Πανεπιστήμιο Πειραιώς
ΕΚΕΦΕ "ΔΗΜΟΚΡΙΤΟΣ"
Δ.Π.Μ.Σ. στην Τεχνητή Νοημοσύνη
Τμήμα Ψηφιακών Συστημάτων Πανεπιστημίου Πειραιώς
Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών ΕΚΕΦΕ
«Δημόκριτος»

Ταξινόμηση δειγμάτων νερού επί τη βάσει της ποσιμότητας

Εργασία στο μάθημα "Μηχανική Μάθηση"

Νικόλαος Μακρής - mtn2208

Διδάσκων: Γιαννακόπουλος Θεόδωρος, Μεταδιδακτορικός Ερευνητής ΕΚΕΦΕ Δημόκριτος

Αθήνα, 2023

Περιεχόμενα

Αρχικά Βήματα

1.1 Περιγραφή προβλήματος

Στη συγκεκριμένη εργασία χρησιμοποιήθηκε ένα σύνολο δεδομένων (dataset) διαθέσιμο στην πλατφόρμα kaggle [?], το οποίο περιλαμβάνει διαφόρους δείκτες ποιότητας νερού, οι οποίοι χρησιμοποιούνται μεταξύ άλλων για την ταξινόμηση (classification) δειγμάτων νερού ως πόσιμου ή μη. Συνολικά το dataset περιλαμβάνει εννιά χαρακτηριστικές ιδιότητες (attributes).

Όπως γίνεται αντιληπτό το πρόβλημα που θα αναλυθεί είναι πρόβλημα δυαδικής ταξινόμησης (binary classification), καθώς με βάση τις τιμές των εννιά διαφορετικών ιδιοτήτων ταξινομούνται δείγματα νερού σε μια από δύο κατηγορίες (0-μη πόσιμο, 1-πόσιμο). Στον πίνακα 1.1 που ακολουθεί παρουσιάζονται αυτές οι ιδιότητες, δίνεται μια επιγραμματική περιγραφή καθώς και η μονάδα μέτρησης κάθε μιας.

Αξίζει να σημειωθεί ότι δεν υπάρχει ισοκατανομή μη πόσιμων και πόσιμων δειγμάτων νερού, καθώς $\sim 61\%$ των δειγμάτων ανήκουν στην κλάση των μη πόσιμων δειγμάτων, ενώ πόσιμα θεωρούνται $\sim 39\%$ των δειγμάτων. Αυτή η παρατήρηση θα φανεί χρήσιμη κατά την ερμηνεία των αποτελεσμάτων στην παράγραφο 2.3.

#	Ιδιότητα (attribute)	Μετάφραση	Περιγραφή
0	pH	pH	Μέτρο οξύτητας H_2O [0-14]
1	Hardness	Σκληρότητα	Περιεκτικότητα H_2O σε Ca & Mg [mg/L]
2	Solids	Στερεά	Ολικά διαλυμένα στερεά (TDS) [ppm]
3	Chloramines	Χλωραμίνες	Ολικά διαλυμένη NH_2Cl [ppm]
4	Sulfate	Θειικά άλατα	Περιεκτικότητα H_2O σε SO_4^{2-} [mg/L]
5	Conductivity	Αγωγιμότητα	Ηλεκτρική αγωγιμότητα H_2O [μS/cm]
6	Organic_Carbon	Οργανικός άνθρακας	Συγκέντρωση οργανικού άνθρακα [ppm]
7	Trihalomethanes	Τριαλομεθάνια	Ομάδα χημικών ουσιών [μg/L]
8	Turbidity	Θολότητα	Μέτρο φωτεινότητας νερού [NTU]

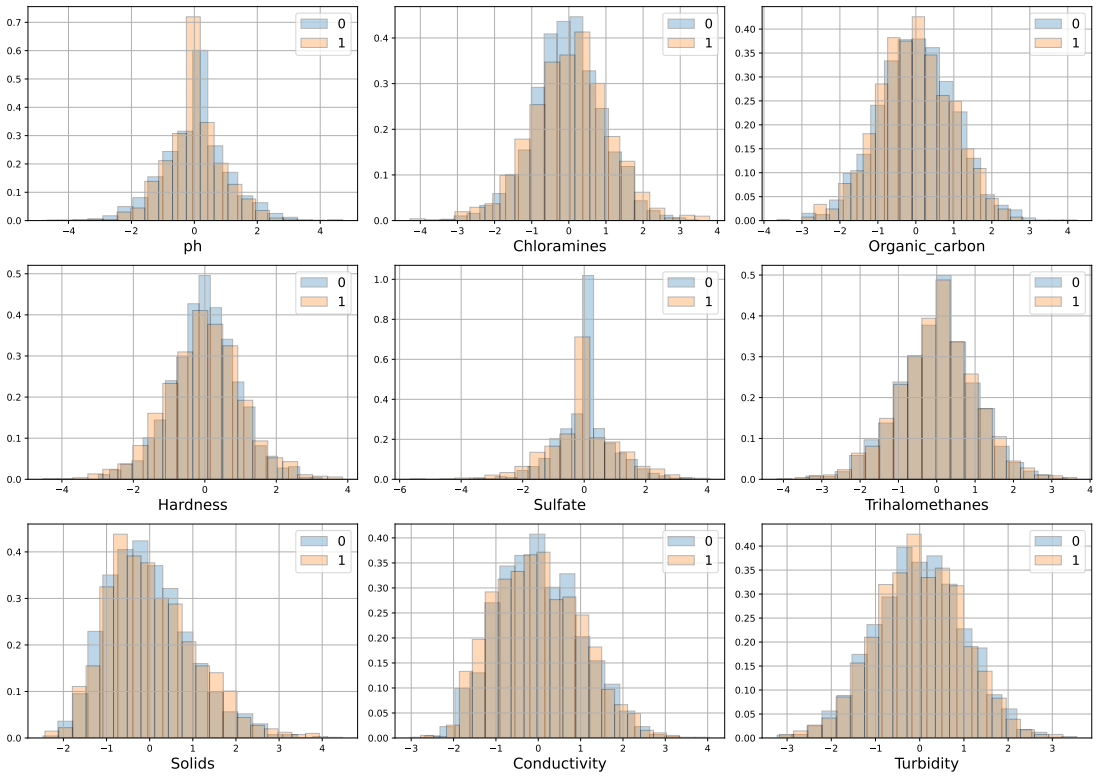
Πίνακας 1.1: Ιδιότητες συνόλου δεδομένων

1.2 Προεπεξεργασία δεδομένων

Το πρώτο βήμα όπως και σε όλα τα προβλήματα μηχανικής μάθησης (machine learning) είναι η διενέργεια μιας προκαταρκτικής ανάλυσης των διαθέσιμων δεδομένων ώστε να γίνει καλύτερα αντιληπτό το τι αντιπροσωπεύει η κάθε μεταβλητή και τον τρόπο με τον οποίο μπορεί να αξιοποιηθεί στην περαιτέρω ανάλυση του προβλήματος. Πιο συγκεκριμένα εξετάζονται τα παρακάτω:

1. Η πληρότητα των δεδομένων (1.2.1)
2. Η ύπαρξη ή μη ακραίων τιμών (outliers) και η επακόλουθη διαχείρισή τους (1.2.2)
3. Ο συντελεστής συσχέτισης μεταξύ των διαφόρων μεταβλητών (1.2.3)

Το σχήμα 1.1 είναι πολύ σημαντικό για την περαιτέρω κατανόηση του προβλήματος, καθώς δίνεται με τη μορφή ιστογράμματος μια εικόνα της κατανομής των διαφόρων μεταβλητών με βάση την κλάση. Φαίνεται λοιπόν ότι οι κατανομές των μετρήσεων κάθε ιδιότητας προσεγγίζουν την κανονική κατανομή, με εξαίρεση την παράμετρο Solids η οποία παρουσιάζει εμφανή θετική λοξότητα (positive skewness). Επιπλέον είναι εμφανής η αλληλοεπικάλυψη των δεδομένων που υφίσταται σε κάθε μεταβλητή, που υποδεικνύει ότι η συσχέτιση της εκάστοτε μεταβλητής με την κλάση είναι πολύ μικρή όπως θα δούμε και στην παράγραφο 1.2.3.



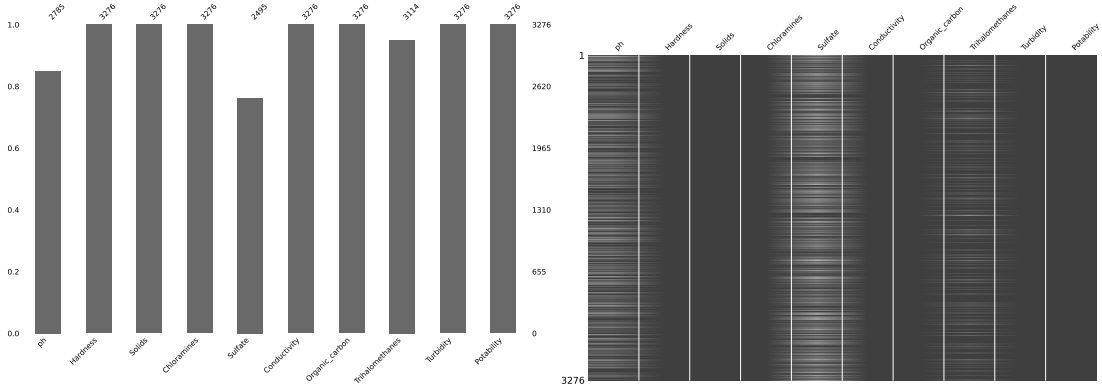
Σχήμα 1.1: Ιστόγραμμα ιδιοτήτων (PDF)

1.2.1 Διαχείριση μη διαθέσιμων τιμών

Αρχικά όπως βλέπουμε στο σχήμα 1.2 το σύνολο των δεδομένων είναι σχετικά πλήρες, ενώ υπάρχουν κάποια κενά στις μετρήσεις των μεταβλητών pH, Sulfate και Trihalomethanes. Η γενική αντιμετώπιση μη διαθέσιμων τιμών (missing values) στα δεδομένα αντιμετωπίζεται είτε με την καθολική παράλειψη της συγκεκριμένης γραμμής δεδομένων είτε με τη χρήση κάποιου καταλογιστή (imputer) που χρησιμοποιείται για την κάλυψη των κενών με κάποια τιμή που βασίζεται σε μια συγκεκριμένη λογική.

Στην παρούσα άσκηση και εφόσον τα δεδομένα είναι σχετικά λίγα, επιλέχθηκε η δεύτερη λύση καθώς σε περίπτωση επιλογής της πρώτης τα διαθέσιμα δεδομένα δεν θα επαρκούσαν για την εκπαίδευση του ταξινομητή. Πιο αναλυτικά σε κάθε μη διαθέσιμη τιμή αποδόθηκε

η τιμή της διαμέσου (median) της συγκεκριμένης στήλης.



Σχήμα 1.2: Μη διαθέσιμες τιμές

1.2.2 Διαχείριση ακραίων τιμών

Αναφορικά με τις ακραίες τιμές, αρχικά έγινε χρήση του λεγόμενου θηκογράμματος, όπως φαίνεται στο σχήμα 1.3, ώστε να υπάρχει μια σαφής εικόνα για την ύπαρξη ή μη ακραίων τιμών. Εύκολα λοιπόν παρατηρείται ότι ακραίες τιμές εμφανίζονται και στις εννιά μεταβλητές, ενώ η διαχείριση τους περιγράφεται παρακάτω.

Ο αλγοριθμικός εντοπισμός των ακραίων τιμών γίνεται με τη χρήση της μεθόδου IQR [?]. Στη συγκεκριμένη μέθοδο γίνεται εντοπισμός των άκρων όπου βρίσκεται το 25%, 50% και 75% των τιμών του δείγματος. Οι τιμές του δείγματος για τις οποίες ισχύει η ανωτέρω συνθήκη ονομάζονται Q_1, median (διάμεσος) και Q_3, ενώ οι τιμές του δείγματος που ικανοποιούν τη σχέση 1.2.1c, χαρακτηρίζονται ως ακραίες τιμές. Με Θ συμβολίζεται η εκάστοτε τυχαία μέτρηση, ενώ με O το σύνολο των ακραίων τιμών.

$$IQR = Q_3 - Q_1 \tag{1.2.1a}$$

$$L_1 = Q_1 - IQR, \quad L_2 = Q_3 + IQR \tag{1.2.1b}$$

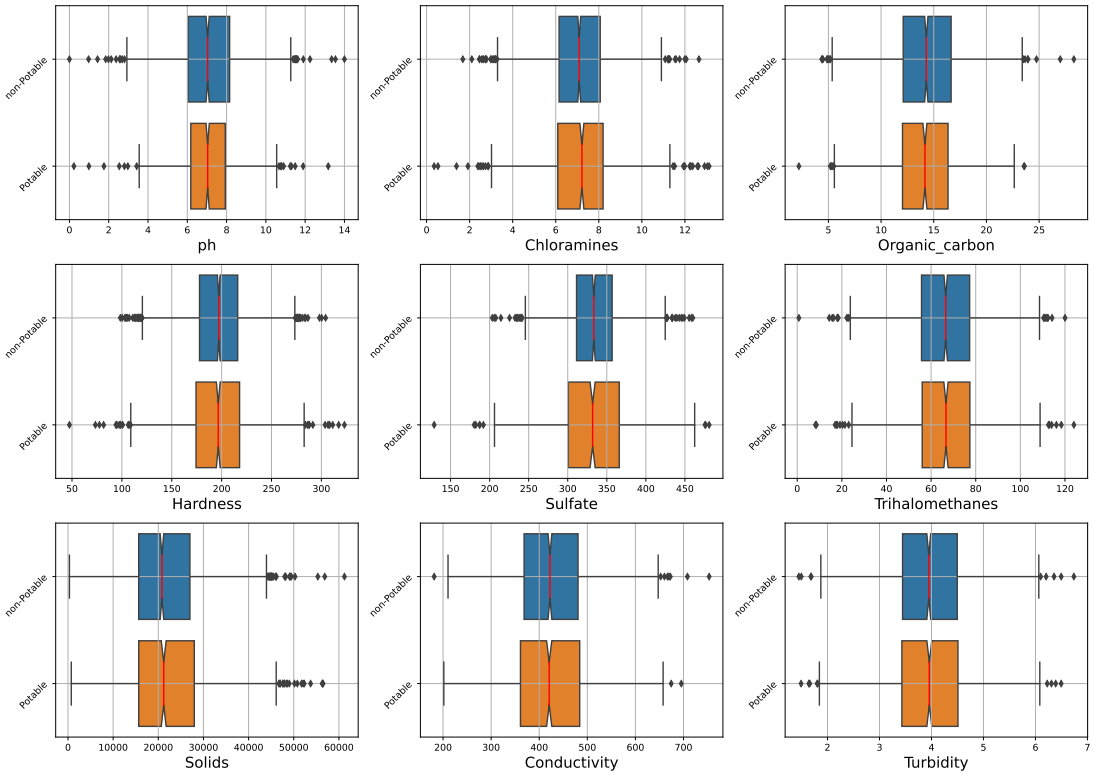
$$if \quad \Theta \leq L_1 \quad OR \quad \Theta \geq L_1 \Rightarrow \Theta \, \epsilon \, O \tag{1.2.1c}$$

Λαμβάνοντας λοιπόν υπόψη το σχήμα 1.3 και το σχήμα 1.1, παρατηρείται ότι η διάμεσος (median) των δειγμάτων που αναπαρίσταται με την κόκκινη γραμμή στο σχήμα 1.3, βρίσκεται σχεδόν στην ίδια κατακόρυφο και για τις δύο κλάσεις του προβλήματος ταξινόμησης. Επιπλέον, αυτή η διάμεσος βρίσκεται σχεδόν σε όλες τις περιπτώσεις στο μέσον του θηκογράμματος, που υποδηλώνει ότι κάθε μεταβλητή ακολουθεί κατανομή που μπορεί να προσεγγιστεί από την κανονική, όπως άλλωστε φαίνεται πιο παραστατικά στο σχήμα 1.1.

Ένα άλλο στοιχείο χρήσιμο στην περαιτέρω ανάλυση, είναι ότι τα Q1, Q3 όρια του θηκογράμματος βρίσκονται και αυτά στην ίδια κατακόρυφο για όλες τις μεταβλητές. Αυτό σε συνδυασμό με το γεγονός ότι οι κατανομές των μετρήσεων και για τις δύο κλάσεις ακολουθούν σχεδόν κανονική κατανομή με ίδια διάμεσο, οδηγεί για άλλη μια φορά στο

συμπέρασμα ότι δεν υπάρχει σαφής συσχέτιση της κλάσης και της εκάστοτε μεταβλητής, που συνεπάγεται σχετικά χαμηλή ακρίβεια πρόγνωσης του τελικού μοντέλου.

Όσον αφορά τον τρόπο αντιμετώπισης των ακραίων τιμών, επιλέχθηκε να αγνοηθούν οι ακραίες μετρήσεις και στη συνέχεια να αντικατασταθούν από τη διάμεση τιμή. Με άλλα λόγια μπορούν να αντιμετωπιστούν ως μη διαθέσιμες τιμές και ύστερα να γίνει χρήση της μεθοδολογίας που παρουσιάστηκε προηγουμένως για την αντικατάστασή τους. Τα αποτελέσματα της εφαρμογής της συγκεκριμένης μεθόδου συγκρίνονται με τα αποτελέσματα της μη εφαρμογής οποιασδήποτε μεθόδου αντιμετώπισης των ακραίων τιμών στον πίνακα 2.1.



Σχήμα 1.3: Θηκόγραμμα με την παρουσία outliers

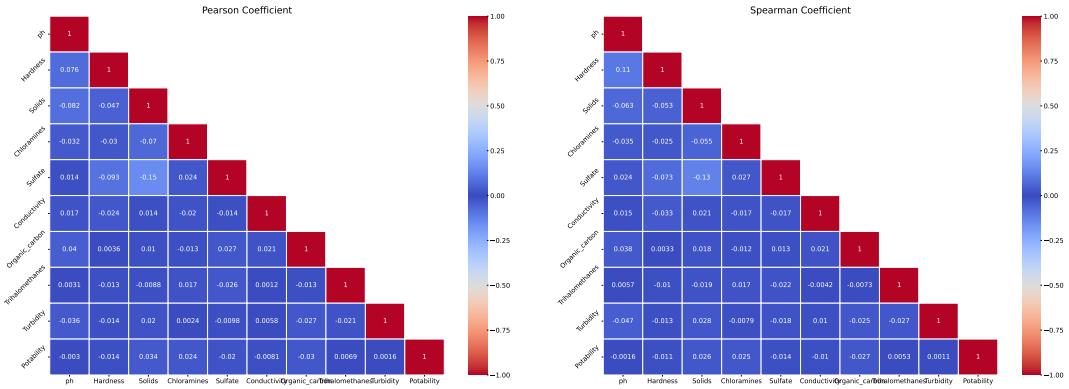
1.2.3 Συντελεστές συσχέτισης

Ο συντελεστής συσχέτισης μεταξύ των διαφόρων μεταβλητών καθώς και η συσχέτιση κάθε μεταβλητής με την μεταβλητή ταξινόμησης είναι πολύ σημαντική για την περαιτέρω επεξεργασία των δεδομένων. Στην περίπτωση που υπάρχει ισχυρή συσχέτιση μεταξύ κάποιων μεταβλητών τότε αυτές μπορούν να συγχωνευθούν σε μια καινούρια η οποία θα εκφράζει την επίδραση και των δύο στην ταξινόμηση. Επιπλέον αν κάποια μεταβλητή έχει πολύ μικρή συσχέτιση με την κατηγορική μεταβλητή ταξινόμησης, σε σύγκριση πάντα με τη συσχέτιση που παρουσιάζουν οι υπόλοιπες μεταβλητές, τότε αυτή θα μπορούσε να παραληφθεί καθώς πιθανότατα δεν έχει θετική επίδραση στην πρόβλεψη.

Στο σχήμα 1.4 παρουσιάζονται οι συντελεστές συσχέτισης κατά Pearson και Spearman. Σε γενικές γραμμές ο πρώτος συντελεστής υποδεικνύει τη γραμμική συσχέτιση μεταξύ δύο μεταβλητών, ενώ στη δεύτερη που είναι πιο γενική σχέση συσχετίζεται η μονοτονία

δύο μεταβλητών. Παρατηρείται λοιπόν η ύπαρξη εξαιρετικά μικρής συσχέτισης μεταξύ των διαφόρων μεταβλητών, ενώ δεν φαίνεται να υπάρχει κάποια μεταβλητή που να έχει μεγάλη συσχέτιση με την μεταβλητή ταξινόμησης.

Αυτός ο συνδυασμός γεγονότων μας δείχνει ότι είναι δύσκολο να δημιουργήσουμε κάποια καινούρια ιδιότητα (feature) συνδυάζοντας κάποια από τις υπάρχουσες ή να αφαιρέσουμε εντελώς κάποια από την ανάλυσή μας. Τα παραπάνω όπως εξηγήθηκε βρίσκονται σε πλήρη αντιστοιχία με τις παρατηρήσεις που έγιναν κατά την ανάλυση των σχημάτων 1.1 και 1.3. Για αυτό τον λόγο στην παρούσα εργασία δεν προχωρήσαμε στη δημιουργία νέων features αλλά έγινε μια τυχαία επιλογή διαφόρων ιδιοτήτων με στόχο όπως θα παρουσιαστεί στην παράγραφο 2.1 να εξεταστεί αν βελτιώνεται ή όχι η ακρίβεια πρόβλεψης των εξεταζόμενων αλγορίθμων ταξινόμησης.



(i) Pearson coefficient (ii) Spearman coefficient

Σχήμα 1.4: Συντελεστές συσχέτισης Pearson & Spearman

1.3 Προετοιμασία δεδομένων

Έχοντας ολοκληρώσει την προεπεξεργασία των δεδομένων στο προηγούμενο βήμα, ακολουθεί η προετοιμασία των δεδομένων πριν αυτά χρησιμοποιηθούν ως είσοδος στα μοντέλα που θα δοκιμαστούν. Είναι γνωστό ότι τα περισσότερα μοντέλα δεν λειτουργούν αποτελεσματικά όταν οι τιμές των διαφόρων features έχουν διαφορετική τάξη μεγέθους, οπότε συνηθίζεται να γίνεται αδιαστατοποίηση των τιμών χρησιμοποιώντας κάποιο scaler. Στην συγκεκριμένη εργασία επιλέχθηκε η κανονικοποίηση (standardization) των δεδομένων χρησιμοποιώντας τη σχέση 1.3.1, όπου με μ αναπαρίσταται ο μέσος όρος του συγκεκριμένου feature και με σ η τυπική απόκλιση. Η συγκεκριμένη τεχνική μπορεί να μην έχει τα επιθυμητά αποτελέσματα εάν τα δεδομένα αποκλίνουν πολύ από την κανονική κατανομή [?], το οποίο όμως όπως ήδη έχει παρουσιαστεί δεν ισχύει για το συγκεκριμένο σύνολο δεδομένων.

$$x^{(i)} = \frac{x^{(i)} - \mu^{(i)}}{\sigma^{(i)}} \tag{1.3.1}$$

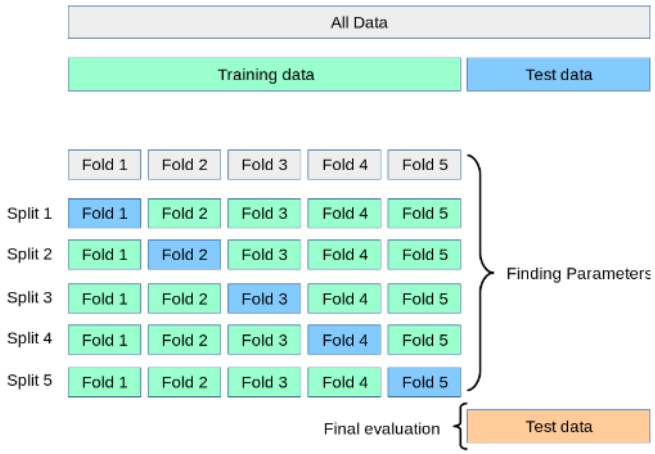
Αυτό είναι και το τελικό εισαγωγικό βήμα και είναι πλέον δυνατό να προχωρήσουμε στην εκπαίδευση και δοκιμή διαφόρων ταξινομητών, ώστε να εξεταστεί η προγνωστική τους ικανότητα για το συγκεκριμένο πρόβλημα, όπως θα δούμε στο επόμενο κεφάλαιο.

Εκπαίδευση μοντέλων

2.1 Εκπαίδευση μοντέλων

Σε αυτό το στάδιο επελέγησαν διάφοροι ταξινομητές από τη βιβλιοθήκη του scikit-learn [?] ώστε να καταδειχθούν οι καταλληλότεροι για το συγκεκριμένο πρόβλημα και να συνεχίσει η διαδικασία εκπαίδευσης με αυτούς. Το κριτήριο το οποίο χρησιμοποιήθηκε για τη σύγκριση των αποτελεσμάτων που παράγονται από τον κάθε ταξινομητή, είναι το κριτήριο της μεγαλύτερης ακρίβειας (accuracy) κατά τη διάρκεια της εκπαίδευσης.

Για τον υπολογισμό της ακρίβειας χρησιμοποιήθηκε το λεγόμενο k-fold cross-validated accuracy. Σύμφωνα με αυτή την τεχνική το σύνολο των δεδομένων που χρησιμοποιούμε κατά την εκπαίδευση X_{train} χωρίζεται σε k τον αριθμό υποσύνολα ίσου μεγέθους. Κατά την διάρκεια της εκπαίδευσης λοιπόν τα $k-1$ τυχαία υποσύνολα χρησιμοποιούνται για την εκπαίδευση του αλγορίθμου, ενώ ένα υποσύνολο χρησιμοποιείται για την εξαγωγή του σφάλματος, παίζοντας κατά κάποιο τρόπο τον ρόλο του test set για την συγκεκριμένη επανάληψη. Η διαδικασία αυτή επαναλαμβάνεται k φορές. Αυτή η διαδικασία περιγράφεται από το σχήμα 2.1.

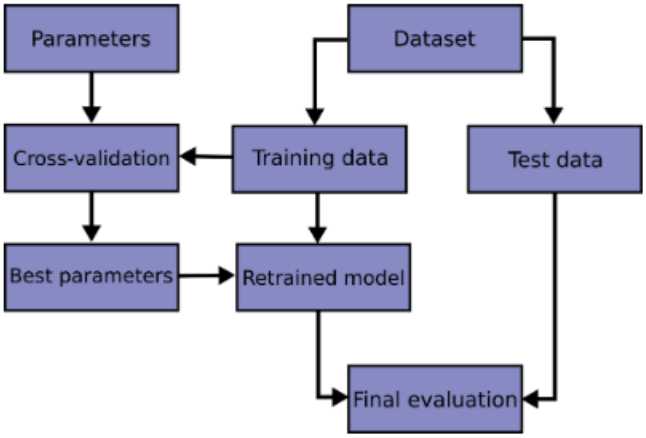


Σχήμα 2.1: k-fold cross-validated [?]

Αναφορικά με το σύνολο των δεδομένων που χρησιμοποιούνται στην εκπαίδευση συμβολίζεται με X_{train} και δίνεται από τη σχέση 2.1.1, όπου με X συμβολίζεται το σύνολο των δεδομένων, ενώ με $testSize$ συμβολίζεται το ποσοστό των μετρήσεων που αποσύρεται από τη διαδικασία της εκπαίδευσης και χρησιμοποιείται ως test set στη τελευταία φάση της εργασίας, όταν θα έχει γίνει επιλογή του καταλληλότερου μοντέλου. Στο σχήμα 2.2 παρουσιάζεται με τη μορφή διαγράμματος ροής η διαδικασία της εκπαίδευσης. Έτσι αρχικά το σύνολο δεδομένων X χωρίζεται σε X_{train} και X_{test}, ενώ ακολουθεί η διαδι-

κασία της εκπαίδευσης των μοντέλων με συγκεκριμένες υπερπαραμέτρους. Σε αυτό το στάδιο η ακρίβεια υπολογίζεται με την τεχνική του cross-validation, ενώ στη συνέχεια γίνεται η εύρεση των καλύτερων υπερπαραμέτρων με τη χρήση μιας συνάρτησης όπως είναι η GridSearchCV ή η RandomizedSearchCV της βιβλιοθήκης του scikit-learn. Τέλος το επανεκπαιδευμένο μοντέλο ελέγχεται ως προς την ακρίβεια ή άλλες παραμέτρους απόδοσης, που θα περιγραφούν στην παράγραφο 2.3, σε σχέση με το test set.

$$X_{train} = (1 - testSize) * X \tag{2.1.1}$$



Σχήμα 2.2: Διαδικασία εκπαίδευσης [?]

Στο συγκεκριμένο παράδειγμα ο χρόνος εκπαίδευσης δεν αποτελεί ιδιαίτερο πρόβλημα καθώς το dataset αποτελείται από λίγα παραδείγματα $\sim$ 3300 και για αυτό τον λόγο δεν έχει έχει συμπεριληφθεί στον πίνακα 2.1, όπου δίνεται η cross-validated ακρίβεια για τον κάθε ταξινομητή. Επιπλέον παρουσιάζονται τα αποτελέσματα εκπαίδευσης για πέντε διαφορετικές περιπτώσεις οι οποίες είναι οι εξής:

- case 1: Χρήση όλων των δεδομένων και μη επεξεργασία ακραίων τιμών
- case 2: Χρήση όλων των δεδομένων εκτός των δεδομένων Conductivity και μη επεξεργασία ακραίων τιμών
- case 3: Χρήση όλων των δεδομένων εκτός των δεδομένων Solids, Turbidity και μη επεξεργασία ακραίων τιμών
- case 4: Χρήση όλων των δεδομένων εκτός των δεδομένων Hardness, Sulfate, Organic_Carbon, Turbidity και μη επεξεργασία ακραίων τιμών
- case 5: Χρήση όλων των δεδομένων και επεξεργασία ακραίων τιμών

Όπως παρατηρούμε λοιπόν από τα συγκεντρωτικά αποτελέσματα του εν λόγω πίνακα, η ακρίβεια πρόβλεψης της κλάσης στην οποία ανήκει το εκάστοτε δείγμα νερού είναι σε όλες τις περιπτώσεις μικρότερη του 70%. Επιπλέον φαίνεται πως η μέθοδος απάλειψης των ακραίων τιμών δεν επηρεάζει θετικά την ακρίβεια των αλγορίθμων και για αυτό τον λόγο επελέγη να μην γίνει χρήση αυτής της τεχνικής στη συνέχεια της εργασίας. Το ίδιο θα μπορούσαμε να ισχυριστούμε και για την παράλειψη ολόκληρων ιδιοτήτων κατά τη διάρκεια της εκπαίδευσης. Βέβαια για τη δεύτερη αυτή περίπτωση που εντάσσεται στα πλαίσια του feature selection/feature engineering μπορούν να γίνουν περαιτέρω δοκιμές,

ενώ σε περίπτωση που ο αναλυτής των δεδομένων διαθέτει επαρκείς γνώσεις επί του συγκεκριμένου θέματος (domain expertise) θα μπορούσε να προχωρήσει στη δημιουργία νέων feature αντιπαρερχόμενος την εικόνα του σχήματος 1.4.

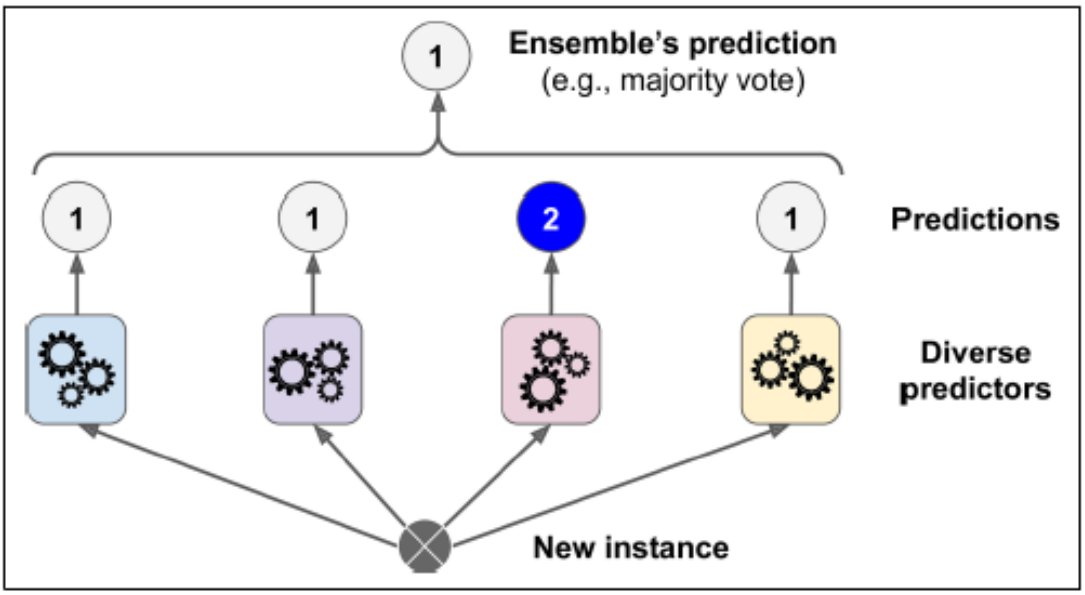
#	Ταξινομητής	case 1	case 2	case 3	case 4	case 5
1	Ridge Regression	0.606	0.606	0.605	0.605	0.606
2	Logistic Regression	0.605	0.606	0.605	0.605	0.605
3	Linear Perceptron	0.505	0.519	0.513	0.518	0.505
4	LDA	0.605	0.606	0.605	0.605	0.605
5	QDA	0.667	0.674	0.664	0.626	0.667
6	Gaussian Process	0.671	0.680	0.662	0.615	0.671
7	Gaussian Naive Bayes	0.621	0.619	0.622	0.602	0.621
8	Linear SVM	0.605	0.605	0.605	0.605	0.605
9	Polynomial SVM	0.605	0.605	0.605	0.605	0.605
10	Gaussian RBF SVM	0.669	0.674	0.666	0.616	0.669
11	k-Neighbors	0.628	0.641	0.623	0.583	0.628
12	Random Forest	0.662	0.661	0.657	0.613	0.661
13	Gradient Boosting	0.662	0.661	0.657	0.613	0.661
14	Ada Boost	0.662	0.661	0.657	0.613	0.661

Πίνακας 2.1: Αποτελέσματα CV accuracy

2.2 Επιλογή μοντέλου & τελειοποίηση

Αφού λοιπόν πραγματοποιήθηκε η παραπάνω ανάλυση και έγινε δοκιμή των διαφόρων μοντέλων καταλήξαμε στην επιλογή των Gaussian RBF SVM, Random Forest, Ada Boost καθώς και του Gradient Boosting ως των καταλληλότερων ταξινομητών για το συγκεκριμένο πρόβλημα. Η λογική στο στάδιο της τελειοποίησης (fine-tuning) είναι να γίνει δοκιμή διαφόρων τιμών,(χρήση της GridSearchCV) για συγκεκριμένες υπερπαραμέτρους (hyperparameters) των μοντέλων ώστε να καταλήξουμε στον συνδυασμό που θα δίνει τη μεγαλύτερη ακρίβεια για κάθε μοντέλο. Η διαδικασία αυτή είναι χρονοβόρα ($\sim 8.5min$ για H/Υ με 8-cores @ 3.9GHz) ακόμα και στην περίπτωση που επιλέγεται η χρησιμοποίηση μόνο λίγων υπερπαραμέτρων ανά μοντέλο όπως έγινε και στη συγκεκριμένη περίπτωση. Παρόλα αυτά εκτελώντας την μια φορά είναι δυνατό να αποθηκευθούν οι συντελεστές και να γίνεται περαιτέρω χρήση τους χωρίς επιπλέον υπολογιστικό και χρονικό κόστος.

Όπως έχει ήδη εξηγηθεί ο τελικός στόχος είναι η επιλογή ενός μοντέλου με όσο το δυνατόν μεγαλύτερη ακρίβεια ταξινόμησης δειγμάτων νερού. Σε αυτό το πλαίσιο δοκιμάστηκε η χρήση των αποτελεσμάτων από περισσότερους του ενός ταξινομητές η οποία επιτυγχάνεται με την εφαρμογή ενός Voting Classifier. Αυτός ο ταξινομητής ανήκει σε μια κατηγορία μεθόδων που λέγεται Ensemble Learning και έχει αποδειχτεί ότι αρκετές φορές βελτιώνει τα αποτελέσματα σε σύγκριση με τις περιπτώσεις όπου χρησιμοποιείται ένας μόνο ταξινομητής [?]. Στη συγκεκριμένη εργασία επιλέξαμε τη χρήση της μεθόδου της σκληρής ψήφου (hard voting) όπου δηλαδή γίνεται άθροιση των προβλέψεων από όλους τους ταξινομητές και στο τέλος επιλέγεται η κλάση με τις περισσότερες ψήφους, όπως φαίνεται παραστατικά και στο σχήμα 2.3. Επιπλέον πρέπει να τονιστεί ότι από τους τέσσερις ταξινομητές που εξετάστηκαν σε αυτό το στάδιο τελικά αποκλείστηκε η χρήση του Ada Boost καθώς παρουσίαζε μικρότερη του αναμενόμενου ακρίβεια κατά τη διαδικασία του fine-tuning.



Σχήμα 2.3: Hard Voting Classifier [?]

2.3 Αξιολόγηση αποτελεσμάτων

Συνοψίζοντας τα βήματα της εκπαίδευσης των ταξινομητών, αρχικά δοκιμάσαμε διάφορους ταξινομητές με τις προεπιλεγμένες υπερπαραμέτρους και καταλήξαμε στα αποτελέσματα CV accuracy του πίνακα 2.1. Πολύ γρήγορα διαπιστώθηκε ότι γραμμικά μοντέλα, όπως τα Ridge Regression, Logistic Regression κτλ, παρουσίαζαν πολύ χαμηλή ακρίβεια. Αξίζει δε σχολιασμού ότι ένας ταξινομητής που θα επέστρεφε σαν αποτέλεσμα πάντα 0 δηλαδή μη πόσιμο νερό θα παρουσίαζε CV accuracy $\sim 61\%$ με βάση την παρατήρηση της παραγράφου 1.1 για την κατανομή των παρατηρήσεων.

Εν συνεχεία επιλέχθηκαν οι τέσσερις επικρατέστεροι ταξινομητές, όπως αναφέρθηκε στην προηγούμενη παράγραφο για τους οποίους εφαρμόστηκε η διαδικασία εύρεσης των καταλληλότερων υπερπαραμέτρων και τελικά συνδυάστηκαν οι τρεις από αυτούς χρησιμοποιώντας τον Voting Classifier.

Έχοντας εκπαιδεύσει τον τελικό ταξινομητή μπορεί να γίνει χρήση του test set το οποίο μέχρι και αυτή τη στιγμή δεν έχει χρησιμοποιηθεί καθόλου στο πρόβλημα. Συγκρίνοντας την ακρίβεια που επιτυγχάνεται στο test set, βλέπε πίνακα 2.2, με το CV accuracy κατά τη διάρκεια της εκπαίδευσης παρατηρείται ότι δεν υπάρχουν σημαντικές αποκλίσεις. Αυτή είναι μια πολύ σημαντική παρατήρηση που μας οδηγεί στο συμπέρασμα ότι ο ταξινομητής δουλεύει σωστά και στην περίπτωση όπου έρχεται αντιμέτωπος με νέα δεδομένα. Σε περίπτωση που το CV accuracy ήταν σημαντικά υψηλότερο του test accuracy θα είχαμε ένδειξη overfitting (high variance), ενώ στην περίπτωση όπου το CV accuracy ήταν πολύ χαμηλό θα είχαμε ένδειξη underfitting (high bias).

Επιπλέον εκτός της ακρίβειας που δίνεται από τη σχέση 2.3.1a και η οποία είναι $\leq 70\%$ για τους λόγους που παρουσιάστηκαν προηγουμένως, είναι σημαντικό να αξιολογηθούν και 3 ακόμα παράμετροι απόδοσης και πιο συγκεκριμένα η λεπτομέρεια (Precision ή Positive Predictive Value) 2.3.1b, η ευαισθησία (sensitivity ή recall) 2.3.1c, καθώς και μια μετα-

βλητή που λέγεται F_1 η οποία συνδυάζει τις δύο προηγούμενες. Τα αποτελέσματα αυτά συνοψίζονται στον πίνακα 2.2, ενώ μπορούν να αποδοθούν με διαφορετικό τρόπο και στο σχήμα 2.4 στο οποίο αναπαρίσταται ο πίνακας σύγχυσης (confusion matrix). Οι παραπάνω παράμετροι απόδοσης ενός ταξινομητή υπολογίζονται με βάση τη σχέση 2.3.1.

$$ACC = \frac{TP + TN}{P + N} \tag{2.3.1a}$$

$$PPV = \frac{TP}{TP + FP} \tag{2.3.1b}$$

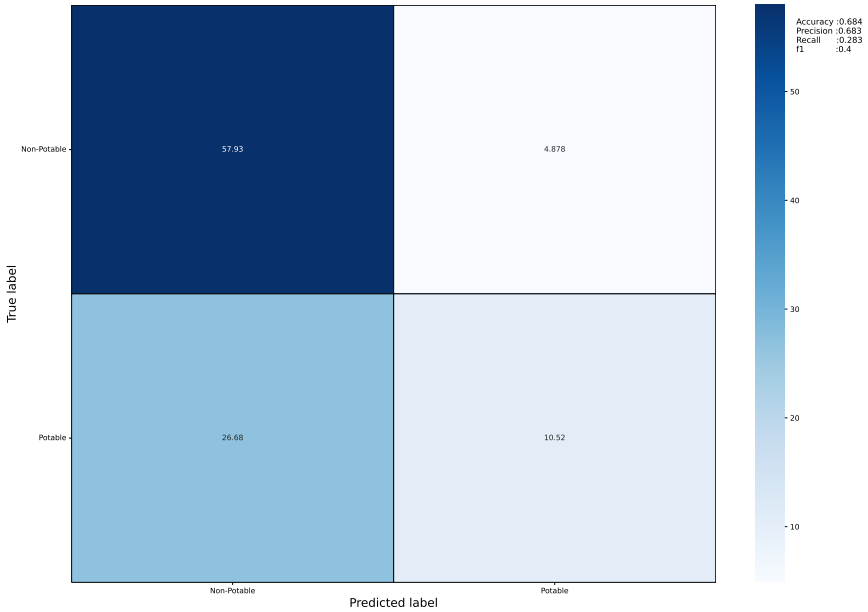
$$SEN = \frac{TP}{TP + FN} \tag{2.3.1c}$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{2.3.1d}$$

$$\begin{cases} TP : True\,Positive,\ FP : False\,Positive \\ TN : True\,Negative,\ FN : False\,Negative \end{cases} \tag{2.3.1e}$$

Ταξινομητής	Ακρίβεια	Λεπτομέρεια	Ευαισθησία	f_1
Gradient Boosting	0.677	0.654	0.279	0.391
Random Forest	0.686	0.698	0.275	0.394
RBF SVM	0.692	0.684	0.320	0.436
Voting Classifier	0.684	0.683	0.283	0.400

Πίνακας 2.2: Αξιολόγηση αποτελεσμάτων (test set)



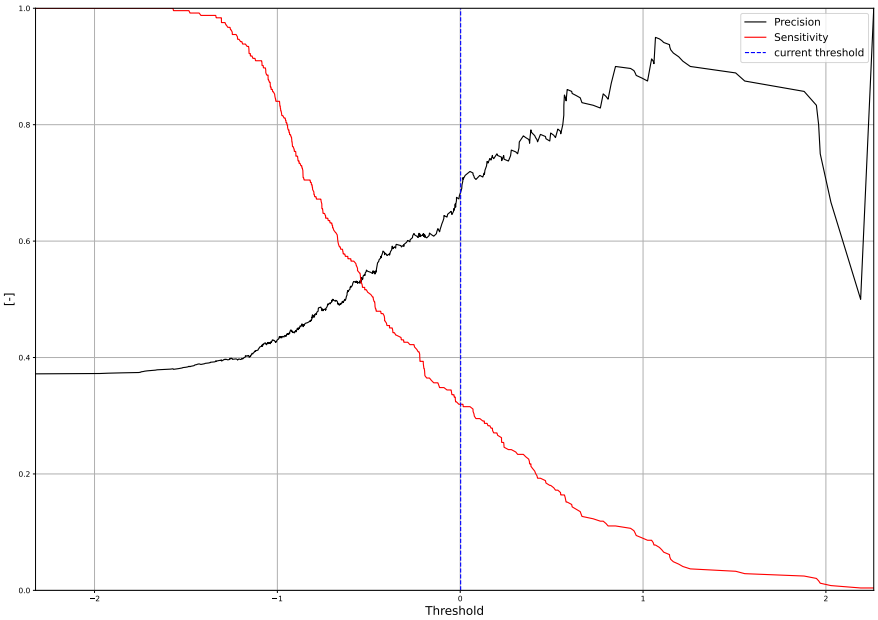
Σχήμα 2.4: Πίνακας σύγχυσης (test set)

Η χρήση της λεπτομέρειας και της ευαισθησίας παίζουν πολύ σημαντικό ρόλο στις περιπτώσεις όπου έχει ιδιαίτερο ενδιαφέρον η μελέτη και αντιμετώπιση των ψευδώς θετικών (False Positive) ή των ψευδώς αρνητικών (False Negative) μετρήσεων αντίστοιχα. Έτσι μπορούμε να φανταστούμε την περίπτωση όπου η ψευδής ταξινόμηση ενός δείγματος νερού

ως πόσιμου μπορεί να είναι απαράδεκτη καθώς μπορεί να έχει αρνητικές συνέπειες για την υγεία ενός ατόμου ή και πληθυσμού που θα το καταναλώσει. Σε αυτή την περίπτωση είναι πολύ πιθανό να απαιτηθεί να υπάρχει πολύ υψηλή λεπτομέρεια, το οποίο όμως θα οδηγήσει σε αντίστοιχη μείωση της ευαισθησίας εξαιτίας της ανταγωνιστικής σχέσης που υπάρχει μεταξύ αυτών των δύο χαρακτηριστικών το οποίο συνοψίζεται στην φράση precision-recall tradeoff.

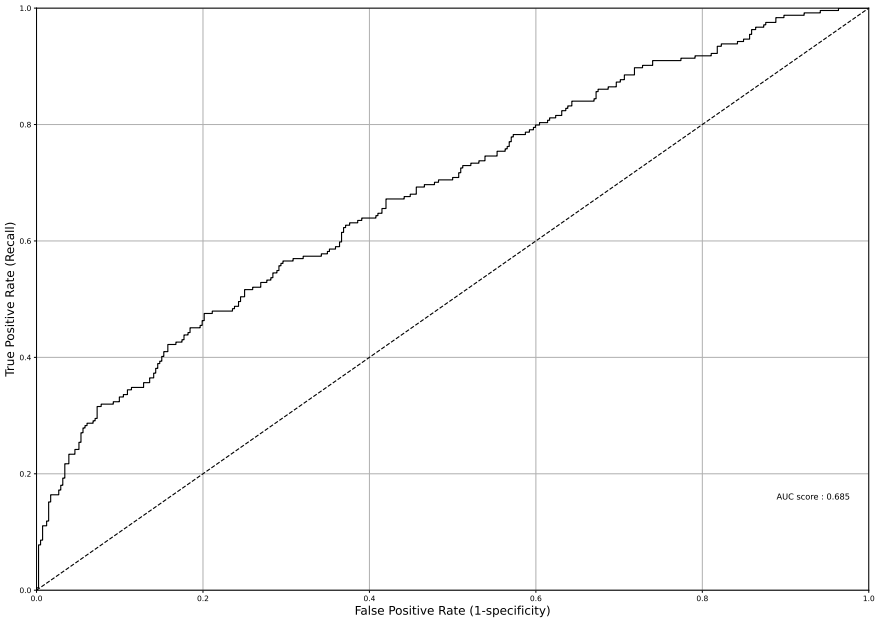
Υπάρχουν λοιπόν συγκεκριμένοι αλγόριθμοι στους οποίους είναι δυνατό να επιλέξουμε το βάρος που θα δοθεί σε μια από αυτές τις χαρακτηριστικές πειράζοντας συγκεκριμένες τιμές κατωφλίου (threshold) στον αλγόριθμο. Για τη μέθοδο που επιλέξαμε, δηλαδή για τον Voting Classifier, κάτι τέτοιο δεν είναι εφικτό να γίνει αλλά θα μπορούσαμε να χρησιμοποιήσουμε ένα ταξινομητή όπως ο Gaussian RBF SVM για τον οποίο είναι δυνατόν να υπολογιστούν τόσο οι πιθανότητες πρόβλεψης κάθε κλάσης.

Για τον συγκεκριμένο αλγόριθμο δίνεται το σχήμα 2.5 στο οποίο παρουσιάζεται η λεπτομέρεια και η ευαισθησία για διάφορες τιμές κατωφλίου. Σε περίπτωση λοιπόν που επί παραδείγματι απαιτηθεί λεπτομέρεια μεγαλύτερη του 90% είναι δυνατό να επιλεγεί η αντίστοιχη τιμή κατωφλίου η οποία θα ικανοποιεί τη συγκεκριμένη συνθήκη.



Σχήμα 2.5: Συμβιβασμός λεπτομέρειας/ακρίβειας Gaussian RBF SVM

Επιπλέον ένα ακόμα διάγραμμα το οποίο χρησιμοποιείται πολλές φορές όταν δουλεύουμε με την λεπτομέρεια και την ευαισθησία είναι το λεγόμενο Receiver Operating Characteristic (ROC), όπως αυτό παρουσιάζεται στο σχήμα 2.6. Η διακεκομμένη γραμμή αναπαριστά την καμπύλη ROC ενός εντελώς τυχαίου ταξινομητή, ενώ ένας καλός ταξινομητής μένει όσο το δυνατόν μακρύτερα από αυτή τη γραμμή. Με αυτό το διάγραμμα λοιπόν μπορούμε να συγκρίνουμε διάφορους ταξινομητές υπολογίζοντας το εμβαδόν υπό της καμπύλης (Area Under the Curve - AUC).



Σχήμα 2.6: Receiving Operating Characteristics (ROC) Gausssian RBF SVM