Κατηγοριοποίηση Συνόλου Δεομένων

Ιονόσφαιρα

Γρηγόριος Μάριος Φραγκάκης
Ψηφιακών Συστημάτων
Πανεπιστήμιο Πειραιώς
Αθήνα Ελλάδα
grigorisfragkakis@gmail.com

Περίληψη

Για την κατηγοριοποίηση του συνόλου δεδομένων «Ιονόσφαιρα» του Uci, χρησιμοποιήθηκαν διάφοροι αλγόριθμοι κατηγοριοποίησης.Συγκεκριμένα:Λογιστική Παλινδρόμηση, Δέντρα Απόφασης, Κ-κοντινότεροι γείτονες,Μηχανές διανυσμάτων υποστήριξης. Χρειάστηκε μια μικρή προεπεξεργασία του συνόλου. Έγιναν βήματα για να αποφευχθεί η υπεπροσαρμογή και έγινε αξιολόγηση των αποτελεσμάτων.

Εισαγωγή

Η κατηγοριοποίηση έγινε με σκοπό,να βρεθεί η ακρίβεια με την οπόια μπορούμε να προβλέψουμε σε ποιά κατηγορία θα ανήκει η κάθε μέτρηση του σύνολου δεδομένων.

Περιγραφή του συνόλου δεδομένων

Το σύνολο δεδομένων αφορά το «Ιονόσφαιρα» του Uci.Τα δεδόμενα έχουν συλλεχθεί απο ραντάρ μέσο ενός συστήματος στο Goose Bay,Labrador.Το σύστημα αυτό αποτελείται από 16 κεραίες υψηλής συχνότητας με συνολική ισχύ 6,4 κιλοβάτ.Οι στόχοι ήταν ελεύθερα ηλεκτρόνια στην ιονόσφαιρα.Οι «καλές» επιστροφές είναι αυτές που δείχνουν στοιχεία κάποιου τύπου δομής στην ιονόσφαιρα.Οι «κακές» επιστροφές είναι αυτες που δεν το κάνουν, διέρχονται απο την ιονόσφαιρα.Τα ληφθέντα σήματα υποβλήθηκαν σε επεξεργασία χρησιμοποιώντας μια συνάρτηση αυτοσυσχέτισης της οποίας τα ορίσματα είναι ο χρόνος ενός παλμού και ο αριθμούς του.Υπήρχαν 17 αριθμοί παλμών.Τα στιγμιότυπα στη βάση δεδομένων περιγράφονται απο 2 χαρακτηριστικά ανα αριθμό παλμού,που αντιστοιχούν στις σύνθετες τιμές που επιστρέφονται από τη συνάρτηση που προκύπτει απο το σύνθετο ηλεκτρομαγνητικό σήμα.

Προ-επεξεργασία δεδομένων

Το σύνολο δεδομένων περιλαμβάνει 34 χαρακτηριστικά τα οποία είναι μετρήσεις, και ακόμα ένα το οποίο έχει τον χαρακτηρισμό «καλό» ή «κακό».Η δεύτερη στήλη έχει μέτρηση 0 για όλα τα στιγμιότυπα.Επομένως επειδή δεν υπάρχει διακύμανση,αφαίρεσα την στήλη απο το σύνολο.

Αλγόριθμοι Κατηγοριοποίησης

Λογιστική Παλινδρόμηση:Επειδή η κατηγοριοποιήση που θα κάνουμε είναι δυαδική(κακή ή καλή μέτρηση),ταιριάζει απόλυτα να χρησιμοποιηθεί η συγκεκριμένη μέθοδος κατηγοριοποίησης. Αυτό διότι η λογιστική παλινδρόμηση παράγει αποτελέσματα σε κλίμακα απο το μηδέν έως το ένα.Κοντά στο μηδέν θεωρείται «κακό» ενώ κοντά στο ένα «καλό».

Δέντρο Απόφασης:Είναι εύκολο να κατασλευαστεί,είναι αρκετά γρήγορο για ένα σύνολο δεδομένων τέτοιου μεγέθους,και μπορούμε να το χρησιμοποιήσουμε για να το συγκρίνουμε με άλλους κατηγοριοποιητές.Το αρνητικό του κατηγοριοποιητή αυτού είναι οτι μπορεί να γίνει εύκολα υπερπροσαρμογή.

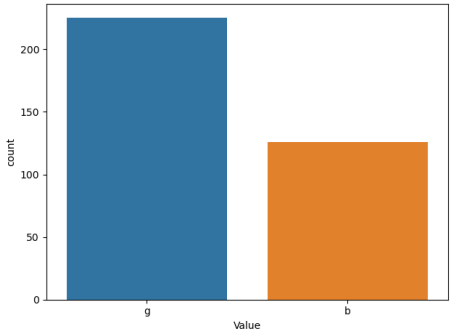
Κ-κοντινότεροι γείτονες: Ο συγκεκριμένος κατηγοριοποιητής χρησιμοποιήθηκε διότι το σύνολο δεδομένων είναι μικρό και οι κλάσεις δεν έχουν θόρυβο.

Μηχανές Διανυσμάτων Υποστήριξης:Παρόμοια με την Λογιστική Παλινδρόμηση, αυτός ο κατηγοριοποιητής,ορίζει το αποτέλεσμα ως «καλό» αν ειναι κοντά στο 1 και «κακό» αν είναι κοντα στο -1.Επίσης το σύνολο δεδομένων είναι μικρό οπότε δεν θα υπάρχει πρόβλημα με την κλιμάκωσει του μοντέλου.

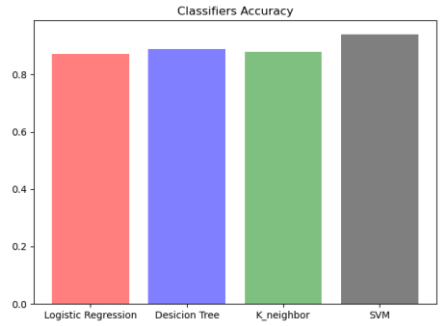
Μεθοδολογία

Έγινε χρήση της βιβλιοθήκης Scikit-Learn και συγκεκριμένα της μεθόδου train_test_split,για να δημιουργήσω τύχαια σύνολα δοκιμής και ελέγχου.Επίσης για να αποφευχθεί η υπερπροσαρμογή χρησιμοποιήθηκε η μέθοδος KFold για να γίνει διασταυρωτική επικύρωση σε όλους τους κατηγοριοποιητές.

Πειραματική Αξιολόγηση



Από το παραπάνω διάγραμμα μπορούμε να καταλάβουμε οτι κατηγορίες δεν έχουν ίσο πληθυσμό.Η διαφορά τους όμως φαίνεται μικρή και δεν επηρεάζει,κάτι που θα αποδειχθεί παρακάτω.

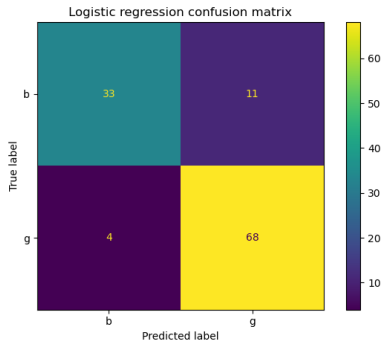


Στο παραπάνω διάγραμμα βλέπουμε την ακρίβεια του κάθε κατηγοριοποιητή , με την χρήση τυχαίων συνόλων δοκιμών και ελέγχων.

Θα προσπαθήσουμε να βγάλουμε αποτελέσματα που θα είναι λιγότερο «τυχαία».

Μήτρα Σύγχυσης

Ας ξεκινήσουμε με την μήτρα σύγχυσης της **λογιστικής παλινδρόμισης.**



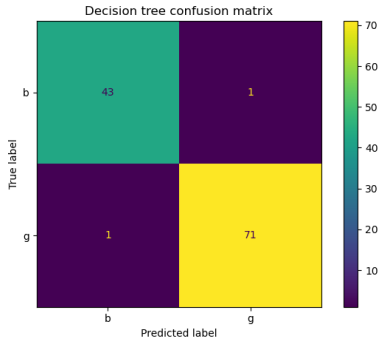
Βλέπουμε οτι τα αποτελέσματα είναι πολύ καλά και μπορούμε να το διαπιστώσουμε με την χρήση της μεθόδου classification_report του Scikit-learn.

Logistic regression classification report:

	precision	recall	f1-score	support
b	0.89	0.75	0.81	44
g	0.86	0.94	0.90	72
accuracy			0.87	116
macro avg	0.88	0.85	0.86	116
weighted avg	0.87	0.87	0.87	116

Βλέπουμε οτι ο σταθμισμένος μέσος όρος του F1 είναι πολύ κοντά στο 1.Επειδή το F1 λαμβάνει υπόψιν τα ψευδής θετικά και τα ψευδής αρνητικά,καταλαβαίνουμε οτι η διαφορά του πληθυσμού των κατηγοριών δεν είναι αρκετά μεγάλη για να επηρεάσει τα αποτελέσματά μας.

Συνεχίζουμε με την μήτρα σύγχυσης του **δέντρου απόφασης**.

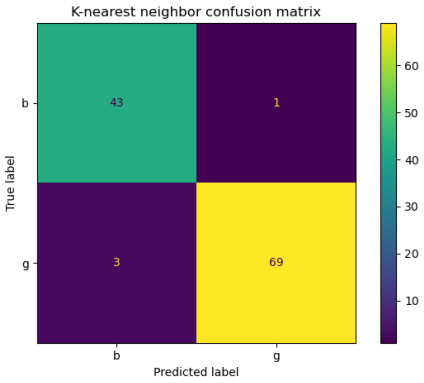


Decision tree classification report:

	precision	recall	f1-score	support
b	0.79	0.77	0.78	44
g	0.86	0.88	0.87	72
accuracy			0.84	116
macro avg	0.83	0.82	0.83	116
weighted avg	0.84	0.84	0.84	116

Διαπιστώνουμε οτι τα αποτελέσματα και για το δέντρο απόφασης είναι πολύ καλά.

Συνεχίζουμε με την μήτρα σύγχυσης του κατηγοριοποιητή **Κ-κοντινότεροι γείτονες.**

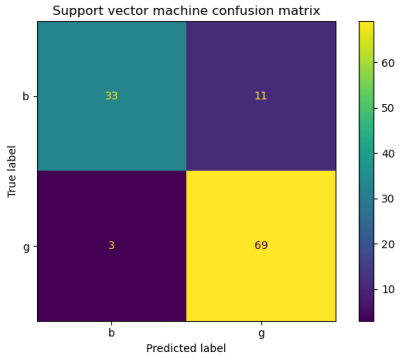


K-nearest neighbor classification report:

	precision	recall	f1-score	support
b	0.92	0.55	0.69	44
g	0.78	0.97	0.86	72
accuracy			0.81	116
macro avg	0.85	0.76	0.77	116
weighted avg	0.83	0.81	0.80	116

Τα αποτελέσματα είναι επίσης καλά.

Τελειώνοντας θα δούμε την μήτρα σύγχυσης του κατηγοριοποιητή **Μηχανές Διανυσμάτων Υποστήριξης.**



Support vector machine classification report:

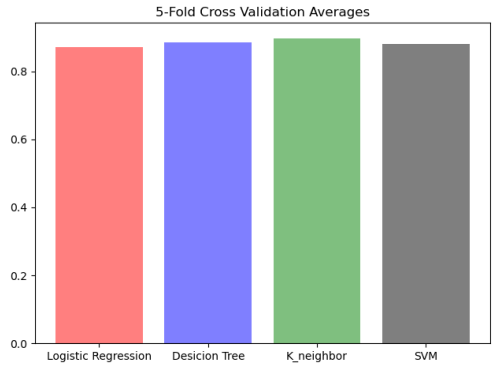
	precision	recall	f1-score	support
b	0.92	0.75	0.83	44
g	0.86	0.96	0.91	72
accuracy			0.88	116
macro avg	0.89	0.85	0.87	116
weighted avg	0.88	0.88	0.88	116

Τα αποτελέσματα είναι και εδώ πολύ καλά.

Διασταυρωτική επικύρωση

Κάνουμε διασταυρωτική επικύρωση με 5 πτυχές για να είμαστε πιο σίγουροι για τα αποτελέσματα μας,σε αντίθεση με τα απλά σύνολα ελέγχου.

Το τελικό διάγραμμα που δημιουργήσαμε είναι το εξής:



Συμπεράσματα

Τελικά αφού έχει ολοκληρωθεί η διασταυρωτική επικύρωση με 5 πτυχές,φαίνεται οτι ο κατηγοριοποιητής K-κοντινότερος γείτονας προσφέρει την μεγαλύτερη ακρίβεια για την πρόβλεψη μας.

Βιβλιογραφικές πηγές

REFERENCES

- Pang-Ning Tan,Michael Steinbach,Anuj Karpatne,Vipin Kumar,Επισημονικη Επιμελεια Βασίλειος Σ. Βερύκιος(2019).ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ,εκδόσεις Τζιολα.
- https://scikit-learn.org/stable/