

Η παρούσα εργασία πραγματεύεται την ανάλυση δεδομένων αναφορικά με την διαγνωστική αξιολόγηση ασθενών με Parkinson, σύμφωνα με την κλίμακα UPDRS, χρησιμοποιώντας ένα σύνολο δεδομένων από το UCI ML Repository, συγκεκριμένα το: **Parkinsons Telemonitoring Data Set**: <http://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring>

Το συγκεκριμένο σύνολο δεδομένων αφορά στη διαγνωστική αξιολόγηση ασθενών με Parkinson, σύμφωνα με την κλίμακα UPDRS. Το σύνολο δεδομένων δημιουργήθηκε από το Πανεπιστήμιο της Οξφόρδης σε συνεργασία με δέκα ιατρικά κέντρα στις ΗΠΑ και την Intel για τον απαραίτητο εξοπλισμό καταγραφής. Τα δεδομένα αφορούν στατιστικά χαρακτηριστικά (16 παραμέτρους) από 5.875 ηχογραφήσεις ομιλίας 42 ατόμων (περίπου 200 από το καθένα) σε μια περίοδο έξι μηνών τηλε-παρακολούθησής τους. Στο σύνολο δεδομένων δίνονται δύο παράμετροι-στόχοι, το `motor_UPDRS` τα οποία είναι τα αντίστοιχα UPDRS scores που προέκυψαν από την ιατρική αξιολόγηση (ground truth). Βασικά, θέλουμε να προβλέψουμε τον βαθμό κινητικότητας με βάση το UPDRS score

1. Περιγραφή Data Set

Η φωνητική παρακολούθηση είναι ένα από τις σημαντικότερες διαδικασίες αναφορικά με την νόσο του Parkinson (Parkinson disease-PD), καθώς παρατηρείται πιθανότερη βλάβη σε περίπου 90% των ασθενών στα πρώιμα στάδια της νόσου. Ως εκ τούτου, υπάρχει αυξανόμενο ενδιαφέρον για την κατασκευή PD διαγνωστικών συστημάτων και συστημάτων τηλεπαρακολούθησης που βασίζονται σε φωνητικά χαρακτηριστικά. Τα συστήματα τηλεδιάγνωσης στοχεύουν στη διάκριση των ασθενών με PD από υγιή άτομα και τα συστήματα τηλεπαρακολούθησης στοχεύουν στην πρόβλεψη των μετρήσεων κλινικής αξιολόγησης για την παρακολούθηση της εξέλιξης της νόσου.

Το συγκεκριμένο σύνολο δεδομένων αφορά στη διαγνωστική αξιολόγηση ασθενών με Parkinson, σύμφωνα με την κλίμακα **unified Parkinson's disease rating scale** (UPDRS). Το αρχείο δεδομένων προς ανάλυση περιλαμβάνει δεδομένα βιοϊατρικών φωνητικών μετρήσεων και δημογραφικά στοιχεία (φύλο, ηλικία) από 42 άτομα που πάσχουν από τη νόσο του Πάρκινσον σε πρώιμο στάδιο. Συνολικά το σύνολο δεδομένων αποτελείται από 5875 εγγραφές, περιλαμβάνοντας περίπου 200 εγγραφές ανά άτομο. Το αρχείο δεδομένων περιλαμβάνει 22 χαρακτηριστικά σε κάθε εγγραφή. Πιο συγκεκριμένα τα χαρακτηριστικά του αντίστοιχου συνόλου δεδομένων είναι τα εξής:

Πίνακας 1: Πεδία συνόλου δεδομένων προς ανάλυση

| <u>ΑΑ</u> | <u>Όνομα πεδίου</u> | <u>Περιγραφή</u> | <u>Μορφή</u> |
|-----------|---------------------|--------------------------------------|--|
| <u>1</u> | <u>Subject</u> | <u>Μοναδικό αναγνωριστικό ασθενή</u> | <u>Αριθμητική - Ακέραιος αριθμός</u> |
| <u>2</u> | <u>Age</u> | <u>Ηλικία ασθενή</u> | <u>Αριθμητική - Ακέραιος αριθμός</u> |
| <u>3</u> | <u>Sex</u> | <u>Φύλο ασθενή</u> | <u>Διαδική (0: Άντρας, 1: Γυναίκα)</u> |

| | | | |
|----------|--------------------|---|-------------------------------|
| <u>4</u> | <u>test time</u> | <u>Χρόνος (ημέρες) από την έναρξη του πειράματος.</u> | <u>Αριθμητική – Δεκαδικός</u> |
| <u>5</u> | <u>motor UPDRS</u> | <u>Η βαθμολογία UPDRS - κινητική</u> | <u>Αριθμητική – Δεκαδικός</u> |
| <u>6</u> | <u>total UPDRS</u> | <u>Η βαθμολογία UPDRS – συνολική</u> | <u>Αριθμητική – Δεκαδικός</u> |

| | | | |
|-----------|-------------------------|--|-----------------------------------|
| <u>7</u> | <u>Jitter(Percent)</u> | <u>Μέτρο μεταβολής βασικής συχνότητας (σε ποσοστό)</u> | <u>Αριθμητική – Δεκαδικός</u> |
| <u>8</u> | <u>Jitter(Absolute)</u> | <u>Μέτρο μεταβολής βασικής συχνότητας</u> | <u>Αριθμητική – Δεκαδικός</u> |
| <u>9</u> | <u>Jitter:RAP</u> | <u>Μέτρο μεταβολής βασικής συχνότητας</u> | <u>Αριθμητική – Δεκαδικός</u> |
| <u>10</u> | <u>Jitter:PPQ5</u> | <u>Μέτρο μεταβολής βασικής συχνότητας</u> | <u>Αριθμητική – Δεκαδικός</u> |

| | | | |
|-----------|---------------------|---|-----------------------------------|
| <u>11</u> | <u>Jitter:DDP</u> | <u>Μέτρο μεταβολής βασικής συχνότητας</u> | <u>Αριθμητική – Δεκαδικός</u> |
| <u>12</u> | <u>Shimmer</u> | <u>Μέτρο μεταβολής εύρους</u> | <u>Αριθμητική – Δεκαδικός</u> |
| <u>13</u> | <u>Shimmer(dB)</u> | <u>Μέτρο μεταβολής εύρους</u> | <u>Αριθμητική – Δεκαδικός</u> |
| <u>14</u> | <u>Shimmer:APQ3</u> | <u>Μέτρο μεταβολής εύρους</u> | <u>Αριθμητική – Δεκαδικός</u> |

| | | | |
|-----------|----------------------|---|---|
| <u>15</u> | <u>Shimmer:APQ5</u> | <u>Μέτρο μεταβολής εύρους</u> | <u>Αριθμητική –</u> <u>Δεκαδικός</u> |
| <u>16</u> | <u>Shimmer:APQ11</u> | <u>Μέτρο μεταβολής εύρους</u> | <u>Αριθμητική –</u> <u>Δεκαδικός</u> |
| <u>17</u> | <u>Shimmer:DDA</u> | <u>Μέτρο μεταβολής εύρους</u> | <u>Αριθμητική –</u> <u>Δεκαδικός</u> |
| <u>18</u> | <u>NHR</u> | <u>Λόγος θορύβου προς αρμονία</u> | <u>Αριθμητική –</u> <u>Δεκαδικός</u> |
| <u>19</u> | <u>HNR</u> | <u>Λόγος αρμονίας προς θόρυβο</u> | <u>Αριθμητική –</u> <u>Δεκαδικός</u> |
| <u>20</u> | <u>RPDE</u> | <u>Μη γραμμικό δυναμικό μέτρο</u> <u>πολυπλοκότητας</u> | <u>Αριθμητική –</u> <u>Δεκαδικός</u> |
| <u>21</u> | <u>DFA</u> | <u>Ένδειξη κλασματικής κλίμακας</u> <u>σήματος</u> | <u>Αριθμητική –</u> <u>Δεκαδικός</u> |
| <u>22</u> | <u>PPE</u> | <u>Μη γραμμικό μέτρο βασικής</u> <u>διακύμανσης συχνότητας</u> | <u>Αριθμητική –</u> <u>Δεκαδικός</u> |

Το UPDRS έχει σχεδιαστεί για να παρακολουθεί την ασθένεια Parkinson, το οποίο σχετίζεται σαφώς με το επίπεδο Parkinson του ασθενούς. Η βαθμολογία UPDRS αποτελείται από 4

διαφορετικά μέρη που αναφέρονται στη συνείδηση και τη συμπεριφορά, τη διάθεση και τις δραστηριότητες της καθημερινότητας του ασθενούς, τις διαπλοκές στο μηχανήμα μέτρησης και γενικά λοιπές επιπλοκές που αφορούν την θεραπεία.

Στο σύνολο δεδομένων δίνονται δύο παράμετροι-στόχοι, το “motor_UPDRS” και “total_UPDRS”, τα οποία είναι τα αντίστοιχα UPDRS scores που προέκυψαν από την ιατρική αξιολόγηση (ground truth).

Το total_UPDRS εκτείνεται από 0 έως 176, όπου το 0 υποδηλώνει απολύτως υγιή άτομα και το 176 ολική αναπηρία. Αντίθετα, το motor_UPDRS είναι ένα υποσύνολο των συνολικών UPDRS και κυμαίνεται από 0 έως 108, όπου το 0 σημαίνει ότι ο ασθενής δεν έχει συμπτώματα ενώ το 108 ότι έχει σοβαρή κινητική βλάβη. Παρόλα αυτά στο dataset μας, έχουμε βρει τα ακόλουθα στατιστικά αναφορικά με τις δύο παραπάνω στήλες:

Πίνακας 2: Στατιστικά σύμφωνα με dataset

| Data Set | MOTOR UPDRS | TOTAL UPDRS |
|----------|-------------|-------------|
| Min | 5.0377 | 7 |
| Max | 39.511 | 54.992 |
| Range | 34.511 | 47.992 |
| Mean | 20.871 | 27.576 |
| Std. | 8.12858964 | 10.6993726 |

Classification/regression

Αλγόριθμοι Ταξινόμησης (classification)

Η κατηγοριοποίηση (classification) είναι η πιο γνωστή και πιο δημοφιλής τεχνική εξόρυξης γνώσης (data mining). Είναι η διαδικασία η οποία απεικονίζει ένα σύνολο δεδομένων σε προκαθορισμένες ομάδες. Τις ομάδες αυτές συχνά τις καλούμε κατηγορίες ή κλάσεις (classes). Δηλαδή, έχοντας δεδομένο ένα σύνολο κλάσεων, επιδιώκουμε να προσδιορίσουμε την κλάση ή τις κλάσεις στις οποίες ανήκει ένα αντικείμενο.

Συχνά μία κλάση αφορά μία πολύ γενικότερη θεματική περιοχή, σ’ αυτήν την περίπτωση ονομάζονται θέματα (topics) και έτσι υφίσταται αντίστοιχη εργασία ταξινόμησης. Μία προσέγγιση στην ταξινόμηση βασίζεται στην μηχανική μάθηση (machine learning). Αφορά δηλαδή, το σύνολο των κανόνων ή γενικότερα, το κριτήριο απόφασης του ταξινομητή, όπου

αυτό μαθαίνεται αυτόματα από τον μηχανισμό του ταξινομητή μέσω δεδομένων εκπαίδευσης (training documents). Παρόλα αυτά, η μη αυτόματη ταξινόμηση εξακολουθεί να υφίσταται, αφού έγγραφα εκπαίδευσης καθορίζονται από κάποιον άνθρωπο που έχει αναλάβει τον χαρακτηρισμό τους (labels). Το labeling είναι ουσιαστικά η διαδικασία της επισημείωσης κάθε εγγράφου με το όνομα της κλάσης του.

Αναφορικά με το σύνολο δεδομένων της εργασίας εργαστήκαμε ως εξής. Αρχικά βρήκαμε το μέσο όρο για κάθε ασθενή (42 στο σύνολο) σύμφωνα με όλες τις μετρήσεις του:

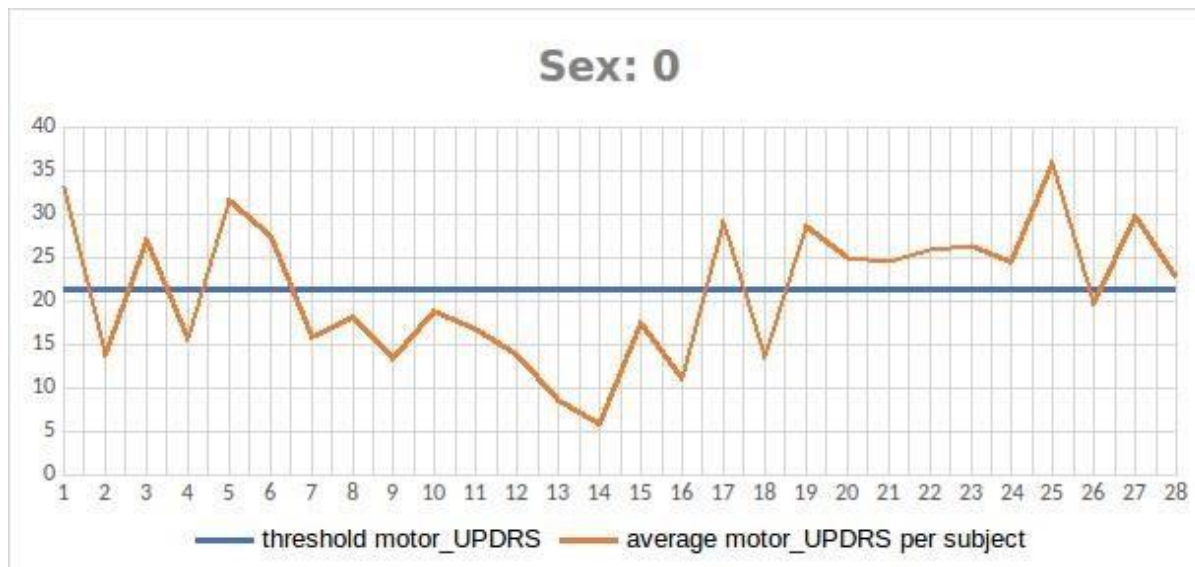
| Subject | average motor_UPDRS per subje | average motor_UPDRS per sub sex | |
|---------|-------------------------------|---------------------------------|---|
| 1 | 33.18075168 | 33.18075168 | 0 |
| 2 | 13.81253793 | 13.81253793 | 0 |
| 3 | 27.12478472 | 27.12478472 | 0 |
| 4 | 15.79082482 | 15.79082482 | 0 |
| 5 | 31.63260256 | 31.63260256 | 0 |
| 6 | 27.53169231 | 27.53169231 | 0 |
| 7 | 16.04706211 | 16.04706211 | 0 |
| 8 | 19.88702 | 19.88702 | 1 |
| 9 | 18.31236184 | 18.31236184 | 0 |
| 10 | 13.42441892 | 13.42441892 | 0 |
| 11 | 18.98756522 | 18.98756522 | 0 |
| 12 | 16.88828037 | 16.88828037 | 0 |
| 13 | 19.51676786 | 19.51676786 | 1 |
| 14 | 13.01445 | 13.01445 | 1 |
| 15 | 13.96458671 | 13.96458671 | 0 |
| 16 | 8.705965942 | 8.705965942 | 0 |
| 17 | 26.43229861 | 26.43229861 | 1 |
| 18 | 5.82345873 | 5.82345873 | 0 |
| 19 | 17.61122481 | 17.61122481 | 0 |
| 20 | 11.18345522 | 11.18345522 | 0 |
| 21 | 29.09265854 | 29.09265854 | 0 |
| 22 | 9.7997125 | 9.7997125 | 1 |
| 23 | 13.47463043 | 13.47463043 | 1 |
| 24 | 13.75973077 | 13.75973077 | 0 |
| 25 | 28.7325625 | 28.7325625 | 0 |
| 26 | 25.04024615 | 25.04024615 | 0 |
| 27 | 10.79183566 | 10.79183566 | 1 |
| 28 | 29.1673806 | 29.1673806 | 1 |
| 29 | 24.63122619 | 24.63122619 | 0 |
| 30 | 25.91511905 | 25.91511905 | 0 |
| 31 | 26.40428462 | 26.40428462 | 0 |
| 32 | 9.944266337 | 9.944266337 | 1 |
| 33 | 26.36831111 | 26.36831111 | 1 |
| 34 | 24.6808882 | 24.6808882 | 0 |
| 35 | 35.98981212 | 35.98981212 | 0 |
| 36 | 23.39494574 | 23.39494574 | 1 |
| 37 | 31.86064286 | 31.86064286 | 1 |
| 38 | 19.7725906 | 19.7725906 | 0 |
| 39 | 29.87755944 | 29.87755944 | 0 |
| 40 | 16.50746479 | 16.50746479 | 1 |
| 41 | 34.40492121 | 34.40492121 | 1 |
| 42 | 22.84406667 | 22.84406667 | 0 |

Εικόνα 4.1: Μέσος όρος motor_UPDRS για κάθε ασθενή

Παρατηρήθηκε σύμφωνα με τα ακόλουθα διαγράμματα πως ο μέσος όρος της τιμής του motor_UPDRS κυμαίνεται στην τιμή '20.9839753915603' για το σύνολο των 42 ασθενών.

Πιο συγκεκριμένα, αν θεωρήσουμε δύο κλάσεις (άνδρες γυναίκες, δηλ. sex 0,1) οι τιμές του motor_UPDRS για τον άνδρα χωρίζονται στην τιμή '20.3260462643673'. Η διαδικασία να για να ορίσουμε το συγκεκριμένο threshold για τους άντρες αλλά και η αντίστοιχη για τις γυναίκες είναι:

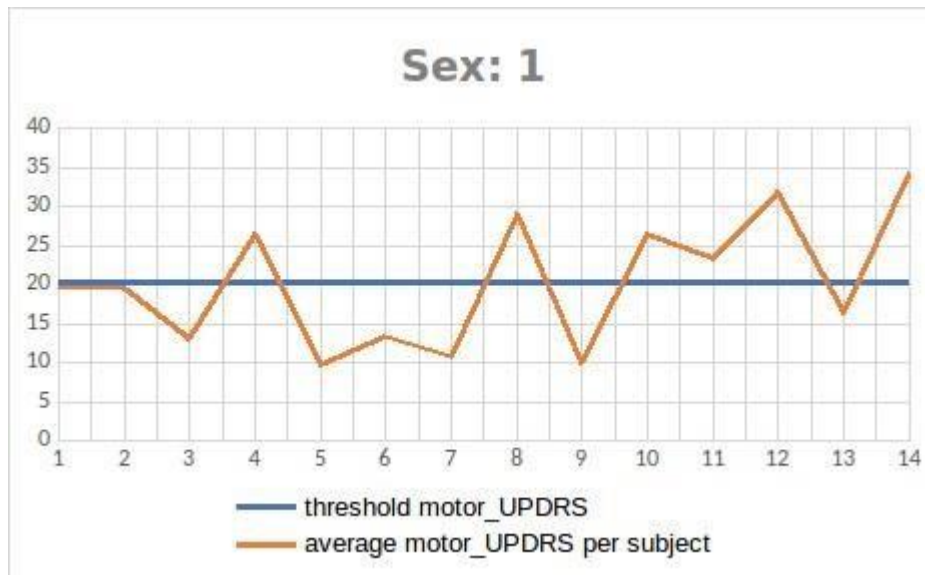
- χωρίσαμε το συνολικό δείγμα σε άντρες και γυναίκες
- επιλέξαμε τις μετρήσεις μόνο των αντρών και για κάθε έναν από τους άντρες ασθενείς βρήκαμε τον μέσο όρο της τιμής motor_UPDRS
- τέλος, υπολογίσαμε τον μέσο όρο του motor_UPDRS όλων των ασθενών/αντρών και ορίσαμε αυτή την τιμή ως threshold για να κάνουμε την "σύγκριση" στην συνέχεια και να βγάλουμε το πόρισμα ("result"), αν δηλαδή ο ασθενής πάσχει ή δεν πάσχει από την νόσο.



Εικόνα 4.2: Διάγραμμα απεικόνισης του μέσου όρου του motor_UPDRS καθενός από τους 28 άντρες συγκριτικά με το threshold

Στο παραπάνω διάγραμμα παρατηρούμε ότι στον άξονα x είναι ο συνολικός αριθμός των subjects (28 άντρες) ενώ στον άξονα y παρατηρούμε τον μέσο όρο της τιμής motor_UPDRS για τον κάθε άντρα. Η μπλε γραμμή αποτελεί το threshold μας.

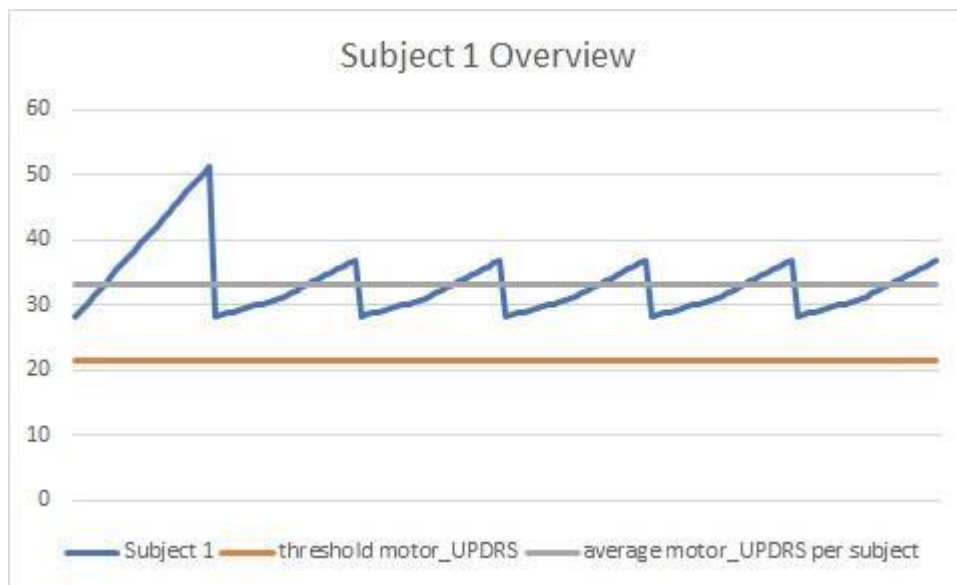
Παράλληλα, οι τιμές του motor_UPDRS για την γυναίκα χωρίζονται στην τιμή '21.3129399551569'. Ομοίως το παρακάτω διάγραμμα και για τις 14 γυναίκες:



Εικόνα 4.3:Διάγραμμα απεικόνισης του μέσου όρου του motor_UPDRS καθενός από τις 14 γυναίκες συγκριτικά με το threshold

Έτσι δημιουργήσαμε μία νέα στήλη στο dataset μας με όνομα 'result'. Αυτή η στήλη είναι τύπου boolean. Με '1' θεωρείται ότι από το σύνολο το δεδομένων μας και σύμφωνα με τη στήλη motor_UPDRS ο ασθενής έχει περισσότερες πιθανότητες να πάσχει από Parkinson, ενώ με '0' το αντίθετο.

Για παράδειγμα και με βάση την παραπάνω παραδοχή που κάναμε, ο πρώτος ασθενής που είναι άντρας πιθανότατα πάσχει από την νόσο, αφού όλες του οι μετρήσεις είναι πάνω από το threshold, όπως φαίνεται και στο παρακάτω διάγραμμα:



Εικόνα 4.4: Διάγραμμα απεικόνισης του αποτελέσματος του *motor_UPDRS* του 1ου ασθενή συγκριτικά με το *threshold*

Μια επιπλέον συνθήκη για το 'labeling', δηλαδή ο προσδιορισμός του *result*, είναι ο περιορισμός του φύλου. Δηλαδή, αν οι μετρήσεις των γυναικών με τιμές άνω της τιμής '21.312939' του *motor_UPDRS* τότε η αντίστοιχη τιμή του *result* θα πάρει την τιμή '1'. Σε αντίθετη περίπτωση θα πάρει την τιμή '0'. Αντίστοιχα και για τους άνδρες με τιμή κατωφλίου '20.32.6046'.

4.2 Support Vector Machine

Στη συνέχεια, ετοιμάζουμε το *train set* (80% των δεδομένων και *label*) και το *test set* (20% *label*) και περνάμε στην διαδικασία το *cross validation*. Αυτή η διαδικασία, αναφέρεται στην ακρίβεια πρόβλεψης σ' ένα σύνολο "μη επισημασμένων"(unlabeled) δεδομένων ώστε να εκτιμήσουμε πόσο καλά είναι τα αποτελέσματά μας μετά την υλοποίηση και απόδοση ενός ταξινομητή. Στόχος, δηλαδή, του *cross validation* είναι να οριστεί ένα σύνολο δεδομένων για να «δοκιμαστεί» το μοντέλο μας στη φάση του *training* προκειμένου να περιοριστούν προβλήματα όπως το *overfitting*.

Εκπαιδεύουμε τον classifier με το μοντέλο '*Support Vector Machine*' που όπως υποδηλώνει και το όνομα του είναι μία διανυσματική μέθοδος μάθησης και υποστηρίζει μηχανική μάθηση σε δεδομένα. Στόχο έχει τον εντοπισμό ενός ορίου απόφασης μεταξύ των κλάσεων, το οποίο να βρίσκεται στη μέγιστη δυνατή απόσταση από οποιοδήποτε σημείο των δεδομένων εκπαίδευσης.

Στο προγραμματιστικό μέρος, έχοντας προεπεξεργαστεί τα δεδομένα μας σύμφωνα με την ενότητα 2 και έχοντας χωρίσει τα δεδομένα μας σε *train* και *test* περνάμε τα αντίστοιχα δεδομένα στον classifier:

Κώδικας σε python

```
#Create a svc Classifier clf =  
  
svm.SVC(kernel='linear') # Linear Kernel  
  
#Train the model using the training sets  
  
clf.fit(X_train, y_train)
```

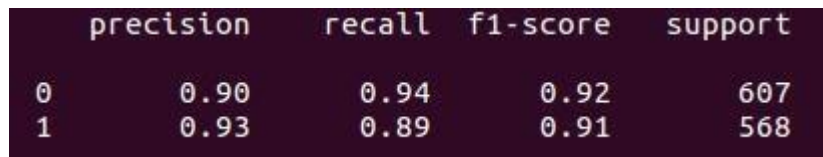
```
#Predict the response for test dataset y_pred
```

```
= clf.predict(X_test)
```

Για την εύρεση της μετρικής 'Accuracy' παρατίθεται ο ακόλουθος κώδικας:

```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

Αρχικά παρατηρείται ο διαχωρισμός των δεδομένων σύμφωνα με το 80% ως train set και 20% ως test set και τέλος η τιμή του accuracy όπου βρέθηκε περίπου '92.6%'. Το accuracy αφορά τον αριθμό του test συνόλου που χαρακτηρίστηκε ως 1 προς τον αριθμό του train συνόλου που χαρακτηρίστηκε ως 0. Όπως φαίνεται και στο παρακάτω screenshot, από το 20% των μετρήσεων που είναι ίσο με 1175 μετρήσεις, τα 607 χαρακτηρίστηκαν ως "0" και τα 568 ως "1" σύμφωνα με το support. Παράλληλα, στον ίδιο πίνακα παρατηρούνται και τα αποτελέσματα και κάποιων σύνηθων μετρικών:



| | precision | recall | f1-score | support |
|---|-----------|--------|----------|---------|
| 0 | 0.90 | 0.94 | 0.92 | 607 |
| 1 | 0.93 | 0.89 | 0.91 | 568 |

Εικόνα 4.6: Screenshot αποτελεσμάτων μετρικών

Ο παρακάτω πίνακας θα μας βοηθήσει να καταλάβουμε το ποσοστό των μετρικών αυτών:

| | |
|----------------|----------------|
| True Positive | False Negative |
| False Positive | True Negative |

Στην περίπτωση μας ο πίνακας αυτός ισούται με τον ακόλουθο, με τις τιμές να αντιστοιχούν στα αντίστοιχα πεδία του πάνω πίνακα:

Πίνακας 4: Πίνακας TP, FP, TN, FN δεδομένων εργασίας

| | |
|-----|----|
| 570 | 37 |
|-----|----|

| | |
|----|-----|
| 62 | 506 |
|----|-----|

- Αληθώς θετικό (**True Positive** -TP): εκτιμάται ότι ανήκει σε μία κατηγορία και πράγματι ανήκει σε αυτήν, στην περίπτωση μας είναι 570 μετρήσεις ασθενών.
- Ψευδώς θετικό (**False Positive** -FP): εκτιμάται ότι ανήκει σε μία κατηγορία ενώ στην πραγματικότητα δεν ανήκει σε αυτήν, στην περίπτωση μας είναι 62 μετρήσεις ασθενών.
- Αληθώς αρνητικό (**True Negative** -TN): εκτιμάται ότι δεν ανήκει σε μία κατηγορία και πράγματι δεν ανήκει σε αυτήν, στην περίπτωση μας είναι 506 μετρήσεις ασθενών.
- Ψευδώς αρνητικό (**False Negative** -FN): εκτιμάται ότι δεν ανήκει σε μία κατηγορία ενώ στην πραγματικότητα ανήκει σε αυτήν, στην περίπτωση μας είναι 37 μετρήσεις ασθενών.

Σε μια ταξινόμηση, το **precision** για μια κλάση είναι ο αριθμός των True-positives (δηλ. ο αριθμός των μετρήσεων που έχουν σωστά επισημανθεί ότι ανήκουν στην positive class) διαιρούμενο με το συνολικό αριθμό των μετρήσεων που επισημάνθηκαν πως ανήκουν στη positive class.

Ενώ, το **recall** ορίζεται ως ο αριθμός των True Positives που διαιρούνται με τον συνολικό αριθμό των στοιχείων που πράγματι ανήκουν στη positive class (δηλαδή το άθροισμα των True Positives και των False Negative, τα οποία δεν έχουν επισημανθεί ότι ανήκουν στη positive class αλλά έπρεπε).

Αναφορές - Βιβλιογραφία για

- [1] https://www.cs.upc.edu/~ayamaui/documents/Report_Busquet_Yamaui.pdf
- [2] https://github.com/NicolasAG/MachineLearningproject4/blob/master/Final_report.pdf
- [3] <https://machinelearningmastery.com/rescaling-data-for-machine-learning-inpythonwith-scikit-learn/>
- [4] <https://github.com/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.09-Principal-Component-Analysis.ipynb>
- [5] <https://github.com/NicolasAG/MachineLearning-project4>
- [6] <https://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/>
- [7] <https://www.techopedia.com/definition/30306/association-rule-mining>
- [8] Raghavan Hinrich Schütze, "An Introduction to Information Retrieval", Cambridge University Press Cambridge, England 2009
- [9] <https://www.datacamp.com/community/tutorials/svm-classification-scikitlearnpython>

- [10] <https://medium.com/@tomernahshon/spectral-clustering-fromscratch38c68968eae0>
- [11] <https://www.kaggle.com/dhanyajothimani/basic-visualization-and-clusteringinpython>
- [12] <https://www.kaggle.com/datatheque/association-rules-mining-market-basketanalysis>