|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Αναλυτική Δεδομένων -  US Accidents Dataset | | | | | |
|  |  |  |  |  |
| Τεχνική Αναφορά της  εργασίας του μαθήματος  «Αναλυτική Δεδομένων».  Ακαδημαϊκό Έτος: 2021 – 2022 | | | Ομάδα:  |  |  | | --- | --- | | Graduation Cap Png Transparent Images - Clip Art Grad Cap, Png Download -  kindpngΜπουμπλίνη Αναστασία (Π19117) | Graduation Cap Png Transparent Images - Clip Art Grad Cap, Png Download -  kindpngΜπριστογιάννης Ιωακείμ (Π19048) | | Mail icon vector png - Pixsectoraboublini@gmail.com | Mail icon vector png - Pixsectorioakeim13@hotmail.gr | | Bxl microsoft teams icon - BoxiconsANASTASIA BOUBLINI (p19117@unipi.gr) | Bxl microsoft teams icon - BoxiconsIOAKEIM EL-KHATTAB-BRISTOGIANNIS (p19048@unipi.gr) | | |



Πίνακας Περιεχομένων

[Αναλυτική Δεδομένων - 1](#_Toc113883753)

[US Accidents Dataset 1](#_Toc113883754)

[Ομάδα: 1](#_Toc113883755)

[Πρόλογος 4](#_Toc113883756)

[Ά Μέρος – Εξοικείωση με τα δεδομένα 5](#_Toc113883757)

[1. Εκκαθάριση Δεδομένων 5](#_Toc113883758)

[1.1 Διαγραφή Περιττών Στηλών 5](#_Toc113883759)

[1.2 Διαγραφή Null Τιμών 6](#_Toc113883760)

[1.3 Εκκαθάριση της στήλης Weather\_Condition 7](#_Toc113883761)

[2. Οπτικοποίηση Δεδομένων 8](#_Toc113883762)

[2.1 Εξαγωγή χρήσιμης πληροφορίας 8](#_Toc113883763)

[2.2 Συσχέτιση και Συνδιακύμανση 10](#_Toc113883764)

[3. Κλιμάκωση Δεδομένων 13](#_Toc113883765)

[΄Β Μέρος – Τεχνικές Ομαδοποίησης 16](#_Toc113883766)

[4. K-means 16](#_Toc113883767)

[4.1 Εφαρμογή του K-means 16](#_Toc113883768)

[4.2 Οπτικοποίηση του K-means 17](#_Toc113883769)

[5. DBSCAN 19](#_Toc113883770)

[5.1 Εφαρμογή του DBSCAN 19](#_Toc113883771)

[5.2 Οπτικοποίηση του DBSCAN 19](#_Toc113883772)

[6. OPTICS 22](#_Toc113883773)

[6.1 Εφαρμογή του OPTICS 22](#_Toc113883774)

[6.2 Οπτικοποίηση του OPTICS 22](#_Toc113883775)

[7. Σύγκριση Τεχνικών Ομαδοποίησης 25](#_Toc113883776)

[΄Γ Μέρος – Μοντέλα Μηχανικής Μάθησης 26](#_Toc113883777)

[8. Απλή Γραμμική Παλινδρόμηση 26](#_Toc113883778)

[8.1 Επιθεώρηση και Προετοιμασία Δεδομένων 27](#_Toc113883779)

[8.2 Κανονικοποίηση 29](#_Toc113883780)

[8.3 Γραμμική Παλινδρόμηση 29](#_Toc113883781)

## Πρόλογος

Η παρούσα εργασία ανάγεται στην μελέτη και την ανάλυση του συνόλου δεδομένων "[US Accidents](https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents)" από το Kaggle, το οποίο αποτελείται από περίπου 2.8 εκατομμύρια εγγραφές, και περιέχει πληροφορίες σχετικά με αυτοκινητιστικά ατυχήματα στις ΗΠΑ κατά το χρονικό διάστημα Φεβρουάριος 2016 – Δεκέμβριος 2021 (ανανεώνεται σε ετήσια βάση).

Η συγκεκριμένη εργασία ανήκει στο πεδίο της Μηχανικής Μάθησης (Machine Learning) και πιο συγκεκριμένα ασχολείται με την ανάλυση των δοθέντων δεδομένων. Στόχος της είναι η εξαγωγή ακριβέστερων συμπερασμάτων, πληροφοριών και αποτελεσμάτων ταξινόμησης των αυτοκινητιστικών ατυχημάτων.

Η παρούσα τεχνική αναφορά χωρίζεται στα παρακάτω βασικά μέρη:

**Ά Μέρος**: Εξοικείωση με το σύνολο δεδομένων

**΄Β Μέρος**: Τεχνικές ομαδοποίησης

**΄Γ Μέρος**: Μοντέλα μηχανικής μάθησης

Η εργασία αναπτύχθηκε σε jupyter notebook, με τη χρήση Python και το περιβάλλον προγραμματισμού που χρησιμοποιήθηκε είναι το Visual Studio Code.

## Ά Μέρος – Εξοικείωση με τα δεδομένα

Η εξοικείωση με τα δεδομένα αποτελεί απαραίτητο βήμα σε οποιοδήποτε έργο που σχετίζεται με την Αναλυτική Δεδομένων και τη Μηχανική Μάθηση. Είναι το πρώτο και το βασικότερο βήμα που πρέπει να γίνει προτού ο ερευνητής προβεί σε οποιαδήποτε άλλη ενέργεια, καθώς περιλαμβάνει τις προπαρασκευαστικές εργασίες που πρέπει να γίνουν στο σύνολο δεδομένων, ώστε να «καθαριστεί» από περιττές ή εσφαλμένες πληροφορίες, να κανονικοποιηθούν τα δεδομένα και να γίνει η οπτικοποίηση τους.

## Εκκαθάριση Δεδομένων

Η εκκαθάριση του συνόλου δεδομένων είναι ζωτικής σημασίας κομμάτι, στην διαδικασία της ανάλυσης, καθώς διασφαλίζεται ότι σύνολο δεδομένων είναι ακριβές και ενήμερο.

Οι ενέργειες που έγιναν σε αυτό το στάδιο αναγράφονται παρακάτω και περιγράφονται λεπτομερώς στη συνέχεια:

* Διαγραφή περιττών στηλών
* Διαγραφή null τιμών
* Εκκαθάριση της στήλης Weather\_Condition

## 1.1 Διαγραφή Περιττών Στηλών

Μετά από την λεπτομερή μελέτη των δεδομένων κρίθηκε αναγκαίο να διαγραφούν κάποιες στήλες του συνόλου δεδομένων, εφόσον δεν προσέφεραν χρήσιμη πληροφορία στην εξαγωγή συμπερασμάτων. Οι, εν λόγω, στήλες είναι οι εξής:

* **Nautical\_Twilight, Astronomical\_Twilight, Civil\_Twilight:** Δίνουν πανομοιότυπη πληροφορία με την στήλη Sunrise\_Sunset.
* **End\_Time:** Η συγκεκριμένη στήλη περιέχει την ώρα τέλους του κάθε ατυχήματος, αλλά παρατηρήθηκε πως, σε σχέση με την ώρα έναρξης, η ώρα λήξης είναι περίπου έξι ώρες αργότερα, επομένως τα δεδομένα της στήλης θεωρήθηκαν μη έγκυρα (Εικόνα 1).
* **Number:** Σε αυτή την στήλη περιλαμβάνεται ένας μεγάλος αριθμός null τιμών (Εικόνα 2).
* **Description:** Η, εν λόγω, στήλη περιλαμβάνει μια σύντομη περιγραφή του ατυχήματος. Παρόλα αυτά οι πληροφορίες που δίνονται μπορούν να εξαχθούν από άλλες στήλες στην πλειοψηφία των δεδομένων.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Εικόνα 2: Μη έγκυρη στήλη Number.

Εικόνα 1: Μη έγκυρες στήλες Start\_Time, End\_Time.

Με την διαγραφή των περιττών στηλών οι διαστάσεις του συνόλου δεδομένων μειώθηκαν από (2845342, 47) σε (2845342, 41).

## 1.2 Διαγραφή Null Τιμών

Παρόλο που διαγράφηκε η στήλη Number, οι null τιμές υπάρχουν σε αρκετά σημεία του συνόλου δεδομένων. Η ύπαρξη τους και κατ ’επέκταση η χρήση τους σε επόμενες ενέργειες της διαδικασίας ανάλυσης, ενδεχομένως να επηρεάσουν και να αλλοιώσουν την τελική εξαγωγή συμπερασμάτων. Με βάση, λοιπόν, τα παραπάνω κρίθηκε αναγκαία η διαγραφή τους, η οποία είχε ως αποτέλεσμα την μείωση των διαστάσεων του συνόλου δεδομένων από (2845342, 41) σε (2207325, 41).

## 1.3 Εκκαθάριση της στήλης Weather\_Condition

Chart, pie chart

Description automatically generatedΗ στήλη Weather\_Condition έχει την μεγαλύτερη συσχέτιση με τη στήλη Severity. Επιπλέον, περιέχει πολλές μοναδικές τιμές, οι οποίες δεν έχουν μεγάλη συχνότητα εμφάνισης και περιγράφουν πολύ συγκεκριμένα καιρικά φαινόμενα. Με βάση τα παραπάνω θεωρήθηκε σημαντικό να «καθαριστεί» η στήλη Weather\_Condition, κατηγοριοποιώντας τις τιμές με μικρή συχνότητα εμφάνισης ως τιμές με μεγάλη συχνότητα εμφάνισης. Δηλαδή τα διάφορα καιρικά φαινόμενα ταξινομήθηκαν σε πιο γενικές κατηγορίες, για παράδειγμα η τιμή «Light Rain» κατηγοριοποιήθηκε ως «Rain». Στις εικόνες που ακολουθούν φαίνεται σε μορφή γραφήματος, η στήλη Weather\_Condition πριν και μετά την «εκκαθάριση» (Εικόνες 3,4).

Εικόνα 4: Οι τιμές της στήλης Weather\_Condition μετά την εκκαθάριση.

Εικόνα 3: Οι τιμές της στήλης Weather\_Condition πριν την εκκαθάριση.

Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generatedΠαρόλο που πραγματοποιήθηκε η παραπάνω εκκαθάριση οι τιμές της στήλης Weather\_Condition ακόμα παραμένουν πυκνές σε πλήθος, όπως φαίνεται και στην Εικόνα 4, επομένως θα πρέπει να πραγματοποιηθεί μια επανάληψη της διαδικασίας που ακολουθήθηκε παραπάνω, ώστε ο αριθμός των τιμών της στήλης να περιοριστεί ακόμα περισσότερο. Το αποτέλεσμα της δεύτερης εκκαθάρισης φαίνεται στην εικόνα που ακολουθεί (Εικόνα 5).

Εικόνα 5: Οι τιμές της στήλης Weather\_Condition μετά την δεύτερη εκκαθάριση.

## 2. Οπτικοποίηση Δεδομένων

Οπτικοποίηση των δεδομένων είναι η προβολή της πληροφορίας είτε σε μορφή γραφήματος, οποιουδήποτε είδους, είτε σε μορφή πίνακα. Η τεχνική της οπτικοποίησης στην Αναλυτική Δεδομένων θεωρείται απαραίτητο στάδιο και η σημασία της έγκειται στο ότι οι ερευνητές μπορούν να αφομοιώσουν πιο γρήγορα μεγάλο όγκο οπτικής πληροφορίας και κατ’ επέκταση να εξοικειωθούν σε σύντομο χρονικό διάστημα με τα δεδομένα τους. Τα γραφήματα πίτας στις Εικόνες 3, 4 και 5 αποτελούν παραδείγματα οπτικοποίησης.

## 2.1 Εξαγωγή χρήσιμης πληροφορίας

Ο όγκος του δοθέντος συνόλου δεδομένων είναι τόσο μεγάλος που δίνεται η δυνατότητα να οπτικοποιηθούν πολλές πτυχές του, οι οποίες ίσως να μην ήταν τόσο κατανοητές από τις αριθμητικές τιμές τους. Τα γραφήματα και οι πίνακες οπτικοποίησης φαίνονται στις εικόνες που ακολουθούν Chart, pie chart

Description automatically generated(Εικόνες 6, 7, 8,και 9).

Εικόνα 6: Ατυχήματα ανα πολιτεία.

Chart, pie chart

Description automatically generated

Εικόνα 7: Σοβαρότητα ατυχημάτων.

Chart, map, scatter chart

Description automatically generated

Εικόνα 8: Ο χάρτης των ατυχημάτων βασισμένος στο γεωγραφικό μήκος και πλάτος.

Graphical user interface

Description automatically generated

Εικόνα 9: Περιγραφικά στατιστικά στοιχεία του συνόλου δεδομένων/

## 2.2 Συσχέτιση και Συνδιακύμανση

Η συσχέτιση (correlation) και η συνδιακύμανση (covariance) είναι δύο μαθηματικές έννοιες που χρησιμοποιούνται αρκετά συχνά στη διαδικασία ανάλυσης δεδομένων. Και οι δύο προσδιορίζουν τη σχέση και μετρούν την εξάρτηση μεταξύ δύο τυχαίων μεταβλητών. Παρόλα αυτά, η σημασία τους δεν είναι η ίδια καθώς έχουν βασικές διαφορές. Αρχικά η συσχέτιση υπολογίζει κατά πόσο δύο τυχαίες μεταβλητές συσχετίζονται μεταξύ τους, δηλαδή τον βαθμό εξάρτησής τους. Σε αντίθεση με τη συσχέτιση, η συνδιακύμανση υπολογίζει κατά πόσο δύο τυχαίες μεταβλητές δεν συσχετίζονται μεταξύ τους, δηλαδή τον βαθμό ανεξαρτησίας τους. Είναι πολύ σημαντικό να υπολογιστούν και να οπτικοποιηθούν και οι δυο μαθηματικές έννοιες, καθώς η συνεισφορά τους στην εξαγωγή συμπερασμάτων είναι μεγάλη. Στις εικόνες που ακολουθούν φαίνονται οι πίνακες συσχέτισης και συνδιακύμανσης και τα αντίστοιχα Graphical user interface, website

Description automatically generatedδιαγράμματά τους (Εικόνες 10, 11, 12 και 13).

Εικόνα 10: Πίνακας συσχέτισης

Chart

Description automatically generated

Εικόνα 11: Διάγραμμα συσχέτισης

Chart, bar chart

Description automatically generatedA computer screen capture

Description automatically generated with medium confidence

Εικόνα 13: Διάγραμμα συνδιακύμανσης

Εικόνα 12: Πίνακας συνδιακύμανσης

3. Κλιμάκωση Δεδομένων

Τελευταίο βήμα στην διαδικασία εξοικείωσης με τα δεδομένα είναι η κλιμάκωση τους. Η τεχνικές κλιμάκωσης δεδομένων (scaling) σχετίζονται μόνο με τα αριθμητικά χαρακτηριστικά και ανάγονται στην αναπαράστασή τους στην ίδια κλίμακα. Για το παρόν σύνολο δεδομένων οι τιμές των χαρακτηριστικών θα αναπαρασταθούν στο διάστημα [0,1]. Η διαδικασία που θα ακολουθηθεί είναι η εξής:

1. Αρχικά θα κλιμακωθούν όλα τα αριθμητικά χαρακτηριστικά.
2. Εν συνεχεία, θα επιλεγούν όλα τα λογικά χαρακτηριστικά (Boolean) και οι τιμές True, False θα αντικατασταθούν με τα λογικά 0 και 1.
3. Οι στήλες Sunrise\_Sunset και Side θα δυαδικοποιηθούν, επίσης, θεωρώντας πως η τιμές “Night” και “R” θα αντικατασταθούν με λογικό 1 και η τιμές “Day” και “L” με λογικό 0, αντίστοιχα.
4. Τέλος, όλα τα παραπάνω θα συνενωθούν σε ένα ενιαίο Data Frame.

A screenshot of a computer

Description automatically generated with medium confidenceΣτις εικόνες που ακολουθούν φαίνεται, ανα βήμα, η διαδικασία που περιγράφηκε παραπάνω (Εικόνες 14, 15, 16, 17, 18, 19).

Εικόνα 14: Αριθμητικά χαρακτηριστικά πριν την κλιμάκωση

A screenshot of a computer

Description automatically generated

Εικόνα 15: Αριθμητικά χαρακτηριστικά μετά την κλιμάκωση (Βήμα 1)

Εικόνα 16: Λογικά χαρακτηριστικά πριν την κλιμάκωση

A computer screen capture

Description automatically generated with medium confidence

A computer screen capture

Description automatically generated with medium confidence

Εικόνα 17: Λογικά χαρακτηριστικά μετά την κλιμάκωση (Βήμα 2)

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Εικόνα 18: Δυαδικοποίηση στηλών Sunrise\_Sunset και Side (Βήμα 3)

A screenshot of a computer

Description automatically generated

Εικόνα 19: Συνένωση όλων των παραπάνω σε ένα ενιαίο Data Frame (Βήμα 4)

Τέλος, εφόσον τα δεδομένα έχουν κλιμακωθεί επιτυχώς μπορεί να οπτικοποιηθεί η μεταξύ τους σχέση σε ένα ενιαίο γράφημα, όπως φαίνεται και στην εικόνα που ακολουθεί.

Εικόνα 20: Οπτικοποιημένη σχέση μεταξύ των χαρακτηριστικών

Diagram

Description automatically generated

## ΄Β Μέρος – Τεχνικές Ομαδοποίησης

Η ομαδοποίηση ή συσταδοποίηση (αγγλικά: clustering), αποτελεί τομέας της μηχανικής μάθησης χωρίς επίβλεψη και της εξόρυξης δεδομένων. Πρόκειται για την διαδικασία κατά την οποία ένας αλγόριθμος χωρίζει ένα δοθέν σύνολο δεδομένων σε ομάδες ομοειδών αντικειμένων. Αντικειμενικός στόχος στην συσταδοποίηση είναι το να δημιουργούνται ομάδες που διαχωρίζουν όσο το δυνατόν γίνεται πιο ορθά τα δεδομένα. Για να είναι επιτυχημένη μια τεχνική ομαδοποίησης, πρέπει τα στοιχεία μιας συστάδας να μοιάζουν όσο γίνεται περισσότερο ενώ τα στοιχεία διαφορετικών συστάδων να διαφέρουν όσο γίνεται περισσότερο. Στην παρούσα εργασία θα εφαρμοστούν οι παρακάτω αλγόριθμοι ομαδοποίησης, με τη χρήση του πακέτου Scikit Learn:

* K-means
* DBSCAN
* OPTICS

## K-means

Ο συγκεκριμένος αλγόριθμος είναι από τους πιο πολυεφαρμοσμένους και ανάγεται στην κατηγορία της επίπεδης ομαδοποίησης διότι παράγει ένα σύνολο συστάδων οι οποίες σχετίζονται μεταξύ τους. Αποτελεί έναν από τους διασημότερους αλγόριθμους ομαδοποίησης χάρη στην απλότητα και την ευελιξία του.

## 4.1 Εφαρμογή του K-means

Η εφαρμογή της μεθόδου K-means στο παρόν σύνολο δεδομένων γίνεται με τη χρήση του πακέτου Scikit Learn. Αρχικά, προτού γίνει οποιαδήποτε άλλη ενέργεια είναι απαραίτητο να χωριστεί το σύνολο σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου (training test, test set).

Με την προϋπόθεση ότι αυτό το βήμα έχει πραγματοποιηθεί επιτυχώς, εφαρμόζεται ο αλγόριθμος K-means. Συγκεκριμένα, υπολογίζονται οι πραγματικές ομάδες στις οποίες ανήκουν τα δείγματα στο σύνολο εκπαίδευσης και στη συνέχεια γίνεται μια πρόβλεψη για τις ομάδες στις οποίες θα ανήκουν τα δείγματα από το σύνολο ελέγχου. Τέλος, υπολογίζεται το σκορ ακριβείας του αλγορίθμου, το οποίο είναι 0.08908566024019807.

## 4.2 Οπτικοποίηση του K-means

Graphical user interface

Description automatically generatedΣτις εικόνες που ακολουθούν φαίνονται τα αποτελέσματα από την εφαρμογή του αλγορίθμου K-means με ένα Scatter Plot και έναν πίνακα σύγχυσης (Εικόνες 21 & 22).

Εικόνα 21: K-means Scatter Plot

Chart

Description automatically generated

Εικόνα 22: K-means Confusion Matrix

## DBSCAN

Ο αλγόριθμος DBSCAN (Density-Based Spatial Clustering of Applications with Noise) ανήκει στην κατηγορία των μεθόδων ομαδοποίησης βασισμένες στην πυκνότητα και δημιουργεί τις συστάδες με βάση τον ελάχιστο αριθμό ομάδων και την πυκνότητα. Η πυκνότητα ορίζεται ως το ελάχιστο πλήθος σημείων που απέχουν συγκεκριμένη απόσταση μεταξύ τους. Αξίζει να σημειωθεί πως για τεχνικούς λόγους και, ιδίως, για να είναι πιο γρήγορος ο χρόνος εκτέλεσης, οι διαστάσεις του συνόλου θα μειωθούν από (2101929, 28) σε (10000, 28).

## 5.1 Εφαρμογή του DBSCAN

Εφόσον έχουν μειωθεί οι διαστάσεις του συνόλου δεδομένων είναι απαραίτητο να ακολουθηθεί η διαδικασία διαχωρισμού δεδομένων εκπαίδευσης και ελέγχου για το νέο, μειωμένο σύνολο. Με την προϋπόθεση ότι αυτό το βήμα έχει πραγματοποιηθεί επιτυχώς, εφαρμόζεται ο αλγόριθμος DBSCAN, με τη χρήση του πακέτου Scikit Learn.

Συγκεκριμένα, υπολογίζονται οι πραγματικές ομάδες στις οποίες ανήκουν τα δείγματα στο σύνολο εκπαίδευσης και στη συνέχεια γίνεται μια πρόβλεψη για τις ομάδες στις οποίες θα ανήκουν τα δείγματα από το σύνολο ελέγχου. Τέλος, υπολογίζεται το σκορ ακριβείας του αλγορίθμου, το οποίο είναι 0.3852.

## 5.2 Οπτικοποίηση του DBSCAN

Στις εικόνες που ακολουθούν φαίνονται τα αποτελέσματα από την εφαρμογή του αλγορίθμου K-means με ένα Scatter Plot και έναν πίνακα σύγχυσης (Εικόνες 23 & 24).

Chart

Description automatically generated

Εικόνα 23: DBSCAN Scatter Plot

A screenshot of a computer

Description automatically generated with low confidence

Εικόνα 24: DBSCAN Confusion Matrix

## OPTICS

Ο αλγόριθμος OPTICS (Ordering Points To Identify Cluster Structure) είναι σχεδόν πανομοιότυπος με τον αλγόριθμο ομαδοποίησης DBSCAN. Προσθέτει, όμως, δύο ακόμη όρους:

* Απόσταση Πυρήνα (Core Distance): πρόκειται για την ελάχιστη τιμή της ακτίνας, που απαιτείται ούτως ώστε ένα σημείο να ταξινομηθεί ως σημείο πυρήνα. Αν το σημείο δεν είναι σημείο πυρήνα τότε η απόσταση πυρήνα του είναι απροσδιόριστη.
* Απόσταση Προσβασιμότητας (Reachability Distance): Ορίζεται σε σχέση με ένα άλλο σημείο q. Η απόσταση προσβασιμότητας μεταξύ ενός σημείου p και ενός σημείου q είναι η μέγιστη απόσταση πυρήνα του p και η Ευκλείδεια απόσταση (ή κάποια άλλη μέτρηση απόστασης) μεταξύ p και q. Η απόσταση προσβασιμότητας δεν ορίζεται εάν το q δεν είναι σημείο πυρήνα.

Αξίζει να σημειωθεί πως και για την εκτέλεση του OPTICS θα χρησιμοποιηθούν τα δεδομένα εκπαίδευσης και ελέγχου από το μειωμένο σύνολο δεδομένων.

## 6.1 Εφαρμογή του OPTICS

Η εφαρμογή της μεθόδου OPTICS στο παρόν σύνολο δεδομένων γίνεται με τη χρήση του πακέτου Scikit Learn. Συγκεκριμένα, υπολογίζονται οι πραγματικές ομάδες στις οποίες ανήκουν τα δείγματα στο σύνολο εκπαίδευσης και στη συνέχεια γίνεται μια πρόβλεψη για τις ομάδες στις οποίες θα ανήκουν τα δείγματα από το σύνολο ελέγχου. Τέλος, υπολογίζεται το σκορ ακριβείας του αλγορίθμου, το οποίο είναι 0.0064.

## 6.2 Οπτικοποίηση του OPTICS

Στις εικόνες που ακολουθούν φαίνονται τα αποτελέσματα από την εφαρμογή του αλγορίθμου OPTICS με ένα Scatter Plot και έναν πίνακα σύγχυσης (Εικόνες 25 & 26).

Chart, box and whisker chart

Description automatically generated

Εικόνα 25: OPTICS Scatter Plot

A picture containing text, tiled, tile

Description automatically generated

Εικόνα 26: OPTICS Confusion Matrix

## Σύγκριση Τεχνικών Ομαδοποίησης

Όπως αναφέρθηκε και σε προηγούμενες παραγράφους, η ομαδοποίηση είναι μια τεχνική στη μηχανική μάθηση (unsupervised machine learning) χωρίς επίβλεψη, η οποία ομαδοποιεί τα σημεία δεδομένων σε μικρά συμπλέγματα, με βάση την ομοιότητα των διαθέσιμων πληροφοριών για τα χαρακτηριστικά του συνόλου δεδομένων. Στην παρούσα ερευνητική εργασία εφαρμόσθηκαν και αναλύθηκαν εκτενώς οι αλγόριθμοι K-means και DBSCAN, οι οποίοι αποτελούν δύο από τους πιο δημοφιλείς αλγόριθμους ομαδοποίησης, όπως επίσης και ο αλγόριθμος OPTICS, ο οποίος προσομοιάζει κατά πολύ στον DBSCAN. Στον πίνακα που ακολουθεί, αναγράφονται κάποιες παρατηρήσεις που έγιναν κατά την εφαρμογή των αλγορίθμων και κάνουν αισθητές τις διαφορές μεταξύ τους.

Πίνακας 1: Σύγκριση Τεχνικών Ομαδοποίησης

|  |  |  |
| --- | --- | --- |
| K-means | DBSCAN | OPTICS |
| * Ο αριθμός των ομάδων καθορίζονται από τον ερευνητή. | * Ο αριθμός των ομάδων δεν χρειάζεται να καθοριστεί. | * Ο αριθμός των ομάδων δεν χρειάζεται να καθοριστεί. |
| * Απαιτεί μια βασική παράμετρο, τον αριθμό ομάδων. | * Απαιτεί δύο βασικές παραμέτρους, την ακτίνα και τον ελάχιστον αριθμό σημείων. | * Απαιτεί δύο βασικές παραμέτρους, την ακτίνα και τον ελάχιστον αριθμό σημείων. |
| * Η πυκνότητα των σημείων δεν επηρεάζει την διαδικασία ομαδοποίησης. | * Η πυκνότητα των σημείων επηρεάζει την διαδικασία ομαδοποίησης. | * Η πυκνότητα των σημείων επηρεάζει την διαδικασία ομαδοποίησης. |
| * Δεν εκτελείται σωστά σε ένα θορυβώδες σύνολο δεδομένων. | * Ορθός χειρισμός θορυβωδών συνόλων δεδομένων. | * Ορθός χειρισμός θορυβωδών συνόλων δεδομένων. |
| * Έχει φυσιολογικές απαιτήσεις μνήμης. | * Έχει φυσιολογικές απαιτήσεις μνήμης. | * Έχει μεγάλες απαιτήσεις μνήμης. |
| * Πολυπλοκότητα: O(n2) | * Πολυπλοκότητα: O(nlogn) | * Πολυπλοκότητα: O(nlogn) + O(n) |
| * Ακρίβεια στο σύνολο US Accidents: 0.29830270436912326 | * Ακρίβεια στο σύνολο US Accidents: 0.2236 | * Ακρίβεια στο σύνολο US Accidents: 0.0064 |

## ΄Γ Μέρος – Μοντέλα Μηχανικής Μάθησης

Η Μηχανική Μάθηση (Machine Learning) αποτελεί ίσως τον πιο ραγδαία αναπτυσσόμενο τομέα της Τεχνητής Νοημοσύνης καθώς τα τελευταία χρόνια, ειδικά μετά την έλευση της Βαθιάς Μάθησης (Deep Learning), έχει προσφέρει πληθώρα μεθόδων με πολύ καλά έως εντυπωσιακά αποτελέσματα σε όλες σχεδόν τις εφαρμογές που απαιτούν ευφυΐα. Ως μηχανική μάθηση ορίζεται η δυνατότητα της «μηχανής» να εκπαιδευτεί, δηλαδή να μάθει, μέσα από μια διαδικασία εκπαίδευσης.Το πεδίο της μηχανικής μάθησης, διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, όπως θα γίνει και στην παρούσα εργασία, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα.

Με βάση το σύνολο δεδομένων US Accidents, θα γίνει μια απόπειρα εκτίμησης της σοβαρότητας ενός ατυχήματος με την εφαρμογή τριών αλγορίθμων μηχανικής μάθησης, οι οποίοι αναγράφονται παρακάτω:

* Απλή Γραμμική Παλινδρόμηση (Basic Linear Regression)
* Νευρωνικά Δίκτυα (Neural Networks)
* Something else (something else)

Αξίζει να σημειωθεί πως για όλο το ΄Γ Μέρος, έχει χρησιμοποιηθεί το πακέτο Tensorflow / Keras. Επιπλέον, υπενθυμίζεται πως οι διαστάσεις του συνόλου δεδομένων έχουν μειωθεί για τεχνικούς λόγους.

## Απλή Γραμμική Παλινδρόμηση

Ο κλάδος της Στατιστικής που εξετάζει τη σχέση μεταξύ δύο ή περισσότερων μεταβλητών με απώτερο σκοπό την πρόβλεψη μιας από αυτές μέσω των άλλων χαρακτηρίζεται με την ονομασία ανάλυση παλινδρόμησης (regression analysis). Η απλούστερη περίπτωση παλινδρόμησης είναι η απλή γραμμική παλινδρόμηση (simple linear regression), κατά την οποία υπάρχει μόνο μια ανεξάρτητη μεταβλητή X (και η εξαρτημένη μεταβλητή Y η οποία μπορεί να προσεγγιστεί ικανοποιητικά από μία γραμμική συνάρτηση του X. Στην προκειμένη περίπτωση θα γίνει εκτίμηση της σοβαρότητας ενός ατυχήματος (εξαρτημένη μεταβλητή), βάσει της ορατότητας (ανεξάρτητη μεταβλητή).

## 8.1 Επιθεώρηση και Προετοιμασία Δεδομένων

Graphical user interface

Description automatically generated with medium confidenceΠρώτο βήμα πριν τη δημιουργία του μοντέλου της απλής γραμμικής παλινδρόμησης είναι να χωριστούν τα δεδομένα σε σύνολα δεδομένων εκπαίδευσης και ελέγχου. Η διαδικασία του διαχωρισμού επαναλαμβάνεται, καθώς οι διαστάσεις των δύο υποσυνόλων πρέπει να ανταποκρίνονται στις απαιτήσεις του μοντέλου, επομένως το σύνολο δεδομένων εκπαίδευσης περιλαμβάνει το 80% του αρχικού συνόλου, ενώ το σύνολο δεδομένων ελέγχου περιλαμβάνει ένα ποσοστό των 20%. Στη συνέχεια, τα υποσύνολα διαχωρίζονται ξανά σε Χαρακτηριστικά - Features (ανεξάρτητες μεταβλητές) και Ετικέτες - Labels (εξαρτημένη μεταβλητή). Στις εικόνες που ακολουθούν φαίνονται κάποια στατιστικά στοιχεία για τα σύνολα δεδομένων εκπαίδευσης και ελέγχου (Εικόνες 27, 28).

Εικόνα 27: Στατιστικά στοιχεία για το σύνολο εκπαίδευσης

Graphical user interface, application

Description automatically generated

Εικόνα 28: Στατιστικά στοιχεία για το σύνολο ελέγχου

## 8.2 Κανονικοποίηση

Graphical user interface, text, application

Description automatically generatedΑρχικά, υποβάλλουμε το σύνολο των δεδομένων εκπαίδευσης σε κανονικοποίηση ώστε η διαδικασία εκμάθησης να είναι πιο σταθερή. Στην εικόνα που ακολουθεί φαίνεται το αποτέλεσμα της κανονικοποίησης για το πρώτο επίπεδο (Εικόνα 28).

Εικόνα 29: Κανονικοποίηση πρώτου επιπέδου

## 8.3 Γραμμική Παλινδρόμηση

Μετά την προετοιμασία και την κανονικοποίηση των δεδομένων είναι πλέον εφικτό να δημιουργηθεί το μοντέλο της απλής γραμμικής παλινδρόμησης, μέσα από το οποίο θα γίνει εκτίμηση της σοβαρότητας ενός ατυχήματος από το χαρακτηριστικό της ορατότητας. Για την εκμάθηση του μοντέλου με το εργαλείο Tensorflow / Keras, χρειάζεται πρώτα να οριστεί η αρχιτεκτονική του μοντέλου. Για την παρούσα ερευνητική εργασία επιλεχθεί το Sequential μοντέλο.

Στη συνέχεια για τη δημιουργία του μοντέλου θα ακολουθηθούν τα παρακάτω βήματα:

1. Κανονικοποίηση του χαρακτηριστικού “Visibility(mi)”.
2. Εφαρμογή του γραμμικού μετασχηματισμού y = ax + b.

Graphical user interface, text, application

Description automatically generated

Εικόνα 30: Sequential μοντέλο

Graphical user interface, text

Description automatically generatedΜετά την δημιουργία του μοντέλου, ακολουθεί το στάδιο της εκπαίδευσης του αντίστοιχου συνόλου. Τα στατιστικά της διαδικασίας εκμάθησης φαίνονται στην Εικόνες 31 και 32.

Εικόνα 31: Στατιστικά διαδικασίας εκπαίδευσης

Graphical user interface

Description automatically generated with medium confidence

Εικόνα 32: Στατιστικά διαδικασίας εκπαίδευσης

Στη συνέχεια, συλλέγονται τα αποτελέσματα της εκμάθησης, όπως επίσης οι προβλέψεις του μοντέλου συναρτήσει της εισόδου. Στις εικόνες που ακολουθούν φαίνονται, οπτικοποιημένες οι προβλέψεις του μοντέλου σε ένα Scatter Plot και σε έναν πίνακα σύγχυσης (Εικόνες 33 & 34).

Chart, scatter chart

Description automatically generated

Εικόνα 33: Simple Linear Regression Scatter Plot

Chart

Description automatically generated

Εικόνα 34: Simple Linear Regression Confusion Matrix

Τέλος, το σκορ του μοντέλου απλής γραμμικής παλινδρόμησης είναι 0.1325664520263672.