01/01/2023

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

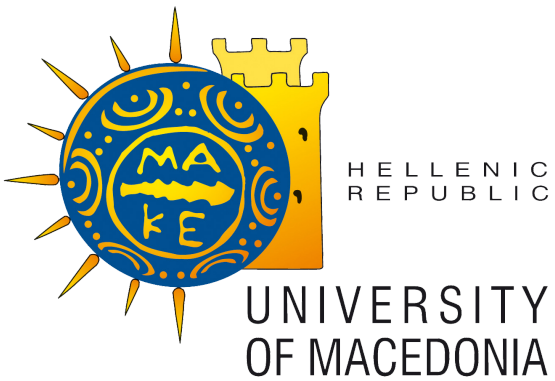
2η Εργασία

Σχολή:
Εφαρμοσμένης Πληροφορικής

Τμήμα:
Εισαγωγή στην Επιστήμη και Τεχνολογία Υπολογιστών

Θέμα: «Classification problems»

Ομάδα:
1. Μιχαήλ Ζέρβας , ics20015
2. Νέστορας Σωτηρίου , ics20012
3. Χρήστος Δρίκος , ics20043



Ακαδημαϊκό έτος 2022-2023

Πίνακας περιεχομένων

Πίνακας περιεχομένων γραφημάτων

Εισαγωγή

Δεδομένων κάποιων χρηματοπιστωτικών δεικτών ελληνικών επιχειρήσεων, καθώς και κάποια επιπλέον δεδομένα που αφορούν αυτές, γίνεται χρήση 4 μοντέλων κατηγοριοποίησης με σκοπό να βρεθεί το καλύτερο από αυτά που προβλέπει, με βάση αυτά τα δεδομένα, πότε μια εταιρία είναι χρεοκοπημένη ή όχι. Τα 4 μοντέλα κατηγοριοποίησης που χρησιμοποιήθηκαν είναι τα: Linear Discriminant Analysis, Decision Trees, k-Nearest Neighbors και Naïve Bayes. Η επιλογή του βέλτιστου μοντέλου έγινε βάση της μετρικής F1 score.

Μέθοδοι που εφαρμόστηκαν

Γλώσσα που χρησιμοποιήθηκε: Python

Έχουμε στην διάθεσή μας δεδομένα διαφόρων επιχειρήσεων σε διάστημα 4 χρόνων από το 2006 έως το 2009. Ο όγκος των δεδομένων ανέρχεται συνολικά στα 10716 (σε γραμμές), για όλα τα έτη και τις επιχειρήσεις.

2ο ερώτημα)
Το ερώτημα αυτό υλοποιείται με την κατάλληλη χρήση του groupby και του aggregate. Υπολογίζεται για κάθε έτος ο αριθμός υγιών και χρεοκοπημένων επιχειρήσεων και για κάθε κατηγορία ξεχωριστά υπολογίζεται το min, max και mean.

3ο ερώτημα)
Τα δεδομένα χωρίζονται σε 80% training data και 20% testing data.

5ο ερώτημα)
Επιλέχθηκαν τα εξής μοντέλα:

- Linear Discriminant Analysis (LDA):
  Ο σκοπός της LDA είναι να επιτύχουμε καλό διαχωρισμό μεταξύ των κατηγοριών (κλάσεων), με ταυτόχρονη ελάττωση διαστάσεων (μεταβλητών από τις οποίες αποτελούνται τα δεδομένα).Για την επίτευξη της καλύτερης διαχωρισιμότητας, τα δεδομένα κατά το μετασχηματισμό τους προβάλλονται σε χώρο μικρότερης διάστασης από τον αρχικό, με αποτέλεσμα να ελαττώνεται το πλήθος των διαστάσεών τους (δηλαδή των μεταβλητών από τις οποίες αποτελούνται), έτσι ώστε στη συνέχεια η ταξινόμηση νέων δεδομένων να είναι ακριβέστερη.
- Decision Trees:
  Κατά την συγκεκριμένη μέθοδο, εφαρμόζεται η τεχνική του «διαίρει και βασίλευε» (Divide and Conquer),διεξάγοντας μια άπληστη αναζήτηση για τον εντοπισμό των βέλτιστων σημείων διάσπασης μέσα σε ένα δέντρο. Ο χώρος του προβλήματος χωρίζεται σε περιοχές από στιγμιότυπα που φέρουν την ίδια τιμή ως προς κάποια μεταβλητή χαρακτηριστικό, και η διαδικασία επαναλαμβάνεται αναδρομικά έως ότου όλες ή η πλειονότητα των εγγραφών, ταξινομηθούν σε συγκεκριμένες κλάσεις, αναπαριστώντας με τον τρόπο αυτό το παραγόμενο μοντέλο ως δένδρο απόφασης. Το αν όλα τα σημεία δεδομένων ταξινομούνται ή όχι ως ομοιογενή σύνολα εξαρτάται σε μεγάλο βαθμό από την πολυπλοκότητα του δέντρου αποφάσεων.

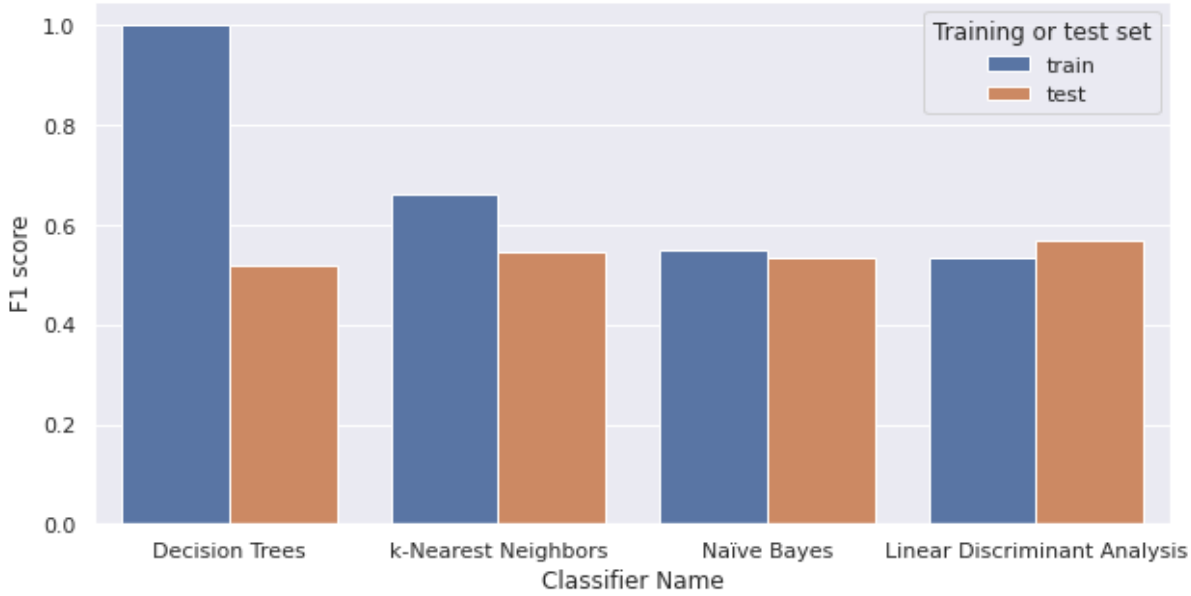
- k-Nearest Neighbors (kNN):
  Ο kNN προσπαθεί να προβλέψει τη σωστή κλάση για τα testing data υπολογίζοντας την απόσταση μεταξύ των testing data και όλων των σημείων εκπαίδευσης. Στη συνέχεια, επιλέγει τον αριθμό k των σημείων που είναι πιο κοντά στα test data. Ο αλγόριθμος kNN υπολογίζει την πιθανότητα τα test data να ανήκουν στις κλάσεις των "k" training data και θα επιλεγεί η κλάση με την υψηλότερη πιθανότητα. Για την επίλυση του προβλήματος της άσκησης χρησιμοποιήθηκε k=3.
- Naïve Bayes:
  Ο Naïve Bayes αλγόριθμος βασίζεται στο θεώρημα πιθανοτήτων Bayes για να υπολογίσει την πιθανότητα το δείγμα εισόδου να ανήκει σε κάθε πιθανή κλάση, με βάση τις τιμές των χαρακτηριστικών. Η κλάση με την υψηλότερη πιθανότητα επιλέγεται στη συνέχεια ως η προβλεπόμενη κλάση.
  Υπάρχουν 3 τύποι αλγορίθμων Naïve Bayes: ο Gaussian Naïve Bayes, ο Multinomial Naïve Bayes και ο Bernoulli Naïve Bayes. Εδω χρησιμοποιήθηκε ο Gaussian Naïve Bayes, σύμφωνα με τον οποίο όταν έχουμε συνεχείς τιμές χαρακτηριστικών, υποθέτουμε ότι οι τιμές που σχετίζονται με κάθε κλάση κατανέμονται σύμφωνα με την κατανομή Gauss ή την κανονική κατανομή.

7^{ο} ερώτημα)
Χρησιμοποιήθηκαν 4 μετρικές για την αξιολόγηση της επίδοσης κάθε εκπαιδευμένου μοντέλου. Αναλυτικά:
Accuracy: Πρόκειται για τον αριθμό των σωστών προβλέψεων που έκανε το μοντέλο, διαιρεμένο με τον συνολικό αριθμό των προβλέψεων που έγιναν. Είναι ένα μέτρο του συνολικού ποσοστού ορθής ταξινόμησης του μοντέλου.
Precision: Πρόκειται για τον αριθμό των αληθώς θετικών προβλέψεων που έκανε το μοντέλο, διαιρεμένο με τον συνολικό αριθμό των θετικών προβλέψεων που έγιναν. Είναι ένα μέτρο του ποσοστού των θετικών προβλέψεων που ήταν πραγματικά σωστές.
Recall: Πρόκειται για τον αριθμό των αληθώς θετικών προβλέψεων που έκανε το μοντέλο, διαιρεμένο με τον συνολικό αριθμό των πραγματικών θετικών περιπτώσεων στα δεδομένα. Είναι ένα μέτρο του ποσοστού των πραγματικών θετικών περιπτώσεων που προβλέφθηκαν σωστά.
F1 score: Πρόκειται για τον αρμονικό μέσο όρο του Precision και του Recall. Είναι ένα μέτρο της ισορροπίας μεταξύ Precision και Recall, με υψηλότερες τιμές να υποδηλώνουν καλύτερη ισορροπία.

Αυτές οι τέσσερις μετρικές αξιολόγησης χρησιμοποιούνται για την αξιολόγηση της απόδοσης ενός ταξινομητή και μπορούν να δώσουν πληροφορίες για διάφορες πτυχές της συμπεριφοράς του. Για παράδειγμα, ένας ταξινομητής με υψηλό Accuracy μπορεί

να μην έχει απαραίτητα υψηλό Precision ή Recall και το αντίστροφο. Το F1 score μπορεί να δώσει μια ενιαία μετρική που αντικατοπτρίζει τη συνολική απόδοση του ταξινομητή.

Παρακάτω φαίνεται μία συνολική γραφική παράσταση της απόδοσης όλων των ταξινομητών που χρησιμοποιήθηκαν όσον αφορά την μετρική F1 score, τόσο στο training set όσο και στο test set.



Στα παραπάνω διάγραμμα παρατηρούμε μεγάλη διαφορά στα αποτελέσματα μεταξύ train και test sets στον Decision Tree Classifier. Αυτό οφείλεται στο γεγονός ότι το Decision Tree έκανε overfitting στο training dataset. Έτσι, όπως ήταν αναμενόμενο, η απόδοσή του στο test set ήταν κατά πολύ μειωμένη συγκριτικά με αυτήν στο training set. Μια ακόμη, μικρότερης τάξεως διαφορά παρατηρείται στον kNN. Εδώ πιθανόν να οφείλεται είτε σε underfitting, όπου το μοντέλο αδυνατεί να βρει σε ικανοποιητικό βαθμό τα διάφορα μοτίβα που μπορεί υπάρχουν στο train set και έτσι δεν κάνει καλές προβλέψεις στο test set, είτε η τιμή του k μπορεί να μην ήταν η ιδανική, είτε ακόμα και στην τυχαία διαχωριοποίηση των δεδομένων σε train και test. Στους υπόλοιπους ταξινομητές δεν παρατηρήθηκε κάποια αξιοσημείωτη διαφορά στην F1, ανάμεσα στο train και στο test set. Τελικά το καλύτερο μοντέλο σύμφωνα με το F1 score, όπως φαίνεται και στο διάγραμμα, είναι το Linear Discriminant Analysis.

Συμπεράσματα

Βάση την μετρική F1-score, το μοντέλο που αποδίδει καλύτερα (αν και οι διαφορές μεταξύ τους είναι μικρές) φαίνεται να είναι το Linear Discriminant Analysis (LDA), μετά ο Naïve Bayes με τον kNN και τέλος το Decision Tree. Για την βελτίωση του Decision Tree ώστε να μην κάνει overfitting, μπορούμε είτε να κλαδέψουμε τμήματα από το πλήρες δέντρο που θα δημιουργηθεί, είτε να δημιουργήσουμε εξ αρχής ένα δέντρο με λιγότερα κλαδιά από όσα ήταν να δημιουργηθούν. Για την βελτίωση όλων των μοντέλων, θα μπορούσαμε, εφόσον είναι εφικτό, να χρησιμοποιήσουμε περισσότερα δεδομένα. Η αύξηση του όγκου των δεδομένων εκπαίδευσης μπορεί συχνά να βοηθήσει ένα μοντέλο να εντοπίσει πιο γενικά πρότυπα, γεγονός που μπορεί να βελτιώσει την απόδοσή του στο σύνολο δοκιμών.

Πηγές

https://medisp.uniwa.gr/
https://www.ibm.com/topics/decision-trees
https://nemertes.library.upatras.gr/items/4391cc25-e726-448a-ab83-85f3a8ffe46f
https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4
https://www.kaggle.com/code/prashant111/naive-bayes-classifier-in-python