ΕΡΓΑΣΙΑ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΑΠΟ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

ΕΜΜΑΝΟΥΗΛ ΜΟΡΦΙΑΔΑΚΗΣ 3150112
ΙΩΑΝΝΗΣ ΜΠΟΥΖΙΟΣ 3150119
ΚΩΝΣΤΑΝΤΙΝΟΣ ΛΟΥΡΙΔΑΣ 3150093

Αρχικά προσπαθήσαμε να προσεγγίσουμε το πρόβλημα με λογιστική παλινδρόμιση χρησιμοποιώντας μόνο τις μεταβλητές Departure,Arrival και DateOfDeparture, αλλά δεν καταφέραμε να πέτυχουμε υψηλό σκορ.

Στην συνέχεια χρησιμοποιήσαμε KNearestNeighborhoodClassifier αναμεσά LogitudeDeparture και LogitudeArrival,επειδή ο συγκεκριμένος αλγόριθμος δεν λειτουργεί καλα σε υψηλές διαστάσεις και τα συγκεκριμένα χαρακτηριστικά έχουν μεγάλο κέρδος πληροφορίας από όσο παρατηρήσαμε και οι τιμες είναι πραγματικοι αριθμοι.Το σκορ μας όμως ανέβηκε λίγο(0,42) , και έπρεπε να προσεγγίσουμε καλυτέρα το πρόβλημα.

Μετά αποφασίσαμε να χρησιμοποιήσουμε DecisionTreeClassifier μεταξύ των δεδομένων Departure , Arrival και DateOfDeparture.Τα κωδικοποιήσαμε με το OneHotEncoder με σκοπό να αυξηθούν οι διαστάσεις των δεδομένων, επειδή ο DecisionTreeClassifier λειτουργεί καλά σε υψηλές διαστάσεις . Παρ όλα αυτά το σκορ δεν ανέβηκε ιδιαίτερα, μόλις στο (0,43), μετά ορίσαμε ως maxdepth το 1000 με σκοπό να αποφύγουμε το overfitting,και ανέβηκε λίγο ακόμα το σκορ μας. Έπειτα προσπαθήσαμε να κάνουμε τροποποιήσεις στον αριθμό των φύλλων και των στοιχείων που θα έχει το κάθε

φύλλο ,όμως αυτό αντί να μας ανεβάσει το σκορ, το μειωσε κατά λίγο.

Οπότε χρησιμοποιήσαμε Νευρωνικά Δίκτυα στα τρία χαρακτηριστικά (DateofDeparture , Departure , Arrival) .Δεν προσθέσαμε τις εβδομάδες επειδή δεν μπορούσαμε να τις κωδικοποιήσουμε αρκετά καλά. Αρχικά, χρησιμοποιήσαμε hidden_layer_sizes(1000,1000,1000) επειδή είχαμε 592 features εξαιτίας του OneHotEncoder και το σκορ μας ανέβηκε στο 0,51.Επειτα αποφασίσαμε να μειώσουμε από (1000,1000,1000) σε (600,600,600) ώστε να μην έχουμε πολλούς περιττούς νευρώνες .Με αποτέλεσμα το σκορ μας να ανεβεί στο 0,57.

Με τους παρακάτω αλγόριθμους PCA (Μείωση Διαστάσεων), Bernoulli,Gaussian, το σκορ που πετύχαμε ήταν πάρα πολύ χαμηλό.
Επίσης με RandomForestClassifier δεν καταφέραμε να προσεγγίσουμε καλυτέρα το πρόβλημα από ότι είχαμε καταφέρει με το DecisionTree.

Ακομη δεν προσθεσαμε τα υπολοιπα χαρακτηριστικα επειδη ηταν παραγωγα των αλλων χαρακτηριστικων