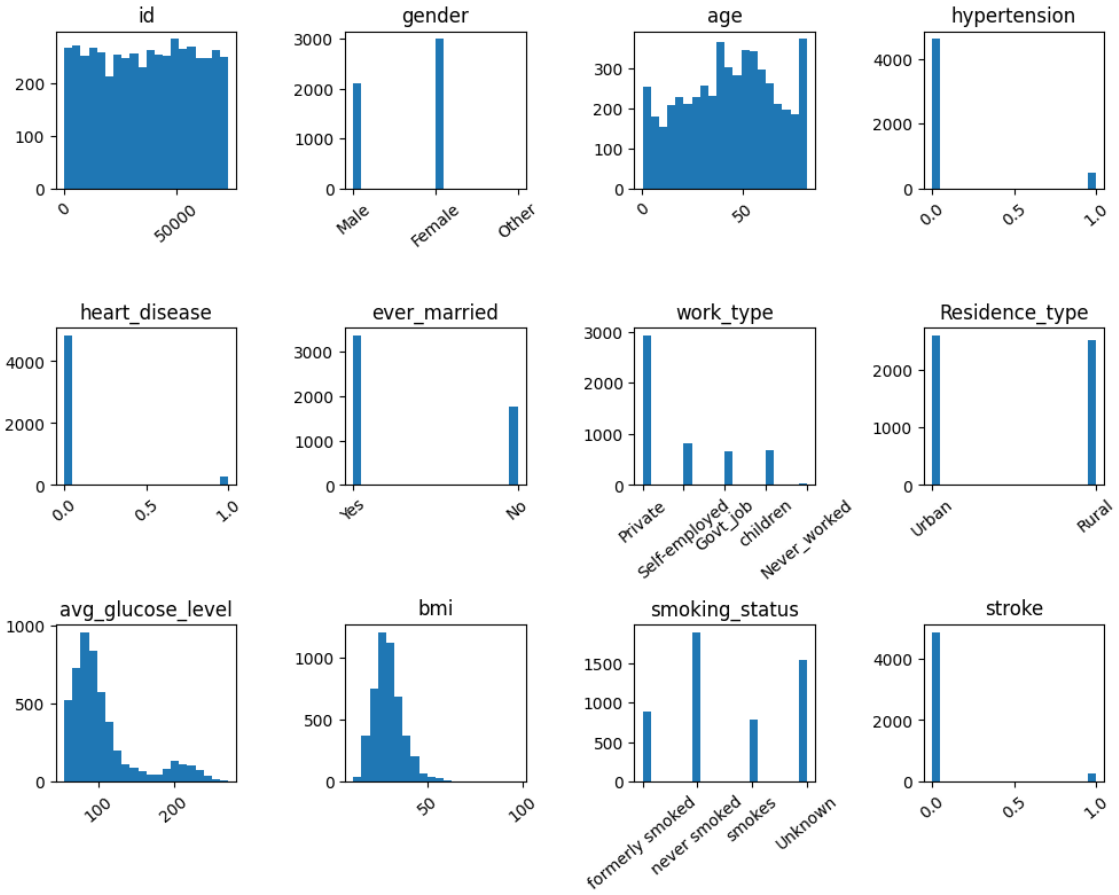
Χαρακτηριστικά	Πληροφορίες
ID	Μοναδικός αριθμός για κάθε ασθενή
Gender	Αρσενικό, Θηλυκό, Άλλο
Age	Ηλικία ασθενή
Hypertension	Όχι: 0, Ναι: 1
Heart disease	Όχι: 0, Ναι: 1
Ever married	Όχι, Ναι
Work type	Παιδί, Δημόσιο, Ιδιωτικό, Αυτοαπασχολούμενος, Δεν έχει εργαστεί ποτέ
Residence type	Αγροτική, Αστική περιοχή
Avg glucose level	Συνεχής τιμή Μ.Ο γλυκόζης στο αίμα
BMI	Δείκτης μάζας σώματος
Smoking status	Πρώην καπνιστής, καπνιστής, δεν κάπνισε ποτέ
Stroke	Δέχτηκε εγκεφαλικό: Ναι: 1, Όχι: 0

Πίνακας 3: Πληροφορίες για τα χαρακτηριστικά του συνόλου δεδομένων

Αρχικά, ο πρώτος έλεγχος που πρέπει να γίνει είναι τα αριθμητικά δεδομένα δηλαδή, το ID, Age, Hypertension, Heart disease, Avg glucose level, BMI και stroke. Ελέγχεται αν οι τιμές που έχει το σύνολο δεδομένων είναι αποδεκτές για το χαρακτηριστικό που υποδηλώνουν και αν αυτές είναι χρήσιμες για την επίλυση του προβλήματος.

Όπως παρουσιάστηκε, ο **Πίνακας 3** με όλες τις πληροφορίες που φέρει το κάθε χαρακτηριστικό, είναι εξίσου σημαντικό να δημιουργηθεί και η γραφική αναπαράσταση για το κάθε χαρακτηριστικό για την καλύτερη κατανόησή τους.

Παρουσιάζονται γραφικά σε μορφή ιστογράμματος, όλα τα χαρακτηριστικά στην **Εικόνα 15**:



Εικόνα 15: Γραφική αναπαράσταση όλων των χαρακτηριστικών

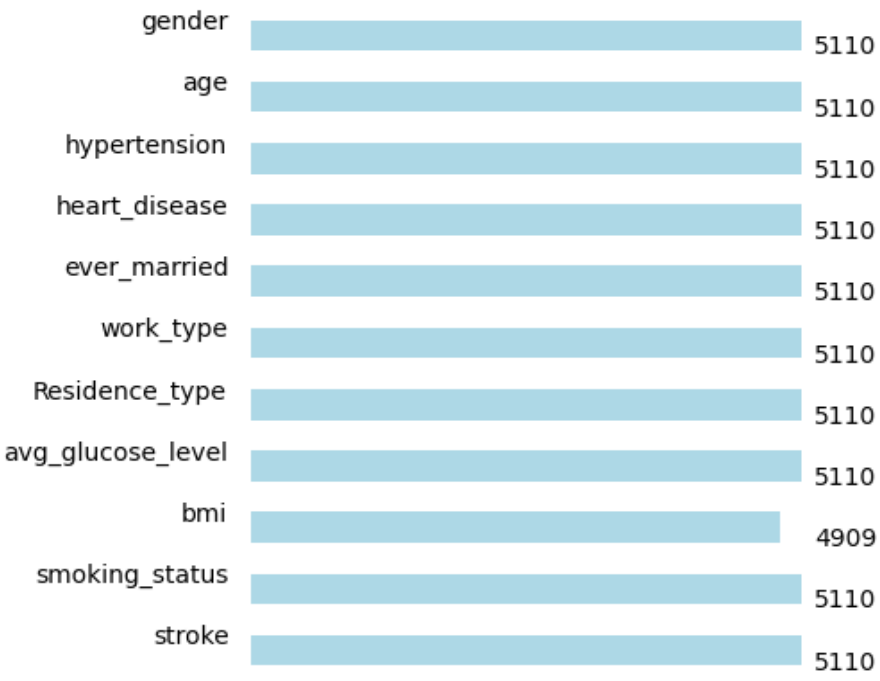
Η ελάχιστη τιμή που έχει το ID είναι 67 και η μέγιστη 72.940 και υποδηλώνει έναν μοναδικό αριθμό για τον κάθε ασθενή. Η χρησιμότητα της ιδιότητας αυτής θα ήταν για την εύρεση του συγκεκριμένου ασθενούς. Άρα, η συγκεκριμένη στήλη θα ήταν άχρηστη για την εκπαίδευση του μοντέλου αργότερα μιας και δεν προσφέρει κάποια περαιτέρω πληροφορία για την πρόβλεψη του εγκεφαλικού επεισοδίου και για αυτό θα πρέπει να διαγραφεί.

Ένα δεύτερο χαρακτηριστικό που αναλύθηκε είναι η ηλικία, βρέθηκε ως ελάχιστη ηλικία το 0.08 και ως μέγιστη 82, το οποίο μπορεί να είναι πραγματικό. Το πρόβλημα είναι τι θα μπορούσε να σημαίνει η ηλικία 0,08; Έπειτα αναλύθηκε όλη η στήλη της ηλικίας και αξίζει να αναλυθεί συγκεκριμένα το εύρος των παρακάτω τιμών:

Εύρος τιμών	Πλήθος
0,08 – 1,88	120
0,08 – 0,88	43
1	5
1,08 – 1,88	72

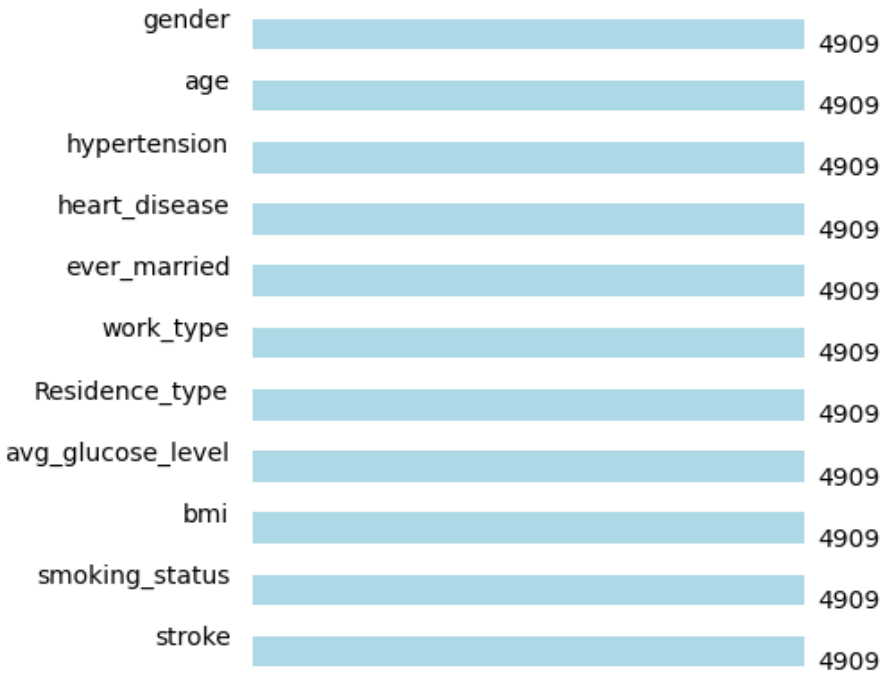
Πίνακας 4: Εύρος τιμών ηλικίας από 0,08 - 1,88

Από 0,08 – 0,88 υπάρχουν 11 διαφορετικές τιμές και εξίσου το ίδιο για το εύρος 1,08 – 1,88. Οι τιμές είναι (0,08), (0,16), (0,32), ... (0,72), (0,80), (0,88) και το αντίστοιχο συμβαίνει για το άλλο εύρος. Όπως φαίνεται η κάθε τιμή αυξάνεται κατά 0,08 από την προηγούμενή της και συνολικά είναι 11. Η πρώτη υπόθεση που μπορεί να γίνει είναι ότι οι συγκεκριμένες τιμές πιθανότατα να αντιπροσωπεύουν την ηλικία σε μήνες για παιδιά μικρότερα των 2 ετών. Το 0,08 υποδηλώνει τον 1ο μήνα, το 0,16 τον 2ο μήνα, το 0,88 τον 11ο μήνα ενώ η ακέραια τιμή 1 υποδηλώνει το 1ο έτος της ηλικίας του βρέφους. Αφού οι τιμές αυτές είναι πραγματικές τότε θα ήταν καλό να παραμείνουν στο σύνολο δεδομένων.



Εικόνα 16: Αναπαράσταση χαρακτηριστικών για ελλιπείς τιμές

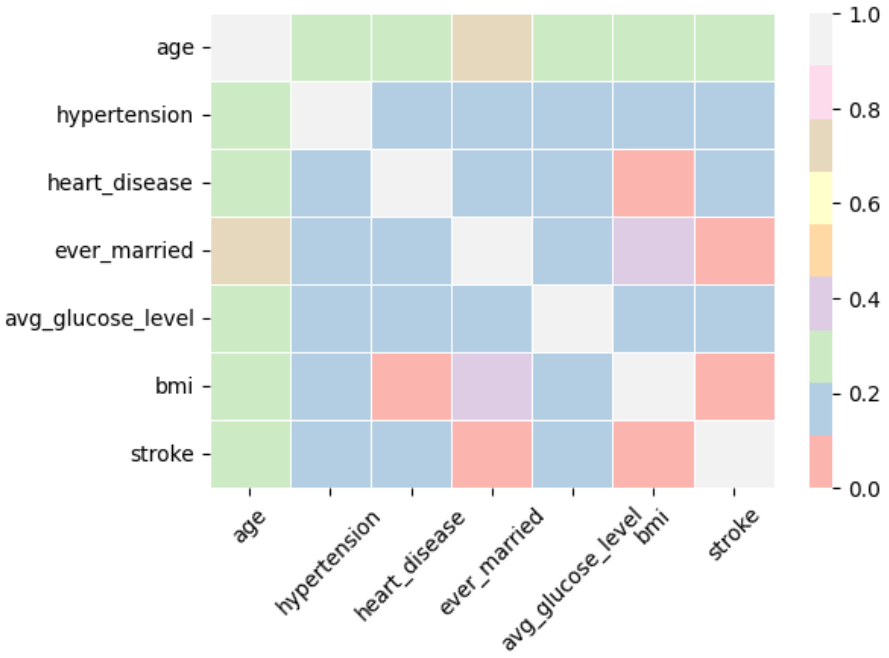
Παρατηρείται, επίσης, ότι το συνολικό πλήθος της στήλης BMI είναι 4.909 ενώ οι συνολικοί ασθενείς είναι 5.110. Η παραπάνω **Εικόνα 16** υποδηλώνει ότι υπάρχουν 201 άγνωστες τιμές NaN, οι τιμές αυτές είναι ελλιπείς και θα μπορούσαν να δημιουργήσουν πολλά σφάλματα κατά την εκπαίδευση των μοντέλων. Επιλέχτηκε η μέθοδος της αφαίρεσης αυτών των τιμών διότι μπορεί αυτές να μην υποδηλώνουν πραγματικές πληροφορίες για τα χαρακτηριστικά και τώρα τα δείγματα των χαρακτηριστικών είναι 4.909. Χρησιμοποιήθηκε η βιβλιοθήκη ανοιχτού κώδικα missingno [59], η οποία βοηθά στην απεικόνιση και ανάλυση ελλιπών τιμών.



Εικόνα 17: Αναπαράσταση χαρακτηριστικών μετά την αφαίρεση ελλιπών τιμών

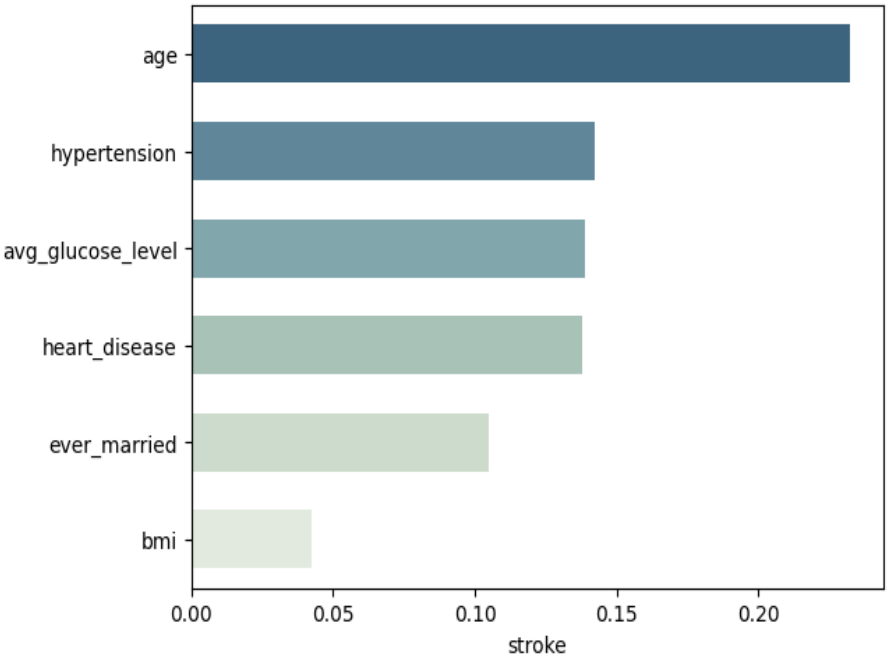
Επόμενη διαδικασία είναι η ανάλυση και αντιμετώπιση των κατηγορηματικών χαρακτηριστικών, παρατηρείται ότι στην στήλη gender υπάρχουν τιμές ‘other’, οι οποίες δεν υποδηλώνουν κάτι για αυτό και αφαιρούνται. Επιπροσθέτως, η στήλη ever married περιλαμβάνει τιμές yes και no, οι οποίες μετατρέπονται για yes = 1 και για no = 0.

Επίσης, πολύ σημαντικό είναι, να ελεγχθεί ο συντελεστής συσχέτισης Pearson, που χρησιμοποιείται για τον υπολογισμό της συσχέτισης μεταξύ στηλών-χαρακτηριστικών **MONO** για αριθμητικές τιμές σε ένα σύνολο δεδομένων κατά ζεύγος. Η συσχέτιση αποτελεί μια στατιστική μέθοδο που περιγράφει τη γραμμική σχέση μεταξύ δύο μεταβλητών. Κυμαίνεται [-1,1], όπου το -1 υποδηλώνει μια ισχυρή αρνητική συσχέτιση, ενώ αντιθέτως το 1 υποδηλώνει μια ισχυρή θετική συσχέτιση και το 0 δεν υποδηλώνει κάποια συσχέτιση. Στην **Εικόνα 18** φαίνεται ο πίνακας συσχέτισης του συνόλου δεδομένων, για τις θετικές συσχετίσεις και για κάθε χαρακτηριστικό:



Εικόνα 18: Πίνακας συσχέτισης Pearson

Στον παραπάνω πίνακα φαίνεται ότι το χαρακτηριστικό της ηλικίας συσχετίζεται σχεδόν με όλα τα υπόλοιπα χαρακτηριστικά και ιδιαίτερα με το χαρακτηριστικό stroke που είναι και το ζητούμενο. Επιπροσθέτως, τα χαρακτηριστικά hypertension, heart_disease και avg_glucose είναι τα αμέσως επόμενα χαρακτηριστικά που συσχετίζονται επίσης με το stroke.



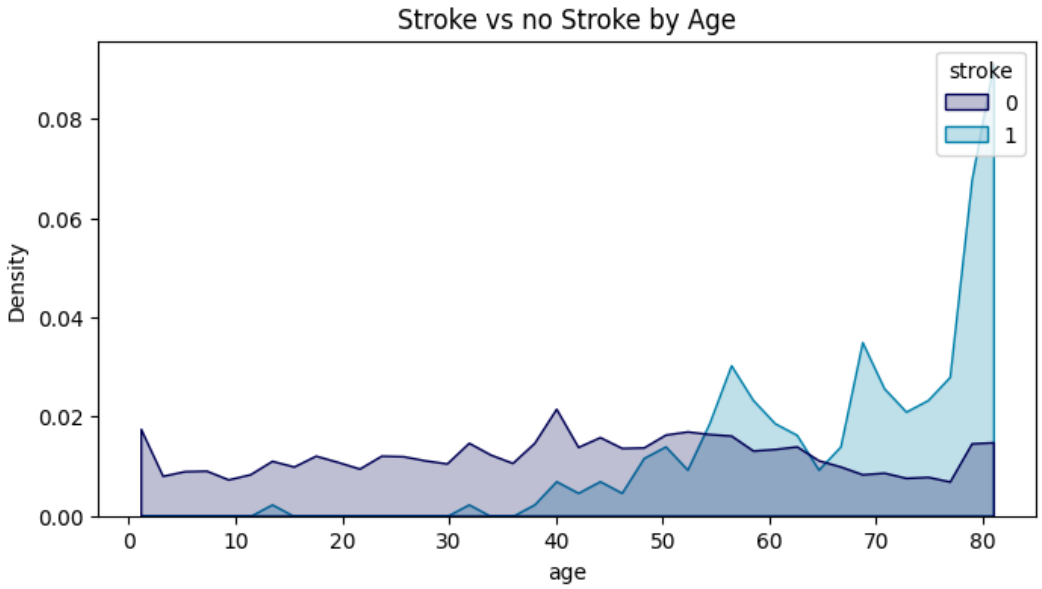
Εικόνα 19: Ραβδόγραμμα αριθμητικών χαρακτηριστικών συσχέτισης με stroke

Παρακάτω γίνεται αναπαράσταση των ποσοστών συσχέτισης με το χαρακτηριστικό stroke, η ηλικία παρουσιάζει περίπου 23% συσχέτιση με το stroke, για αυτό και θα ήταν καλό να αναλυθεί περισσότερο.

Χαρακτηριστικά	**Συσχέτιση**
age	0,232
hypertension	0,142
avg_glucose_level	0,139
heart_disease	0,138
ever_married	0,105
bmi	0,042

Πίνακας 5: Ποσοστά συσχέτισης για κάθε αριθμητικό χαρακτηριστικό με το stroke

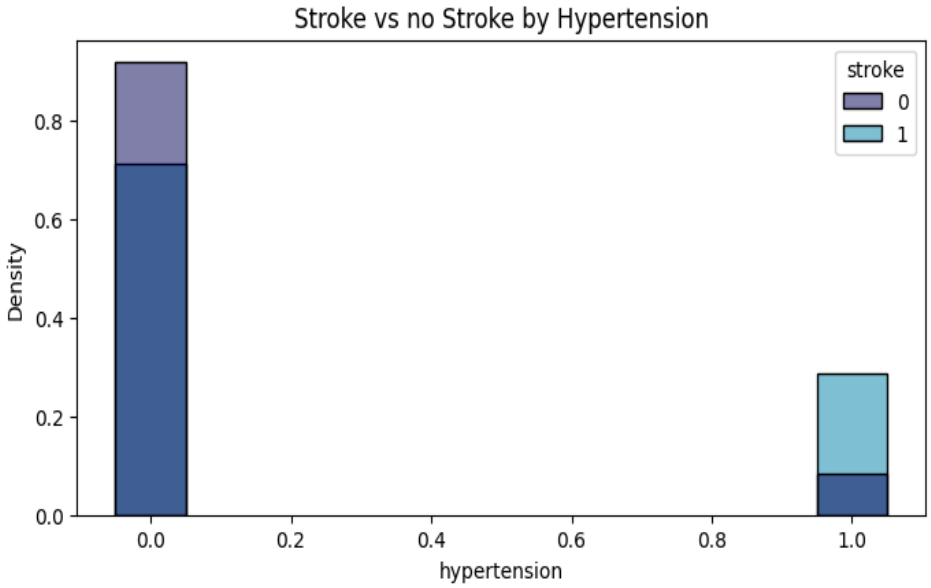
Τώρα, θα ήταν χρήσιμο να πραγματοποιηθεί η αναπαράσταση της πυκνότητα της ηλικίας ανάλογα με τους ασθενείς που έχουν δεχτεί εγκεφαλικό.



Εικόνα 20: Γραφική αναπαράσταση ηλικίας ανάλογα με το εγκεφαλικό

Στην παραπάνω εικόνα μπορεί κανείς να παρατηρήσει ότι οι περισσότεροι ασθενείς που έχουν δεχθεί εγκεφαλικό επεισόδιο έχουν ηλικία άνω των 40, σε αντίθεση με τους μικρότερους σε ηλικία που είναι πιο σπάνια η εμφάνιση του εγκεφαλικού. Όμως, φαίνεται ότι όσο μεγαλύτερη είναι ηλικία τόσο πιο πιθανό είναι να έχουν δεχτεί εγκεφαλικό, κάνοντας τις κρίσιμες ηλικίες αυτές των 65 και άνω.

Επόμενο χαρακτηριστικό που πρέπει να μελετηθεί για την συσχέτιση με το χαρακτηριστικό stroke είναι το hypertension δηλαδή η υπέρταση.



Εικόνα 21: Ραβδόγραμμα πυκνότητας υπέρτασης ανάλογα με το stroke

Στο παραπάνω ραβδόγραμμα φαίνεται ότι, η πυκνότητα των ασθενών που δεν έχουν εμφανίσει υπέρταση είναι μεγαλύτερη από την πυκνότητα των ασθενών που έχουν δεχτεί εγκεφαλικό, ενώ σε αντίθεση με αυτούς που έχουν εμφανίσει υπέρταση τα εγκεφαλικά επεισόδια έχουν μεγαλύτερη πυκνότητα. Με άλλα λόγια αν υπήρχαν περισσότερα δείγματα ασθενών που έχουν εμφανίσει υπέρταση τότε πιθανότατα θα είχαν παρουσιαστεί και πολύ περισσότερες περιπτώσεις εγκεφαλικών επεισοδίων.

Επειδή οι ταξινομητές (classifiers) ή αλγόριθμοι για να μπορούν να χρησιμοποιηθούν θα πρέπει όλες οι τιμές των χαρακτηριστικών να είναι αριθμητικές. Χρησιμοποιήθηκε η μέθοδος get_dummies ή οποία κωδικοποιεί το σύνολο κατηγορηματικών στηλών που του δίνεται ως είσοδος και η έξοδος είναι στήλες με αριθμητικές τιμές που περιλαμβάνουν λογικές τιμές 0 και 1, για το λάθος και σωστό αντίστοιχα. Οι στήλες που επιλέχθηκαν για την κωδικοποίηση είναι gender, work_type, residence_type και smoking_status, παρακάτω φαίνεται ένα παράδειγμα για το gender και το work type πριν και μετά την κωδικοποίηση.

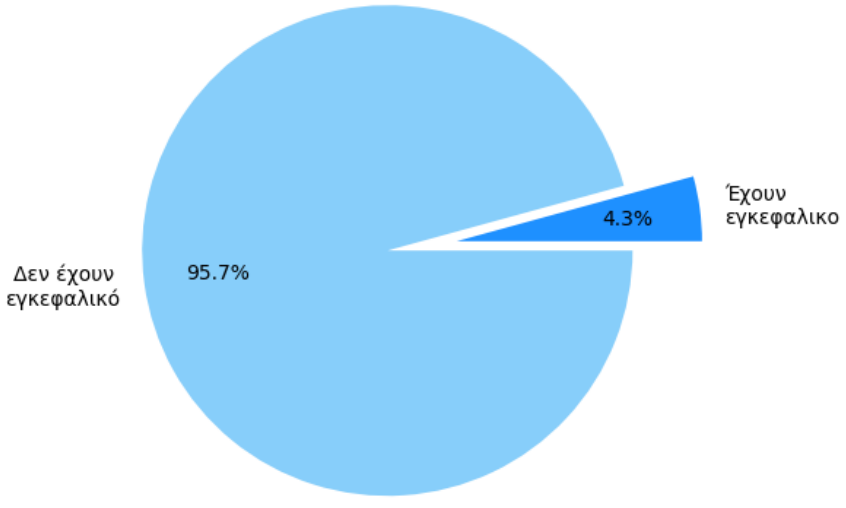
gender	age	hypertension	heart_disease	ever_married	work_type
Male	67.0	0	1	1	Private
Male	80.0	0	1	1	Private
Female	49.0	0	0	1	Private
Female	79.0	1	0	1	Self-employed
Male	81.0	0	0	1	Private

Πίνακας 6: Dataframe πριν την κωδικοποίηση

gender_Female	gender_Male	work_type_Govt_job	work_type_Never_worked	work_type_Private	work_type_Self-employed	work_type_children
0	1	0	0	1	0	0
0	1	0	0	1	0	0
1	0	0	0	1	0	0
1	0	0	0	0	1	0
0	1	0	0	1	0	0

Πίνακας 7: Dataframe μετά την κωδικοποίηση

Πολύ σημαντικό είναι, επίσης, να ελεγχθεί η αναλογία του χαρακτηριστικού stroke για τον αν είναι ισορροπημένα τα δεδομένα, όσοι δεν έχουν εγκεφαλικό ανήκουν στην κλάση 0, ενώ όσοι έχουν εγκεφαλικό ανήκουν στην κλάση 1, όπως φαίνεται στην παρακάτω εικόνα:



Εικόνα 22: Αναλογία όσων έχουν εγκεφαλικό ή όχι στα δεδομένα

Όπως φαίνεται παραπάνω, μπορεί κανείς να παρατηρήσει ότι τα περιστατικά της κλάσης 0 είναι 4.699 (95.7%) ενώ στην περίπτωση της κλάσης 1 είναι μόλις 209 (4.3%) ασθενείς και αυτό θα μπορούσε να αποτελέσει πρόβλημα στην εκπαίδευση του μοντέλου καθώς τα δεδομένα είναι ανομοιογενή. Παρακάτω θα γίνει εξονυχιστική ανάλυση για το πρόβλημα που δημιουργείται και πως μπορεί να αντιμετωπιστεί.

3.3.2 Προετοιμασία δεδομένων για την εκπαίδευση

Αφού, έγινε η ανάλυση και η επεξεργασία των δεδομένων επόμενο βήμα είναι ο διαχωρισμός δεδομένων σε σετ δεδομένων εκπαίδευσης και σε δεδομένων δοκιμής. Η μέθοδος που χρησιμοποιήθηκε για το βέλτιστο αποτέλεσμα είναι η συνάρτηση train test split με αναλογία 70% δεδομένα εκπαίδευσης προς 30% δεδομένα δοκιμής και random state = 42. Το random state είναι μια μεταβλητή που όταν δέχεται τιμή τότε για κάθε φορά που χρησιμοποιείται θα επιφέρει το ίδιο αποτέλεσμα. Συνήθως οι αναλογίες που χρησιμοποιούνται για την μέθοδο αυτή είναι 85/15, 80/20, 75/25, 70/30 και 65/35. Αυτός ο διαχωρισμός γίνεται ώστε το μοντέλο να εκπαιδευτεί στο σετ δεδομένων εκπαίδευσης και έπειτα να δοκιμαστεί στο σετ δεδομένων δοκιμής το οποίο είναι τελείως άγνωστο στο μοντέλο. Ένα εξαιρετικά σημαντικό λάθος που μπορεί να γίνει στην μέθοδο αυτή και πρέπει να αναφερθεί, είναι όταν το μοντέλο εκπαιδεύεται σε **ΟΛΟ** το σύνολο δεδομένων και έπειτα δοκιμάζεται στο σύνολο δεδομένων δοκιμής, τότε το μοντέλο έχει ήδη δει τα δεδομένα δοκιμής και έχει ήδη μάθει πράγματα για αυτό οπότε όταν θα γίνει η δοκιμή τότε το μοντέλο θα παρουσιάσει άριστη ευστοχία και εκεί είναι που μπορεί κάποιος να νομίσει ότι το μοντέλο είναι εξαιρετικά ικανό να προβλέψει, αλλά αυτό είναι λάθος.

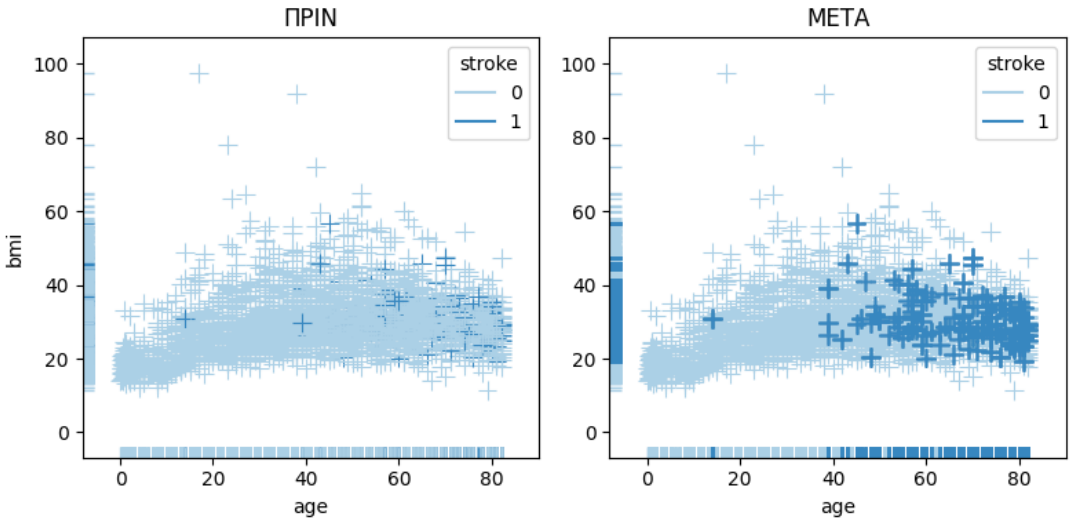
	Δείγματα	**Ποσοστό**
Σετ δεδομένων εκπαίδευσης	3.435	70 %
Σετ δεδομένων δοκιμής	1.473	30 %
Σύνολο	4.908	100 %

Πίνακας 8: Πλήθος δειγμάτων σε σύνολο εκπαίδευσης και δοκιμής

Υπάρχουν πολλοί τρόποι για να αντιμετωπιστούν τα ανομοιογενή δεδομένα, ένας απ' αυτούς είναι η μέθοδος της Επαναδειγματοληψίας (Resampling). Με την μέθοδο αυτή μπορεί να αλλάξει η αναλογία της κλάσης 0 και κλάσης 1 στην επιθυμητή, ωστόσο υπάρχουν 2 είδους μέθοδοι, η μέθοδος της Υπερδειγματοληψίας (Oversampling) και η μέθοδος της Υποδειγματοληψίας (Downsampling). Η πρώτη μέθοδος μπορεί να προσθέσει δείγματα στην μειονοτική κλάση μέχρι και στον ίδιο αριθμό δειγμάτων με την κυρίαρχη. Αντιθέτως, στην περίπτωση που χρησιμοποιηθεί η μέθοδος Downsample τότε

αυτή μπορεί να αφαιρέσει δείγματα από την κυρίαρχη κλάση μέχρι και ίσα με την μειονοτική. Τονίζεται ότι οι μέθοδοι αυτοί χρησιμοποιούνται **ΜΟΝΟ** στο σύνολο δεδομένων εκπαίδευσης. Για παράδειγμα, έστω ότι η κυρίαρχη κλάση είναι η 0, με 1000 δείγματα, ενώ η κλάση 1 είναι η μειονοτική και έχει 100 δείγματα, στην περίπτωση που γίνει υπερδειγματοψία στο 50 % της κυρίαρχης κλάσης τότε η μειονοτική κλάση θα αυξηθεί μέχρι το μισό της κυρίαρχης, δηλαδή στα 500 δείγματα και η κυρίαρχη θα παραμείνει στα 1000.

Στο συγκεκριμένο πρόβλημα χρησιμοποιήθηκε η μέθοδος υπερδειγματοληψίας Random Ovesampling με 100% και random state = 46. Η υπερδειγματοληψία έγινε σε όλο το σετ δεδομένων εκπαίδευσης και δίνεται ένα παράδειγμα του διαγράμματος διασποράς όσον αφορά την ηλικία και το BMI.



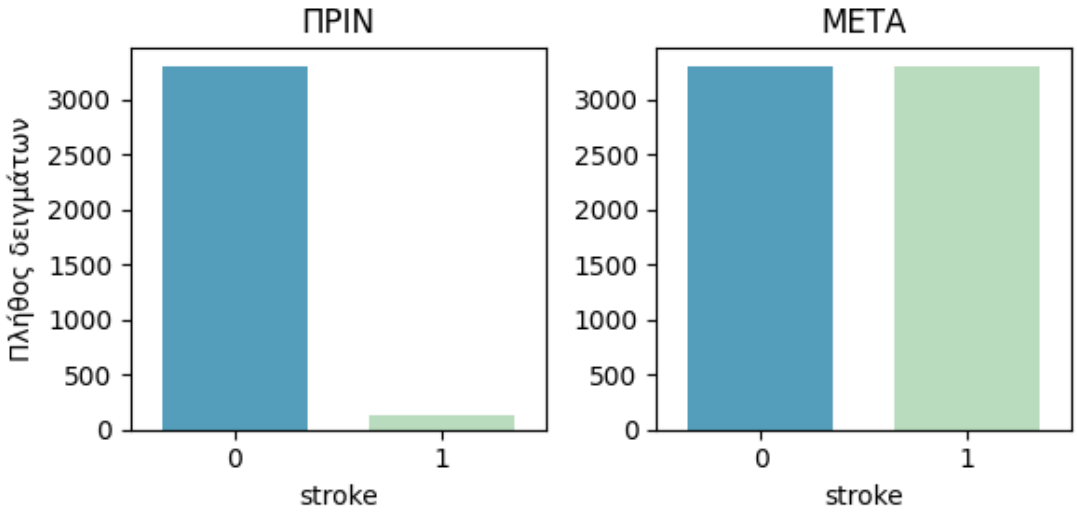
Εικόνα 23: Διάγραμμα διασποράς Age και BMI πριν και μετά την υπερδειγματοληψία

Οι γαλάζιοι σταυροί απεικονίζουν τα δείγματα σε συνάρτηση ηλικίας και BMI για όσους δεν έχουν εγκεφαλικό (Κλάση 0), ενώ οι μπλε σταυροί απεικονίζουν όσους έχουν δεχτεί εγκεφαλικό (Κλάση 1). Πριν την υπερδειγματοληψία, μπορεί κανείς να αντιληφθεί ότι οι περιπτώσεις της κλάσης 1 είναι λιγότερες από αυτές μετά. Ο **Πίνακας 9** παρουσιάζει το πλήθος δειγμάτων ανά κλάση πριν και μετά την υπερδειγματοληψία:

	ΔΕΙΓΜΑΤΑ	
	Πριν το Oversampling	Μετά το Oversampling
Κλάση 0	3.298	3.298
Κλάση 1	137	3.298

Πίνακας 9: Πλήθος δειγμάτων πριν και μετα το Oversampling

Πριν την υπερδειγματοληψία η διαφορά στο πλήθος δειγμάτων της κλάσης 0 και κλάσης 1 είναι χαοτικά μεγάλη αλλά μετά τα η αναλογία έφτασε ένα προς ένα.



Εικόνα 24: Ραβδόγραμμα δειγμάτων stroke πριν και μετά την υπερδειγματοληψία

3.3.3 Εκπαίδευση και έλεγχος επίδοσης αλγορίθμου

Εφόσον, τα δεδομένα είναι έτοιμα και κατάλληλα για την εκπαίδευση, το επόμενο βήμα είναι η επιλογή του κατάλληλου αλγορίθμου. Επειδή, το είδος του προβλήματος είναι της ταξινόμησης, επιλέχθηκαν οι αλγόριθμοι Decision Tree, Random Forest, KNeighbors και XGBoost, οι οποίοι είναι όλοι ταξινομητές-classifiers. Αρχικά, η εκπαίδευση έγινε στο σύνολο δεδομένων πριν το Oversampling και έπειτα στο νέο σύνολο δεδομένων μετά το Oversampling. Η πρόβλεψη έγινε στο σύνολο δεδομένων δοκιμής, που αφορά 1.473 περιπτώσεις ασθενών και αξιολογώντας πρώτα την ακρίβεια, αφού έγινε η παραμετροποίηση των αλγορίθμων για το βέλτιστο αποτέλεσμα.

Classifier	Ακρίβεια
Decision Tree	95,11 %
Random Forest	95,11 %
KNeighbors	95,11 %
XGBoost	94,84 %

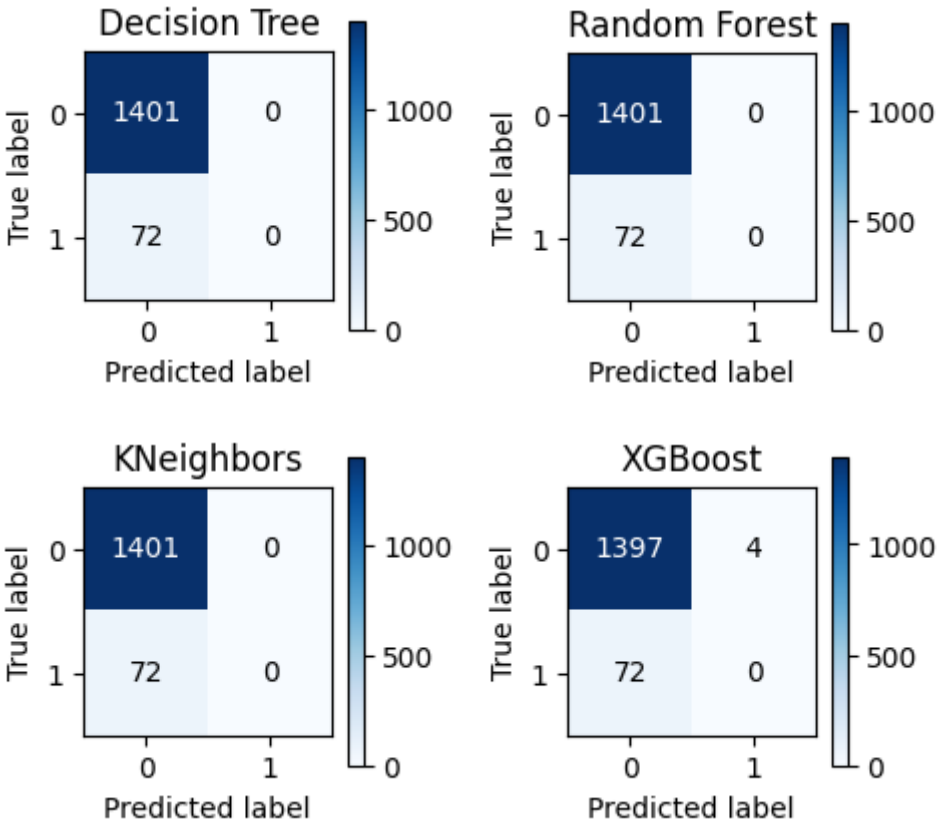
Πίνακας 10: Ποσοστό ακρίβειας για κάθε μοντέλο πριν το Oversampling

Όπως φαίνεται όλα τα μοντέλα παρουσιάζουν υψηλή ακρίβεια στο σύνολο δεδομένων δοκιμής, όμως αυτό θα ήταν καλό να ερευνηθεί περαιτέρω και για αυτό μια άλλη μετρική είναι η ανάκληση (Recall) της κάθε κλάσης, αυτό φαίνεται παρακάτω:

Classifier	**Recall**	
	Κλάση 0	Κλάση 1
Decision Tree	100 %	0 %
Random Forest	100 %	0 %
KNeighbors	100 %	0 %
XGBoost	100 %	0 %

Πίνακας 11: Ποσοστά Recall για όλα τα μοντέλα ανά κλάση πριν το Oversampling

Με την έρευνα της ανάκλησης παρατηρείται ότι το μοντέλο προέβλεψε όλες τις περιπτώσεις σαν κλάση 0. Στην πραγματικότητα αυτό σημαίνει ότι το μοντέλο προέβλεψε για όλους τους 1.473 ασθενείς ότι δεν έχουν δεχτεί εγκεφαλικό επεισόδιο, το οποίο είναι λάθος επειδή υπάρχουν 72 περιπτώσεις ασθενών που έχουν δεχτεί εγκεφαλικό επεισόδιο. Αυτό μπορεί να γίνει πιο εύκολα κατανοητό παρακολουθώντας τους πίνακες σύγχυσης για κάθε μοντέλο, τα δύο πρώτα είναι Decision Tree και Random Forest, ενώ τα 2 τελευταία οι υπόλοιποι αλγόριθμοι.



Εικόνα 25: Πίνακες σύγχυσης ανά μοντέλο πριν το Oversampling

Μπορεί κανείς να παρατηρήσει από τον πίνακα σύγχυσης ότι υπάρχουν 72 False positives και 0 True positives, όπως προαναφέρθηκε, το μοντέλο δεν μπορεί να προβλέψει ασθενείς που έχουν δεχτεί εγκεφαλικό. Αυτός είναι και ο κύριος λόγος που χρησιμοποιήθηκε η μέθοδος του Oversampling.

Επόμενο βήμα είναι η εκπαίδευση του μοντέλου μετά το Oversampling, χρησιμοποιήθηκαν οι ίδιοι αλγόριθμοι με προηγουμένως και αφού έγινε η παραμετροποίησή τους για το βέλτιστο αποτέλεσμα, ελέγχθηκε η μετρική της ακρίβειας, όπως φαίνεται στο παρακάτω πίνακα:

Classifier	**Ακρίβεια**
Decision Tree	72,44 %
Random Forest	72,84 %
KNeighbors	70,67 %
XGBoost	**73,86 %**

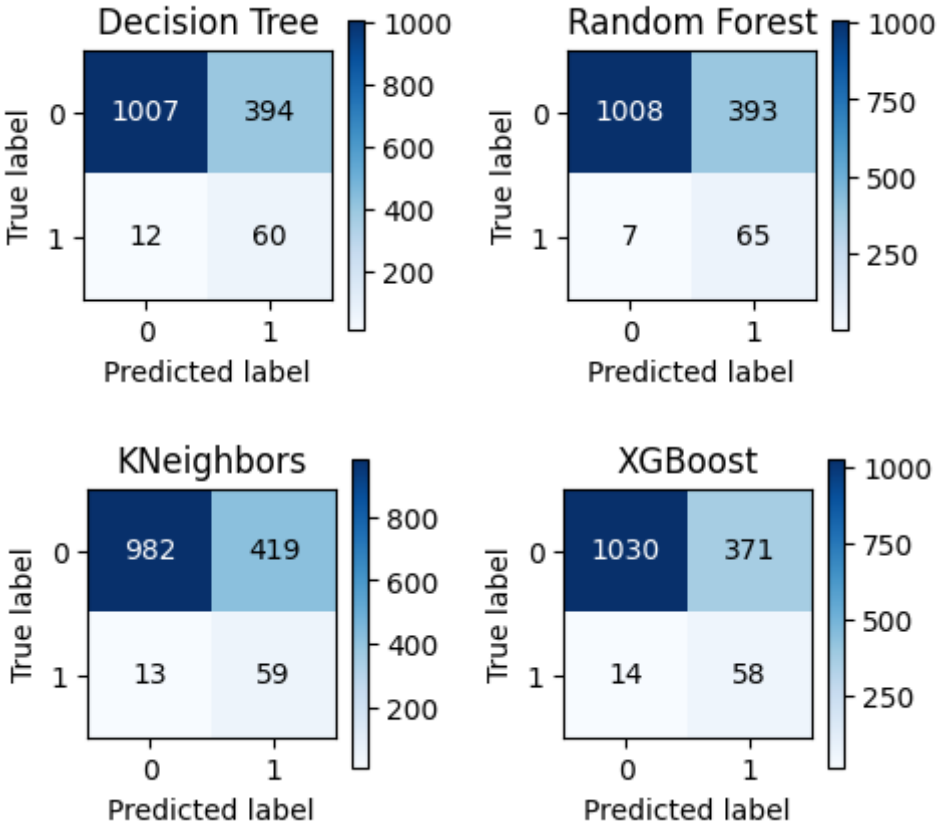
Πίνακας 12: Ποσοστό ακρίβειας για κάθε μοντέλο μετά το Oversampling

Με μια γρήγορη ματιά φαίνεται πως την καλύτερη επίδοση την έχει ο αλγόριθμος XGBoost. Σίγουρα όμως θα πρέπει να ελεγχθεί και η μετρική του Recall για να υπάρχει μια καλύτερη εικόνα, βλέπε παρακάτω:

Classifier	**Recall**		
	Κλάση 0	Κλάση 1	Μ.Ο
Decision Tree	72 %	83 %	77,61 %
Random Forest	**72 %**	**90 %**	**81,11 %**
KNeighbors	70 %	82 %	76,02 %
XGBoost	74 %	81 %	77,04 %

Πίνακας 13: Ποσοστά Recall για όλα τα μοντέλα ανά κλάση μετά το Oversampling

Πριν τον έλεγχο της μετρικής του Recall, ο αλγόριθμος XGBoost είχε την καλύτερη μετρική ακρίβειας. Στον παραπάνω πίνακα όμως δεν παρουσιάζει την καλύτερη μετρική Recall. Παρατηρείται ότι την καλύτερη μετρική recall παρουσιάζει ο αλγόριθμος Random Forest, ωστόσο θα ήταν καλό να ελεγχθεί επίσης και ο πίνακας σύγχυσης για το κάθε μοντέλο, τα δύο πρώτα είναι Decision Tree και Random Forest, ενώ τα 2 τελευταία KNeighbors και XGBoost, βλέπε παρακάτω:



Εικόνα 26: Πίνακες σύγχυσης ανά μοντέλο μετά το Oversampling

Στον αλγόριθμο **Decision Tree**, υπάρχουν 1007 TN, 60 TP, 394 FN και 12 FP.

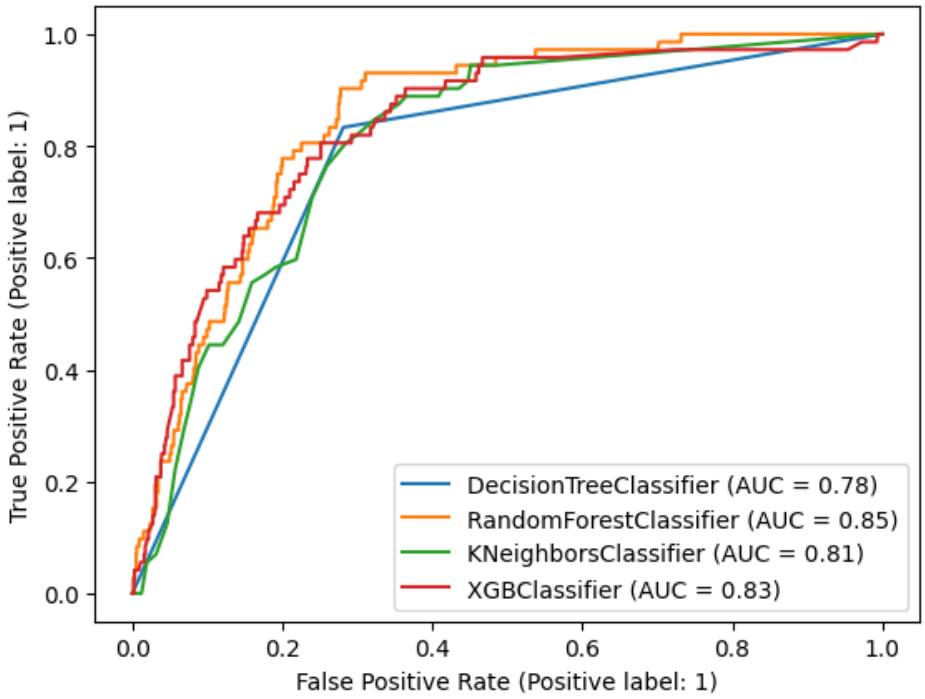
Στον αλγόριθμο **Random Forest**, υπάρχουν 1008 TN, 65 TP, 393 FN και 7 FP.

Στον αλγόριθμο **KNeighbors**, υπάρχουν 982 TN, 59 TP, 419 FN και 13 FP.

Στον αλγόριθμο **XGBoost**, υπάρχουν 1030 TN, 58 TP, 371 FN και 14 FP.

Όπως φαίνεται όλα τα μοντέλα είχαν καλή επίδοση αλλά πρέπει να επιλεχτεί μόνο ένα μοντέλο για να δοκιμαστεί και σε άλλα καινούργια άγνωστα δεδομένα. Γενικά, το καλύτερο μοντέλο θεωρείται αυτό που έχει μεγαλύτερες τιμές στα TP και TN, που μπορεί να προβλέψει σωστά και τις δύο κλάσεις. Στην περίπτωση αυτή όμως, ως καλύτερο και κυρίαρχο μοντέλο πρέπει να θεωρηθεί εκείνο το οποίο έχει την ικανότητα να προβλέψει τους περισσότερους ασθενείς που δέχτηκαν εγκεφαλικό επεισόδιο – την κλάση 1, δηλαδή πρέπει το TP να λαμβάνει την μεγαλύτερη τιμή. Ο αλγόριθμος Random Forest που χρησιμοποιήθηκε για το μοντέλο, προέβλεψε σωστά 65 περιστατικά ασθενών θετικά σε σύγκριση με τα 7 λανθασμένα. Επιπλέον προέβλεψε 393 περιπτώσεις ασθενών θετικά που στην πραγματικότητα δεν είχαν δεχτεί εγκεφαλικό, ενώ ο XGBoost, προέβλεψε λανθασμένα 371 περιπτώσεις αλλά παρουσίασε μόλις 58 σωστές προβλέψεις για την κλάση 1.

Καλό θα ήταν επίσης να παρουσιαστεί και η χαρακτηριστική καμπύλη λειτουργίας ή αλλιώς το ROC curve.

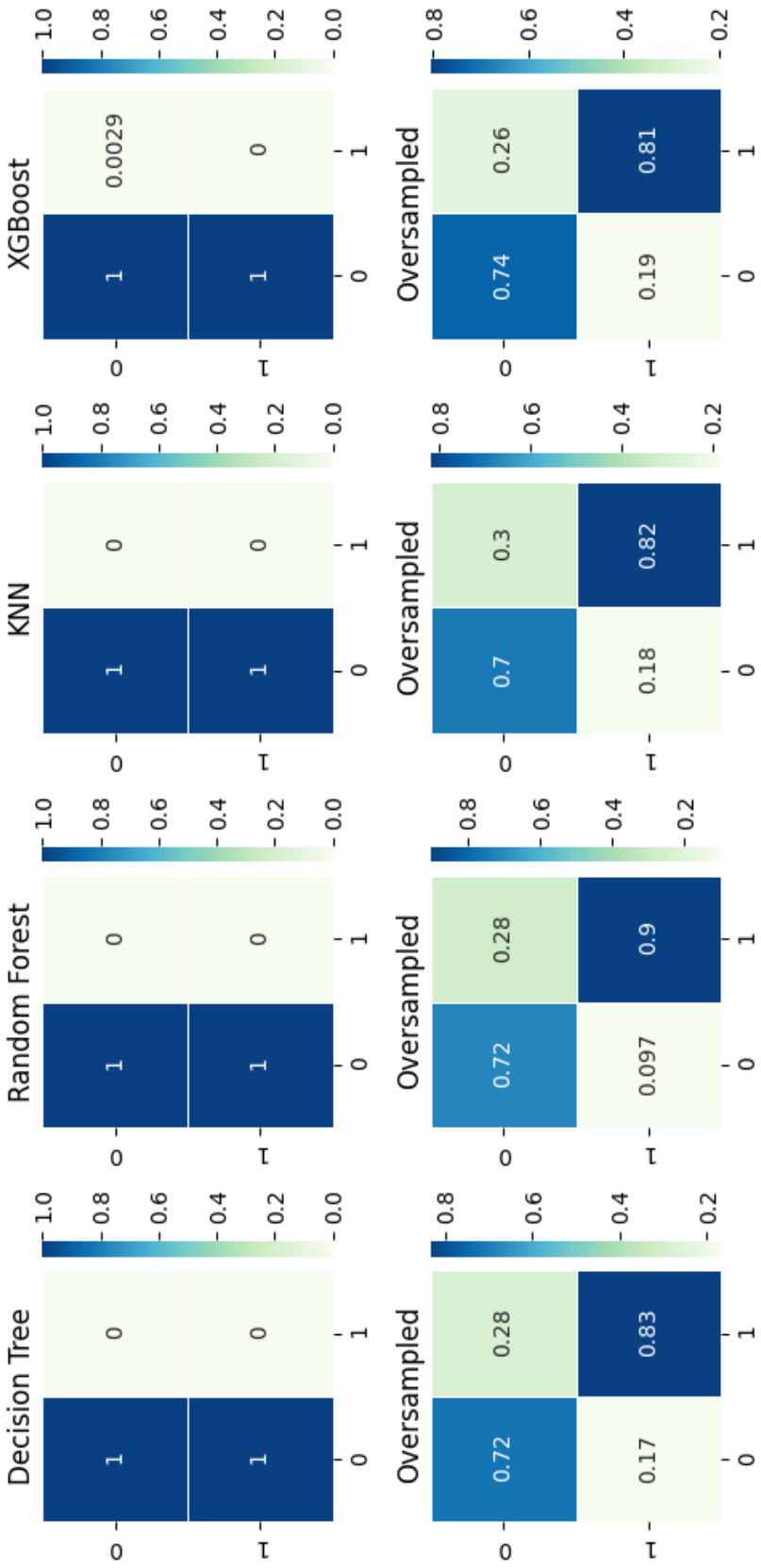


Πίνακας 14: Χαρακτηριστική καμπύλη λειτουργίας για όλα τα μοντέλα

Όπως προαναφέρθηκε ο πιο σημαντικός παράγοντας για την επιλογή του μοντέλου στο συγκεκριμένο πρόβλημα είναι η καλύτερη πρόβλεψη των ασθενών που δέχτηκαν το εγκεφαλικό. Το ερώτημα είναι θα άξιζε το μοντέλο να προβλέπει ελάχιστα χειρότερα το FN και να προβλέπει καλύτερα το TP. Η απάντηση βεβαίως και είναι ναι, διότι π.χ. είναι καλύτερο για έναν γιατρό να μπορέσει να προστατέψει όσο το δυνατόν περισσότερους ασθενείς που είναι όντως άρρωστοι και ας έχει κάποιους παραπάνω υποτιθέμενους άρρωστους ασθενείς, Δηλαδή η κλάση 1 έχει μεγαλύτερη βαρύτητα στο πρόβλημα αυτό.

Από την χαρακτηριστική καμπύλη ο Random Forest επιτυγχάνει 85 % σε σύγκριση με τον XGBoost που έχει 83 % επιτυχία. Ως καλύτερο μοντέλο επιλέγεται αυτό του Random Forest, επειδή είχε καλύτερη επίδοση πρώτον στην εξαιρετικά ικανοποιητική πρόβλεψη στις περιπτώσεις των εγκεφαλικών επεισοδίων, δεύτερων η κλάση 1 παρουσίασε εξαιρετικό Recall φτάνοντας στο 90 % και τρίτον η χαρακτηριστική καμπύλη ήταν ικανοποιητική.

Συμπεραίνοντας, τα ανομοιογενή δεδομένα πριν την αντιμετώπισή τους δεν παρουσίασαν καλά αποτελέσματα για την σωστή πρόβλεψη, ενώ έπειτα η μέθοδος του Resampling βοήθησε σε μεγάλο βαθμό τα δεδομένα ώστε το μοντέλο να επιτύχει το καλύτερο και βέλτιστο αποτέλεσμα. Παρακάτω στην **Εικόνα 27**, παρουσιάζονται οι πίνακες σύγχυσης πριν και μετα το Resampling για κάθε μοντέλο, με παράμετρο το ‘Normalization = True’, που σημαίνει το ποσοστό επιτυχίας για το κάθε TP, TN, FP και FN.



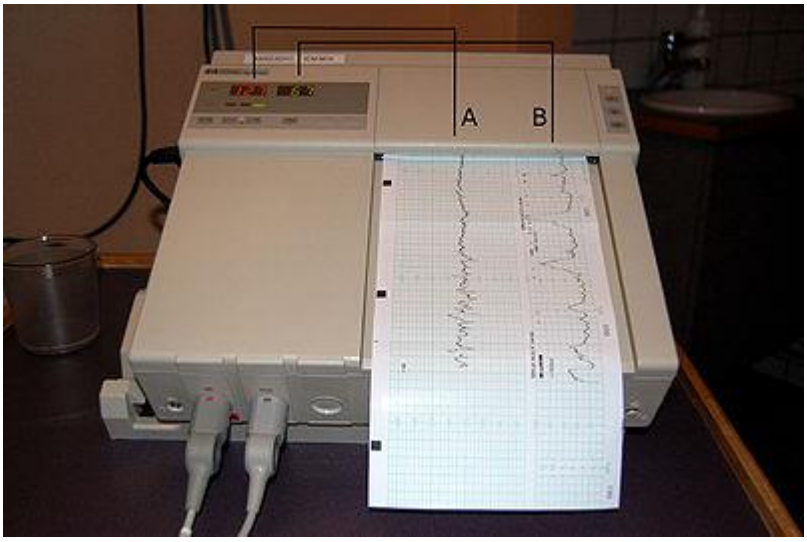
Εικόνα 27: Πίνακες σύγχυσης ανά μοντέλο πριν και μετά το Oversampling

Κεφάλαιο 4: Πρόβλεψη κατάστασης εμβρυϊκής καρδιακής λειτουργίας

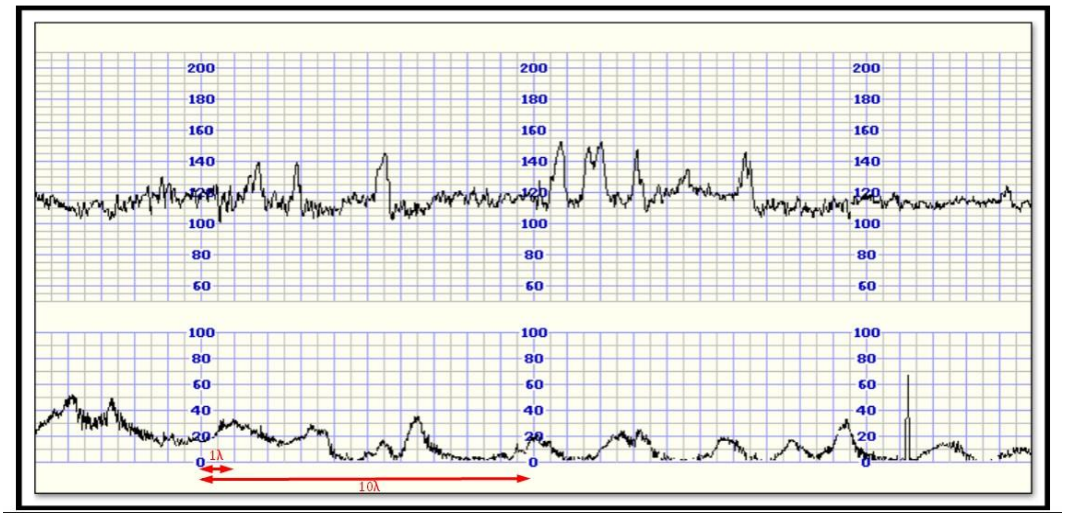
Ένας από τους σημαντικότερους παράγοντες για μια ασφαλή και φυσιολογική εγκυμοσύνη, που θα προφυλάσσει την υγεία της μητέρας και θα εξασφαλίζει την υγεία του μωρού, είναι η περιγεννητική παρακολούθηση του εμβρύου. Η περιγεννητική περίοδος ξεκινάει από την 22η ολοκληρωμένη εβδομάδα κύησης και τελειώνει την 7η ημέρα μετά τη γέννηση [60]. Ο θάνατος του εμβρύου αποτελεί πλέον περίπου το 50% των θανάτων μεταξύ 20 εβδομάδων κυήσεως έως την ηλικία ενός έτους, ενώ οι θάνατοι νεογνών αποτελούν μόνο το 25% των θανάτων αυτών [61]. Σε αυτή την περίοδο μια τεχνική για την παρακολούθηση της υγείας του εμβρύου και την αντιμετώπιση των νεογνικών θανάτων είναι η καρδιοτοκογραφία.

4.1 Η καρδιοτοκογραφία

Η καρδιοτοκογραφία είναι μια καταγραφή του εμβρυϊκού καρδιακού ρυθμού (Fetal Health Rate – FHR) και των συσπάσεων της μήτρας (Uterine Contractions - UC). Αποτελεί μία από τις συνηθέστερες διαγνωστικές τεχνικές για την εκτίμηση της μητρότητας και της της ευεξίας του εμβρύου κατά τη διάρκεια της εγκυμοσύνης και πριν από τον τοκετό. Ο γυναικολόγος μπορεί να εκτιμήσει την κατάσταση του εμβρύου παρατηρώντας τα δείγματα του καρδιοτοκογράφου [62]. Ο καρδιοτοκογράφος είναι το ιατρικό μηχάνημα με το οποίο πραγματοποιείται η συγκεκριμένη τεχνική, όπου οι συσπάσεις καταγράφονται από ένα πιεσόμετρο που στηρίζεται στην κοιλιά της μητέρας με έναν ελαστικό ιμάντα ενώ ο καρδιακός ρυθμός του εμβρύου καταγράφεται είτε με έναν υπέρηχο πομπού-δέκτη Doppler στην κοιλιά, ο οποίος ανιχνεύει τις καρδιακές κινήσεις του εμβρύου κατόπιν και τον καρδιακό ρυθμό, είτε εάν η συγκεκριμένη τεχνική δεν είναι ικανοποιητική, τότε χρησιμοποιείται ένα κλιπ, που είναι γνωστό ως το ηλεκτρόδιο κεφαλής του εμβρύου (Fetal Scalp Electrode - FSE), το οποίο ενσωματώνεται στην περιοχή του δέρματος του μωρού για να ανιχνεύσει το κύμα RR του εμβρυϊκού ηλεκτροκαρδιογραφήματος (Electrocardiogram – ECG) [62, 63]. Στην **Εικόνα 28** απεικονίζεται το μηχάνημα της καρδιοτοκογραφίας, όπου Α: καρδιακός ρυθμός του εμβρύου και Β: οι συσπάσεις της μήτρας, ενώ στην **Εικόνα 29** φαίνεται ένα παράδειγμα καρδιοτοκογραφήματος.



Εικόνα 28: Το μηχάνημα της καρδιοτοκογραφίας [65]



Εικόνα 29: Παράδειγμα γραφικής απεικόνισης της καρδιοτοκογραφίας [66]

Στο καρδιοτοκογράφημα – ΚΤΓ όπως φαίνεται παραπάνω υπάρχουν 2 γραφικές, όπου η πάνω απεικονίζει το ρυθμό του εμβρύου ενώ η κάτω απεικονίζει τις συσπάσεις τις μήτρας. Ο οριζόντιος άξονας παρουσιάζει τον ρυθμό καταγραφής, όπου κάθε 2 μικρά κουτάκια υποδηλώνουν το 1 λεπτό [67].

4.2 Χαρακτηριστικά καρδιοτοκογραφήματος

Ο μνημονικός κανόνας DR C BRAVADO βοηθάει τους γιατρούς να μπορέσουν να βγάλουν συμπέρασμα για το αν το έμβρυο είναι υγιές ή όχι. Το συμπέρασμα αυτό ονομάζεται εμβρυϊκή κατάσταση (NSP) και χωρίζεται σε 3 κατηγορίες το Φυσιολογικό

- Normal, το ύποπτο - Suspicious και το παθολογικό - Pathologic. Τα βήματα του κανόνα χωρίζονται όπως φαίνεται παρακάτω [68]:

DR	Define risk
C	Contractions
BRa	Baseline rate
V	Variability
A	Accelerations
D	Decelerations
O	Overall impression

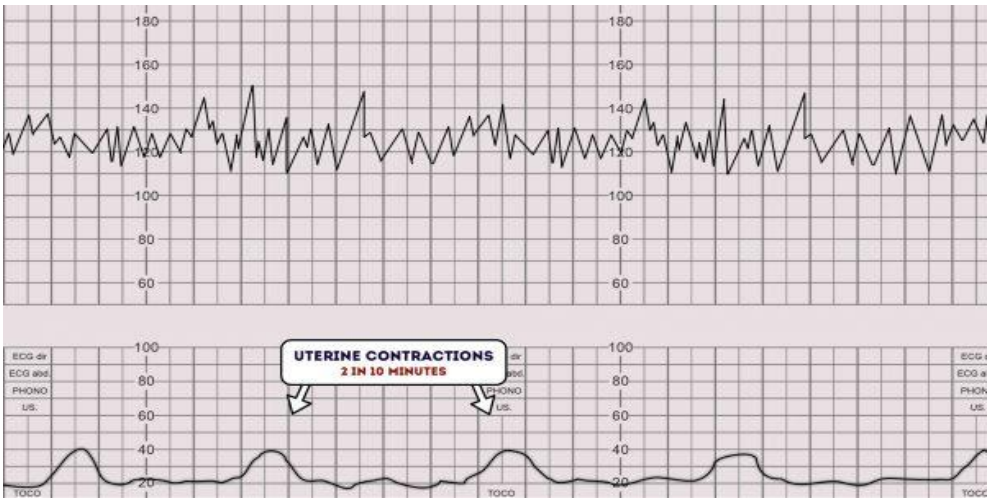
Πίνακας 15: Μνημονικός κανόνας DR C BRAVADO

Define risk

Πρώτο και πολύ σημαντικό είναι ο προσδιορισμός του κινδύνου της γέννας, αν η γυναίκα που πρόκειται να γεννήσει παρουσιάζει διάφορες επιπλοκές στην υγεία της ή ίσως έχει κάποια αρρώστια.

Contractions

Επόμενο βήμα είναι η καταγραφή των συσπάσεων που εμφανίζονται σε διάστημα 10 λεπτών και σημαντικό είναι να γίνει η αξιολόγηση της έντασης αλλά και της διάρκειας.

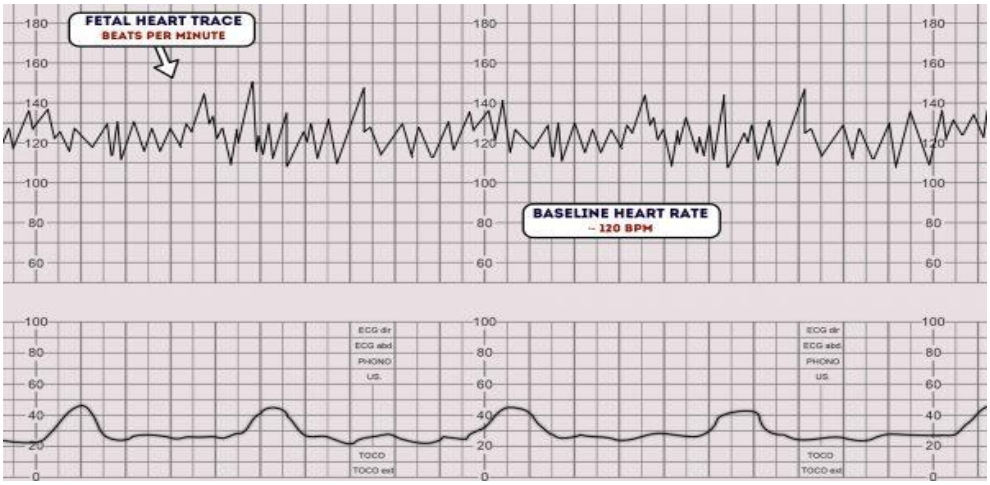


Εικόνα 30: Συσπάσεις της μήτρας [68]

Baseline rate - FHR

Ο βασικός ρυθμός ή ο καρδιακός ρυθμός του εμβρύου είναι ο μέσος καρδιακός ρυθμός του εμβρύου σε διάστημα 10 λεπτών και η φυσιολογική τιμή είναι από 110-160 beats per

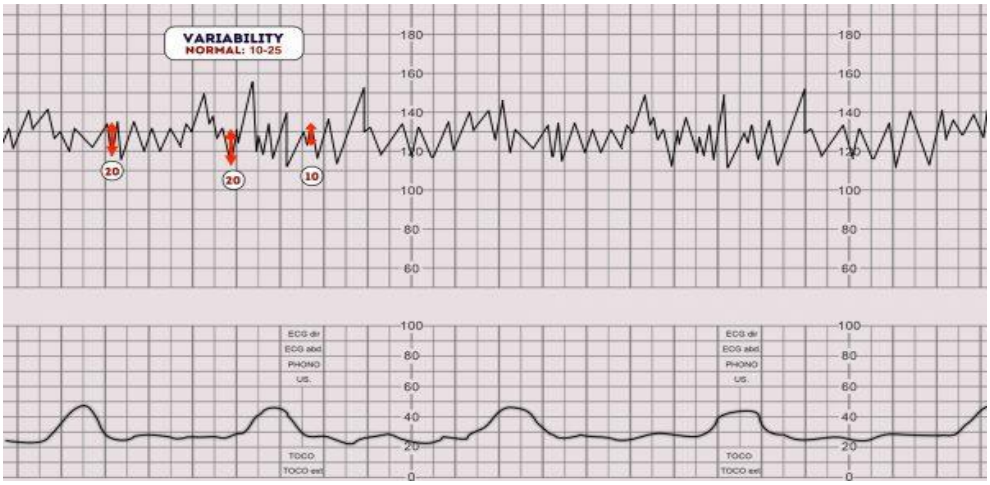
minute – bpm. Σε περίπτωση που για πάνω από 10 λεπτά ο μέσος καρδιακός ρυθμός είναι πάνω από 160 bpm τότε παρουσιάζεται η ταχυκαρδία στο έμβρυο, ενώ αν ο καρδιακός ρυθμός είναι κάτω από 110 bpm για περισσότερο από 10 λεπτά τότε το έμβρυο παρουσιάζει βραδυκαρδία [67].



Εικόνα 31: Βασικός εμβρυϊκός καρδιακός ρυθμός [68]

Variability

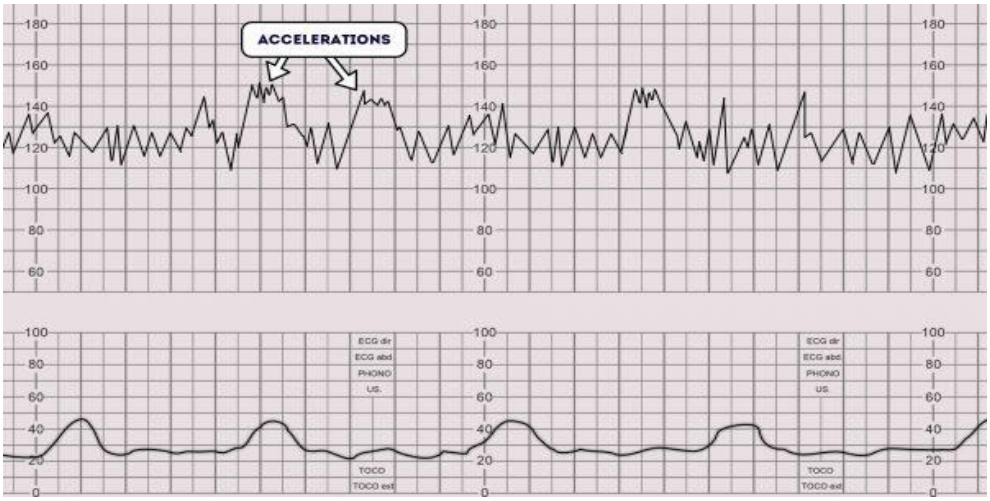
Η μεταβλητότητα μπορεί να μετρηθεί με την ανάλυση σε περίοδο ενός λεπτού του ΚΤΓ και την αξιολόγηση της διαφοράς μεταξύ του υψηλότερου και χαμηλότερου βασικού καρδιακού ρυθμού κατά τη διάρκεια αυτής της περιόδου [69]. Η μεταβλητότητα μπορεί να χαρακτηριστεί ως Καθησυχαστική – Reassuring από 5-15 bpm, Μη καθησυχαστική – Non-reassuring, αν είναι μικρότερη από 5 bpm μεταξύ 30-50 λεπτών και αν είναι μεγαλύτερη από 25 bpm μεταξύ 15-25 λεπτών και Ασυνήθιστη – Abnormal αν είναι μικρότερη από 5 bpm για περισσότερο από 50 λεπτά και αν είναι μεγαλύτερη από 25 bpm για περισσότερο από 25 λεπτά [68].



Εικόνα 32: Μεταβλητότητα [68]

Accelerations

Όταν ο βασικός εμβρυϊκός καρδιακός ρυθμός αυξάνεται απότομα λαμβάνοντας τιμή μεγαλύτερη από 15 bpm για περισσότερο από 15 δευτερόλεπτα ονομάζεται επιτάχυνση και συχνά σχετίζεται με την εμβρυϊκή δραστηριότητα και θεωρείται ως ένδειξη ότι το έμβρυο είναι υγιές.



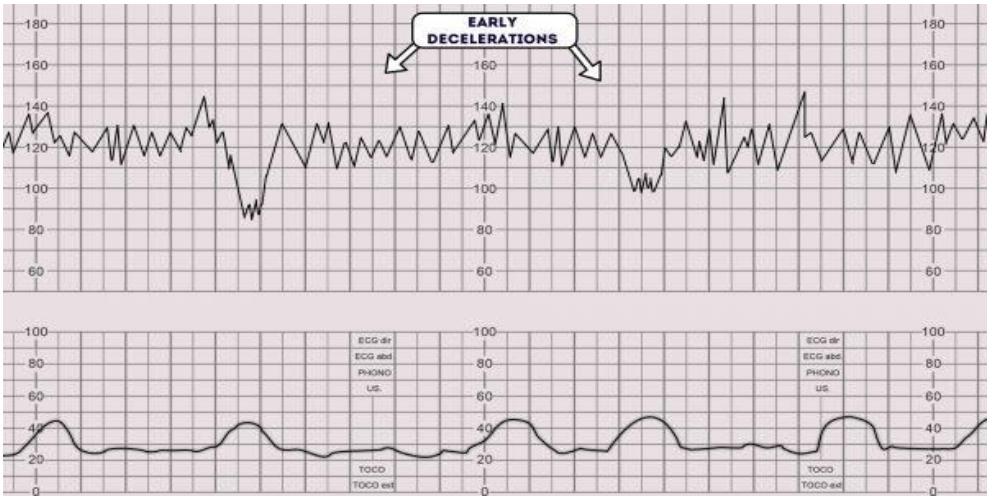
Εικόνα 33: Επιταχύνσεις [68]

Decelerations

Οι επιβραδύνσεις είναι περιοδικές, παροδικές μειώσεις του καρδιακού ρυθμού του εμβρύου, που συνήθως συνδέονται με τις συσπάσεις της μήτρας. Οι επιβραδύνσεις χωρίζονται κυρίως σε τέσσερις τύπους ανάλογα με τη μορφή και το χρόνο τους σε σύγκριση με τις συσπάσεις της μήτρας. Οι τύποι είναι οι Πρώιμες – Early, οι Όψιμες – Late, οι Μεταβαλλόμενες – Variable και οι Παρατεταμένες – Prolonged. Οι συσπάσεις της μήτρας πρέπει να παρακολουθούνται επαρκώς προκειμένου να ταξινομηθεί σωστά μια επιβράδυνση [69].

Πρώιμες επιβραδύνσεις:

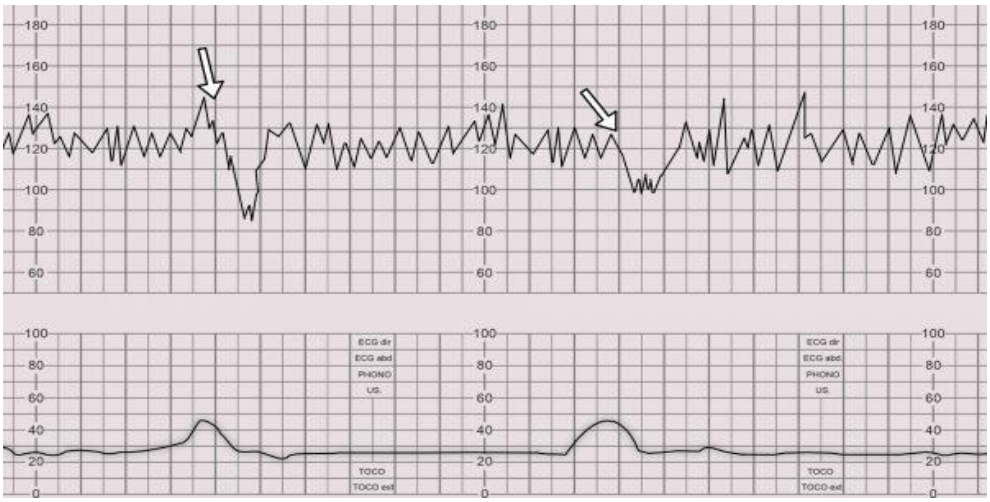
Οι πρώιμες επιβραδύνσεις εμφανίζονται όταν ξεκινά η σύσπαση της μήτρας και επανέρχονται όταν η σύσπαση της μήτρας διακόπτεται και προκαλούνται από τη συμπίεση του κεφαλιού του εμβρύου κατά τη διάρκεια μιας σύσπασης [69].



Εικόνα 34: Πρώιμες επιβραδύνσεις [68]

Όψιμες επιβραδύνσεις:

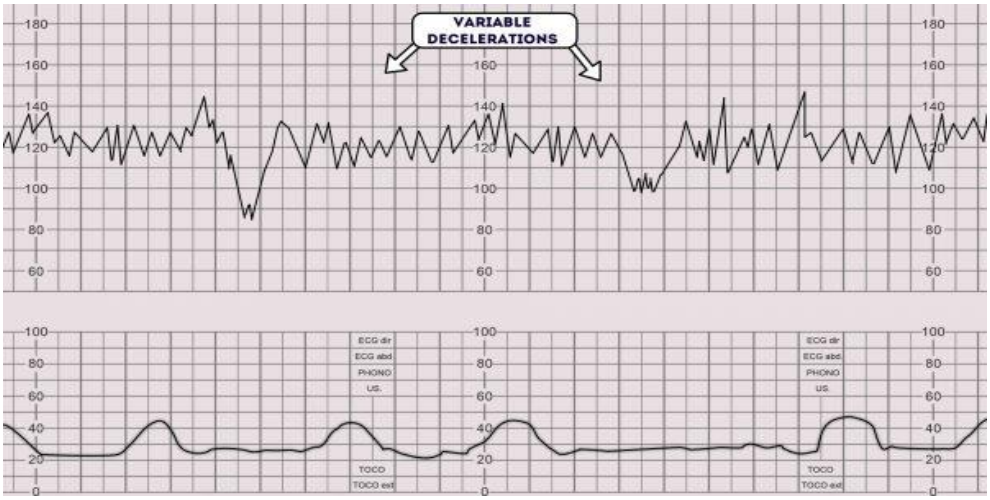
Οι όψιμες επιβραδύνσεις εμφανίζονται στο ανώτατο σημείο της σύσπασης της μήτρας και επιστρέφουν στην φυσιολογική τιμή αφού τελειώσει η σύσπαση. Αυτός ο τύπος επιβράδυνσης υποδεικνύει ότι δεν υπάρχει επαρκής ροή αίματος στη μήτρα και τον πλακούντα.



Εικόνα 35: Όψιμες επιβραδύνσεις [68]

Μεταβαλλόμενες επιβραδύνσεις:

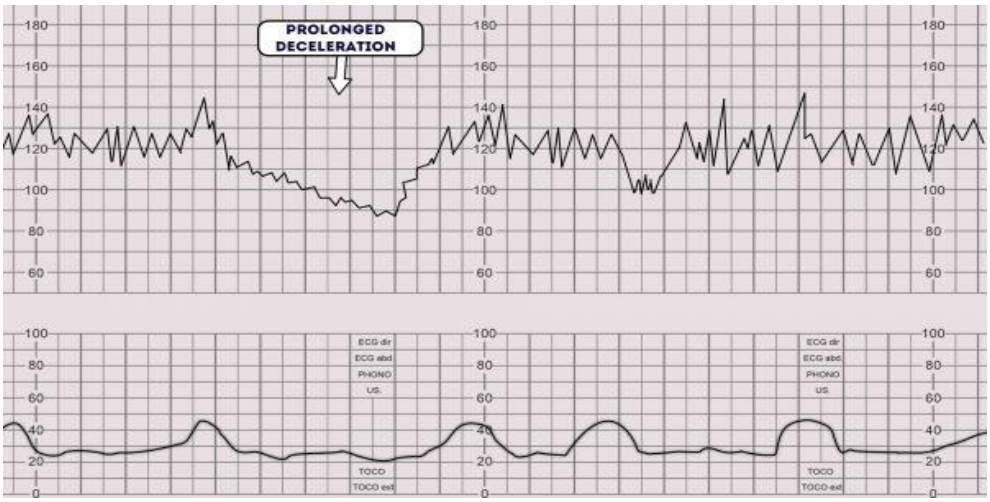
Οι μεταβαλλόμενες επιβραδύνσεις αποτελούν μια περιγραφή των επιβραδύνσεων του εμβρυϊκού καρδιακού ρυθμού που μεταβάλλονται τόσο ως προς το χρόνο όσο και ως προς το μέγεθος και ενδέχεται να παρουσιάσουν αυξημένη μεταβλητότητα του εμβρυϊκού καρδιακού ρυθμού [69].



Εικόνα 36: Μεταβαλλόμενες επιβραδύνσεις [68]

Παρατεταμένες επιβραδύνσεις:

Μια επιβράδυνση που διαρκεί περισσότερο από 2 λεπτά ονομάζεται παρατεταμένη επιβράδυνση, αν η επιβράδυνση συνεχιστεί για 2-3 λεπτά, πρόκειται για μη καθησυχαστική κατάσταση, αλλά αν συνεχιστεί περισσότερο από 3 λεπτά, τότε πρόκειται για μια ασυνήθιστη κατάσταση.



Εικόνα 37: Παρατεταμένες επιβραδύνσεις [68]

Overall Impression

Τέλος αφού ο γιατρός αξιολογήσει όλες τις παραμέτρους του καρδιοτοκογραφήματος τότε θα είναι ικανός να συμπεράνει ποια είναι η κατάσταση του εμβρύου και να την κατηγοριοποιήσει σε μια από τις τρεις καταστάσεις, αν είναι φυσιολογική, ύποπτη ή παθολογική [68].

4.3 Εφαρμογή της Μηχανικής Μάθησης

Όπως προαναφέρθηκε προηγουμένως, η υγεία του εμβρύου είναι ένα εξαιρετικά σημαντικό πρόβλημα την σήμερον ημέρα, το οποίο μπορεί σίγουρα να αντιμετωπιστεί. Στην ενότητα αυτή θα παρουσιαστεί μια λύση - μέθοδος με την βοήθεια της μηχανικής μάθησης, για την πρόβλεψη της καρδιακής εμβρυϊκής λειτουργίας ώστε να μειωθεί ο αριθμός των θανάτων των εμβρύων ή των νεογνών. Για την μελέτη αυτή χρησιμοποιήθηκε ένα συγκεκριμένο σύνολο δεδομένων που αφορά πληροφορίες για περιπτώσεις γυναικών κατά την γέννα από το *UCI Machine Learning Repository* [70]. Να σημειωθεί ότι το σύνολο των δεδομένων αυτό χρησιμοποιήθηκε για την πρόβλεψη τριών κλάσεων. Πιο συγκεκριμένα το σύνολο δεδομένων αφορά πληροφορίες από καρδιοτοκογραφήματα που έχουν συλλεχθεί, επεξεργαστεί σαν σύνολο δεδομένων. Οι πληροφορίες αυτές αναλύθηκαν από ειδικούς της επιστήμης και κατηγοριοποιήθηκαν σε τρεις κλάσεις, οι οποίες υποδηλώνουν την κατάσταση της εμβρυϊκής καρδιακής λειτουργίας.

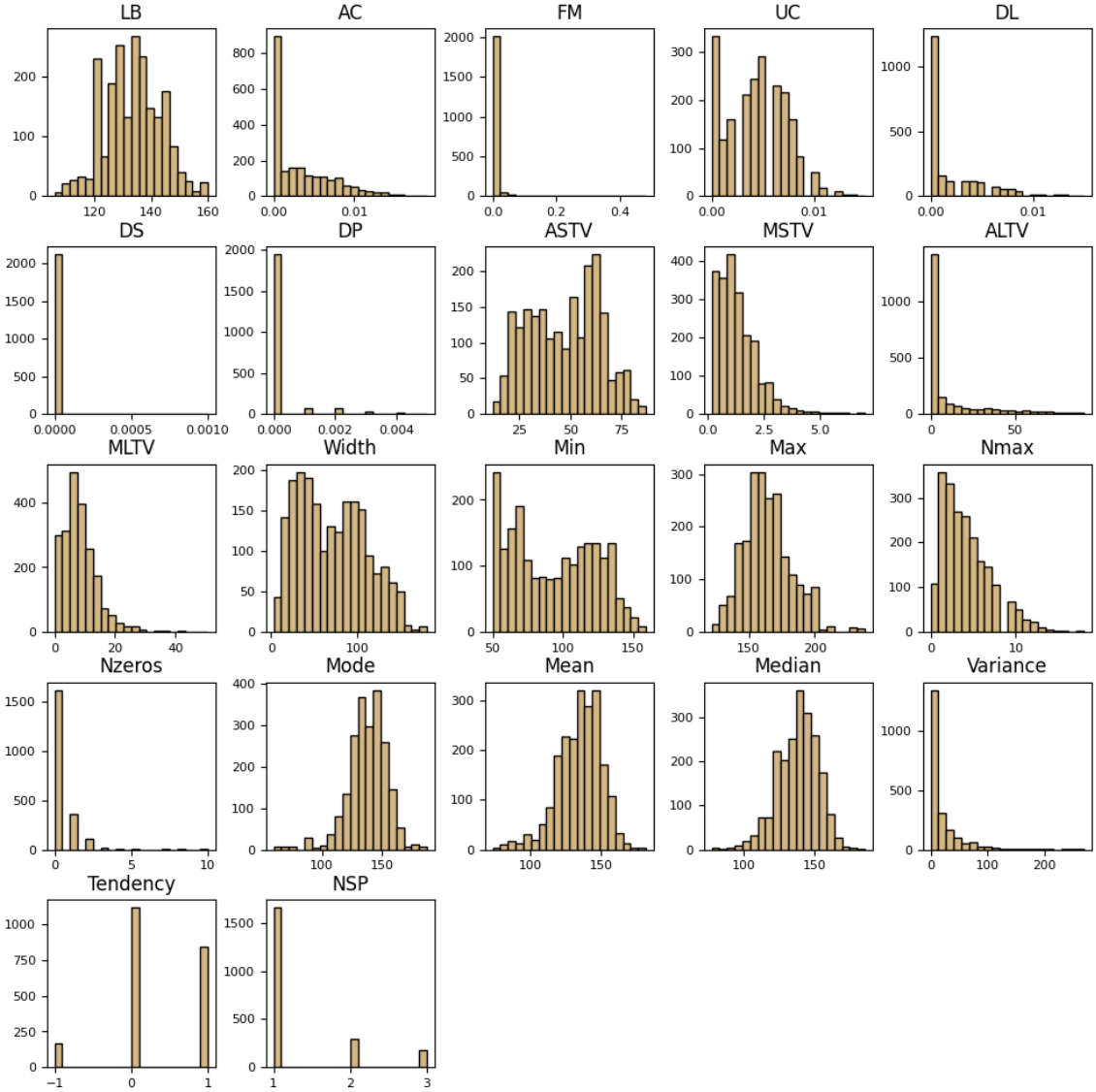
4.3.1 Ανάλυση και επεξεργασία συνόλου δεδομένων

Αρχικά, θα ήταν πολύ χρήσιμο να γίνει μια απεικόνιση των χαρακτηριστικών των δεδομένων που χρησιμοποιήθηκαν στο πρόβλημα αυτό. Το σύνολο δεδομένων αφορά 2.126 περιπτώσεις εγκύων γυναικών με 22 διαφορετικά χαρακτηριστικά – πληροφορίες από καρδιοτοκογραφήματα με ένα από αυτά, το NSP, να παρουσιάζει την κατάσταση της καρδιακής λειτουργίας του εμβρύου η οποία είναι απαραίτητο να προβλεφθεί και αλλιώς μπορεί να πάρει την ονομασία ως στόχος. Το χαρακτηριστικό αυτό λαμβάνει 3 τιμές, την φυσιολογική κατάσταση – Normal, ύποπτη κατάσταση – Suspect και παθολογική – Pathologic. Όπως μπορεί κανείς να αντιληφθεί η πρώτη δηλώνει ότι η κατάσταση του εμβρύου είναι φυσιολογική, η δεύτερη να σημαίνει ότι η έγκυος χρειάζεται περαιτέρω παρακολούθηση της κατάστασής της, ενώ στην τελευταία και η πιο σημαντική κατάσταση ο γιατρός πρέπει να αναλύσει και εξετάσει καλύτερα διότι υπάρχει μεγάλη πιθανότητα το βρέφος να αντιμετωπίζει κάποιο σοβαρό πρόβλημα. Ο παρακάτω **Πίνακας 16** απεικονίζει τις πληροφορίες για τα χαρακτηριστικά των δεδομένων:

Χαρακτηριστικά	Πληροφορίες
LB	Μεταβλητότητα εμβρυϊκού καρδιακού ρυθμού σε κτύπους/λεπτό (FHR)
AC	Επιταχύνσεις ανά δευτερόλεπτα
FM	Κινήσεις του εμβρύου ανά δευτερόλεπτο
UC	Συσπάσεις της μήτρας ανά δευτερόλεπτο
DL	Πρώιμες επιβραδύνσεις ανά δευτερόλεπτο
DS	Αργές επιβραδύνσεις ανά δευτερόλεπτο
DP	Παρατεταμένες επιβραδύνσεις ανά δευτερόλεπτο
ASTV	Ποσοστό χρόνου με ασυνήθιστη βραχυπρόθεσμη μεταβλητότητα
MSTV	Μέση τιμή της βραχυπρόθεσμης μεταβλητότητας
ALTV	Ποσοστό χρόνου με ασυνήθιστη μακροχρόνια μεταβλητότητα
MLTV	Μέση τιμή της μακροχρόνιας μεταβλητότητας
Width	Πλάτος του ιστογράμματος FHR
Min	Ελάχιστη τιμή του ιστογράμματος FHR
Max	Μέγιστη τιμή του ιστογράμματος FHR
Nmax	Αριθμός κορυφών στο ιστόγραμμα
Nzeros	Αριθμός μηδενικών στο ιστόγραμμα
Mode	Ιστόγραμμα επικρατούσας τιμής
Mean	Ιστόγραμμα μέσης τιμής
Median	Ιστόγραμμα διαμέσου τιμής
Variance	Ιστόγραμμα διασποράς
Tendency	Ιστόγραμμα τάσης
NSP	Κατάσταση του εμβρύου

Πίνακας 16: Χαρακτηριστικά συνόλου δεδομένων καρδιοτοκογραφίας

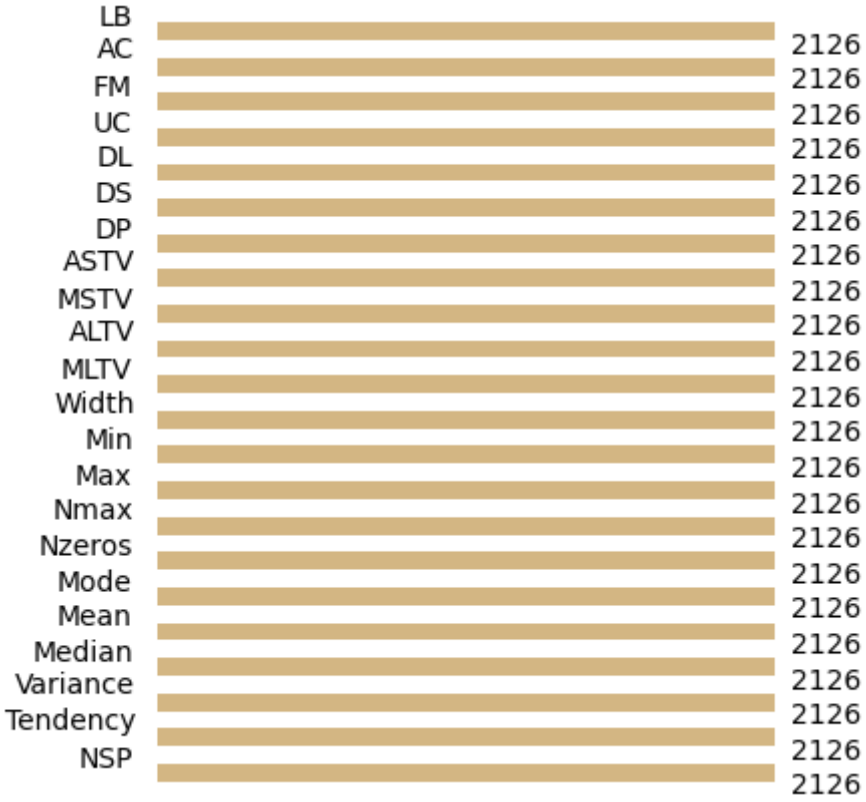
Μια υπόθεση που μπορεί να γίνει είναι ότι μάλλον το συγκεκριμένο σύνολο αποτελείται **μόνο** από αριθμητικές τιμές. Επίσης φαίνεται ότι τα βασικά χαρακτηριστικά είναι η μεταβλητότητα του εμβρυϊκού καρδιακού ρυθμού, οι κινήσεις του εμβρύου, οι επιταχύνσεις και οι επιβραδύνσεις, ενώ τα υπόλοιπα χαρακτηριστικά εξαρτώνται από αυτά και ειδικότερα από το FHR. Στο επόμενο βήμα θα πρέπει να υπάρξει μια εικόνα για το τι τιμές λαμβάνει το κάθε χαρακτηριστικό σε μορφή ιστογράμματος, αυτό φαίνεται στην **Εικόνα 38**:



Εικόνα 38: Γραφική αναπαράσταση για κάθε χαρακτηριστικό

Η υπόθεση που έγινε προηγουμένως για τις αριθμητικές τιμές είναι σωστή και αυτό είναι προφανές με την παρατήρηση της παραπάνω εικόνας. Με μια πρώτη ματιά φαίνεται τα χαρακτηριστικά Mode, Mean και Median ταυτίζονται σε μεγάλο βαθμό και αυτό είναι λογικό αφού υποδηλώνουν μέσες τιμές, όπως επίσης μοιάζουν τα χαρακτηριστικά Nmax και Nzeros αφού προσδιορίζουν τα πλήθη κορυφών και μηδενικών του ιστογράμματος FHR. Φυσιολογικά ιστογράμματα παρουσιάζουν τα ASTV και ALTV αφού το εύρος τιμών τους είναι από μηδέν έως εκατό επειδή υποδηλώνουν ποσοστά. Παρατηρείται επίσης ότι τα χαρακτηριστικά LB, Min και Max παρουσιάζουν ίδιο εύρος τιμών αφού πρόκειται για ίδια ιστογράμματα που απεικονίζουν διαφορετικές καταστάσεις, άρα είναι φυσιολογικά. Το χαρακτηριστικό NSP φαίνεται ότι λαμβάνει τιμές ‘1’, ‘2’ και ‘3’

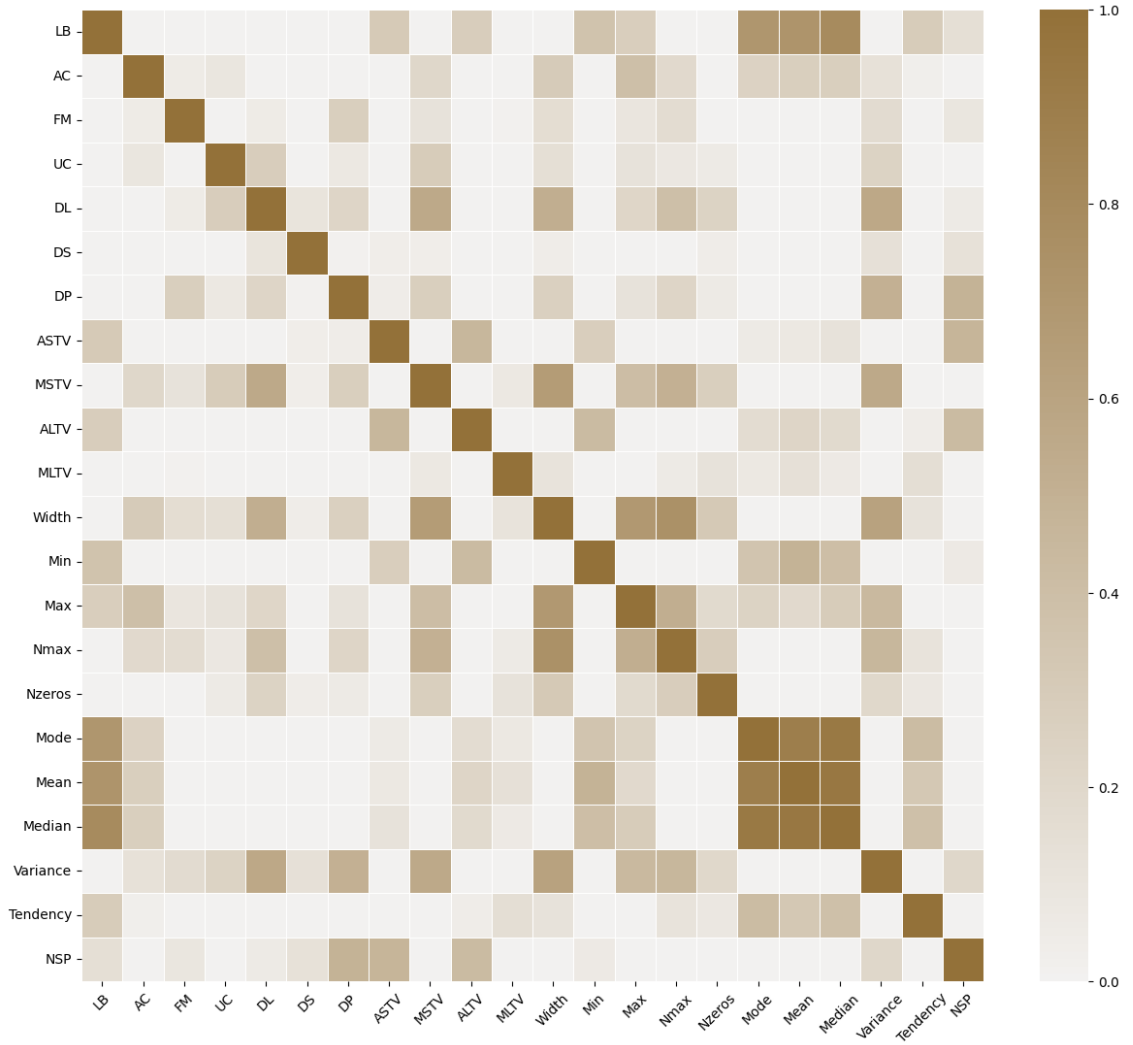
πράγμα που δεν είναι πρόβλημα αφού στην συνέχεια όταν εκπαιδευτεί το μοντέλο οι τιμές πρέπει να είναι όλες αριθμητικές. Το θα '1' συμβολίζει την φυσιολογική κατάσταση 'Normal', το '2' την ύποπτη κατάσταση 'Suspect' ενώ το '3' την παθολογική κατάσταση 'Pathologic'. Αφού όλο το σύνολο δεδομένων φαίνεται ότι όλα τα χαρακτηριστικά έχουν ιδανικές τιμές, επόμενο βήμα είναι να γίνει έλεγχος αν υπάρχουν ελλιπείς τιμές στο σύνολο αυτό αφού τα ιστογράμματα δεν βοηθούν στο συγκεκριμένο έλεγχο.



Εικόνα 39: Ραβδόγραμμα ελλιπών τιμών για κάθε χαρακτηριστικό

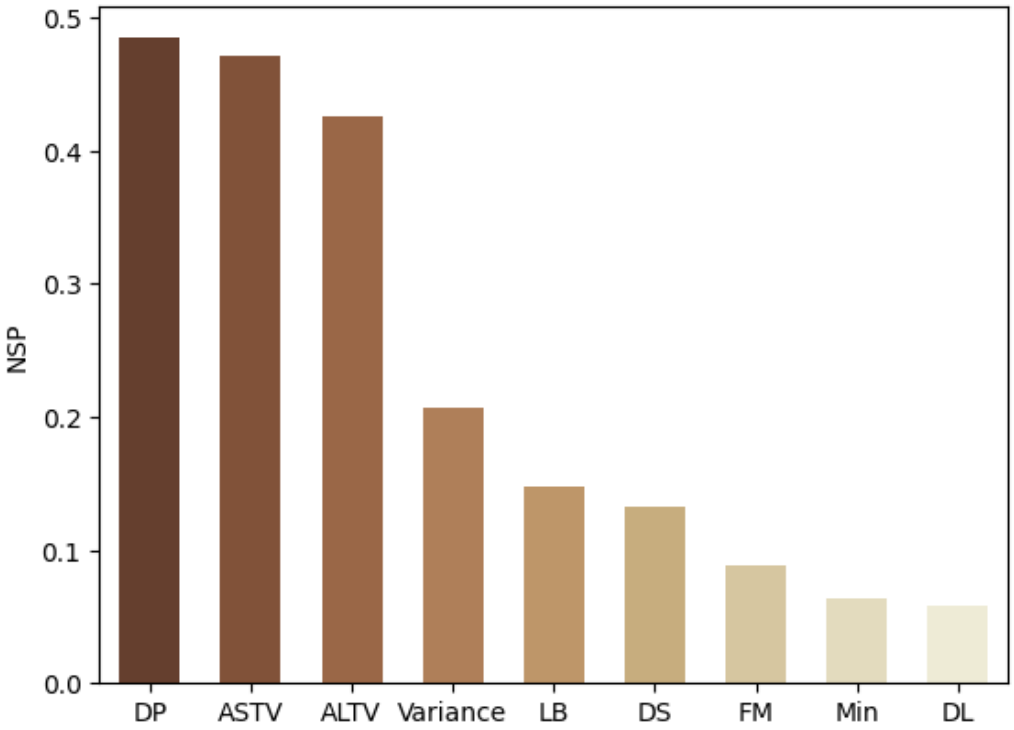
Όπως φαίνεται στο παραπάνω ραβδόγραμμα δεν υπάρχει καμία ελλιπής τιμή, άρα τα δεδομένα είναι εντάξει σε αυτή την φάση.

Στο επόμενο σημαντικό βήμα που πρέπει να γίνει είναι ο έλεγχος της συσχέτισης όλων των χαρακτηριστικών μεταξύ τους διότι βοηθάει στην καλύτερη κατανόηση των χαρακτηριστικών αν υπάρχουν σοβαρές συσχετίσεις που πρέπει να ληφθούν υπόψιν. Παρακάτω απεικονίζεται στην ο χάρτης θερμότητας του πίνακα συσχέτισης Pearson για κάθε χαρακτηριστικό:



Εικόνα 40: Πίνακας συσχέτισης Pearson

Παρατηρείται ότι τα Mode, Mean και Median έχουν υψηλή συσχέτιση μεταξύ τους όπως και αναφέρθηκε προηγουμένως στα ιστογράμματα. Πολύ σημαντικό όμως είναι η συσχέτιση όλων των χαρακτηριστικών σε σχέση με το χαρακτηριστικό της κατάστασης του εμβρύου για αυτό και στην **Εικόνα 41** παρουσιάζεται ένα ραβδόγραμμα με τις υψηλότερες συσχετίσεις καθώς και ο **Πίνακας 17** τα ποσοστά συσχέτισης.



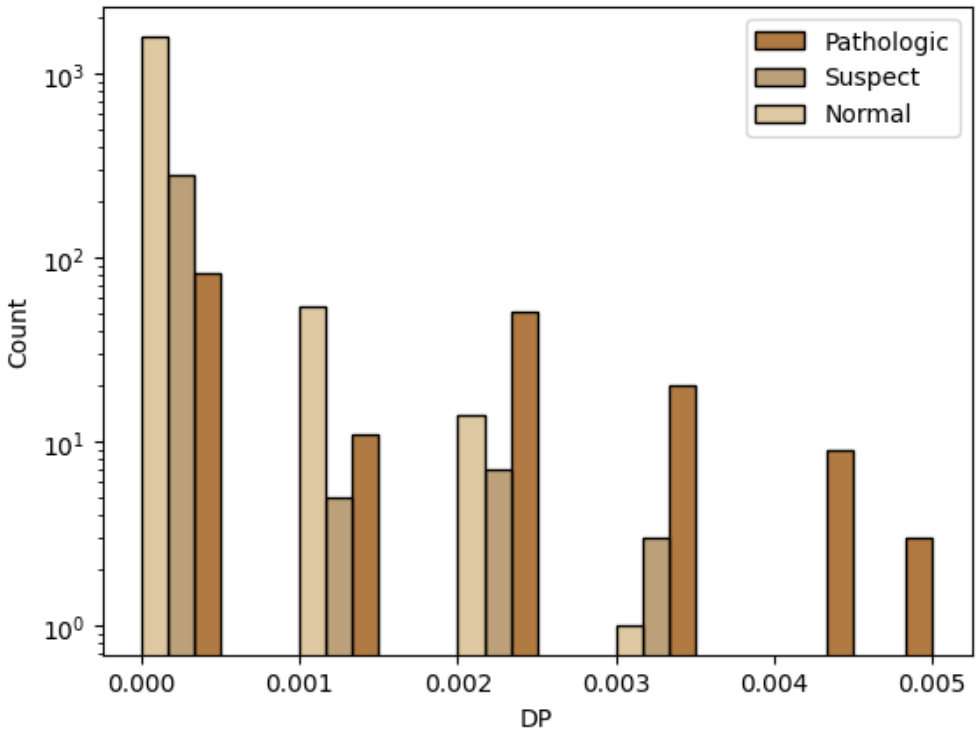
Εικόνα 41: Ραβδόγραμμα χαρακτηριστικών συσχέτισης με NSP

Χαρακτηριστικά	Συσχέτιση
DP	0,485
ASTV	0,471
ALTV	0,426
Variance	0,207
LB	0,181
DS	0,132
FM	0,089
Min	0,063
DL	0,059

Πίνακας 17: Ποσοστά συσχέτισης για κάθε χαρακτηριστικό με το NSP

Όπως φαίνεται πολύ ξεκάθαρα οι παρατεταμένες επιβραδύνσεις, τα ποσοστά χρόνου με ασυνήθιστη βραχυπρόθεσμη μεταβλητότητα και τα ποσοστά χρόνου με ασυνήθιστη μακροπρόθεσμη μεταβλητότητα παρουσιάζουν τα υψηλότερα ποσοστά συσχέτισης με το χαρακτηριστικό της κατάστασης του εμβρύου.

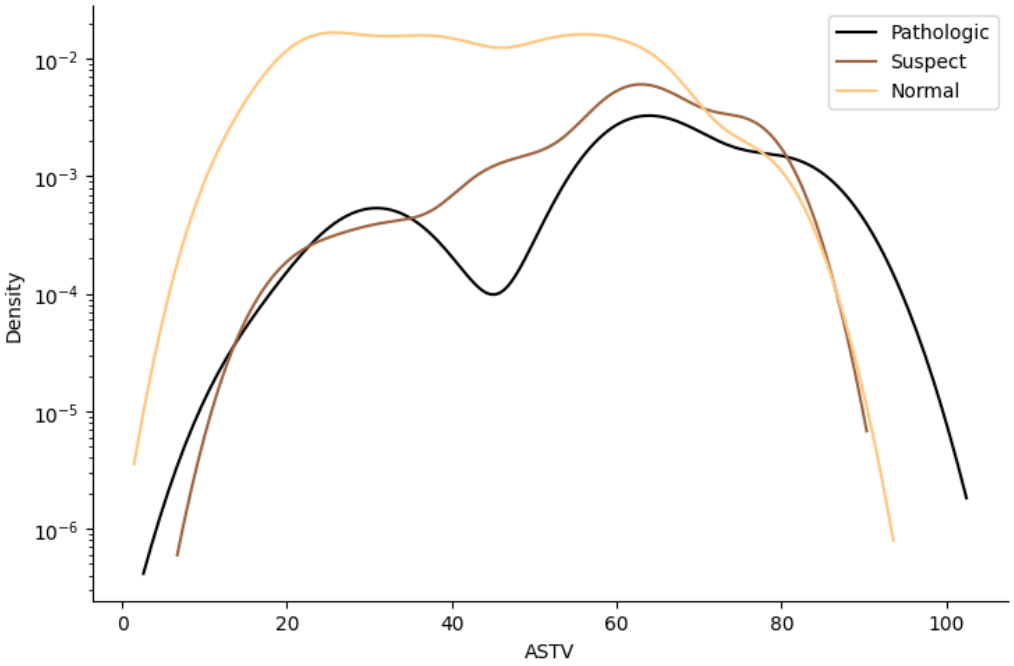
Για αυτό τον λόγο είναι εξαιρετικά σημαντικό να αναλυθούν παραπάνω τα συγκεκριμένα χαρακτηριστικά σε σχέση πάντα με το χαρακτηριστικό NSP. Αρχικά, θα παρουσιαστεί στην παρακάτω **Εικόνα 42**, το ιστόγραμμα σε λογαριθμική κλίμακα που αφορά το πλήθος των παρατεταμένων επιβραδύνσεων ανάλογα με την κατάσταση του εμβρύου.



Εικόνα 42: Πλήθος τιμών DP ανά NSP

Αρχικά, παρατηρείται ότι όσο η τιμή των παρατεταμένων επιβραδύνσεων στην φυσιολογική κατάσταση αυξάνεται τόσο περισσότερο μειώνεται το πλήθος των τιμών αυτό. Στην περίπτωση της ύποπτης κατάστασης παρουσιάζεται σχεδόν το ίδιο, όσο αυξάνεται η τιμή τόσο μειώνεται και πάλι το πλήθος, ενώ στην παθολογική κατάσταση δεν φαίνεται κάποια συγκεκριμένη χρήσιμη πληροφορία από το πλήθος των παρατεταμένων επιβραδύνσεων παρά μόνο ότι μπορεί να θεωρηθεί σίγουρα κρίσιμη τιμή υψηλότερη από 0,003 διότι εκεί τα πλήθη είναι μεγαλύτερα από ότι της φυσιολογικής και ύποπτης κατάστασης.

Στην συνέχεια σειρά έχει ο έλεγχος των ποσοστών χρόνου με ασυνήθιστη βραχυπρόθεσμη μεταβλητότητα, όπως παρουσιάζεται σε λογαριθμική κλίμακα με καμπύλη kde, η πυκνότητα των ποσοστών χρόνου με ασυνήθιστη βραχυπρόθεσμη μεταβλητότητα ανά κατάσταση στην **Εικόνα 43**:



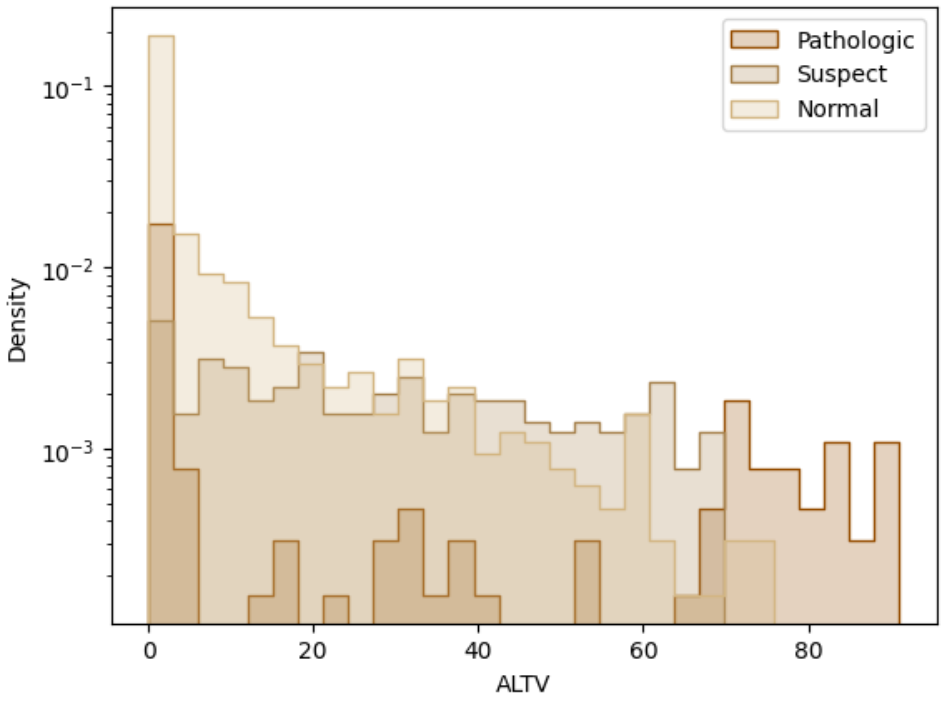
Εικόνα 43: Γραφική αναπαράσταση πυκνότητας ASTV ανά NSP

Μπορεί κανείς να αντιληφθεί ότι η ανοιχτόχρωμη γραμμή της φυσιολογικής κατάστασης στα χαμηλά ποσοστά ξεκινά απότομα και συνεχίζει να παρουσιάζει υψηλή πυκνότητα μέχρι και περίπου το 70% όπου και μετά παρουσιάζει απότομη πτώση. Η ύποπτη και η παθολογική κατάσταση παρουσιάζουν σχεδόν την ίδια καμπύλη μόνο που η καφέ καμπύλη έχει λίγο υψηλότερη πυκνότητα από την μαύρη παρουσιάζει την ίδια απότομη πτώση με την ανοιχτόχρωμη.

Άρα ένα συμπέρασμα θα μπορούσε να ήταν ότι τα πολύ υψηλά ποσοστά ασυνήθιστης βραχυπρόθεσμης μεταβλητότητας μπορεί να αποτελέσουν πρόβλημα στην υγεία του εμβρύου.

Έπειτα, επόμενο βήμα είναι ο έλεγχος και η ανάλυση του ποσοστού ασυνήθιστη μακροπρόθεσμης μεταβλητότητας σε σχέση με την κατάσταση του εμβρύου.

Όπως φαίνεται στην **Εικόνα 44** το ιστόγραμμα σε λογαριθμική κλίμακα:

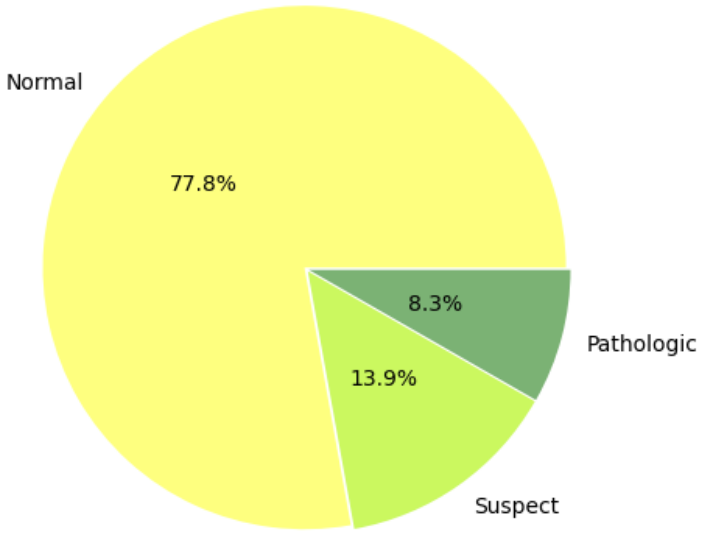


Εικόνα 44: Γραφική αναπαράσταση πυκνότητας ALTV ανά NSP

Στην παραπάνω εικόνα φαίνεται ότι στις χαμηλές τιμές των ποσοστών ασυνήθιστης μακροπρόθεσμης μεταβλητότητας της φυσιολογικής κατάστασης παρουσιάζεται υψηλή πυκνότητα ενώ όσο αυξάνονται τα ποσοστά η πυκνότητα μειώνεται σταδιακά. Το ίδιο φαίνεται να συμβαίνει και στην ύποπτη κατάσταση ενώ στην παθολογική κατάσταση φαίνεται ότι στα πολύ υψηλά ποσοστά παρουσιάζεται μεγαλύτερη πυκνότητα σε σύγκριση με τις άλλες δύο καταστάσεις.

Εφόσον έγινε ο έλεγχος των χαρακτηριστικών της μεγαλύτερης συσχέτισης με την κατάσταση του εμβρύου, τώρα επόμενο βήμα είναι να ελεγχθεί εξονυχιστικά το χαρακτηριστικό NSP. Από την **Εικόνα 38**, που περιέχει ραβδόγραμμα για κάθε χαρακτηριστικό με μια πρώτη ματιά μπορεί να γίνει αντιληπτό ότι το πλήθος των περιστατικών της κάθε κατάστασης διαφέρει κατά πολύ από τις αντίπαλες καταστάσεις τους.

Με την βοήθεια γραφήματος πίτας μπορεί να απεικονιστεί στην **Εικόνα 45** η διαφορά σε ποσοστά για κάθε μια κατάσταση του χαρακτηριστικού NSP.



Εικόνα 45: Ποσοστό ανά NSP σε γράφημα πίτας

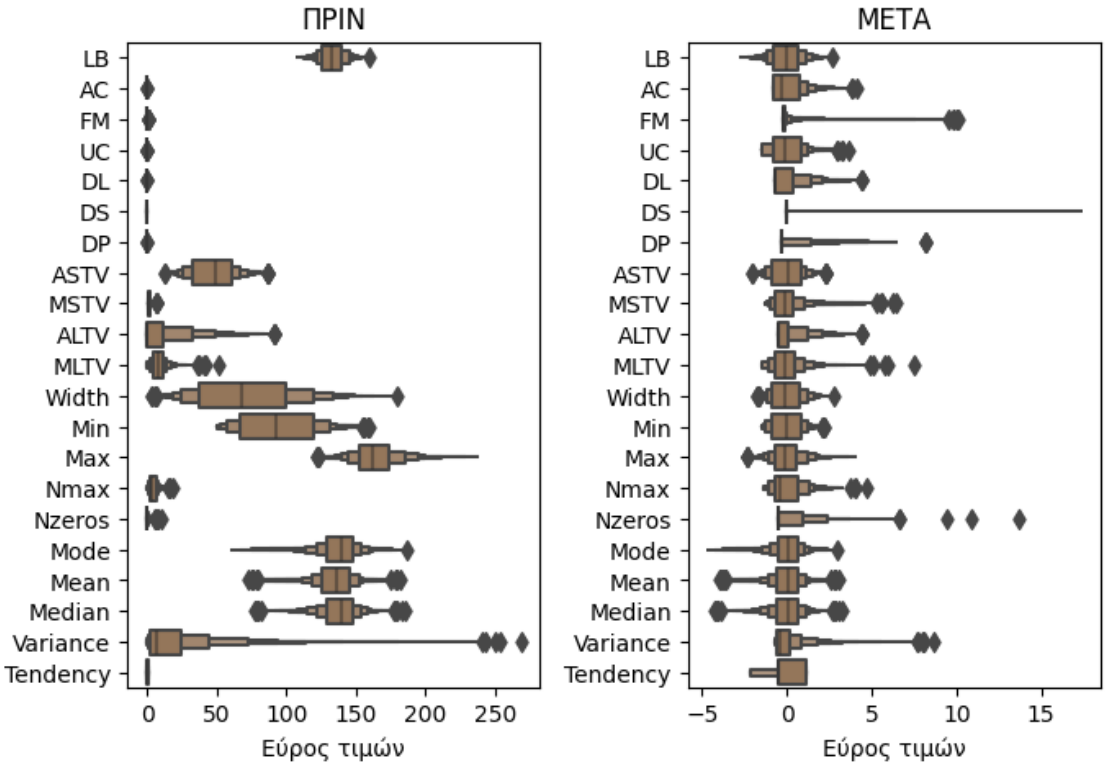
Στο συγκεκριμένο παράδειγμα είναι προφανές ότι το χαρακτηριστικό NSP παρουσιάζει ανομοιογένεια, όπως προαναφέρθηκε και εξετάστηκε στο Κεφάλαιο 3, αυτό μπορεί να προκαλέσει πρόβλημα όταν θα πραγματοποιηθεί η εκπαίδευση του μοντέλου, οπότε θα πρέπει σίγουρα να αντιμετωπιστεί. Στην επόμενη ενότητα θα γίνει η ανάλυση και ο τρόπος με τον οποίο αντιμετωπίστηκε.

4.3.2 Προετοιμασία δεδομένων για την εκπαίδευση

Το σύνολο δεδομένων αναλύθηκε και ελέγχθηκε εξονυχιστικά οπότε επόμενο βήμα είναι ο διαχωρισμός του συνόλου δεδομένων σε δεδομένα εισόδου και εξόδου – στόχου για την ευκολότερη χρήση των δεδομένων αυτών.

Στην συνέχεια εφόσον παρατηρήθηκε ότι τα κάποια χαρακτηριστικά δεν παρουσιάζουν ίδιο εύρος τιμών σε σύγκριση τα άλλα, καλό θα ήταν τα δεδομένα να μετατραπούν σε ίδιας κλίμακας τιμές. Στην περίπτωση αυτή χρησιμοποιήθηκε η μέθοδος Standard scaler, η οποία μετατρέπει τα δεδομένα έτσι ώστε η κατανομή τους να έχει μέση τιμή 0 και τυπική απόκλιση 1 [71] και αυτό είναι πολύ χρήσιμο να γίνεται πριν την εκπαίδευση συγκεκριμένων αλγορίθμων όπως π.χ. Linear Regression κ.λπ.

Στην **Εικόνα 46** απεικονίζεται γραφικά όλο το σύνολο δεδομένων ανά χαρακτηριστικό πριν και μετά την κλιμάκωση:



Εικόνα 46: Θηκόγραμμα για κάθε χαρακτηριστικό πριν και μετα την κλιμάκωση

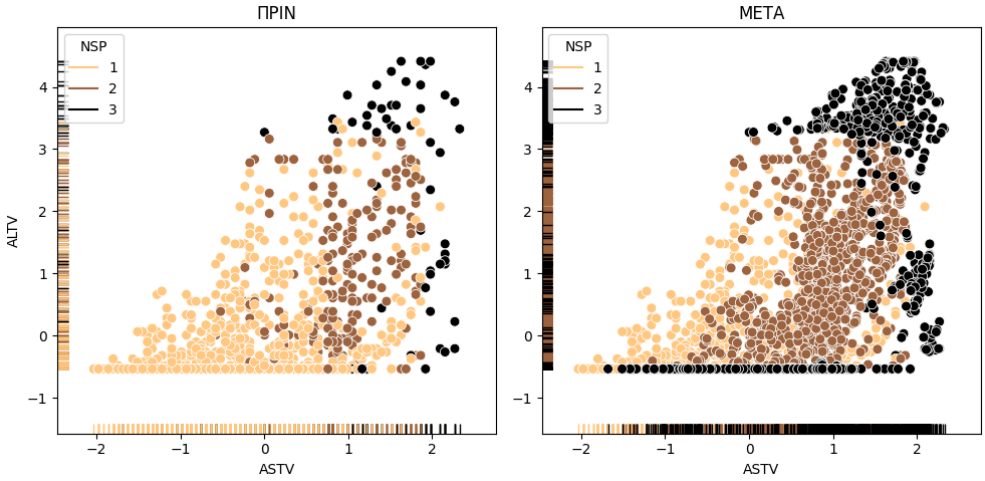
Προφανώς παρατηρείται ότι το εύρος των τιμών των χαρακτηριστικών πριν την μετατροπή είναι διάσπαρτο, στην συνέχεια αφού έγινε η μετατροπή το εύρος τιμών άλλαξε σε κλίμακα περίπου (-5,15) με μέση τιμή το μηδέν.

Επόμενο βήμα τώρα είναι ο διαχωρισμός του συνόλου δεδομένων σε σύνολα εκπαίδευσης και σύνολα δοκιμής. Χρησιμοποιήθηκε η μέθοδος train_test_split, με σύνολο δοκιμής να είναι το 25 % όλων των δεδομένων και το υπόλοιπο 75 % να αφορά το σύνολο εκπαίδευσης και με random state = 42 , όπως φαίνεται στον παρακάτω πίνακα η αναλογία των δειγμάτων σε ποσοστά:

	Δείγματα	**Ποσοστό**
Σετ δεδομένων εκπαίδευσης	1.594	75 %
Σετ δεδομένων δοκιμής	532	25 %
Σύνολο	2.126	100 %

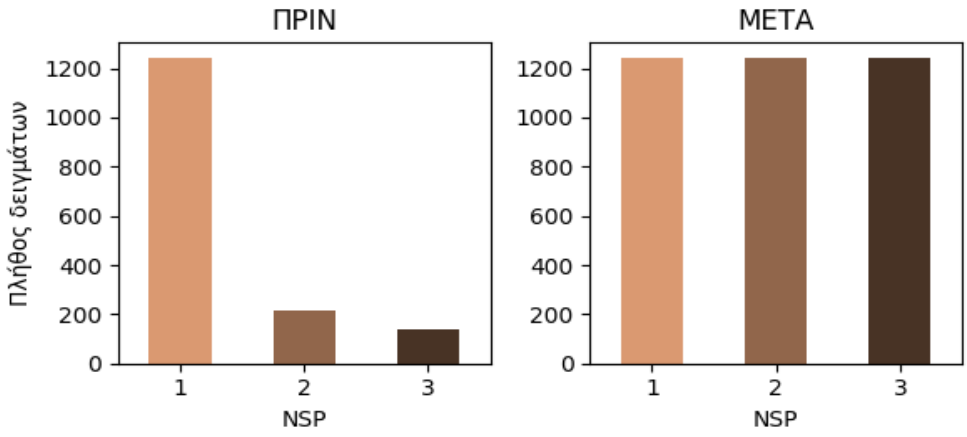
Πίνακας 18: Πλήθος δειγμάτων σε σύνολο εκπαίδευσης και δοκιμής

Προηγουμένως, παρατηρήθηκε ότι το σύνολο δεδομένων παρουσίασε ανομοιογένεια για αυτό και πρέπει να γίνει Επαναδειγματοληψία (Resampling). Χρησιμοποιήθηκε η μέθοδος υπερδειγματοληψίας SMOTE Ovesampling με υπερδειγματοληψία στο 100% και random state = 42. Η μέθοδος SMOTE (Synthetic Minority Over-sampling Technique) είναι μέθοδος υπερδειγματοληψίας που δημιουργεί συνθετικά παραδείγματα της τάξης της μειονότητας, ενώ αυτά τα συνθετικά παραδείγματα που παράγονται από την συγκεκριμένη μέθοδο βασίζονται στους πλησιέστερους γείτονες των περιπτώσεων της κλάσης μειονότητας Η υπερδειγματοληψία έγινε σε όλο το σετ δεδομένων εκπαίδευσης και δίνεται ένα παράδειγμα του διαγράμματος διασποράς όσον αφορά το ALTV και το ASTV.



Εικόνα 47: Διάγραμμα διασποράς ASTV και ALTV πριν και μετά το SMOTE

Οι ανοιχτόχρωμοι κύκλοι απεικονίζουν την φυσιολογική κατάσταση, τα καφέ την ύποπτη κατάσταση ενώ οι μαύροι κύκλοι απεικονίζουν την παθολογική κατάσταση. Πριν την υπερδειγματοληψία φαίνεται το πλήθος της παθολογικής κατάστασης είναι πολύ μικρό ενώ στην συνέχεια είναι ίσο με την φυσιολογική κατάσταση. Το ίδιο βέβαια συμβαίνει και με την ύποπτη κατάσταση. Στην παρακάτω εικόνα φαίνεται η διαφορά των τριών καταστάσεων:



Εικόνα 48: Ραβδόγραμμα δειγμάτων NSP πριν και μετά το SMOTE

Ο **Πίνακας 19** παρουσιάζει το πλήθος δειγμάτων για την κατάσταση του εμβρύου ανά κλάση πριν και μετά την μέθοδο SMOTE:

	ΔΕΙΓΜΑΤΑ	
	Πριν το SMOTE	Μετά το SMOTE
Κλάση 1	1.242	1.242
Κλάση 2	213	1.242
Κλάση 3	139	1.242

Πίνακας 19: Πλήθος δειγμάτων πριν και μετα το SMOTE

Τέλος, τα δεδομένα είναι έτοιμα για την εκπαίδευση αλγορίθμου και την επιλογή του ως κατάλληλο μοντέλο με στόχο την καλύτερη επίδοση στο σύνολο δεδομένων που μελετήθηκε.

4.3.3 Εκπαίδευση και έλεγχος επίδοσης αλγορίθμου

Στο συγκεκριμένο πρόβλημα, για την επιλογή του κατάλληλου αλγορίθμου χρησιμοποιήθηκε η μέθοδος της διασταυρωμένης επικύρωσης – cross validation. Σκοπός της είναι η εκτίμησης της απόδοσης ενός μοντέλου ανάλογα με το κριτήριο που έχει επιλεχτεί, δηλαδή πόσο καλά θα ανταποκριθεί σε άγνωστα δεδομένα και συνήθως το κριτήριο που επιλέγεται είναι αυτό της ακρίβειας. Αρχικά, όλο το σύνολο δεδομένων διαχωρίζεται ανά Κ πειράματα (K-folds) σε σύνολο εκπαίδευσης και σε σύνολο δοκιμής ή επικύρωσης, το μοντέλο εκπαιδεύεται στο σύνολο εκπαίδευσης και κρίνεται βάση της ακρίβειας στο σύνολο δοκιμής. Έπειτα η ακρίβεια του κάθε πειράματος αθροίζεται και τέλος διαιρείται με τον αριθμό των πειραμάτων όπου προκύπτει ο μέσος όρος αυτής, ενώ αυτός ο αριθμός υποδηλώνει την επίδοση του αλγορίθμου [17]. Η πιο συχνή μέθοδος που χρησιμοποιείται είναι το K – fold cross – validation, όπου χρησιμοποιήθηκε παρακάτω, ενώ το Κ υποδηλώνει τον αριθμό των πειραμάτων που θα χωριστούν τα δεδομένα. Για παράδειγμα, έστω ότι K = 5, το σύνολο δεδομένων θα χωριστεί σε πέντε πειράματα, έπειτα το κάθε πείραμα θα χωριστεί ισάξια σε πέντε τμήματα όπου το ένα από τα πέντε αποτελεί το σύνολο δοκιμής, ενώ τα υπόλοιπα τέσσερα τα σύνολα εκπαίδευσης. Στο 1^{ο} πείραμα, το 1^{ο} τμήμα θα είναι το σύνολο δοκιμής, ενώ από το 2^{ο} τμήμα και μετά θα είναι τα τέσσερα σύνολα εκπαίδευσης για κάθε ένα τμήμα. Στο 2^{ο} πείραμα, το 1^{ο} τμήμα αποτελεί το σύνολο εκπαίδευσης, το 2^{ο} τμήμα αποτελεί το σύνολο δοκιμής και τα υπόλοιπα τρία τμήματα είναι τα σύνολα εκπαίδευσης για το πείραμα αυτό. Στο 3^{ο} πείραμα, το 1^{ο} και το 2^{ο} τμήμα αποτελούν δύο σύνολα εκπαίδευσης για το κάθε τμήμα, το 3^{ο} τμήμα αποτελεί το σύνολο δοκιμής και το 4^{ο} και 5^{ο} τμήμα αποτελούν τα δύο υπόλοιπα σύνολα εκπαίδευσης για το πείραμα αυτό. Η ίδια διαδικασία συμβαίνει

μέχρις ότου το σύνολο δοκιμής στο 5ο πείραμα να είναι το 5ο τμήμα [17, 72]. Για την καλύτερη κατανόηση θα ήταν χρήσιμο να γίνει αναπαράσταση του παραδείγματος αυτού στην **Εικόνα 49**:

Εικόνα 49: Διαχωρισμός δεδομένων σε πέντε K-Folds

Πιο συγκεκριμένα στο πρόβλημα αυτό χρησιμοποιήθηκε η μέθοδος cross_val_score, το οποίο διαχωρίστηκε σε 5 πειράματα όπως επεξηγήθηκε παραπάνω. Ο λόγος που χρησιμοποιήθηκε είναι για την καλύτερη επιλογή αλγορίθμου μεταξύ των Logistic Regression, Decision Tree, Random Forest, Ada Boost και Gradient Boosting. Μια σημαντική λεπτομέρεια που πρέπει να αναφερθεί είναι ότι το σύνολο δεδομένων που χρησιμοποιήθηκε για τον διαχωρισμό σε 5 πειράματα είναι το σύνολο δεδομένων δοκιμής που προέκυψε από την train_test_split. Επειδή χρησιμοποιήθηκε η μέθοδος της υπερδειγματοληψίας για να μπορέσει το μοντέλο να μάθει από το σύνολο εκπαίδευσης τις καλύτερες και πιο χρήσιμες πληροφορίες για τα δεδομένα **ΔΕΝ** πρέπει για κανέναν λόγο το σύνολο δοκιμής να έχει επαφή με τα δεδομένα που εκπαιδεύτηκε το μοντέλο. Εάν συμβεί αυτό τότε η εκτίμηση που θα ληφθεί θα είναι υπερβολικά αισιόδοξη διότι προκύπτει πρόβλημα υπερπροσαρμογής, με πιο απλά λόγια το μοντέλο θα παρουσιάσει πολύ υψηλότερη απόδοση από ότι θα παρουσίαζε σε άγνωστα δεδομένα [73]. Ο **Πίνακας 20** παρουσιάζει τις μέσες τιμές της ακρίβειας για κάθε αλγόριθμο στο σύνολο δοκιμής που προέκυψε από τα 5 πειράματα σε σύγκριση με την ακρίβεια πάνω στα άγνωστα δεδομένα της train_test_split.

	Μέση τιμή ακρίβειας	**Ακρίβεια στα άγνωστα δεδομένα**
Logistic Regression	84,5 %	85 %
Decision Tree	94,1 %	93 %
Random Forest	96,7 %	94,9 %
Ada Boost	89,2 %	87,6 %
Gradient Boosting	95,8 %	94 %

Πίνακας 20: Σύγκριση ακρίβειας ανά αλγόριθμο

Μπορεί κανείς να παρατηρήσει και να συμπεράνει ότι οι μέσες τιμές σύμφωνα με τον διαχωρισμό των 5 πειραμάτων και η επίδοση πάνω στα σύνολα δοκιμής των πειραμάτων παρουσιάζουν υψηλότερη ακρίβεια σε σχέση με την ακρίβεια στα άγνωστα δεδομένα σε όλους τους αλγορίθμους εκτός του Logistic Regression. Αυτό το συμπέρασμα πρέπει να θεωρηθεί λανθασμένο, διότι τα δεδομένα που χρησιμοποιούνται στον πραγματικό κόσμο είναι επίσης και αυτά άγνωστα, οπότε δεν μπορεί να ληφθεί στα σοβαρά αυτή η επίδοση. Άρα, η πρώτη στήλη με τις μέσες τιμές **ΔΕΝ** πρέπει να ληφθεί υπόψιν και παρουσιάστηκε για την καλύτερη κατανόηση του προβλήματος.

Τότε, ποιό είναι το νόημα της χρησιμότητας αυτής της μεθόδου; Η απάντηση σε αυτό το ερώτημα είναι απλή, η επιλογή της μεθόδου αυτή έγινε επειδή μπορεί να προσφέρει πληροφορίες για την καλύτερη επιλογή αλγορίθμου. Στην περίπτωση αυτή, ο **Πίνακας 20** δείχνει ότι**,** οι αλγόριθμοι που χρειάζεται να γίνει παραμετροποίηση ρυθμίσεων και ανάλυση περισσότερο είναι αυτοί που παρουσίασαν την μεγαλύτερη ακρίβεια, αυτοί είναι οι αλγόριθμοι Decision Tree, Random Forest και Gradient Boosting.

Αφού επιλέχθηκαν οι καταλληλότεροι αλγόριθμοι, επόμενο βήμα είναι η παραμετροποίησή τους και η επιλογή ενός από αυτούς με την καλύτερη επίδοση στα άγνωστα δεδομένα.

Decision Tree

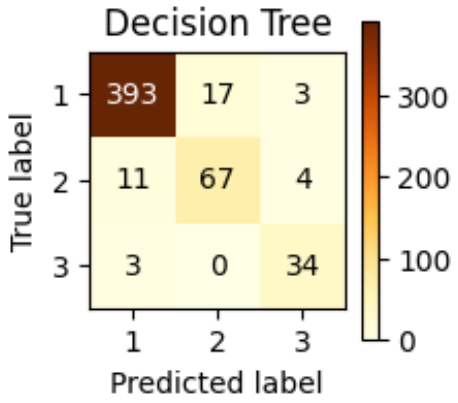
Σε αυτόν τον αλγόριθμο οι καλύτερες ρυθμίσεις που παρουσίασαν το καλύτερο αποτέλεσμα είναι min samples split = 10, ο ελάχιστος αριθμός δειγμάτων που απαιτείται για τη διάσπαση ενός εσωτερικού κόμβου, max_depth = 16 το μέγιστο βάθος του δέντρου και random state = 42. Ο αλγόριθμος αυτός παρουσίασε 92,86 % ακρίβεια σε 532 δείγματα δοκιμής.

Όπως προαναφέρθηκε και στο κεφάλαιο 3, σε προβλήματα δύο ή περισσοτέρων κλάσεων αξίζει να αναλυθεί το κριτήριο Recall για κάθε κλάση:

	Recall
Κλάση 1	95 %
Κλάση 2	82 %
Κλάση 3	92 %
Μέση	90 %

Πίνακας 21: Ποσοστά Recall για κάθε κλάση στο Decision Tree

Με μια πρώτη ματιά η κλάση 2 παρουσιάζει το χαμηλότερο ποσοστό, ενώ οι υπόλοιπες δύο παρουσιάζουν πολύ υψηλά ποσοστά. Καλό είναι να αναλυθεί και ο πίνακας σύγχυσης για την κάθε κλάση όπως φαίνεται στην **Εικόνα 50**:



Εικόνα 50: Πίνακας σύγχυσης για Decision Tree

Οι διαστάσεις του πίνακα σύγχυσης είναι 3x3 όπως φαίνεται παραπάνω και αυτό συμβαίνει διότι το πρόβλημα αυτό είναι τριών κλάσεων. Σε αυτή την περίπτωση το καλύτερο αποτέλεσμα είναι όταν η διαγώνιος από αριστερά προς τα δεξιά έχει τις υψηλότερες τιμές σε σχέση με τις υπόλοιπες οριζόντιες τιμές. Με τον πίνακα αυτό μπορεί να κατανοηθεί καλύτερα το πως προβλέπει τις κλάσεις το μοντέλο που εκπαιδεύτηκε.

Στην ‘κλάση 1’ το μοντέλο προέβλεψε σωστά 393 περιπτώσεις, ενώ προέβλεψε εσφαλμένα ως ‘κλάση 2’ τις 17 περιπτώσεις και 3 περιπτώσεις ως ‘κλάση 3’.

Στην ‘κλάση 2’ το μοντέλο κατηγοριοποίησε σωστά 67 περιπτώσεις, ενώ κατηγοριοποίησε εσφαλμένα ως ‘κλάση 1’ τις 11 περιπτώσεις και ως ‘κλάση 3’ τις 4 περιπτώσεις.

Στην ‘κλάση 3’ το μοντέλο προέβλεψε σωστά 34 περιπτώσεις από τις 37 συνολικά με μόλις 3 περιπτώσεις να προβλέπονται εσφαλμένα ως ‘κλάση 1’.

Να σημειωθεί ότι η πιο κρίσιμη και κυρίαρχη κλάση που πρέπει να προβλεφθεί καλύτερα είναι η ‘κλάση 3’, ενώ σαν δεύτερη πιο σημαντική είναι η ‘κλάση 2’.

Random Forest

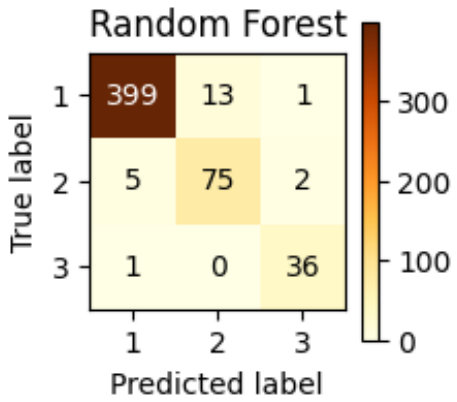
Η βέλτιστες ρυθμίσεις που χρησιμοποιήθηκαν στον αλγόριθμο αυτό είναι n estimators = 50 που σημαίνει το πλήθος του δέντρου απόφασης, κριτήριο την εντροπία, max depth = 16 το μέγιστο βάθος του δέντρου και random state = 42. Ο αλγόριθμος αυτός παρουσίασε 95,86 % ακρίβεια στο σύνολο δεδομένων δοκιμής ξεπερνώντας τον προηγούμενο αλγόριθμο. Αυτό είναι φυσιολογικό διότι ο αλγόριθμος Random Forest είναι μια βελτίωση του Decision Tree.

Όπως και στον προηγούμενο αλγόριθμο θα αναλυθεί το κριτήριο της ανάκλησης, παρουσιάζεται παρακάτω:

	Recall
Κλάση 1	97 %
Κλάση 2	91 %
Κλάση 3	97 %
Μέση	95 %

Πίνακας 22: Ποσοστά Recall για κάθε κλάση στον Random Forest

Το κριτήριο της ανάκλησης αποδεικνύει ότι όντως ο Random Forest έχει καλύτερη επίδοση σε όλες τις κλάσεις από τον προηγούμενο. Ειδικά τα ποσοστά της κλάσης 1 και 3 είναι εξαιρετικά υψηλά. Ας ριχτεί όμως και μια ματιά στον πίνακα σύγχυσης για την καλύτερη κατανόηση των προβλέψεων, όπως φαίνεται στην παρακάτω εικόνα:



Εικόνα 51: Πίνακας σύγχυσης για Random Forest

Στην ‘κλάση 1’ το μοντέλο προέβλεψε σωστά 399 περιπτώσεις, ενώ προέβλεψε εσφαλμένα ως ‘κλάση 2’ τις 13 περιπτώσεις και 1 περίπτωση ως ‘κλάση 3’.

Στην ‘κλάση 2’ το μοντέλο προέβλεψε σωστά 75 περιπτώσεις, ενώ προέβλεψε εσφαλμένα ως ‘κλάση 1’ τις 5 περιπτώσεις και ως ‘κλάση 3’ τις 2 περιπτώσεις.

Στην ‘κλάση 3’ το μοντέλο προέβλεψε σωστά 36 περιπτώσεις, ενώ προέβλεψε εσφαλμένα μόνο 1 περίπτωση ως ‘κλάση 1’.

Προς το παρών το ο αλγόριθμος αυτός παρουσιάζει εξαιρετικό αποτέλεσμα όμως σημαντικό είναι να ελεγχθεί και ο αλγόριθμος Gradient Boosting.

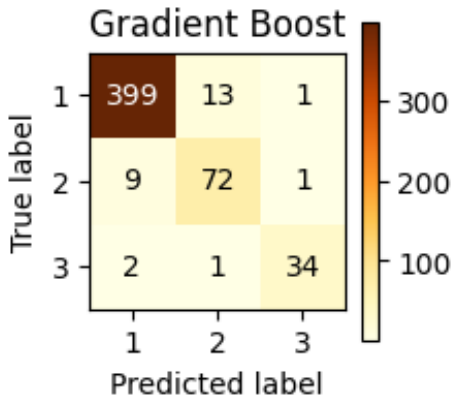
Gradient Boosting

Σε αυτόν τον αλγόριθμο οι καλύτερες ρυθμίσεις που χρησιμοποιήθηκαν είναι n estimators = 20, max depth = 10, learning rate = 0.8 και είναι ο ρυθμός εκπαίδευσης που περιορίζει τη συμμετοχή κάθε δέντρου και Random state = 12. Με την συγκεκριμένη παραμετροποίηση ο αλγόριθμος κατάφερε να σημειώσει 94,92 % ακρίβεια.

	Recall
Κλάση 1	97 %
Κλάση 2	88 %
Κλάση 3	92 %
Μέση	92 %

Πίνακας 23: Ποσοστά Recall για κάθε κλάση στον Gradient Boosting

Ο Gradient Boosting φαίνεται και αυτός με την σειρά του να παρουσιάζει πολύ καλές επιδόσεις στο κριτήριο της ανάκλησης. Σε μεγαλύτερο ποσοστό προβλέπει την ‘κλάση 1’, τα πάει βέβαια πολύ καλά και στην ‘κλάση 3’ ενώ στην ‘κλάση 2’ όχι και τόσο ικανοποιητικά. Ας παρουσιαστεί και ο πίνακας σύγχυσης παρακάτω:



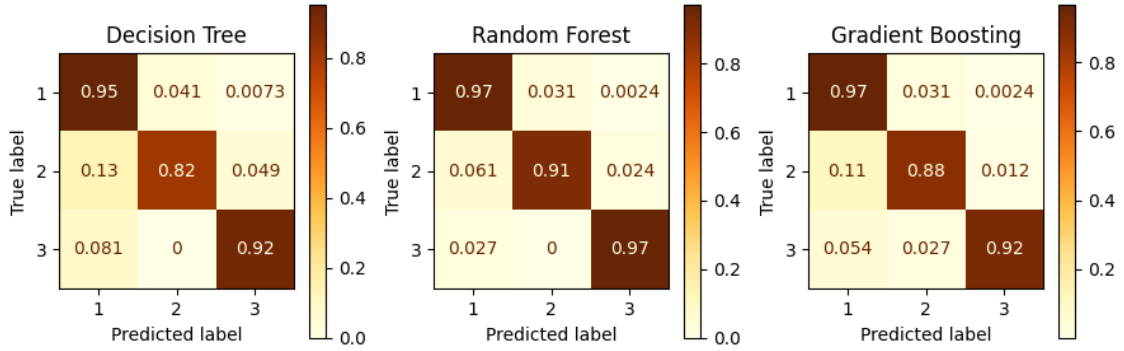
Εικόνα 52: Πίνακας σύγχυσης για Gradient Boosting

Στην ‘κλάση 1’ το μοντέλο προέβλεψε σωστά 399 περιπτώσεις, ενώ προέβλεψε εσφαλμένα ως ‘κλάση 2’ τις 13 περιπτώσεις και 1 περίπτωση ως ‘κλάση 3’.

Στην ‘κλάση 2’ το μοντέλο προέβλεψε σωστά 72 περιπτώσεις, ενώ προέβλεψε εσφαλμένα ως ‘κλάση 1’ τις 9 περιπτώσεις και ως 1 περίπτωση ως ‘κλάση 3’.

Στην ‘κλάση 3’ το μοντέλο προέβλεψε σωστά 34 περιπτώσεις, ενώ προέβλεψε εσφαλμένα ως ‘κλάση 1’ 2 περιπτώσεις και 1 περίπτωση ως ‘κλάση 2’.

Όπως φαίνεται όλοι οι αλγόριθμοι παρουσιάζουν υψηλές επιδόσεις και η επιλογή του καλύτερου πρέπει να γίνει με την σύγκριση μεταξύ τους. Για αυτό και θα γίνει η αναπαράσταση στην **Εικόνα 53** των πινάκων συσχέτισης για κάθε κλάση σύμφωνα με την μέση τιμή της ανάκλησης για κάθε μοντέλο.



Εικόνα 53: Πίνακες σύγχυσης για κάθε κλάση και για κάθε μοντέλο

Αρχικά ας συγκριθούν τα ποσοστά ανά κλάση, ξεκινώντας από την ‘κλάση 1’, τα μοντέλα Random Forest και Gradient Boosting παρουσιάζουν μέση ανάκληση 97 %, ενώ ο τελευταίος μόλις 95 %. Στην ‘κλάση 2’ ο Random Forest είναι πρωτοπόρος με 91 % ακολουθώντας δεύτερος ο Gradient Boosting με 3% λιγότερο, ενώ ο Decision Tree δεν έχει τόση καλή επίδοση στο κριτήριο αυτό. Τελευταία η ‘κλάση 3’ και η πιο σημαντική φαίνεται και πάλι ξεκάθαρα ότι ο Random Forest προβλέπει την ‘κλάση 3’ με ακρίβεια 97% ενώ οι υπόλοιποι δύο να προβλέπουν με 5 % λιγότερο.

Προφανώς και ως καταλληλότερο μοντέλο επιλέγεται ο Random Forest, διότι στην ύποπτη κατάσταση – ‘κλάση 2’ και στην παθολογική κατάσταση – ‘κλάση 3’ καταφέρνει να προβλέψει περισσότερες περιπτώσεις αυτών, αφού βέβαια προβλέπει επίσης εξαιρετικά την ‘κλάση 1’. Φυσικά θα μπορούσε να χρησιμοποιηθεί και το μοντέλο Gradient Boosting ως δεύτερο κατάλληλο αφού και αυτό με τη σειρά του παρουσιάζει εξαιρετική επίδοση. Αυτό έχει ως αποτέλεσμα ότι το μοντέλο είναι ικανό να προβλέψει προβλήματα για την καρδιακή κατάσταση του εμβρύου και θα μπορούσε να αποτελέσει σημαντικό εργαλείο για τον γιατρό από το να μην το έχει στην κατοχή του.

Βιβλιογραφικές αναφορές

[1] 'What Is Machine Learning (ML)?' https://ischoolonline.berkeley.edu/blog/what-is-machine-learning/

[2] Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας, και Η. Σακελλαρίου, *Τεχνητή Νοημοσύνη*, Γ' Έκδοση. : Εκδόσεις Πανεπιστημίου Μακεδονίας, 2011.

[3] 'Deep Learning: The Effects of Artificial Intelligence in Business'. https://www.salesforce.com/eu/blog/2021/11/deep-learning.html.

[4] 'Data science - Wikipedia'. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://en.wikipedia.org/wiki/Data_science

[5] J. D. Kelleher και B. Tierney, *Data Science*. The MIT Press, 2018. doi: 10.7551/mitpress/11140.001.0001.

[6] 'Data set - Wikipedia'. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://en.wikipedia.org/wiki/Data_set

[7] 'Statistics - Wikipedia'. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://en.wikipedia.org/wiki/Statistics

[8] 'Algorithm Definition & Meaning - Merriam-Webster'. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://www.merriam-webster.com/dictionary/algorithm

[9] 'What Is a Decision Tree?' https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/

[10] B. Charbuty και A. Abdulazeez, 'Classification Based on Decision Tree Algorithm for Machine Learning', *J. Appl. Sci. Technol. Trends*, τ. 2, τχ. 01, σσ. 20–28, Μαρτίου 2021, doi: 10.38094/jastt20165.

[11] Tin Kam Ho, 'Random Decision Forests', στο *Proceedings of 3rd International Conference on Document Analysis and Recognition*, σσ. 278–282. doi: 10.1109/ICDAR.1995.598994.

[12] 'What is a Random Forest?' https://www.tibco.com/reference-center/what-is-a-random-forest

[13] 'Gradient Boosting – What You Need to Know — Machine Learning — DATA SCIENCE'. https://datascience.eu/machine-learning/gradient-boosting-what-you-need-to-know/

[14] '7 Steps to Machine Learning: How to Prepare for an Automated Future | by Dr Mark van Rijmenam | DataSeries | Medium'. https://medium.com/dataseries/7-steps-to-machine-learning-how-to-prepare-for-an-automated-future-78c7918cb35d

[15] Αναστασία Δήμητρα Λυπιτάκη, 'Μηχανική Μάθηση σε ανομοιογενή δεδομένα', Μεταπτυχιακή Διατριβή, Πανεπιστήμιο Πατρών, Τμήμα Μαθηματικών, ΠΑΤΡΑ, 2014. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://nemertes.library.upatras.gr/server/api/core/bitstreams/3d96ea08-998b-4b8a-820f-8218b650796e/content

[16] E. Zhang και Y. Zhang, 'F-Measure', *Encycl. Database Syst.*, σσ. 1147–1147, 2009, doi: 10.1007/978-0-387-39940-9_483.

[17] Κ. Διαμανταράς και Δ. Μπότσης, *Μηχανική Μάθηση*, 1η έκδ. Κλειδάριθμος, 2019.

[18] 'Compare Deep Learning Models Using ROC Curves'. https://www.mathworks.com/help//deeplearning/ug/compare-deep-learning-models-using-ROC-curves.html

[19] 'Genomics', *Wikipedia*. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://en.wikipedia.org/w/index.php?title=Genomics&oldid=1130891905

[20] W. P. Blackstock και M. P. Weir, 'Proteomics: quantitative and physical mapping of cellular proteins', *Trends Biotechnol.*, τ. 17, τχ. 3, σσ. 121–127, Μαρτίου 1999, doi: 10.1016/S0167-7799(98)01245-1.

[21] K. Shailaja, B. Seetharamulu, και M. A. Jabbar, 'Machine Learning in Healthcare: A Review', στο *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Μαρτίου 2018, σσ. 910–914. doi: 10.1109/ICECA.2018.8474918.

[22] G. Parthiban και S. K. Srivatsa, 'Applying machine learning methods in diagnosing heart disease for diabetic patients', *Int. J. Appl. Inf. Syst.*, τ. 3, τχ. 7, σσ. 25–30, 2012.

[23] 'Robot-assisted minimally invasive surgery (RMIS) schematics...', *ResearchGate*. https://www.researchgate.net/figure/Robot-assisted-minimally-invasive-surgery-RMIS-schematics-with-the-proposed-MRP-tactile_fig1_339574322

[24] Mahtab J. Fard, Sattar Ameri, Ratna B. Chinnam, Abhilash K. Pandya, Michael D. Klein, και R. D. Ellis, 'Machine Learning Approach for Skill Evaluation in Robotic-Assisted Surgery'. arXiv, 15 Νοέμβριος 2016. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: http://arxiv.org/abs/1611.05136

[25] A. U. Haq *κ.ά.*, 'Intelligent Machine Learning Approach for Effective Recognition of Diabetes in E-Healthcare Using Clinical Data', *Sensors*, τ. 20, τχ. 9, σ. 2649, Μαΐου 2020, doi: 10.3390/s20092649.

[26] N. Ahmidi *κ.ά.*, 'A Dataset and Benchmarks for Segmentation and Recognition of Gestures in Robotic Surgery', *IEEE Trans. Biomed. Eng.*, τ. 64, τχ. 9, σσ. 2025–2041, Σεπτεμβρίου 2017, doi: 10.1109/TBME.2016.2647680.

[27] 'Emergency department', *Wikipedia*. 16 Δεκέμβριος 2022. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://en.wikipedia.org/w/index.php?title=Emergency_department&oldid=1127662073

[28] J. S. Peck, J. C. Benneyan, D. J. Nightingale, και S. A. Gaehde, 'Predicting Emergency Department Inpatient Admissions to Improve Same-day Patient Flow: PREDICTING ED INPATIENT ADMISSIONS', *Acad. Emerg. Med.*, τ. 19, τχ. 9, σσ. E1045–E1054, Σεπτεμβρίου 2012, doi: 10.1111/j.1553-2712.2012.01435.x.

[29] 'Social media', *Wikipedia*. 17 Ιανουάριος 2023. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://en.wikipedia.org/w/index.php?title=Social_media&oldid=1134171022

[30] 'Depression', *National Institute of Mental Health (NIMH)*. https://www.nimh.nih.gov/health/topics/depression

[31] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, και H. Ohsaki, 'Recognizing Depression from Twitter Activity', στο *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul Republic of Korea, Απριλίου 2015, σσ. 3187–3196. doi: 10.1145/2702123.2702280.

[32] M. De Choudhury, M. Gamon, S. Counts, και E. Horvitz, 'Predicting Depression via Social Media', *Proc. Int. AAAI Conf. Web Soc. Media*, τ. 7, τχ. 1, σσ. 128–137, Αυγούστου 2021, doi: 10.1609/icwsm.v7i1.14432.

[33] 'Coronavirus'. https://www.who.int/health-topics/coronavirus

[34] S. Lalmuanawma, J. Hussain, και L. Chhakchhuak, 'Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review', *Chaos Solitons Fractals*, τ. 139, σ. 110059, 2020, doi: 10.1016/j.chaos.2020.110059.

[35] C. Warlow, 'Epidemiology of stroke', *The Lancet*, τ. 352, σσ. S1–S4, Οκτωβρίου 1998, doi: 10.1016/S0140-6736(98)90086-1.

[36] 'WHO EMRO | Stroke, Cerebrovascular accident | Health topics', *World Health Organization - Regional Office for the Eastern Mediterranean*. http://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html

[37] 'Εγκεφαλικό: Τι είναι - Συμπτώματα - Αντιμετώπιση', *Euroclinic*. https://www.euroclinic.gr/article/egkefaliko-ti-einai-symptomata-antimetopisi/

[38] 'World Stroke Day'. https://www.who.int/southeastasia/news/detail/28-10-2021-world-stroke-day

[39] J. Khan, A. A. Shah, A. Jielani, και Attique-ur-Rehman, 'FREQUENCY OF HYPERTENSION IN STROKE PATIENTS PRESENTING AT AYUB TEACHING HOSPITAL', *J Ayub Med Coll Abbottabad*.

[40] J. Shou, L. Zhou, S. Zhu, και X. Zhang, 'Diabetes is an Independent Risk Factor for Stroke Recurrence in Stroke Patients: A Meta-analysis', *J. Stroke Cerebrovasc. Dis.*, τ. 24, τχ. 9, σσ. 1961–1968, Σεπτεμβρίου 2015, doi: 10.1016/j.jstrokecerebrovasdis.2015.04.004.

[41] S. A. E. Peters, R. R. Huxley, και M. Woodward, 'Diabetes as a risk factor for stroke in women compared with men: a systematic review and meta-analysis of 64 cohorts, including 775 385 individuals and 12 539 strokes', *The Lancet*, τ. 383, τχ. 9933, σσ. 1973–1980, Ιουνίου 2014, doi: 10.1016/S0140-6736(14)60040-4.

[42] J. A. Staessen, J. Wang, G. Bianchi, και W. H. Birkenhäger, 'Essential hypertension', *The Lancet*, τ. 361, τχ. 9369, σσ. 1629–1641, Μαΐου 2003, doi: 10.1016/S0140-6736(03)13302-8.

[43] 'Hypertension'. https://www.who.int/news-room/fact-sheets/detail/hypertension

[44] P. K. Whelton *κ.ά.*, '2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults', *J. Am. Coll. Cardiol.*, τ. 71, τχ. 19, σσ. e127–e248, Μαΐου 2018, doi: 10.1016/j.jacc.2017.11.006.

[45] Maria, 'HIGH BLOOD PRESSURE HYPERTENSION - Diagnosis, Symptoms, Hypertension Diet and 12 Supplements to Help Treat Hypertension Naturally',

Ecosh, 30 Ιούλιος 2020. https://ecosh.com/high-blood-pressure-hypertension-diagnosis-symptoms-hypertension-diet-and-12-supplements-to-help-treat-hypertension-naturally/

[46] 'Diabetes'. https://www.who.int/news-room/fact-sheets/detail/diabetes

[47] 'What is Diabetes? | NIDDK', *National Institute of Diabetes and Digestive and Kidney Diseases*. https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes

[48] N. G. Forouhi και N. J. Wareham, 'Epidemiology of diabetes', *Medicine (Baltimore)*, τ. 38, τχ. 11, σσ. 602–606, Νοεμβρίου 2010, doi: 10.1016/j.mpmed.2010.08.007.

[49] 'Obesity and overweight'. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[50] WHO Consultation on Obesity (1999: Geneva S. και Organization W. H., 'Obesity : preventing and managing the global epidemic : report of a WHO consultation', World Health Organization, 2000. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://apps.who.int/iris/handle/10665/42330

[51] 'Body mass index | Definition, Formula, Chart, & Facts | Britannica'. https://www.britannica.com/science/body-mass-index

[52] 'Tobacco'. https://www.who.int/news-room/fact-sheets/detail/tobacco

[53] R. S. Shah και J. W. Cole, 'Smoking and stroke: the more you smoke the more you stroke', *Expert Rev. Cardiovasc. Ther.*, τ. 8, τχ. 7, σσ. 917–932, Ιουλίου 2010, doi: 10.1586/erc.10.56.

[54] K. K. Andersen, T. S. Olsen, C. Dehlendorff, και L. P. Kammersgaard, 'Hemorrhagic and Ischemic Strokes Compared: Stroke Severity, Mortality, and Risk Factors', *Stroke*, τ. 40, τχ. 6, σσ. 2068–2072, Ιουνίου 2009, doi: 10.1161/STROKEAHA.108.540112.

[55] J. Gomes και A. M. Wachsman, 'Types of Strokes', στο *Handbook of Clinical Nutrition and Stroke*, M. L. Corrigan, A. A. Escuro, και D. F. Kirby, Επιμ. Totowa, NJ: Humana Press, 2013, σσ. 15–31. doi: 10.1007/978-1-62703-380-0_2.

[56] J. Fields και A. Bhardwaj, 'Cerebral Blood Flow and Metabolism: Physiology and Monitoring', στο *Handbook of Neurocritical Care*, A. Bhardwaj και M. A. Mirski, Επιμ. New York, NY: Springer New York, 2010, σσ. 51–60. doi: 10.1007/978-1-4419-6842-5_4.

[57] 'What is Stroke? – The Singapore National Stroke Association (SNSA)'. https://snsa.org.sg/resources/what-is-stroke/

[58] 'Stroke Prediction Dataset'. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[59] 'Tool missingno for NaN values'. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://github.com/ResidentMario/missingno

[60] R. H. N. Nguyen, 'Terms in reproductive and perinatal epidemiology: 2. Perinatal terms', *J. Epidemiol. Community Health*, τ. 59, τχ. 12, σσ. 1019–1021, Δεκεμβρίου 2005, doi: 10.1136/jech.2004.023465.

[61] C. Gribbin και D. James, ‘Assessing fetal health’, *Best Pract. Res. Clin. Obstet. Gynaecol.*, τ. 18, τχ. 3, σσ. 411–424, Ιουνίου 2004, doi: 10.1016/j.bpobgyn.2004.02.004.

[62] Sundar. C, M. C. M.Chitradevi, και G. Geetharamani, ‘Classification of Cardiotocogram Data using Neural Network based Machine Learning Technique’, *Int. J. Comput. Appl.*, τ. 47, τχ. 14, σσ. 19–25, Ιουνίου 2012, doi: 10.5120/7256-0279.

[63] ‘Cardiotocography - an overview | ScienceDirect Topics’. https://www.sciencedirect.com/topics/medicine-and-dentistry/cardiotocography

[64] Θ. Λάμπρος, ‘Ανάλυση εμβρυακού ηλεκτροκαρδιογραφήματος με χρήση προσομοιωμένων σημάτων’, Μεταπτυχιακή Εργασία, Πανεπιστήμιο Ιωαννίνων, Τμήμα Πληροφορικής και Τηλεπικοινωνιών, Άρτα, 2018.

[65] ‘Cardiotocography’, *Wikipedia*. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://en.wikipedia.org/w/index.php?title=Cardiotocography&oldid=1136126097

[66] ‘Obstetrician Gynaecologist Limassol - Dr Efterpi Tingi’, *Δρ Ευτέρπη Τίγγη*, 12 Μάιος 2021. https://obgynae.com.cy/el/καρδιοτοκογράφημα/

[67] Θ. Ι. Δαγκλής, ‘Καρδιοτοκογράφημα: Τεχνικά χαρακτηριστικά – Ερμηνεία’. 2019. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://www.hellenic-embryomitriki.gr/wp-content/uploads/2021/12/NTS_1.pdf

[68] Dr Lewis Potter·Data Interpretation, ‘How to Read a CTG | CTG Interpretation | Geeky Medics’, 29 Μάρτιος 2011. https://geekymedics.com/how-to-read-a-ctg/

[69] C. Dr Todd, M. Dr Rucklidge, και T. Kay, ‘FETAL HEART RATE MONITORING – PRINCIPLES AND INTEPRETATION OF CARDIOTOCOGRAPHY’. 2013. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://resources.wfsahq.org/wp-content/uploads/294_english.pdf

[70] ‘UCI Machine Learning Repository: Cardiotocography Data Set’. https://archive.ics.uci.edu/ml/datasets/cardiotocography

[71] user6903745, ‘Answer to 'Can anyone explain me StandardScaler?'’, *Stack Overflow*, 23 Νοέμβριος 2016. https://stackoverflow.com/a/40767144

[72] A. C. Müller και S. Guido, *Introduction to machine learning with Python: a guide for data scientists*, First edition. Sebastopol, CA: O’Reilly Media, Inc, 2016.

[73] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, και J. Santos, ‘Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier]’, *IEEE Comput. Intell. Mag.*, τ. 13, τχ. 4, σσ. 59–76, Νοεμβρίου 2018, doi: 10.1109/MCI.2018.2866730.