ARISTOTLE UNIVERSITY OF THESSALONIKI

Διάγνωση της νόσου του Πάρκινσον με τη χρήση μηχανικής εκμάθησης.

Άσκηση Εξαμήνου 2020-2021
«Τεχνικές Μηχανικής Μάθησης»

Καλομιτσίνης Γεώργιος
Α.Ε.Μ: 485
Α.Π.Μ: 1183148938140713

«Τεχνικές Μηχανικής Μάθησης»

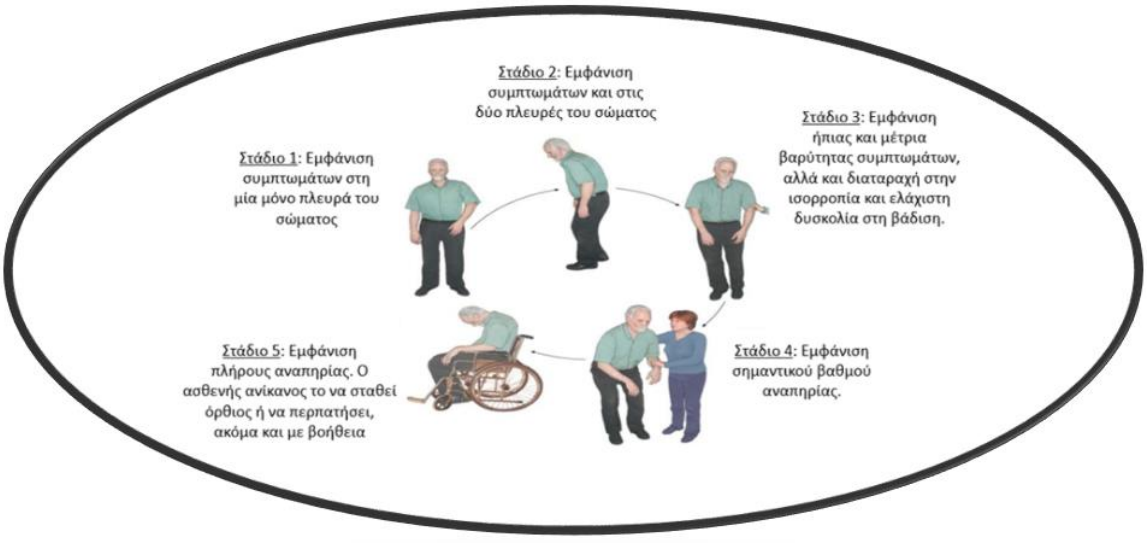
Επ. Καθηγητής: κ. Σαράφης Ιωάννης

I.Περίληψη

Η διάγνωση της νόσου του Πάρκινσον μέσω μηχανικής μάθησης παρέχει καλύτερη εποπτεία της νόσου. Η παρούσα εργασία αναπτύχθηκε σε εργασιακό περιβάλλον με γνώμονα τη γλώσσα Python (version 3.8.2) για περαιτέρω στατιστική ανάλυση, ταξινόμηση και αξιολόγηση των αλγορίθμων που λήφθηκαν υπόψιν. Το σύνολο δεδομένων φωνής για τη νόσο του Πάρκινσον ανακτήθηκε από το αποθετήριο μηχανικής μάθησης UCI [1]. Συγκεκριμένα, περιέχει μετρήσεις φωνής από υγιείς ανθρώπους και από ανθρώπους με Πάρκινσον με τις εξής μετρήσεις: *MDVP: Fo (Hz), MDVP: Fhi (Hz), MDVP: Flo (Hz), MDVP: Jitter (%), MDVP: Jitter (Abs), MDVP: RAP, MDVP: PPQ, Jitter: DDP, MDVP: Shimmer, MDVP: Shimmer (dB), Shimmer: APQ3, Shimmer: APQ5, MDVP: APQ, Shimmer: DDA, NHR, HNR, κατάσταση, RPDE, DFA, spread1, spread2, D2 και PPE*. Ο Logistic Regression αλγόριθμος εμφάνισε την καλύτερη επίδοση και ο Support Vector Machine (SVM) τη χειρότερη, σε σύγκριση με τους υπόλοιπους.

II.Εισαγωγή

Η νόσος του Πάρκινσον είναι μια από τις πιο επώδυνες, επικίνδυνες και μη θεραπευτικές ασθένειες που εμφανίζονται σε μεγαλύτερες ηλικίες (κυρίως άνω των 50 ετών) στους ανθρώπους. Εκδηλώνεται ως ο θάνατος των ντοπαμινικών νευρώνων στον εγκέφαλο. Αυτός ο νευροεκφυλισμός οδηγεί σε ένα εύρος συμπτωμάτων, όπως θέματα συντονισμού, βραδυκινησία, φωνητικές αλλαγές, δυσκαμψία ακόμα και προοδευτική αναπηρία. Τα συμπτώματα και η πορεία της νόσου ποικίλλουν, οπότε συχνά δεν διαγιγνώσκεται για πολλά χρόνια. Μέχρι στιγμής, δεν υπάρχει θεραπεία, αν και υπάρχει φαρμακευτική αγωγή που προσφέρει σημαντική επιβράδυνση των συμπτωμάτων, ειδικά στα αρχικά στάδια της νόσου [2]. Επομένως, υπάρχει ανάγκη για δημιουργία πιο ευαίσθητων διαγνωστικών εργαλείων για την ανίχνευση της νόσου, καθώς όσο η ασθένεια εξελίσσεται, τόσο περισσότερα συμπτώματα προκύπτουν που την καθιστούν δύσκολα να αντιμετωπιστεί.



1. Τα 5 στάδια της νόσου Πάρκινσον.

Οι βιοδείκτες που προέρχονται από την ανθρώπινη φωνή μπορούν να προσφέρουν σημαντικές πληροφορίες σχετικά με νευρολογικές διαταραχές, όπως η νόσος του Πάρκινσον, λόγω των υποκείμενων γνωστικών και νευρομυϊκών τους λειτουργιών. Με τις εξελίξεις στην τεχνολογία, αξιόπιστα μοντέλα μπορούν να μεταφράσουν αυτά τα δεδομένα ήχου όπου θα μπορούσαν ενδεχομένως να παρέχουν διαγνώσεις για τη νόσο, τα οποία είναι σε χαμηλό κόστος και υψηλής ακρίβειας. Σε αυτή την εργασία, αξιοποιήθηκε το συγκεκριμένο σύνολο δεδομένων [1] για τη νόσο του Πάρκισον, για την ανάπτυξη σχετικών μοντέλων για τη διάγνωση της νόσου και σύγκριση μεταξύ τους για το ποιο εν τέλει θα κριθεί ακριβέστερο. Οι αλγόριθμοι είναι οι εξής:

- Logistic-Regression: Η ταξινόμηση Logistic Regression πραγματοποιείται με βάση τη συνάρτηση σιγμοειδούς (sigmoid function), γνωστή ως συνάρτηση logistic που λαμβάνει μια πραγματική είσοδο και δίνει μια τιμή μεταξύ 0 και 1 [6].

- DecisionTree: Στην περίπτωση του DecisionTree ταξινομητή, η είσοδος χωρίζεται σε υποδιαστήματα βάσει συγκεκριμένων λειτουργιών. Βοηθά στην επίτευξη ενός συμπεράσματος που βασίζεται σε δηλώσεις υπό όρους ελέγχου [7].

- GaussianNB: Ο ταξινομητής Naive Bayes υποθέτει ότι το αποτέλεσμα ενός συγκεκριμένου χαρακτηριστικού σε μια κλάση, είναι ανεξάρτητο από τα υπόλοιπα χαρακτηριστικά [8].

- Random Forest: Δημιουργεί τυχαία δέντρα αποφάσεων και υπολογίζει το αποτέλεσμα κατά μέσο όρο, μειώνοντας έτσι την υπερβολική εφαρμογή (overfitting) του μοντέλου [9]. Επιλέγει μόνο ένα υποσύνολο χαρακτηριστικών τυχαία και το καλύτερο χαρακτηριστικό χρησιμοποιείται για διαχωρισμό στον κάθε κόμβο.

- Support-Vector-Machine: Ο αλγόριθμος αυτός στοχεύει στην εύρεση ενός υπερπλάνου σε έναν N-διαστατικό χώρο (N - ο αριθμός των χαρακτηριστικών) που ταξινομεί με σαφήνεια τα σημεία δεδομένων [10].

- XGB classifier: Ο ταξινομητής XGBoost είναι μια υλοποίηση ενισχυμένων δέντρων λήψης αποφάσεων (gradient boosted), σχεδιασμένα για ταχύτητα και απόδοση [11].

III.Ανάλυση των Χαρακτηριστικών

Το εύρος δεδομένων των βιοϊατρικών φωνητικών μετρήσεων προέρχεται από 31 άτομα, εκ των οποίων τα 23 άτομα πάσχουν από τη νόσο του Πάρκινσον. Συγκεκριμένα, το σύνολο δεδομένων αποτελείται από 24 χαρακτηριστικά τα οποία είναι:

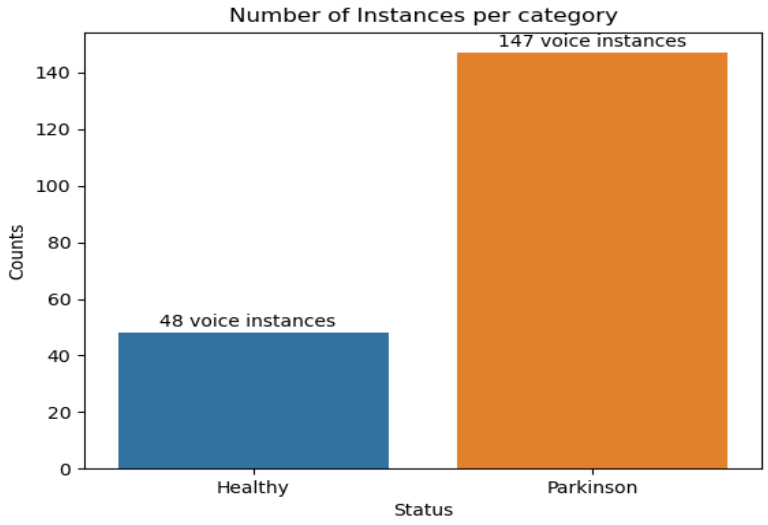
- τα ονόματα *(name)* των μετρήσεων κάθε ατόμου,

- τρία είδη θεμελιωδών συχνοτήτων (*MDVP:Fo(Hz)* μέσο όρο, *MDVP:Fhi(Hz)* ελάχιστο και *MDVP:Flo(Hz)* μέγιστο,
- μετρήσεις σχετικά με τη διακύμανση στη θεμελιώδη συχνότητα *(jitter)*, και στο πλάτος *(shimmer)*.
- δύο μη γραμμικές δυναμικής πολυπλοκότητας μετρήσεις *(RPDE,D2)*,
- έναν εκφραστή κλιμάκωσης φράκταλ *(DFA)*
- τρία μη γραμμικά μέτρα θεμελιώδους διακύμανσης συχνότητας (*spread1,spread2,PPE).*
- και τέλος, το σύνολο δεδομένων περιέχει και πληροφορίες σχετικά με την κατάσταση των ατόμων *(status).*

Επιπλέον, στη συνέχεια παρουσιάζεται ο τύπος κάθε χαρακτηριστικού:

```
#   Column           Non-Null Count  Dtype
---  ------          --------------  -----
 0   name            195 non-null    object
 1   MDVP:Fo(Hz)     195 non-null    float64
 2   MDVP:Fhi(Hz)    195 non-null    float64
 3   MDVP:Flo(Hz)    195 non-null    float64
 4   MDVP:Jitter(%)  195 non-null    float64
 5   MDVP:Jitter(Abs)  195 non-null  float64
 6   MDVP:RAP        195 non-null    float64
 7   MDVP:PPQ        195 non-null    float64
 8   Jitter:DDP      195 non-null    float64
 9   MDVP:Shimmer    195 non-null    float64
 10  MDVP:Shimmer(dB)  195 non-null  float64
 11  Shimmer:APQ3    195 non-null    float64
 12  Shimmer:APQ5    195 non-null    float64
 13  MDVP:APQ        195 non-null    float64
 14  Shimmer:DDA     195 non-null    float64
 15  NHR             195 non-null    float64
 16  HNR             195 non-null    float64
 17  status          195 non-null    int64
 18  RPDE            195 non-null    float64
 19  DFA             195 non-null    float64
 20  spread1         195 non-null    float64
 21  spread2         195 non-null    float64
 22  D2              195 non-null    float64
 23  PPE             195 non-null    float64
dtypes: float64(22), int64(1), object(1)
memory usage: 36.7+ KB
```

Η στήλη “status” αντιπροσωπεύει "κατάσταση" που έχει οριστεί σε 0 για υγιή και 1 για ασθενείς με Πάρκινσον. Τα δεδομένα είναι σε μορφή ASCII CSV. Στη συνέχεια, ακολουθεί η παρουσίαση σχετικών γραφημάτων για την περιγραφή των χαρακτηριστικών των δεδομένων, καθώς και για την κατάσταση κάθε μέτρησης φωνής.

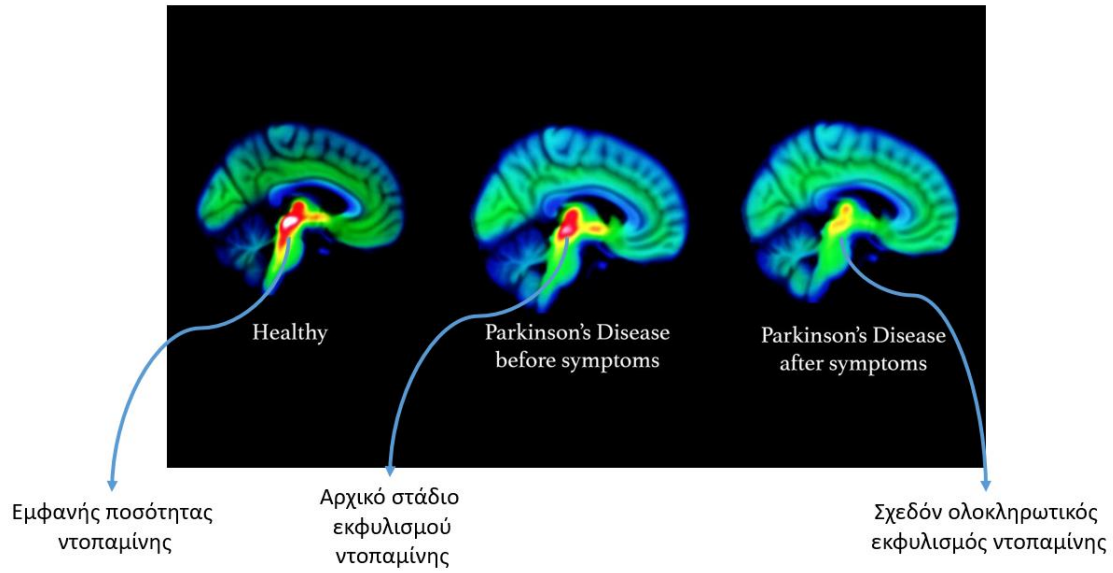


2. Απεικόνιση της κατάστασης των ασθενών.

	count	mean	std	min	25%	50%	75%	max
MDVP:Fo(Hz)	195.0	154.228641	41.390065	88.333000	117.572000	148.790000	182.769000	260.105000
MDVP:Fhi(Hz)	195.0	197.104918	91.491548	102.145000	134.862500	175.829000	224.205500	592.030000
MDVP:Flo(Hz)	195.0	116.324631	43.521413	65.476000	84.291000	104.315000	140.018500	239.170000
MDVP:Jitter(%)	195.0	0.006220	0.004848	0.001680	0.003460	0.004940	0.007365	0.033160
MDVP:Jitter(Abs)	195.0	0.000044	0.000035	0.000007	0.000020	0.000030	0.000060	0.000260
MDVP:RAP	195.0	0.003306	0.002968	0.000680	0.001660	0.002500	0.003835	0.021440
MDVP:PPQ	195.0	0.003446	0.002759	0.000920	0.001860	0.002690	0.003955	0.019580
Jitter:DDP	195.0	0.009920	0.008903	0.002040	0.004985	0.007490	0.011505	0.064330
MDVP:Shimmer	195.0	0.029709	0.018857	0.009540	0.016505	0.022970	0.037885	0.119080
MDVP:Shimmer(dB)	195.0	0.282251	0.194877	0.085000	0.148500	0.221000	0.350000	1.302000
Shimmer:APQ3	195.0	0.015664	0.010153	0.004550	0.008245	0.012790	0.020265	0.056470
Shimmer:APQ5	195.0	0.017878	0.012024	0.005700	0.009580	0.013470	0.022380	0.079400
MDVP:APQ	195.0	0.024081	0.016947	0.007190	0.013080	0.018260	0.029400	0.137780
Shimmer:DDA	195.0	0.046993	0.030459	0.013640	0.024735	0.038360	0.060795	0.169420
NHR	195.0	0.024847	0.040418	0.000650	0.005925	0.011660	0.025640	0.314820
HNR	195.0	21.885974	4.425764	8.441000	19.198000	22.085000	25.075500	33.047000
status	195.0	0.753846	0.431878	0.000000	1.000000	1.000000	1.000000	1.000000
RPDE	195.0	0.498536	0.103942	0.256570	0.421306	0.495954	0.587562	0.685151
DFA	195.0	0.718099	0.055336	0.574282	0.674758	0.722254	0.761881	0.825288
spread1	195.0	-5.684397	1.090208	-7.964984	-6.450096	-5.720868	-5.046192	-2.434031
spread2	195.0	0.226510	0.083406	0.006274	0.174351	0.218885	0.279234	0.450493
D2	195.0	2.381826	0.382799	1.423287	2.099125	2.361532	2.636456	3.671155
PPE	195.0	0.206552	0.090119	0.044539	0.137451	0.194052	0.252980	0.527367

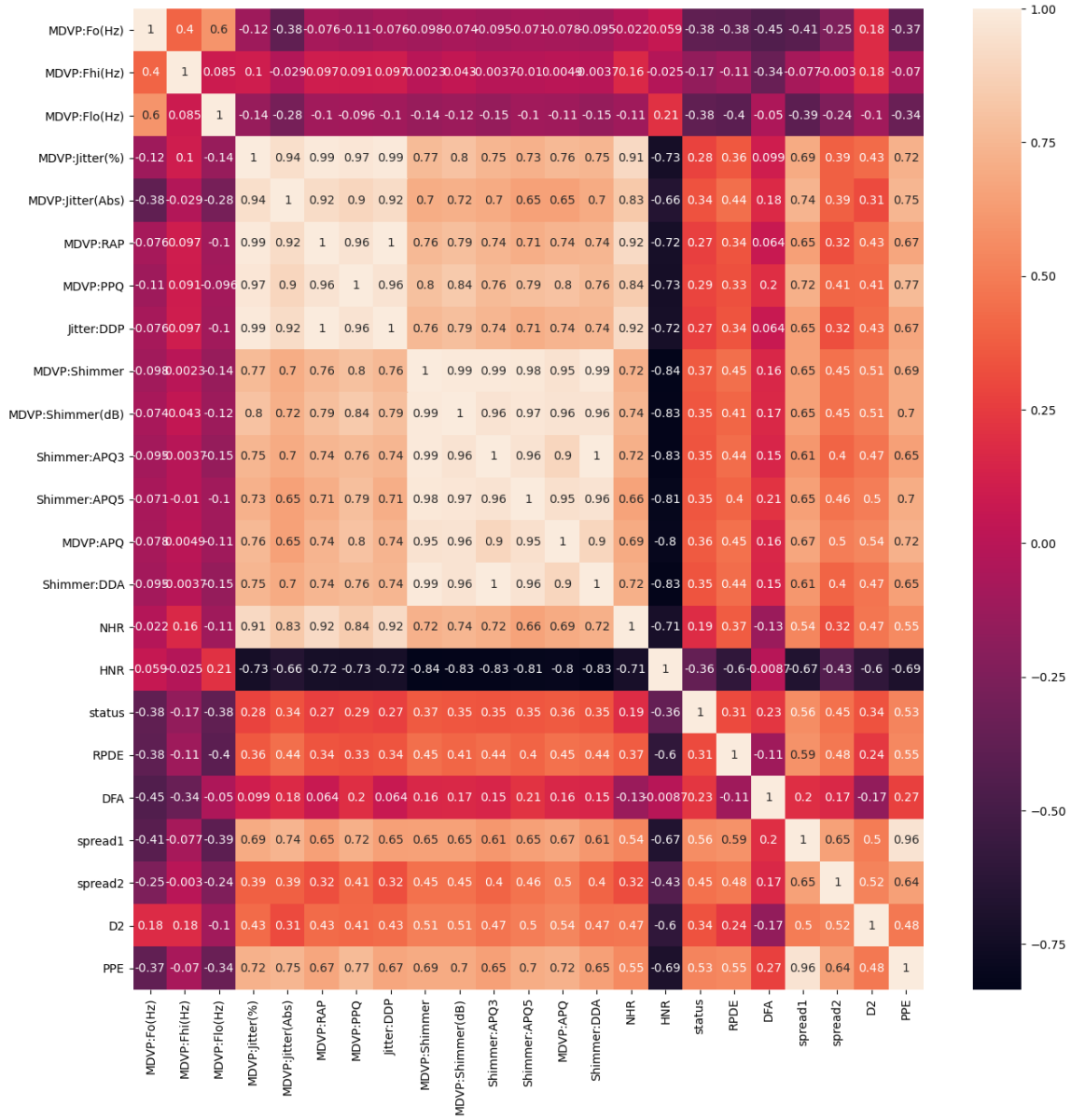
3. Στατιστικά στοιχεία για τα δεδομένα.

Από τη γραφική αναπαράσταση 1, ο αριθμός των φωνητικών μετρήσεων από ανθρώπους με Parkinson είναι 147 , σε ποσοστό 75.38 % και ο αριθμός των υγιή ατόμων είναι 48 με ποσοστό 24.62 %.



4. Η πορεία εξέλιξης του εγκεφάλου.

Στη συνέχεια, παρουσιάζεται η συσχέτιση των χαρακτηριστικών του συνόλου δεδομένων.



5. Heatmap για την απεικόνιση της συσχέτισης των χαρακτηριστικών.

Όπως διακρίνεται και από το γράφημα 5, ορισμένα χαρακτηριστικά έχουν υψηλό βαθμό συσχέτισης και άλλα όχι. Γι' αυτό το λόγο, θα γίνει επιλογή χαρακτηριστικών ως είσοδο στους αλγόριθμους, με στόχο τη μείωση του υπολογιστικού κόστους της μοντελοποίησης αλλά και ταυτόχρονα τη βελτίωση των αποδόσεων των μοντέλων. Άσχετα ή εν μέρει σχετικά χαρακτηριστικά μπορούν να επηρεάσουν αρνητικά την απόδοση του μοντέλου [3].

IV.Επιλογή Χαρακτηριστικών (Feature Selection)

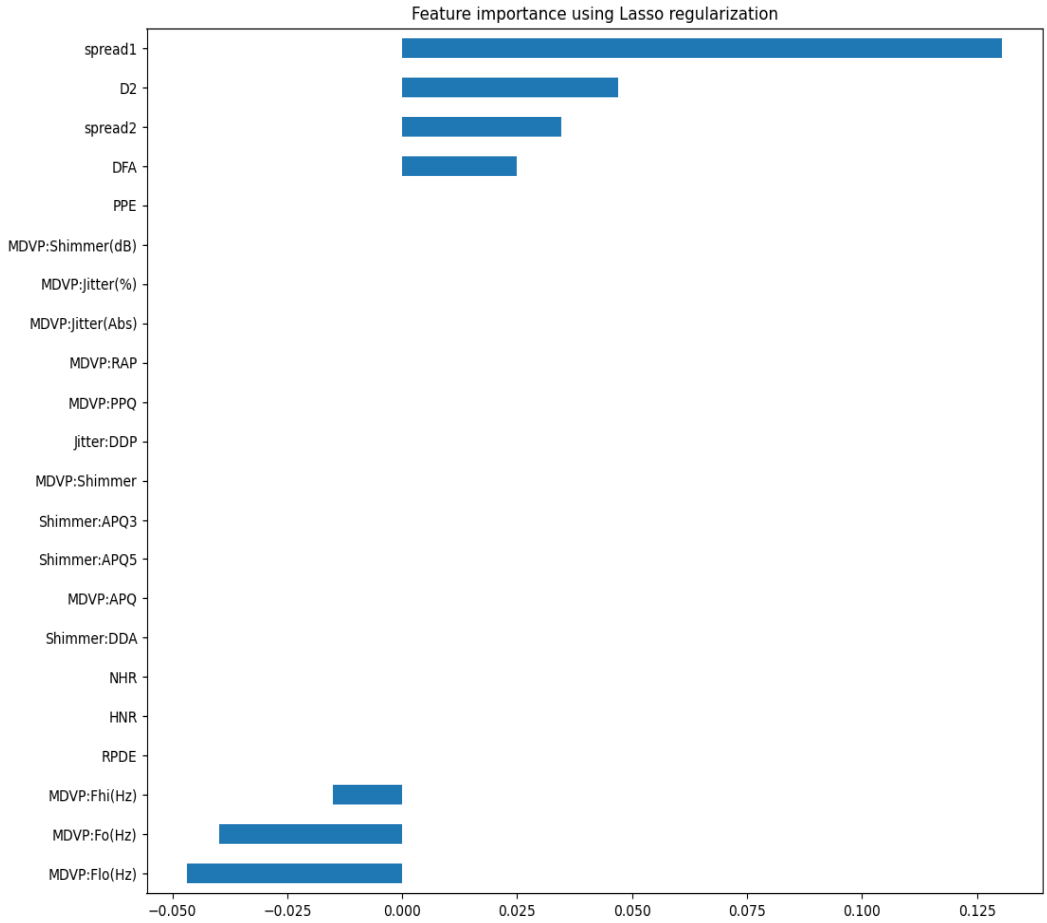
Η επιλογή χαρακτηριστικών είναι η διαδικασία που επιλέγει έναν μειωμένο αριθμό επεξηγηματικής μεταβλητής για να περιγράψει μια μεταβλητή απόκρισης. Οι κύριοι λόγοι για τους οποίους χρησιμοποιείται η επιλογή χαρακτηριστικών είναι:

- Να διευκολύνει την ερμηνεία του μοντέλου, αφαιρώντας μεταβλητές που είναι περιττές και δεν προσθέτουν πληροφορίες.
- Μείωση του μεγέθους του συνόλου δεδομένων με στόχο οι αλγόριθμοι να λειτουργούν πιο γρήγορα, καθιστώντας δυνατή τη διαχείριση δεδομένων υψηλής διάστασης,
- Μείωση κινδύνου *overfitting* .

Στο παρόν πρόβλημα, το σύνολο δεδομένων περιέχει 22 μεταβλητές, οπότε είναι αναγκαία η διαδικασία της επιλογής των χαρακτηριστικών. Όπως αναφέρθηκε και προηγουμένως, ορισμένα χαρακτηριστικά έχουν υψηλό βαθμό συσχέτισης και άλλα όχι. Στη βιβλιογραφία υπάρχουν διάφοροι τύποι μεθόδων για την εξαγωγή των πιο σημαντικών χαρακτηριστικών. Στη συγκεκριμένη εργασία, πραγματοποιήθηκε Lasso ομαλοποίηση των χαρακτηριστικών, όπου έχει την ιδιότητα της συρρίκνωσης μη σημαντικών συντελεστών (coefficients) στο μηδέν. Πριν τη διαδικασία επιλογής των χαρακτηριστικών, τα δεδομένα κλιμακώθηκαν, με *StandardScaler(),* αφού οι τιμές των στιγμιοτύπων κάθε χαρακτηριστικού ποικίλουν.

Ειδικότερα, η μέθοδος Lasso θέτει έναν περιορισμό στο άθροισμα των απόλυτων τιμών των παραμέτρων, όπου το άθροισμα πρέπει να είναι μικρότερο από μια σταθερή τιμή (άνω όριο). Για να πραγματοποιηθεί αυτό, η μέθοδος εφαρμόζει μια διαδικασία συρρίκνωσης (κανονικοποίηση) όπου «τιμωρεί» τους συντελεστές των μεταβλητών, συρρικνώνοντας μερικούς στο μηδέν. Κατά τη διάρκεια της διαδικασίας επιλογής χαρακτηριστικών, οι μεταβλητές που εξακολουθούν να έχουν μη μηδενικό συντελεστή μετά τη διαδικασία συρρίκνωσης επιλέγονται ως μέρος του μοντέλου. Ο στόχος αυτής της διαδικασίας είναι να ελαχιστοποιηθεί το σφάλμα πρόβλεψης. Πρακτικά, η παράμετρος λ, που ελέγχει την ισχύ της ποινής, έχει μεγάλη σημασία. Όταν το λ είναι αρκετά μεγάλο, οι συντελεστές αναγκάζονται να είναι ίσοι με το μηδέν, όπου με αυτόν τον τρόπο μειώνεται η διάσταση του προβλήματος. Όσο μεγαλύτερη είναι η παράμετρος λ, τόσο περισσότερος αριθμός συντελεστών συρρικνώνεται στο μηδέν. Όταν $\lambda = 0$, έχουμε παλινδρόμηση OLS (Ordinary Least Sqaure). Το κύριο πλεονέκτημα στη χρήση της μεθόδου Lasso είναι ότι η συρρίκνωση και η αφαίρεση των συντελεστών μειώνει τη διακύμανση του συνόλου δεδομένων, χωρίς σημαντική αύξηση του *bias* [4].

Τα αποτελέσματα ομαλοποίησης των χαρακτηριστικών είναι:



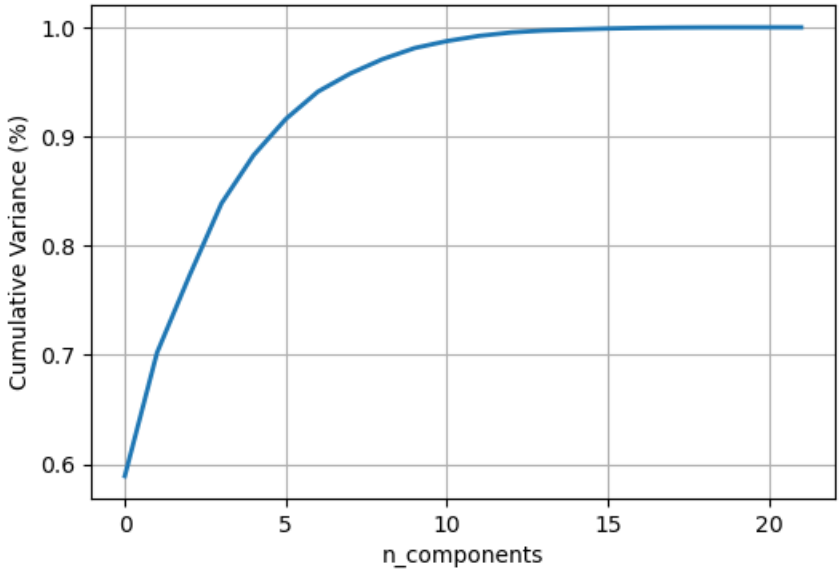
6. Lasso regularization.

Οπότε από την ομαλοποίηση Lasso, τα σημαντικά χαρακτηριστικά όπως προκύπτουν, με τους αντίστοιχους συντελεστές τους είναι:

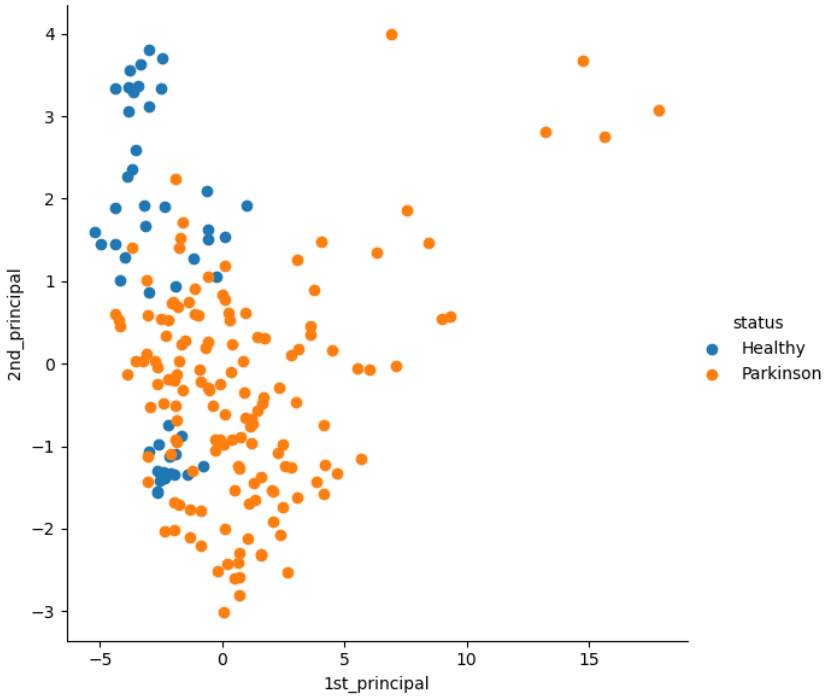
MDVP:Fo(Hz)	-0.039797
MDVP:Fhi(Hz)	-0.015180
MDVP:Flo(Hz)	-0.046725
DFA	0.024883
spread1	0.130428
spread2	0.034689
D2	0.046906

Επιπλέον, εφαρμόστηκε και η μέθοδος PCA (Pricipal Component Analysis) στο σύνολο δεδομένων. Η PCA μέθοδος είναι μια τεχνική προβολής δεδομένων από χώρο υψηλότερης διάστασης σε χώρο χαμηλότερης, μεγιστοποιώντας τη διακύμανση κάθε διάστασης [5]. Όπως παρατηρείται και από τα παρακάτω γραφήματα, εύκολα διαπιστώνει

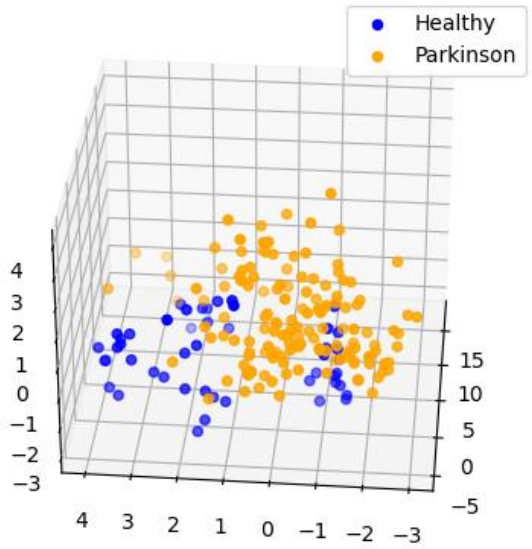
κανείς ότι η συνολική πολυπλοκότητα του συνόλου δεδομένων μπορεί να περιγραφθεί και από λιγότερα χαρακτηριστικά από ό,τι είναι. Επομένως, και από τις δυο μεθόδους προκύπτει ότι η μείωση του πλήθους των χαρακτηριστικών είναι ικανή για την συνολική διακύμανση του συνόλου δεδομένων, και μάλιστα μ' αυτό τον τρόπο, μειώνεται και το υπολογιστικό κόστος των αλγορίθμων.



7. Συνολική πολυπλοκότητα του dataset, ως προς το πλήθος των χαρακτηριστικών του.



8. Αναπαράσταση των στιγμιότυπων του συνόλου δεδομένων σε 2 διαστάσεις.



9. Αναπαράσταση των στιγμιότυπων του συνόλου δεδομένων σε 3 διαστάσεις.

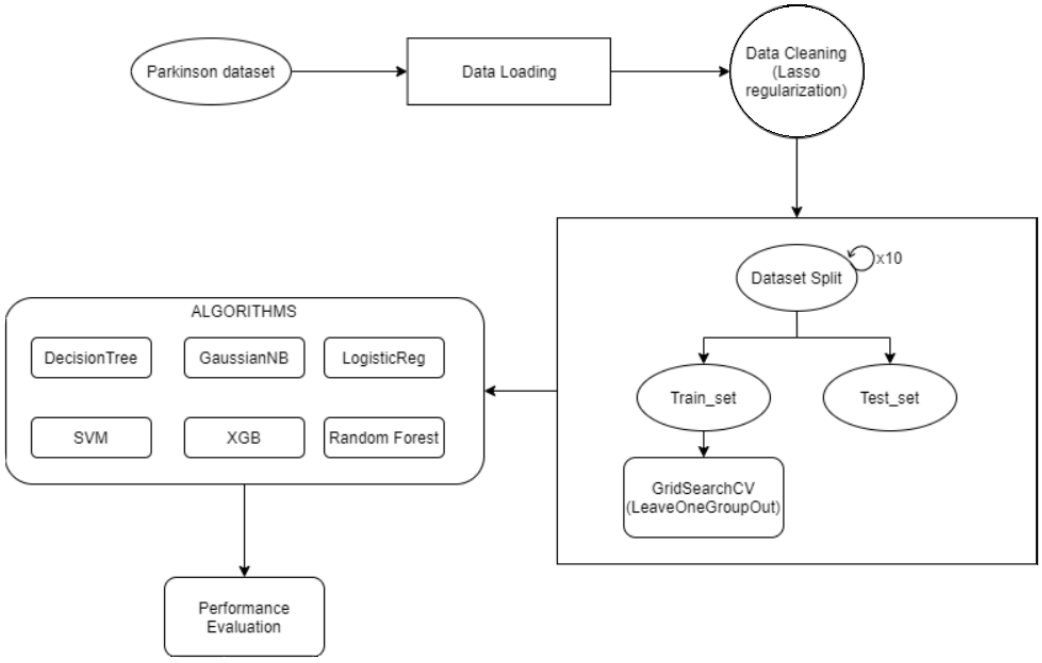
Επομένως, από τα 22 χαρακτηριστικά, έπειτα από την Lasso ομαλοποίηση, 7 προκύπτουν να είναι τα σημαντικά, που θα αποτελέσουν την είσοδο στους αλγορίθμους μηχανικής μάθησης για τη διάγνωση της νόσου.

V.Μεθοδολογία

Κάθε άτομο στο σύνολο δεδομένων έχει 6 ή 7 φωνητικές μετρήσεις που αντιστοιχούν σε αυτό. Για την αξιολόγηση του κάθε αλγόριθμου που λήφθηκε υπόψιν, το σύνολο δεδομένων χωρίστηκε ανά άτομα και όχι σε επίπεδο μετρήσεων φωνής. Με αυτό τον τρόπο, επιτυγχάνεται ο «σωστός» διαχωρισμός του συνόλου δεδομένων, και λέγοντας σωστός, εννοείται ότι το σύνολο δεδομένων διαχωρίζεται με βάση τα άτομα. Διαφορετικά άτομα στο training set και διαφορετικά στο test set. Έτσι, ο κάθε αλγόριθμος αξιολογείται σε διαφορετικούς ανθρώπους από ό,τι εκπαιδεύτηκε, με αποτέλεσμα η αξιολόγηση τους να γίνεται σε ανθρώπους που «βλέπει» πρώτη φορά.

Επιπλέον, ο διαχωρισμός του συνόλου των δεδομένων πραγματοποιήθηκε 10 φορές, με διαφορετικούς ανθρώπους στο train set και test set, με train_size = 0,8, όπου ισοδυναμεί με 25 ανθρώπους. Για τη εύρεση των καλύτερων υπερπαραμέτρων του κάθε αλγόριθμου εφαρμόστηκε η διαδικασία GridSearchCV με τη μέθοδο LeaveOneGroupOut. Σε κάθε διαχωρισμό του συνόλου δεδομένων, στο train set εφαρμόστηκε η προαναφερθείσα μέθοδος, αφήνοντας ένα άτομο εκτός, δηλαδή 6 ή 7 μετρήσεις για την εύρεση των

καλύτερων υπερπαραμέτρων. Το pipeline της εργασίας συνοψίζεται στο παρακάτω γράφημα:



10. Μεθοδολογία εργασίας.

Για την αξιολόγηση κάθε αλγορίθμου, υπολογίσθηκαν οι εξής μετρικές αξιολόγησης:

- accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- recall:

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

- precision:

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

- f1-score:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad , \quad \text{όπου:}$$

- *TP*: True Positives
- *TN*: True Negatives
- *FP*: False Positives και
- *FN*: False Negatives

Εξαιτίας της φύσης του προβλήματος, ως ιατρικού, στόχος είναι η μείωση θετικών λανθασμένων δειγμάτων στον υπολογισμό. Είτε το μέτρο precision, είτε το recall δεν καλύπτουν πλήρως το σκοπό αυτό, όπως και το accuracy. Οπότε, για καλύτερα αποτελέσματα λαμβάνεται υπόψιν το μέτρο f1-score, όπου επιδιώκεται μια ισορροπία μεταξύ των precison και recall ακόμα και σε imbalanced κλάσεις, όπως και στην περίπτωση μας. Επιπλέον, τιμωρούνται οι μεγάλες διαφορές ανάμεσα στα precison και recall, με αποτέλεσμα να υπάρχει μία «δίκαιη ανταλλαγή» μεταξύ τους. Έτσι, ως καλύτερος ταξινομητής, είναι εκείνος με τη μέγιστη τιμή του f1-score.

VI.Παρουσίαση αποτελεσμάτων αλγορίθμων

Ακολουθεί η παρουσίαση των αποτελεσμάτων των μετρικών αξιολόγησης των αλγορίθμων στο train set και στο test set, ανά διαχωρισμό:

1ος Διαχωρισμός:

Train_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	84.87%	92.76%	91.45%	23.68%	92.76%	94.08%
Precision	89.74%	92.00%	89.92%	0.00%	93.39%	94.96%
Recall	90.52%	99.14%	100.00%	0.00%	97.41%	97.41%
F1-score	90.13%	95.44%	94.69%	0.00%	95.36%	96.17%

Test_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	69.77%	72.09%	72.09%	27.91%	76.74%	69.77%
Precision	71.43%	72.09%	72.09%	0.00%	75.61%	73.68%
Recall	96.77%	100.00%	100.00%	0.00%	100.00%	90.32%
F1-score	82.19%	83.78%	83.78%	0.00%	86.11%	81.16%

2ος Διαχωρισμός:

Train_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	79.74%	80.39%	80.39%	99.35%	97.39%	92.16%
Precision	80.67%	80.39%	80.39%	99.19%	96.85%	91.11%
Recall	98.37%	100.00%	100.00%	100.00%	100.00%	100.00%
F1-score	88.64%	89.13%	89.13%	99.60%	98.40%	95.35%

Test_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	54.76%	57.14%	57.14%	57.14%	69.05%	83.33%
Precision	56.10%	57.14%	57.14%	61.54%	70.37%	77.42%
Recall	95.83%	100.00%	100.00%	66.67%	79.17%	100.00%
F1-score	70.77%	72.73%	72.73%	64.00%	74.51%	87.27%

3ος Διαχωρισμός:

Train_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	94.12%	94.77%	94.12%	50.00%	99.35%	100.00%
Precision	94.12%	94.16%	93.48%	46.15%	99.23%	100.00%
Recall	99.22%	100.00%	100.0%	100.00%	100.00%	100.00%
F1-score	96.60%	96.99%	96.63%	63.16%	99.61%	100.00%

Test_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	42.86%	52.38%	50.00%	94.12%	57.14%	47.62%
Precision	41.67%	47.37%	46.15%	93.48%	50.00%	44.44%
Recall	83.33%	100.00%	100.00%	100.0%	88.89%	88.89%
F1-score	55.56%	64.29%	63.16%	96.63%	64.00%	59.26%

4ος Διαχωρισμός:

Train_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	81.70%	27.45%	85.62%	72.55%	90.85%	98.69%
Precision	83.20%	0.00%	85.60%	72.55%	90.76%	100.00%
Recall	93.69%	0.00%	96.40%	100.00%	97.30%	98.20%
F1-score	88.14%	0.00%	90.68%	84.09%	93.91%	99.09%

Test_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	92.86%	14.29%	80.95%	85.71%	88.10%	80.95%
Precision	100.00%	0.00%	100.00%	85.71%	100.00%	93.75%
Recall	91.67%	0.00%	77.78%	100.00%	86.11%	83.33%
F1-score	95.65%	0.00%	87.50%	92.31%	92.54%	88.24%

5ος Διαχωρισμός:

Train_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	80.92%	27.63%	82.24%	84.21%	90.79%	88.82%
Precision	80.00%	0.00%	80.29%	82.58%	91.38%	86.61%
Recall	98.18%	0.00%	100.00%	99.09%	96.36%	100.00%
F1-score	88.16%	0.00%	89.07%	90.08%	93.81%	92.83%

Test_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	83.72%	13.95%	100.00%	97.67%	86.05%	90.70%
Precision	87.50%	0.00%	100.00%	100.00%	100.00%	100.00%
Recall	94.59%	0.00%	100.00%	97.30%	83.78%	89.19%
F1-score	90.91%	0.00%	100.00%	98.63%	91.18%	94.29%

6ος Διαχωρισμός:

Train_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	84.31%	100.00%	85.62%	76.47%	88.89%	88.24%
Precision	87.80%	100.00%	86.82%	76.47%	89.68%	86.67%
Recall	92.31%	100.00%	95.73%	100.00%	96.58%	100.00%
F1-score	90.00%	100.00%	91.06%	86.67%	93.00%	92.86%

Test_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	88.10%	83.33%	85.71%	71.43%	88.10%	85.71%
Precision	87.88%	87.10%	85.29%	71.43%	87.88%	83.33%
Recall	96.67%	90.00%	96.67%	100.00%	96.67%	100.00%
F1-score	92.06%	88.52%	90.62%	83.33%	92.06%	90.91%

7ος Διαχωρισμός:

Train_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	84.97%	79.08%	86.93%	76.47%	94.12%	87.58%
Precision	85.07%	78.52%	85.40%	76.47%	93.55%	86.03%
Recall	97.44%	100.00%	100.00%	100.00%	99.15%	100.00%
F1-score	90.84%	87.97%	92.13%	86.67%	96.27%	92.49%

Test_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	78.57%	71.43%	83.33%	71.43%	76.19%	80.95%
Precision	76.92%	71.43%	81.08%	71.43%	83.33%	78.95%
Recall	100.00%	100.00%	100.00%	100.00%	83.33%	100.00%
F1-score	86.96%	83.33%	89.55%	83.33%	83.33%	88.24%

8ος Διαχωρισμός:

Train_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	81.46%	100.00%	84.11%	83.44%	84.09%	97.35%
Precision	82.05%	100.00%	81.10%	81.97%	98.04%	96.26%
Recall	93.20%	100.00%	100.00%	97.09%	97.09%	100.00%
F1-score	87.27%	100.00%	89.57%	88.89%	97.56%	98.10%

Test_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	77.27%	88.64%	97.73%	95.45%	96.69%	100.00%
Precision	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Recall	77.27%	88.64%	97.73%	95.45%	84.09%	100.00%
F1-score	87.18%	93.98%	98.85%	97.67%	91.36%	100.00%

9ος Διαχωρισμός:

Train_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	84.87%	99.34%	88.82%	23.68%	100.00%	89.47%
Precision	83.45%	100.00%	87.22%	0.00%	100.00%	87.88%
Recall	100.00%	99.14%	100.00%	0.00%	100.00%	100.00%
F1-score	90.98%	99.57%	93.17%	0.00%	100.00%	93.55%

Test_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	72.09%	86.05%	83.72%	27.91%	88.37%	72.09%
Precision	72.09%	83.78%	81.58%	0.00%	90.62%	72.09%
Recall	100.00%	100.00%	100.00%	0.00%	93.55%	100.00%
F1-score	83.78%	91.18%	89.86%	0.00%	92.06%	83.78%

10ος Διαχωρισμός:

Train_set results:

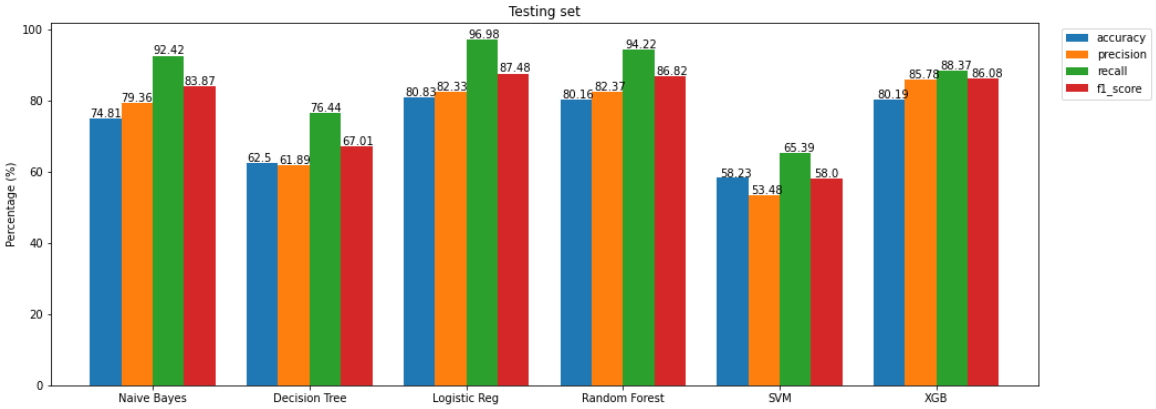
	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	82.35%	100.00%	84.97%	31.37%	99.35%	100.00%
Precision	81.45%	100.00%	82.54%	0.00%	100.00%	100.00%
Recall	96.19%	100.00%	99.05%	0.00%	99.05%	100.00%
F1-score	88.21%	100.00%	90.04%	0.00%	99.52%	100.00%

Test_set results:

	GaussianNB	DecisionTree	Logistic	SVM	XGB	RandomForest
Accuracy	88.10%	85.71%	97.62%	0.00%	88.10%	90.48%
Precision	100.00%	100.00%	100.00%	0.00%	100.00%	100.00%
Recall	88.10%	85.71%	97.62%	0.00%	88.10%	90.48%
F1-score	93.67%	92.31%	98.80%	0.00%	93.67%	95.00%

VII.Συμπεράσματα

Έπειτα από τον υπολογισμό των μετρικών αξιολογήσεων σε κάθε διαχωρισμό, υπολογίστηκε ο μέσος όρος αυτών, για την εύρεση του καλύτερου ταξινομητή σε κάθε σύνολο αξιολόγησης (test_set). Ο μέσος όρος των μετρικών είναι:



11. Αποτελέσματα μέσων όρων των μετρικών των ταξινομητών.

Επιπλέον, στη συνέχεια παρουσιάζονται οι αλγόριθμοι που σημείωσαν την καλύτερη και τη χειρότερη επίδοση τόσο στο train_set, όσο και στο test_set:

<u>Training set</u>

	accuracy	precision	recall	f1_score
XGB	95.02	95.29	98.29	96.74
	accuracy	precision	recall	f1_score
SVM	66.53	58.27	69.62	63.26

<u>Testing set</u>

	accuracy	precision	recall	f1_score
Logistic Reg	80.83	82.33	96.98	87.48
	accuracy	precision	recall	f1_score
SVM	58.23	53.48	65.39	58.0

Όπως διακρίνεται από το γράφημα 11, ο αλγόριθμος Logistic Regression παρουσίασε τα καλύτερα αποτελέσματα και τα χειρότερα ο SVM. Επιπλέον, αναλυτικά αποτελέσματα κάθε αλγορίθμου σε καθένα από τους διαχωρισμούς βρίσκεται στο αρχείο "*ΑΝΑΛΥΤΙΚΑ_ΑΠΟΤΕΛΕΣΜΑΤΑ_ΜΟΝΤΕΛΩΝ.docx"* και όλα τα γραφήματα που έχουν υλοποιηθεί στο πλαίσιο της εργασίας.

Η επιστήμη απαιτεί συνεχή αναπαραγωγή και επιβεβαίωση σε νέα δεδομένα, για τα οποία η κοινή χρήση είναι βασική. Γνωρίζοντας όλο και περισσότερα για τη νόσο του Πάρκινσον μπορεί να βελτιωθεί η απόδοση των αλγορίθμων περαιτέρω και να παρέχεται ολοένα πιο εκτεταμένα στοιχεία κλινικομετρικής επικύρωσης, καθώς μέχρι σήμερα αποτελεί μία νόσο για την οποία δεν έχει ακόμα βρεθεί θεραπεία. Η αντιμετώπιση με τα φάρμακα είναι συμπωματική μορφή θεραπείας και όχι αιτιολογική, δηλαδή αντιμετωπίζει τα συμπτώματα της νόσου και όχι την αιτία. Η μηχανική μάθηση συνεισφέρει στην έγκαιρη ανίχνευση κάθε είδους ασθένειας, που είναι ουσιαστικός παράγοντας και μάλιστα επιτρέπει στρατηγικές διαχείρισης που χορηγούνται από τα πρώτα κιόλας στάδια της νόσου.

VIII.Βιβλιογραφία

[1] https://archive.ics.uci.edu/ml/datasets/parkinsons.

[2] Singh, N., Pillay, V., & Choonara, Y. E. (2007). Advances in the treatment of Parkinson's disease. *Progress in neurobiology*, *81*(1), 29-44.

[3] Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, *1*(1-4), 131-156.

[4] Fonti, V., & Belitser, E. (2017). Feature selection using lasso. *VU Amsterdam Research Paper in Business Analytics*, *30*, 1-25.

[5] Alpaydin, E. (2018). Classifying multimodal data. In *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2* (pp. 49-69).

[6] Stylianou, N., Akbarov, A., Kontopantelis, E., Buchan, I., & Dunn, K. W. (2015). Mortality risk prediction in burn injury: Comparison of logistic regression with machine learning approaches. *Burns*, *41*(5), 925-934.

[7] Mashat, A. F., Fouad, M. M., Philip, S. Y., & Gharib, T. F. (2012). A decision tree classification model for university admission system. *Editorial Preface*, *3*(10).

[8] Zhang, M. L., Peña, J. M., & Robles, V. (2009). Feature selection for multi-label naive Bayes classification. *Information Sciences*, *179*(19), 3218-3229.

[9] Kumari, G. P. (2012). A Study of Bagging and Boosting approaches to develop meta-classifier. *Engineering Science and Technology: An International Journal (ESTIJ)*, *2*(5), 850-855.

[10] Tax, D. M., & Duin, R. P. (2004). Support vector data description. *Machine learning*, *54*(1), 45-66.

[11] Kropf, M., Hayn, D., Morris, D., Radhakrishnan, A. K., Belyavskiy, E., Frydas, A., ... & Schreier, G. (2018). Cardiac anomaly detection based on time and frequency domain features using tree-based classifiers. *Physiological measurement*, *39*(11), 114001.