Text

Description automatically generated

ΑΝΑΛΥΤΙΚΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΙΩΑΝΝΗΣ ΚΩΝΣΤΑΝΤΑΚΟΣ

mpked2215

ΓΕΩΡΓΙΟΣ ΛΕΒΑΝΤΗΣ

mpked2216

ΑΝΑΛΥΤΙΚΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Το dataset που επιλέχτηκε ώστε να κάνουμε Προπαρασκευή Δεδομένων, Συσταδοποίηση και Ομαδοποίηση σε γλώσσα Python είναι το US\_Accidents dataset το οποίο περιέχει δεδομένα για 2.845.342 ατυχήματα στις Η.Π.Α συνολικά για τα έτη 2016 - 2021.

Αρχικά, εγκαταστήσαμε τις βιβλιοθήκες που χρησιμοποιήσαμε για την υλοποίηση και φορτώσαμε τα αρχεία και τα αποθηκεύσαμε με τη βοήθεια της βιβλιοθήκης pandas σε DataFrames.

Στη συνέχεια, κάναμε περιγραφική στατιστική, χρησιμοποιώντας εργαλεία οπτικοποίησης (seaborn, matplotlib), για να γνωρίσουμε καλύτερα τα δεδομένα μας σε βάθος και προχωρήσαμε στον καθαρισμό τους αφαιρώντας διπλότυπες και κενές τιμές και αφού έτσι διασφαλίσαμε ότι τα δεδομένα είναι ακριβή και έτοιμα για ανάλυση.

Στο παρακάτω διάγραμμα φαίνεται η συσχέτιση της μεταβλητής Severity με τις υπόλοιπες μεταβλητές σε ένα matrix heatmap διάγραμμα.



Στο διάγραμμα αυτό οπτικοποιούμε για κάθε στήλη το ποσοστο κενών (null) τιμών.

Chart

Description automatically generated

Σε παρακάτω διάγραμμα περιγράφεται η μεταβλητή Severity. Διακρίνουμε πως η συντριπτική πλειοψηφία του πεδίου που θα μελετήσουμε παίρνει την τιμή «2».

Graphical user interface, application

Description automatically generated

Εδώ βλέπουμε top 10 States σε ατυχήματα. Παρατηρούμε πως μεγάλα ποσοστά έχουν η πολιτεία της Καλιφόρνια και της Φλόριντα και στο αμέσως επόμενο διάγραμμα φαίνονται οι 20 πόλεις με τα περισσότερα ατυχήματα

Chart, pie chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart

Description automatically generated

Στο συγκεκριμένο διάγραμμα βλέπουμε πως κατανέμεται η μεταβλητή Severity στα χρόνια που θα την εξετάσουμε.

A picture containing graphical user interface

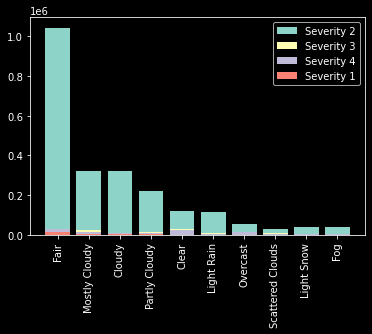
Description automatically generated

Ενώ στο διάγραμμα αυτό βλέπουμε κατά μέσο όρο ποιες ώρες της μέρας και τι επικινδυνότητας ατύχημα πραγματοποιήθηκε. Παρατηρούμε πως οι ώρες που συμβαίνουν πιο συχνά ατυχήματα είναι μεταξύ 05:00 – 08:00 αλλά κυρίως 16:00 – 19:00.

Chart, histogram

Description automatically generated

Στην συνέχεια δημιουργήθηκαν κάποια διαγράμματα για να καταλάβουμε τι έιδους ατυχήματα γίνονται σε σχέση με τα καιρικά φαινόμενα και με τι συχνότητα.



Chart

Description automatically generated

Σειρά είχε ο μετασχηματισμός του συνόλου δεδομένων μας, δηλαδή η διαδικασία τροποποίησης των δεδομένων με κάποιο τρόπο πριν αναλυθούν περαιτέρω ή χρησιμοποιηθούν για μηχανική μάθηση. Συγκεκριμένα, αφαιρέσαμε αρκετά πεδία και γραμμές, δημιουργήσαμε νέα χαρακτηριστικά όπως να μετασχηματίσουμε την ώρα και μέρα του ατυχήματος ώστε να μπορέσουμε να εξετάσουμε για όλο το dataset πως κατανέμονται τα ατυχήματα και να κανονικοποιήσουμε τις τιμές των μεταβλητών που επιλέξαμε ώστε να προχωρήσουμε σε τεχνικές machine learning.

Η Συσταδοποίηση (Clustering) που ακολούθησε είναι μια τεχνική μηχανικής εκμάθησης χωρίς επίβλεψη που περιλαμβάνει τη διαίρεση ενός συνόλου δεδομένων σε ομάδες (clusters) με βάση την ομοιότητα των δεδομένων σε κάθε ομάδα και καταλήξαμε στο ότι καλύτερη συσταδοποίηση πραγματοποιείται με τον k-means. Αν και όπως είναι αναμενόμενο τα αποτελέσματα δεν είναι ακριβή γιατί το συντριπτικό ποσοστό των ατυχημάτων ανήκουν στο “severity” = 2. Επίσης συμπεραίνουμε ότι τα features δεν είναι αρκετά ισχυρά ώστε να κατατάξουν με ασφάλεια ένα case στη σωστή τάξη-κλάση.

Chart, scatter chart

Description automatically generated

A picture containing text, monitor, screen, display

Description automatically generated

Τέλος, πραγματοποιήσαμε Classification στο μοντέλο μας δοκιμάζοντας πέντε διαφορετικές προσεγγίσεις αλλά δυστυχώς μόνο οι δύο δούλεψαν και έφεραν αποτελέσματα. Συγκεκριμένα δοκιμάσαμε Logistic Regression, Decision Trees, Support Vector Machine (SVM), K-Nearest και Random Forest και δούλεψαν οι δύο πρώτες προσεγγίσεις.

**Logistic Regression**

precision recall f1-score support

1 0.00 0.00 0.00 8383

2 0.90 1.00 0.95 788324

3 0.23 0.00 0.00 45063

4 0.38 0.00 0.01 37553

accuracy 0.90 879323

macro avg 0.38 0.25 0.24 879323

weighted avg 0.83 0.90 0.85 879323

[[ 0 8378 0 5]

[ 0 788075 19 230]

[ 0 45035 6 22]

[ 0 37396 1 156]]

**Decision Trees**

precision recall f1-score support

1 0.35 0.40 0.37 8383

2 0.92 0.92 0.92 788324

3 0.23 0.23 0.23 45063

4 0.26 0.28 0.27 37553

accuracy 0.85 879323

macro avg 0.44 0.46 0.45 879323

weighted avg 0.85 0.85 0.85 879323

[[ 3342 4561 398 82]

[ 5391 724376 31515 27042]

[ 708 30656 10180 3519]

[ 124 23664 3125 10640]]

Το Logistic Regression μας δίνει αρκετά σχεδόν απόλυτη ακριβεια αλλά μόνο σε σχέση με την πρόβλεψη της κλάσης 2. Ενώ στο Decision Trees έχουμε καλύτερα ποσοστά για όλες τις κλάσεις και ελαφρώς χειρότερα για τη 2. Εδώ όμως πρέπει να αναρωτηθούμε τι μας δίνει συνολικα καλύτερη εικόνα. Αυτό γιατί τα κλάση «2» είναι σχεδόν 2.5 εκ. οπότε αν έχουμε μια απόλυτη ακρίβεια σε αυτά θα έχω καλή ακριβεια σχεδόν σε όλο το σετ. Ενώ αν αυξήσω λίγο την ακρίβεια των άλλων 3 κλάσεων δεν θα κερδίσω συνολικά γιατί αυτές οι κλάσεις είναι μικρό ποσόστο του σετ.