Αναλυτική Δεδομένων - US Accidents Dataset

Τεχνική Αναφορά της εργασίας του μαθήματος «Αναλυτική Δεδομένων».

Ακαδημαϊκό Έτος: 2021 – 2022

Ομάδα:

Μπουμπλίνη Αναστασία (Π19117)
aboublini@gmail.com
ANASTASIA BOUBLINI (p19117@unipi.gr)

Μπριστογιάννης Ιωακείμ (Π19048)
ioakeim13@hotmail.gr
IOAKEIM EL-KHATTAB-BRISTOGIANNIS (p19048@unipi.gr)

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

Πίνακας Περιεχομένων

Πρόλογος

Η παρούσα εργασία ανάγεται στην μελέτη και την ανάλυση του συνόλου δεδομένων "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.

1. 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:

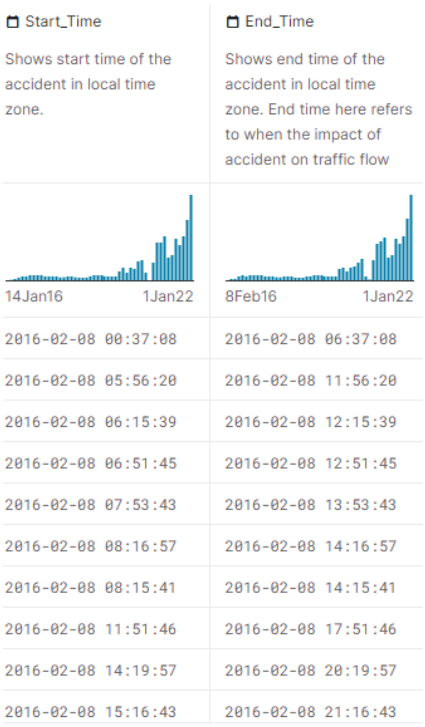
- Διαγραφή περιττών στηλών
- Διαγραφή null τιμών
- Εκκαθάριση της στήλης Weather_Condition

1.1 Διαγραφή Περιττών Στηλών

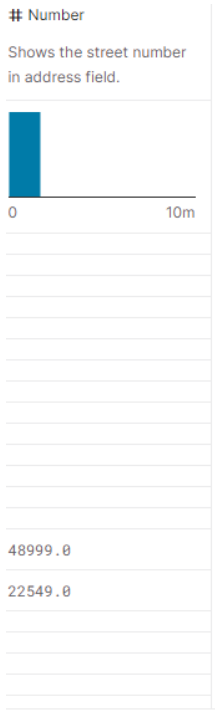
Μετά από την λεπτομερή μελέτη των δεδομένων κρίθηκε αναγκαίο να διαγραφούν κάποιες στήλες του συνόλου δεδομένων, εφόσον δεν προσέφεραν χρήσιμη πληροφορία στην εξαγωγή συμπερασμάτων. Οι, εν λόγω, στήλες είναι οι εξής:

- 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.



Εικόνα 1: Μη έγκυρες στήλες Start_Time, End_Time.



Εικόνα 2: Μη έγκυρη στήλη Number.

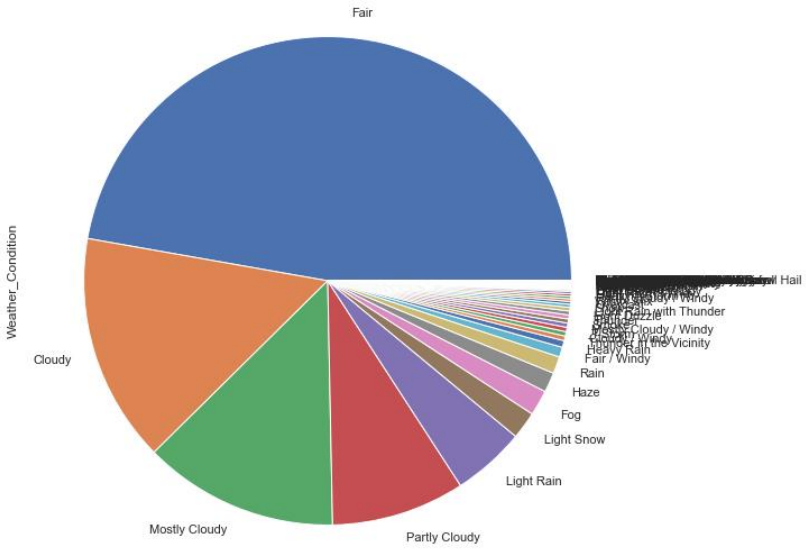
Με την διαγραφή των περιττών στηλών οι διαστάσεις του συνόλου δεδομένων μειώθηκαν από (2845342, 47) σε (2845342, 41).

1.2 Διαγραφή Null Τιμών

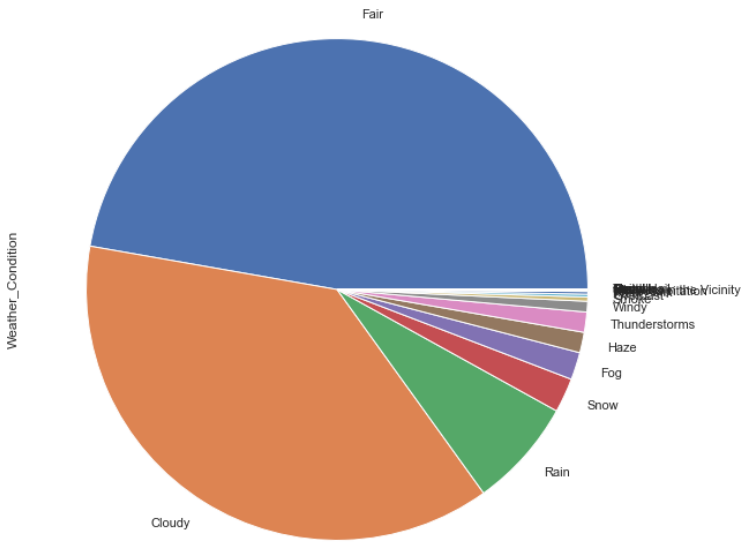
Παρόλο που διαγράφηκε η στήλη Number, οι null τιμές υπάρχουν σε αρκετά σημεία του συνόλου δεδομένων. Η ύπαρξη τους και κατ 'επέκταση η χρήση τους σε επόμενες ενέργειες της διαδικασίας ανάλυσης, ενδεχομένως να επηρεάσουν και να αλλοιώσουν την τελική εξαγωγή συμπερασμάτων. Με βάση, λοιπόν, τα παραπάνω 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

Η 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).

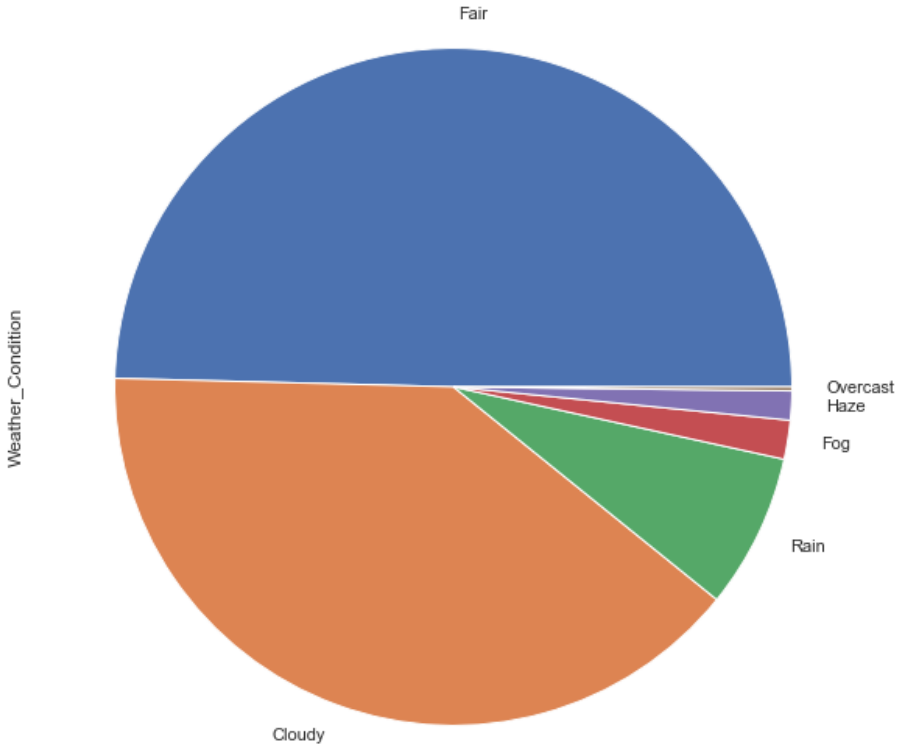


Εικόνα 3: Οι τιμές της στήλης Weather_Condition πριν την εκκαθάριση.



Εικόνα 4: Οι τιμές της στήλης Weather_Condition μετά την εκκαθάριση.

Παρόλο 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.

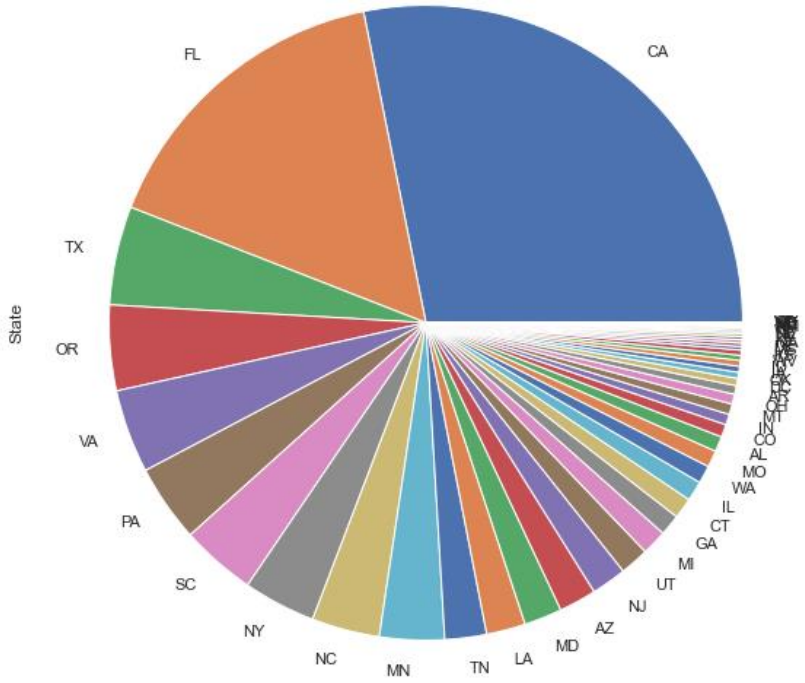
2. Οπτικοποίηση Δεδομένων

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.

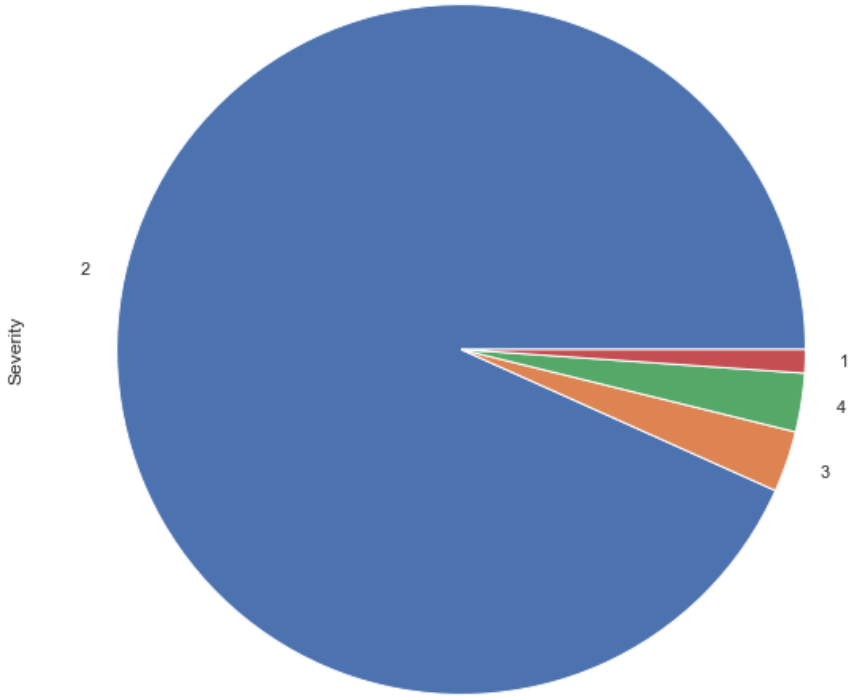
2.1 Εξαγωγή χρήσιμης πληροφορίας

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

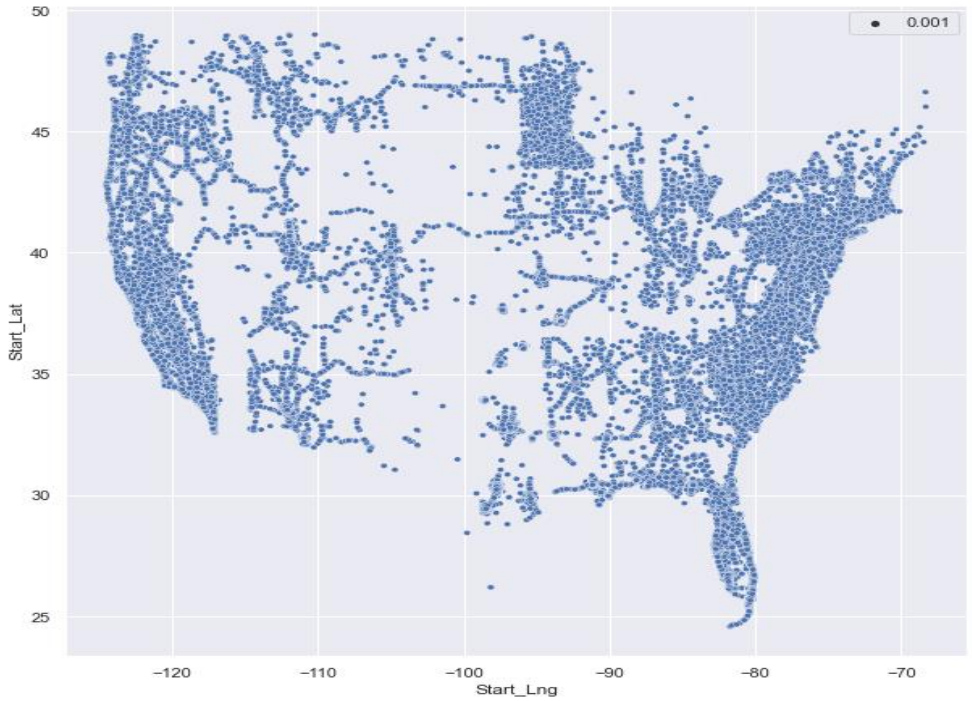
τιμές τους. Τα γραφήματα και οι πίνακες οπτικοποίησης φαίνονται στις εικόνες που ακολουθούν (Εικόνες 6, 7, 8,και 9).



Εικόνα 6: Ατυχήματα ανα πολιτεία.



Εικόνα 7: Σοβαρότητα ατυχημάτων.



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

	Severity	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)
count	2.101929e+06	2.101929e+06	2.101929e+06	2.101929e+06	2.101929e+06	2.101929e+06	2.101929e+06	2.101929e+06	2.101929e+06	2.101929e+06	2.101929e+06
mean	2.072569e+00	3.587314e+01	-9.672859e+01	3.587328e+01	-9.672832e+01	6.776976e-01	6.255000e+01	6.162088e+01	6.416557e+01	2.940105e+01	9.231672e+00
std	3.795681e-01	5.393932e+00	1.834432e+01	5.394036e+00	1.834408e+01	1.438928e+00	1.779366e+01	1.941785e+01	2.261024e+01	1.046924e+00	2.366669e+00
min	1.000000e+00	2.456603e+01	-1.245481e+02	2.456601e+01	-1.245457e+02	0.000000e+00	-3.300000e+01	-5.010000e+01	1.000000e+00	1.672000e+01	0.000000e+00
25%	2.000000e+00	3.292400e+01	-1.180044e+02	3.292392e+01	-1.180037e+02	4.500000e-02	5.000000e+01	5.000000e+01	4.800000e+01	2.923000e+01	1.000000e+01
50%	2.000000e+00	3.539927e+01	-9.103596e+01	3.539954e+01	-9.103523e+01	2.040000e-01	6.400000e+01	6.400000e+01	6.600000e+01	2.975000e+01	1.000000e+01
75%	2.000000e+00	3.991150e+01	-8.033532e+01	3.991000e+01	-8.033562e+01	7.660000e-01	7.600000e+01	7.600000e+01	8.300000e+01	2.997000e+01	1.000000e+01
max	4.000000e+00	4.900058e+01	-6.748413e+01	4.907500e+01	-6.748413e+01	1.551860e+02	1.960000e+02	1.960000e+02	1.000000e+02	5.890000e+01	1.000000e+02

Εικόνα 9: Περιγραφικά στατιστικά στοιχεία του συνόλου δεδομένων/

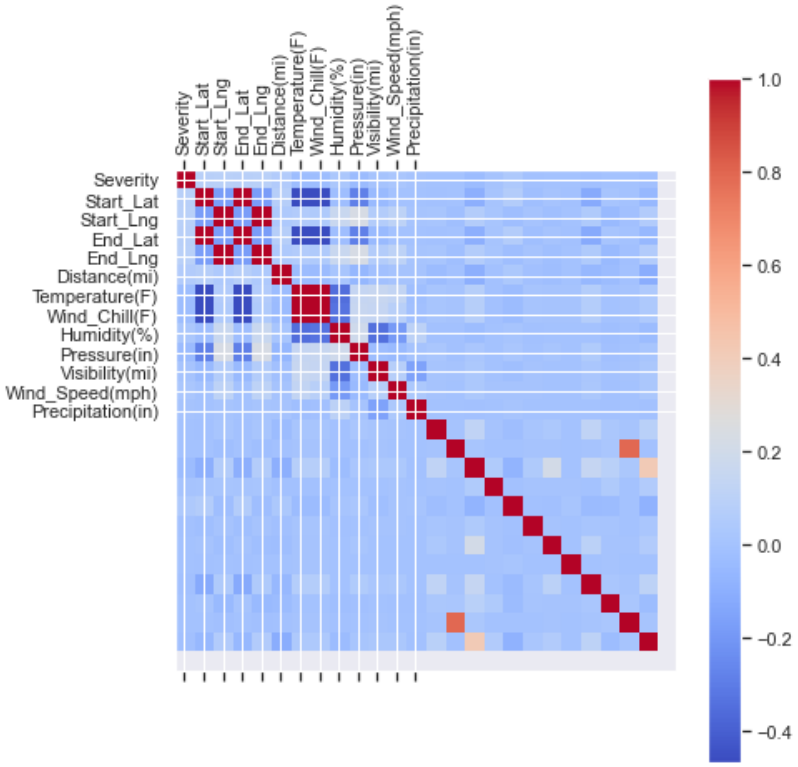
2.2 Συσχέτιση και Συνδιακύμανση

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 διαγράμματά τους (Εικόνες 10, 11, 12 και 13).

	Severity	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	...	Give_Way	Junction
Severity	1.000000	0.084747	0.090828	0.084750	0.090831	0.059298	-0.029568	-0.032065	0.028896	-0.042994	...	0.006578	0.040396
Start_Lat	0.084747	1.000000	-0.180964	0.999996	-0.180960	0.075544	-0.462861	-0.468223	-0.024004	-0.290402	...	0.012208	0.061543
Start_Lng	0.090828	-0.180964	1.000000	-0.180961	0.999999	0.038353	0.058058	0.043426	0.162430	0.225997	...	0.016832	-0.026095
End_Lat	0.084750	0.999996	-0.180961	1.000000	-0.180957	0.075529	-0.462856	-0.468218	-0.024009	-0.290404	...	0.012194	0.061529
End_Lng	0.090831	-0.180960	0.999999	-0.180957	1.000000	0.038356	0.058062	0.043429	0.162425	0.225999	...	0.016831	-0.026090
Distance(mi)	0.059298	0.075544	0.038353	0.075529	0.038356	1.000000	-0.037084	-0.040191	0.016258	-0.066651	...	-0.007292	0.031190
Temperature(F)	-0.029568	-0.462861	0.058058	-0.462856	0.058062	-0.037084	1.000000	0.994450	-0.353158	0.144717	...	-0.008291	-0.034139
Wind_Chill(F)	-0.032065	-0.468223	0.043426	-0.468218	0.043429	-0.040191	0.994450	1.000000	-0.336111	0.153995	...	-0.008460	-0.034099
Humidity(%)	0.028896	-0.024004	0.162430	-0.024009	0.162425	0.016258	-0.353158	-0.336111	1.000000	0.187201	...	-0.000576	0.008684
Pressure(in)	-0.042994	-0.290402	0.225997	-0.290404	0.225999	-0.066651	0.144717	0.153995	0.187201	1.000000	...	-0.002578	0.021621
Visibility(mi)	0.002591	-0.047017	0.062383	-0.047009	0.062386	-0.010259	0.152016	0.143983	-0.347146	-0.032033	...	0.001418	-0.016885
Wind_Speed(mph)	0.021898	-0.001362	0.108915	-0.001359	0.108919	0.002098	0.133145	0.085398	-0.187341	-0.043941	...	0.003092	0.002425
Precipitation(in)	0.004597	0.001459	0.012860	0.001459	0.012859	0.003392	-0.015199	-0.014665	0.102115	0.006371	...	-0.001357	0.010596
Amenity	-0.003593	-0.006963	0.016177	-0.006964	0.016177	-0.036146	0.013639	0.013886	-0.005183	0.020693	...	0.002145	-0.025661
Bump	-0.002299	0.000140	-0.016052	0.000140	-0.016052	-0.006036	0.004166	0.004384	-0.008623	-0.004522	...	-0.000080	-0.000002
Crossing	-0.036036	-0.100479	0.059682	-0.100486	0.059678	-0.100693	0.072712	0.071275	-0.031312	0.021725	...	0.052891	-0.076618

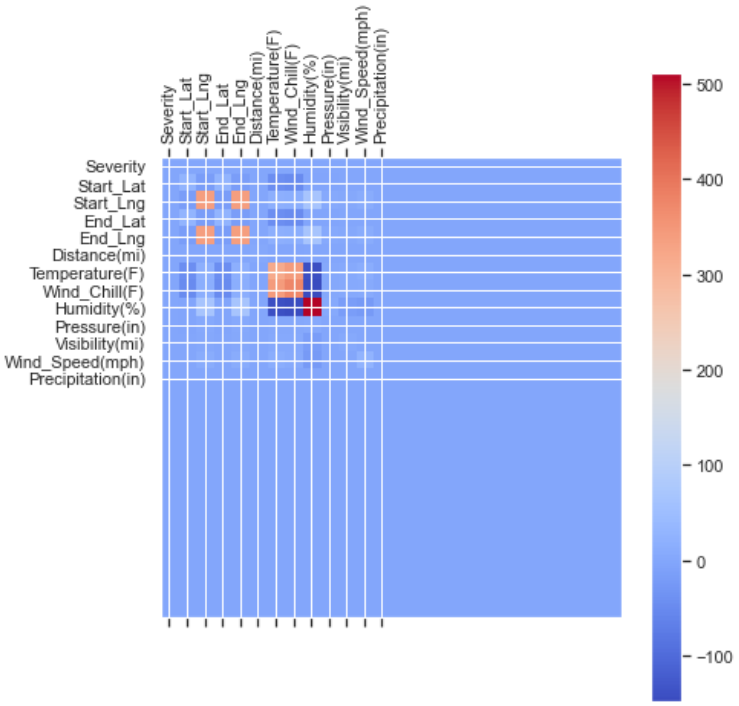
Εικόνα 10: Πίνακας συσχέτισης



Εικόνα 11: Διάγραμμα συσχέτισης

	Severity	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	...	Give_Way
Severity	0.144072	0.173508	0.632425	0.173516	0.632439	0.032387	-0.199698	-0.236332	0.247991	-0.017085	...	1.229311e-04
Start_Lat	0.173508	29.094501	-17.906050	29.094941	-17.905388	0.586332	-44.424350	-49.040974	-2.927446	-1.639912	...	3.241907e-03
Start_Lng	0.632425	-17.906050	336.514002	-17.906064	336.509337	1.012382	18.950953	15.468511	67.370953	4.340297	...	1.520093e-02
End_Lat	0.173516	29.094941	-17.906064	29.095626	-17.905425	0.586228	-44.424736	-49.041479	-2.928178	-1.639956	...	3.238292e-03
End_Lng	0.632439	-17.905388	336.509337	-17.905425	336.505147	1.012440	18.951780	15.469420	67.368143	4.340276	...	1.520067e-02
Distance(mi)	0.032387	0.586332	1.012382	0.586228	1.012440	2.070513	-0.949483	-1.122969	0.528938	-0.100406	...	-5.165785e-04
Temperature(F)	-0.199698	-44.424350	18.950953	-44.424736	18.951780	-0.949483	316.614429	343.597011	-142.082079	2.695877	...	-7.263451e-03
Wind_Chill(F)	-0.236332	-49.040974	15.468511	-49.041479	15.469420	-1.122969	343.597011	377.053068	-147.566908	3.130560	...	-8.087112e-03
Humidity(%)	0.247991	-2.927446	67.370953	-2.928178	67.368143	0.528938	-142.082079	-147.566908	511.222944	4.431285	...	-6.411079e-04
Pressure(in)	-0.017085	-1.639912	4.340297	-1.639956	4.340276	-0.100406	2.695877	3.130560	4.431285	1.096051	...	-1.328945e-04
Visibility(mi)	0.002328	-0.600203	2.708360	-0.600114	2.708459	-0.034938	6.401647	6.616832	-18.576134	-0.079368	...	1.652634e-04
Wind_Speed(mph)	0.043145	-0.038145	10.371227	-0.038065	10.371485	0.015673	12.297864	8.607757	-21.987691	-0.238796	...	7.902517e-04
Precipitation(in)	0.000086	0.000387	0.011606	0.000387	0.011605	0.000240	-0.013305	-0.014010	0.113591	0.000328	...	-3.286360e-06
Amenity	-0.000141	-0.003880	0.030651	-0.003880	0.030651	-0.005372	0.025068	0.027851	-0.012104	0.002238	...	1.090510e-05
Bump	-0.000018	0.000016	-0.006068	0.000016	-0.006068	-0.000179	0.001528	0.001754	-0.004018	-0.000098	...	-8.073442e-08
Crossing	-0.003674	-0.145585	0.294088	-0.145597	0.294067	-0.038920	0.347539	0.371765	-0.190172	0.006110	...	6.994564e-04

Εικόνα 12: Πίνακας συνδιακύμανσης



Εικόνα 13: Διάγραμμα 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.

Στις εικόνες που ακολουθούν φαίνεται, ανα βήμα, η διαδικασία που περιγράφηκε παραπάνω (Εικόνες 14, 15, 16, 17, 18, 19).

	Severity	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)
0	3	40.108910	-83.092860	40.112060	-83.031870	3.230	42.1	36.1	58.0	29.76	10.0	10.4
4	3	39.172393	-84.492792	39.170476	-84.501798	0.500	37.0	29.8	93.0	29.69	10.0	10.4
56	2	38.178100	-85.719460	38.181040	-85.721160	0.223	17.1	5.8	68.0	30.12	10.0	9.2
57	2	38.185770	-85.806780	38.206480	-85.827850	1.832	17.1	5.8	68.0	30.12	10.0	9.2
58	3	38.271910	-85.808380	38.271910	-85.808380	0.000	17.1	5.8	68.0	30.12	10.0	9.2

Εικόνα 14: Αριθμητικά χαρακτηριστικά πριν την κλιμάκωση

	Severity	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)
0	0.666667	0.636103	0.726469	0.634300	0.727527	0.020814	0.327948	0.350264	0.575758	0.309151	0.10	0.009568
1	0.666667	0.597775	0.701937	0.595882	0.701767	0.003222	0.305677	0.324665	0.929293	0.307492	0.10	0.009568
2	0.333333	0.557083	0.680440	0.555512	0.680398	0.001437	0.218777	0.227143	0.676768	0.317686	0.10	0.008464
3	0.333333	0.557397	0.678910	0.556550	0.678528	0.011805	0.218777	0.227143	0.676768	0.317686	0.10	0.008464
4	0.666667	0.560922	0.678882	0.559219	0.678869	0.000000	0.218777	0.227143	0.676768	0.317686	0.10	0.008464
...	...	...	...	...	...	...	...	...	...	...	...	...
2101924	0.333333	0.386193	0.125626	0.384874	0.125738	0.003499	0.519651	0.553027	0.393939	0.289237	0.10	0.011960
2101925	0.333333	0.335629	0.129679	0.334552	0.129546	0.002178	0.449782	0.488013	0.727273	0.300379	0.10	0.005520
2101926	0.333333	0.376902	0.117417	0.375837	0.117215	0.003615	0.462882	0.500203	0.636364	0.308677	0.10	0.009200
2101927	0.333333	0.385783	0.107687	0.384230	0.107780	0.004975	0.454148	0.492076	0.808081	0.305832	0.10	0.007360
2101928	0.333333	0.391573	0.128227	0.390524	0.128044	0.003460	0.489083	0.524584	0.464646	0.282361	0.07	0.006440

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

	Amenity	Bump	Crossing	Give_Way	Junction	No_Exit	Railway	Roundabout	Station	Stop	Traffic_Calming	Traffic_Signal	Turning_Loop
0	False	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False	False
56	False	False	False	False	False	False	False	False	False	False	False	False	False
57	False	False	False	False	False	False	False	False	False	False	False	False	False
58	False	False	False	False	False	False	False	False	False	False	False	False	False

Εικόνα 16: Λογικά χαρακτηριστικά πριν την κλιμάκωση

	Amenity	Bump	Crossing	Give_Way	Junction	No_Exit	Railway	Roundabout	Station	Stop	Traffic_Calming	Traffic_Signal	Turning_Loop
0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0
56	0	0	0	0	0	0	0	0	0	0	0	0	0
57	0	0	0	0	0	0	0	0	0	0	0	0	0
58	0	0	0	0	0	0	0	0	0	0	0	0	0

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

	Sunrise_Sunset
0	1
1	0
2	0
3	0
4	0

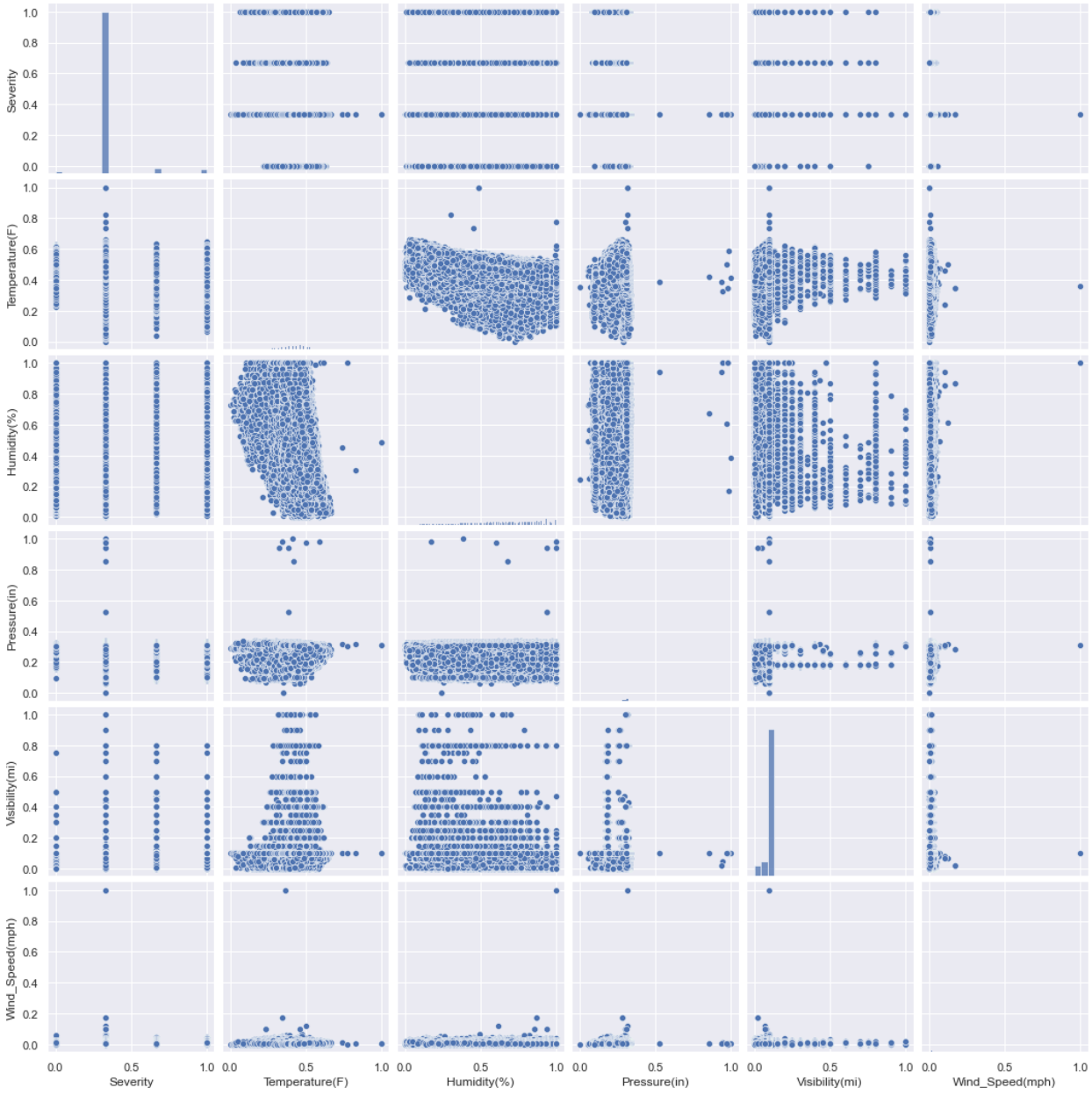
	Side
0	0
1	0
2	0
3	0
4	0

Εικόνα 18: Δυαδικοποίηση στηλών Sunrise_Sunset και Side (Βήμα 3)

Start_Lng	End_Lat	End_Lng	Distance(mi)	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	...	Give_Way	Junction	No_Exit	Railway	Roundabout	Station	Stop
0.726469	0.634300	0.727527	0.020814	0.327948	0.350264	0.575758	0.309151	...	0	0	0	0	0	0	0
0.701937	0.595882	0.701767	0.003222	0.305677	0.324665	0.929293	0.307492	...	0	0	0	0	0	0	0
0.680440	0.555512	0.680398	0.001437	0.218777	0.227143	0.676768	0.317686	...	0	0	0	0	0	0	0
0.678910	0.556550	0.678528	0.011805	0.218777	0.227143	0.676768	0.317686	...	0	0	0	0	0	0	0
0.678882	0.559219	0.678869	0.000000	0.218777	0.227143	0.676768	0.317686	...	0	0	0	0	0	0	0

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

΄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

4. 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.

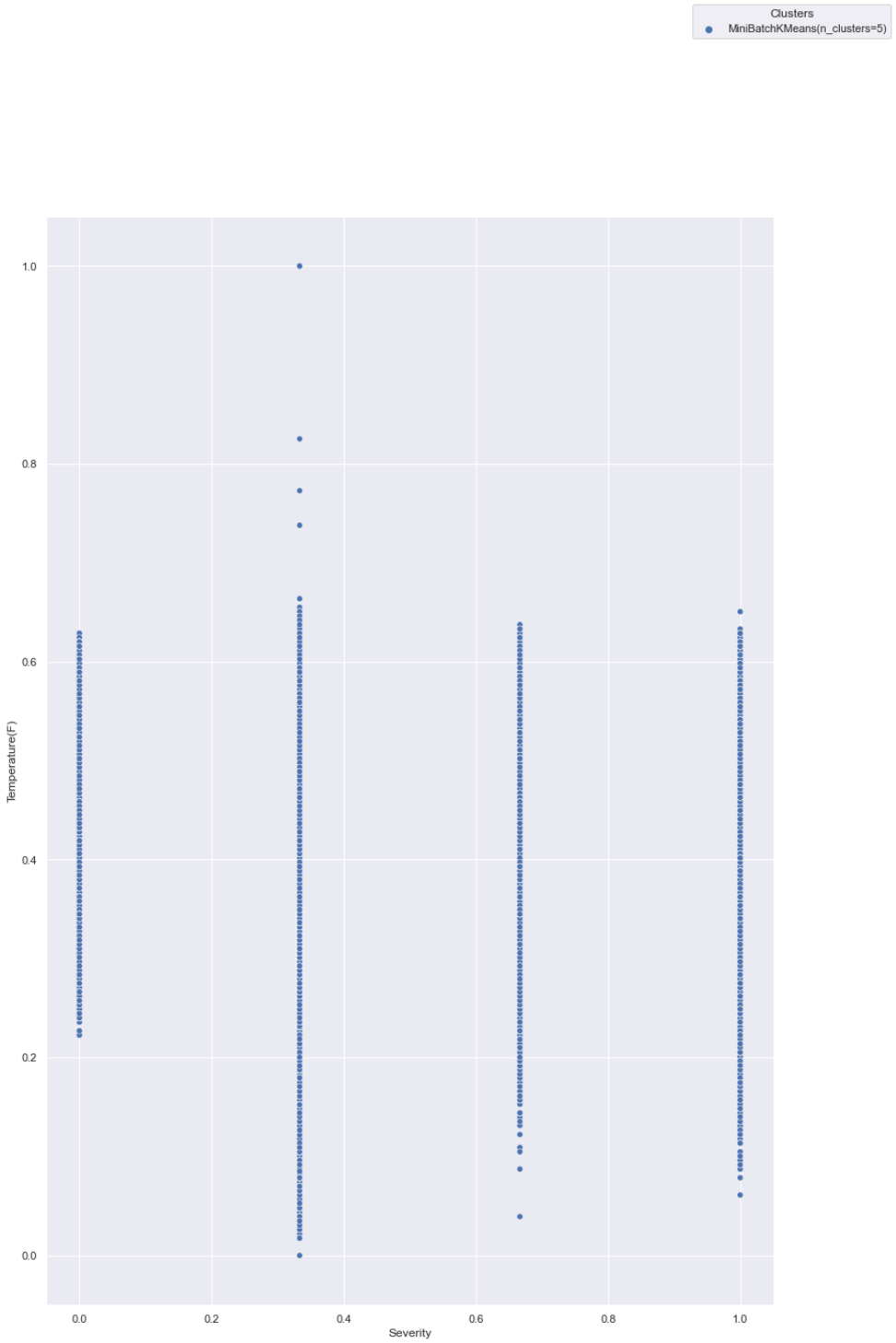
4.1 Εφαρμογή του K-means

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).

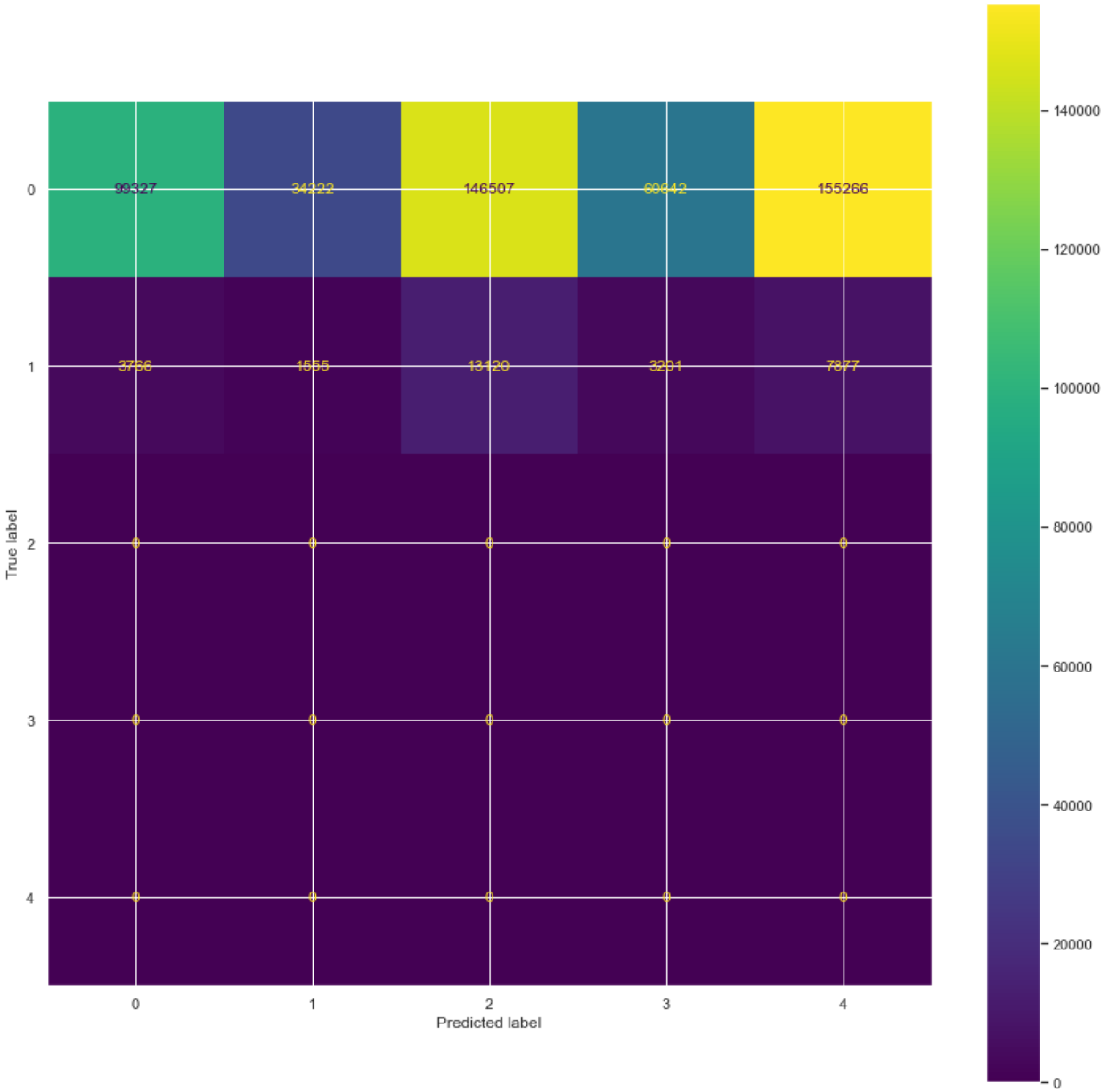
Με την προϋπόθεση ότι αυτό το βήμα έχει πραγματοποιηθεί επιτυχώς, εφαρμόζεται ο αλγόριθμος K-means. Συγκεκριμένα, υπολογίζονται οι πραγματικές ομάδες στις οποίες ανήκουν τα δείγματα στο σύνολο εκπαίδευσης και στη συνέχεια γίνεται μια πρόβλεψη για τις ομάδες στις οποίες θα ανήκουν τα δείγματα από το σύνολο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

4.2 Οπτικοποίηση του K-means

Στις 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



Εικόνα 22: K-means Confusion Matrix

5. 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 οποία 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).

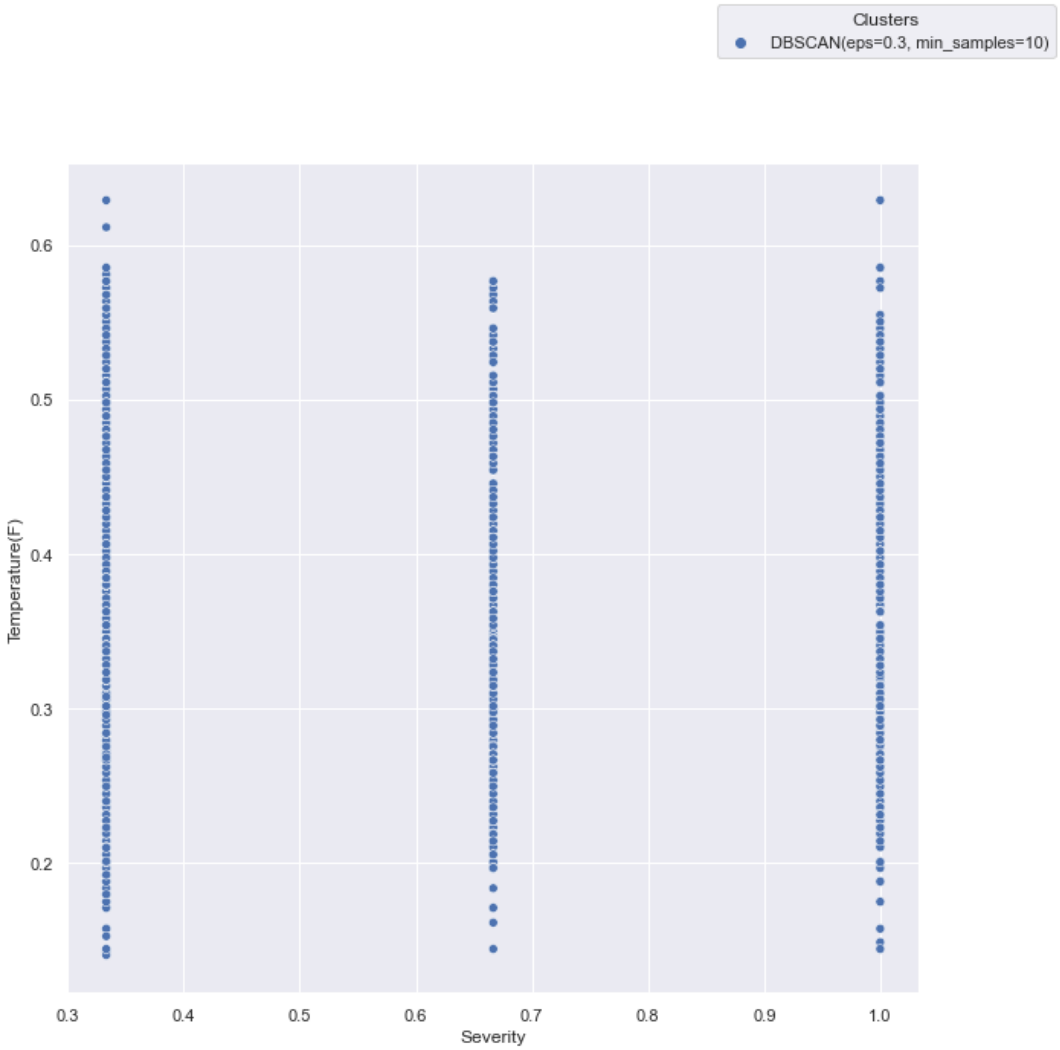
5.1 Εφαρμογή του DBSCAN

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.

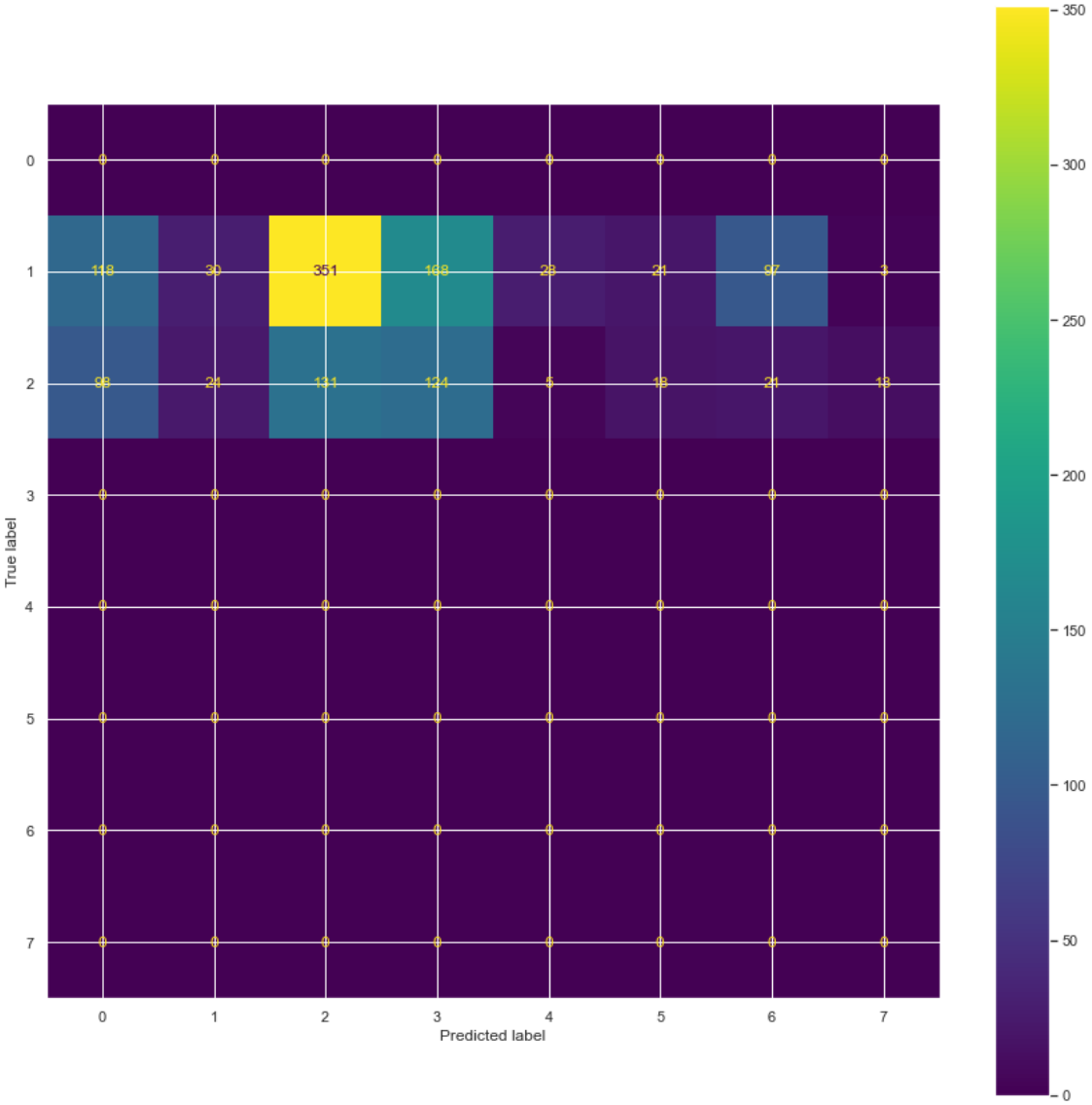
Συγκεκριμένα, υπολογίζονται οι πραγματικές ομάδες στις οποίες ανήκουν τα δείγματα στο σύνολο εκπαίδευσης και στη συνέχεια γίνεται μια πρόβλεψη για τις ομάδες στις οποίες θα ανήκουν τα δείγματα από το σύνολο ελέγχου. Τέλος, υπολογίζεται το σκορ ακριβείας του αλγορίθμου, το οποίο είναι 0.3852.

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).



Εικόνα 23: DBSCAN Scatter Plot



Εικόνα 24: DBSCAN Confusion Matrix

6. 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.

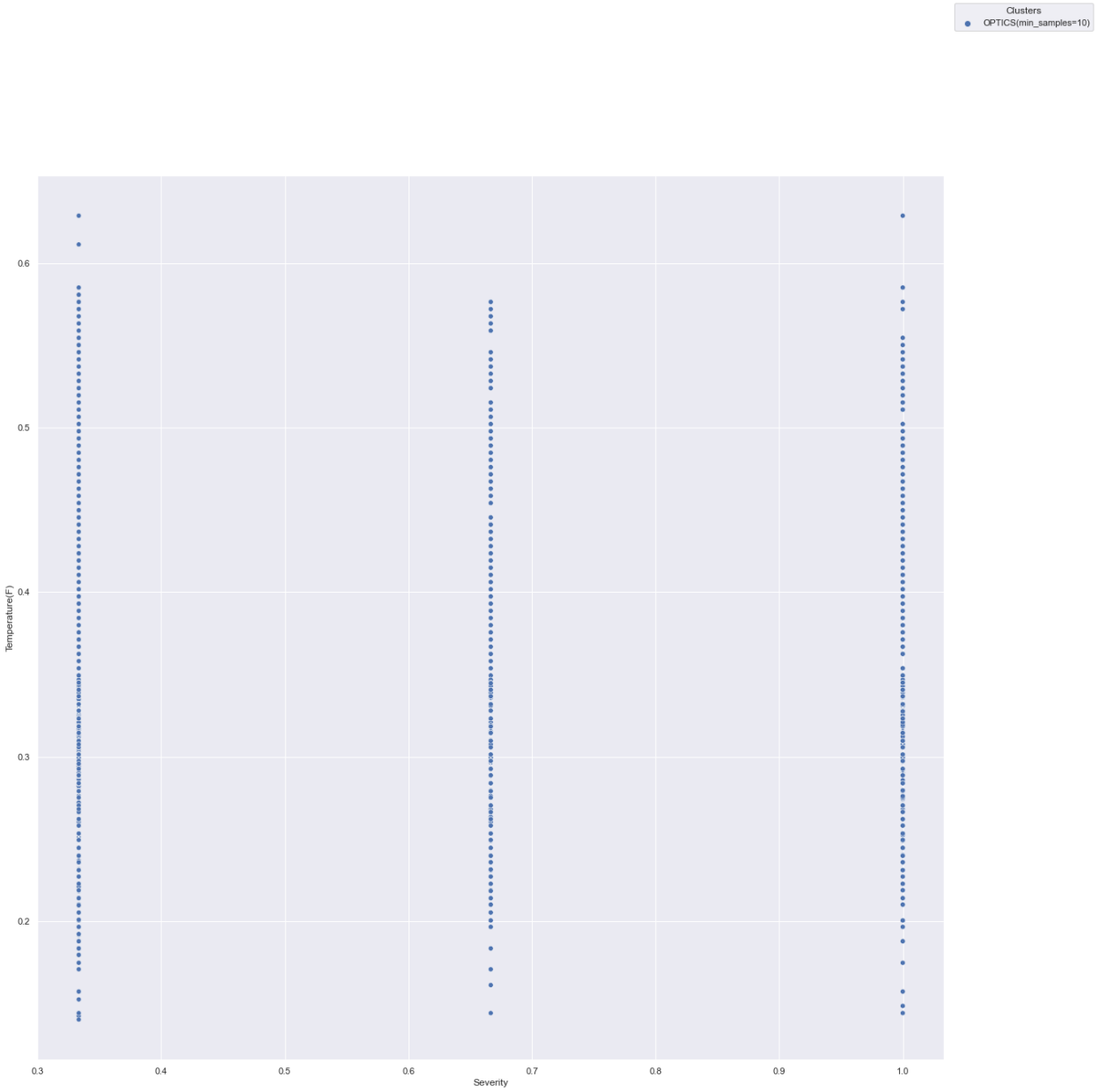
Αξίζει να σημειωθεί πως και για την εκτέλεση του OPTICS θα χρησιμοποιηθούν τα δεδομένα εκπαίδευσης και ελέγχου από το μειωμένο σύνολο δεδομένων.

6.1 Εφαρμογή του OPTICS

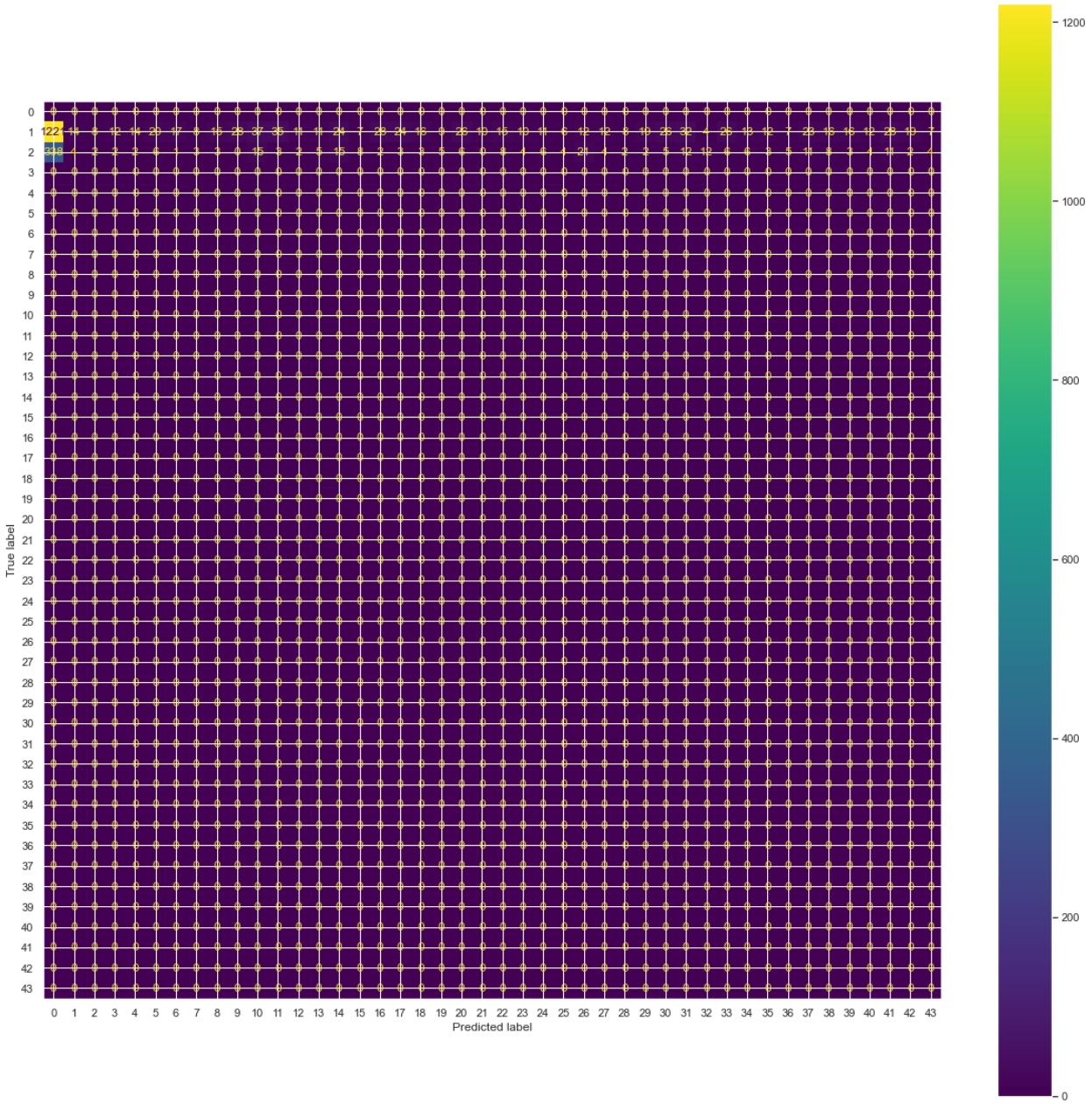
Η εφαρμογή της μεθόδου OPTICS στο παρόν σύνολο δεδομένων γίνεται με τη χρήση του πακέτου Scikit Learn. Συγκεκριμένα, υπολογίζονται οι πραγματικές ομάδες στις οποίες ανήκουν τα δείγματα στο σύνολο εκπαίδευσης και στη συνέχεια γίνεται μια πρόβλεψη για τις ομάδες στις οποίες θα ανήκουν τα δείγματα από το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



Εικόνα 25: OPTICS Scatter Plot



Εικόνα 26: OPTICS Confusion Matrix

7. Σύγκριση Τεχνικών Ομαδοποίησης

Όπως αναφέρθηκε και σε προηγούμενες παραγράφους, η συσταδοποίηση 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.

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: $O(n^2)$	• 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

Πίνακας 1: Σύγκριση Τεχνικών Ομαδοποίησης

΄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)
- Νευρωνικό Δίκτυο (Neural Network)
- Γραμμική Παλινδρόμηση Ελάχιστων Τετραγώνων (Linear Least Squares Regression)

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.

8. 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

Πρώτο 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).

	count	mean	std	min	25%	50%	75%	max
Severity	8000.0	0.445500	0.224122	0.333333	0.333333	0.333333	0.333333	1.000000
Start_Lat	8000.0	0.548471	0.214751	0.034886	0.397249	0.592566	0.706055	1.000000
Start_Lng	8000.0	0.543468	0.306601	0.002295	0.164453	0.645224	0.790158	0.982199
End_Lat	8000.0	0.546822	0.214099	0.034932	0.396049	0.590629	0.703902	1.000000
End_Lng	8000.0	0.543462	0.306608	0.002315	0.164792	0.645244	0.789267	0.983457
Distance(mi)	8000.0	0.006067	0.014127	0.000000	0.000857	0.002587	0.005690	0.418472
Temperature(F)	8000.0	0.382970	0.087158	0.065502	0.314410	0.362445	0.458515	0.637555
Wind_Chill(F)	8000.0	0.412861	0.094395	0.106054	0.338074	0.406745	0.496140	0.662739
Humidity(%)	8000.0	0.720544	0.228217	0.040404	0.565657	0.787879	0.898990	1.000000
Pressure(in)	8000.0	0.302734	0.024773	0.088431	0.299431	0.309625	0.315078	0.335467
Visibility(mi)	8000.0	0.081149	0.031147	0.000000	0.060000	0.100000	0.100000	0.800000
Wind_Speed(mph)	8000.0	0.007919	0.005271	0.000000	0.004600	0.007360	0.011040	0.034959
Precipitation(in)	8000.0	0.001102	0.014779	0.000000	0.000000	0.000000	0.000417	0.416250
Side	8000.0	0.140875	0.347914	0.000000	0.000000	0.000000	0.000000	1.000000
Sunrise_Sunset	8000.0	0.327875	0.469468	0.000000	0.000000	0.000000	1.000000	1.000000
Amenity	8000.0	0.009000	0.094446	0.000000	0.000000	0.000000	0.000000	1.000000
Bump	8000.0	0.000500	0.022356	0.000000	0.000000	0.000000	0.000000	1.000000
Crossing	8000.0	0.046875	0.211384	0.000000	0.000000	0.000000	0.000000	1.000000
Give_Way	8000.0	0.003250	0.056920	0.000000	0.000000	0.000000	0.000000	1.000000
Junction	8000.0	0.149250	0.356357	0.000000	0.000000	0.000000	0.000000	1.000000
No_Exit	8000.0	0.001375	0.037058	0.000000	0.000000	0.000000	0.000000	1.000000
Railway	8000.0	0.008625	0.092475	0.000000	0.000000	0.000000	0.000000	1.000000
Roundabout	8000.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Station	8000.0	0.021125	0.143810	0.000000	0.000000	0.000000	0.000000	1.000000
Stop	8000.0	0.014250	0.118527	0.000000	0.000000	0.000000	0.000000	1.000000
Traffic_Calming	8000.0	0.000750	0.027378	0.000000	0.000000	0.000000	0.000000	1.000000
Traffic_Signal	8000.0	0.074250	0.262194	0.000000	0.000000	0.000000	0.000000	1.000000
Turning_Loop	8000.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Εικόνα 27: Στατιστικά στοιχεία για το σύνολο εκπαίδευσης

	mean	std
Severity	0.445500	0.224122
Start_Lat	0.548471	0.214751
Start_Lng	0.543468	0.306601
End_Lat	0.546822	0.214099
End_Lng	0.543462	0.306608
Distance(mi)	0.006067	0.014127
Temperature(F)	0.382970	0.087158
Wind_Chill(F)	0.412861	0.094395
Humidity(%)	0.720544	0.228217
Pressure(in)	0.302734	0.024773
Visibility(mi)	0.081149	0.031147
Wind_Speed(mph)	0.007919	0.005271
Precipitation(in)	0.001102	0.014779
Side	0.140875	0.347914
Sunrise_Sunset	0.327875	0.469468
Amenity	0.009000	0.094446
Bump	0.000500	0.022356
Crossing	0.046875	0.211384
Give_Way	0.003250	0.056920
Junction	0.149250	0.356357
No_Exit	0.001375	0.037058
Railway	0.008625	0.092475
Roundabout	0.000000	0.000000
Station	0.021125	0.143810
Stop	0.014250	0.118527
Traffic_Calming	0.000750	0.027378
Traffic_Signal	0.074250	0.262194
Turning_Loop	0.000000	0.000000

Εικόνα 28: Στατιστικά στοιχεία για το σύνολο ελέγχου

8.2 Κανονικοποίηση

Αρχικά, 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).

```
First example: [[0.06 0.78 0.06 0.78 0.   0.47 0.51 0.82 0.32 0.1  0.01 0.   0.   1.
  0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.   0.  ]]

Normalized: [[-2.29  0.76 -2.29  0.76 -0.1   1.02  1.01  0.43  0.57  0.61  0.24 -0.07
  -0.4   1.43 -0.1  -0.02 -0.22 -0.06 -0.42 -0.04 -0.09  0.   -0.15 -0.12
  -0.03 -0.28  0.  ]]
```

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. 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.

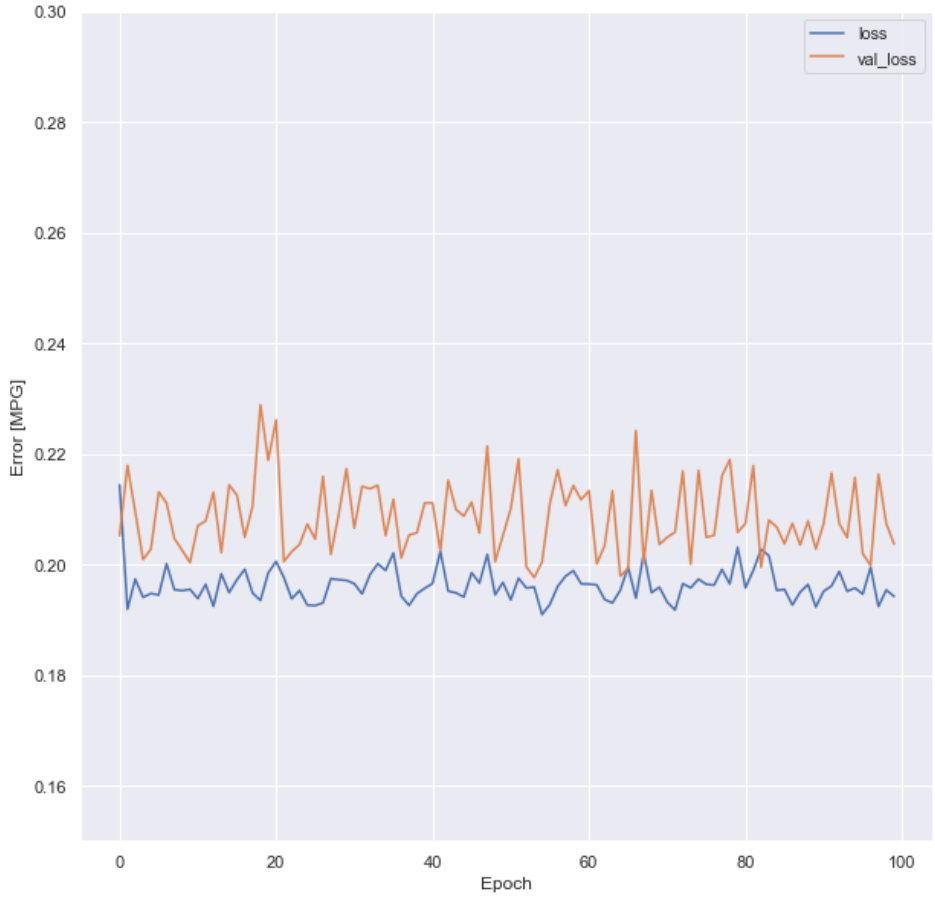
```
Model: "sequential_4"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 normalization_7 (Normalizat  (None, 1)                3
 ion)

 dense_8 (Dense)             (None, 1)                 2

=================================================================
Total params: 5
Trainable params: 2
Non-trainable params: 3
_________________________________________________________________
```

Εικόνα 30: Sequential μοντέλο

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

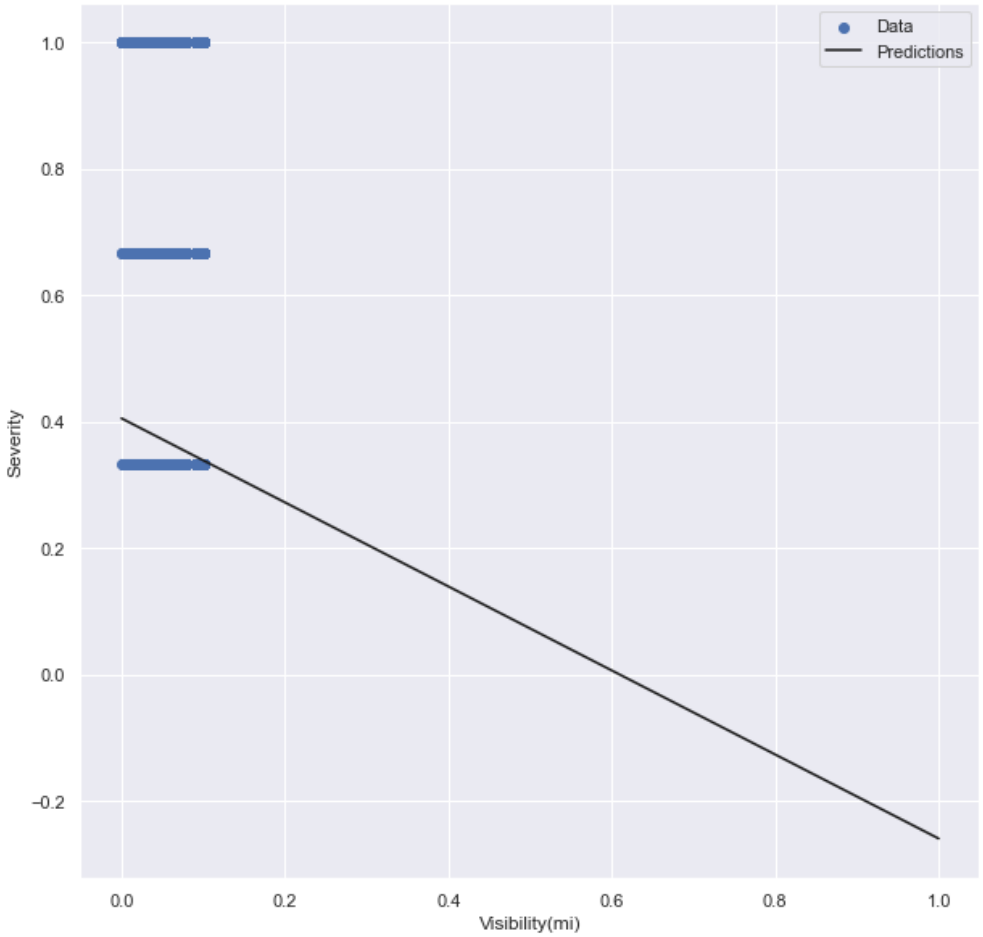
	loss	val_loss	epoch
95	0.124376	0.142807	95
96	0.124223	0.153065	96
97	0.121955	0.139598	97
98	0.124788	0.140469	98
99	0.124595	0.143245	99

Εικόνα 31: Στατιστικά διαδικασίας εκπαίδευσης

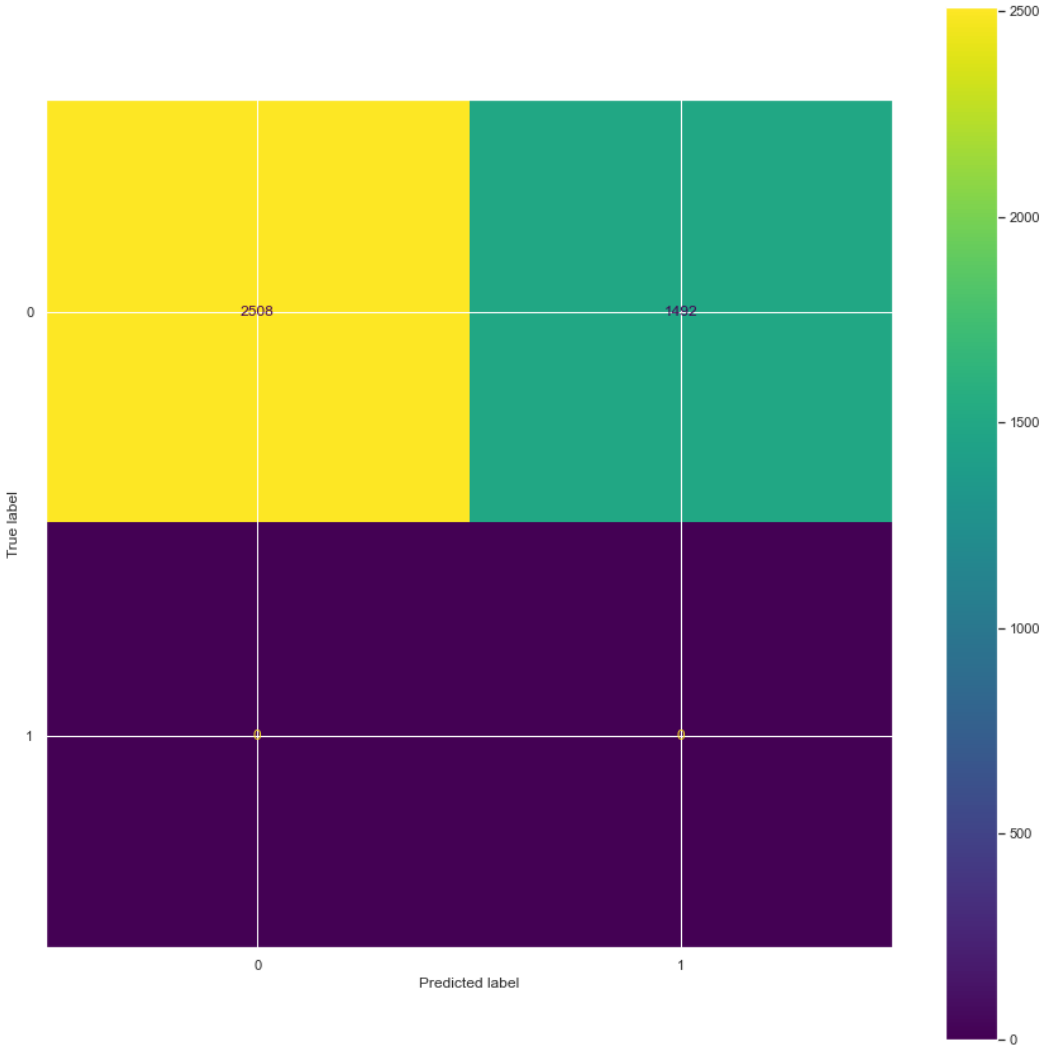


Εικόνα 32: Στατιστικά διαδικασίας εκπαίδευσης

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).



Εικόνα 33: Simple Linear Regression Scatter Plot



Εικόνα 34: Simple Linear Regression Confusion Matrix

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

9. Πολλαπλό Νευρωνικό Δίκτυο

Τεχνητό νευρωνικό δίκτυο είναι ένας όρος που χρησιμοποιείται στην τεχνητή νοημοσύνη και, στην ουσία, προσομοιάζει την λειτουργία του ανθρώπινου εγκεφάλου. Ένα τέτοιο δίκτυο είναι ένα σύστημα επεξεργασίας δεδομένων, το οποίο αποτελείται από κάποιες δομικές μονάδες, δηλαδή τους τεχνητούς νευρώνες. Χωρίζεται σε διάφορα επίπεδα, συγκεκριμένα στο επίπεδο εισόδου όπου εισέρχονται τα δεδομένα, στο επίπεδο εξόδου όπου εξάγονται τα αποτελέσματα και ενδιάμεσα τους στα κρυμμένα επίπεδα όπου γίνεται η επεξεργασία των δεδομένων. Αποτελούν ισχυρά εργαλεία για την επίλυση ιδιαίτερα σύνθετων και δύσκολων προβλημάτων, καθώς μπορούν να εκπαιδευτούν εύκολα με τη διαδικασία της μηχανικής μάθησης.

9.1 Δημιουργία Μοντέλου

Πρώτο βήμα στην εκμάθηση του μοντέλου με το εργαλείο Tensorflow / Keras, να οριστεί η αρχιτεκτονική του. Για την παρούσα ερευνητική εργασία επιλεχθεί το Sequential μοντέλο. Στη συνέχεια ορίζονται τα επίπεδα του νευρωνικού δικτύου, μαζί με κάποιες βασικές παραμέτρους όπως ο αριθμός εισόδων τους, η συνάρτηση ενεργοποίησης και οι διαστάσεις τους. Στην παρούσα εργασία τα επίπεδα του νευρωνικού δικτύου είναι έξι.

9.2 Εκπαίδευση και Αξιολόγηση

Προτού γίνει οποιαδήποτε ενέργεια σχετική με την εκπαίδευση του μοντέλου, είναι απαραίτητο να γίνει η μεταγλώττιση του, σε περίπτωση σφάλματος. Ακολουθεί η διαδικασία εκμάθησης, της οποίας τα αποτελέσματα φαίνονται στις εικόνες που ακολουθούν (Εικόνες 35 & 36).

```
Epoch 1/8
800/800 [==============================] - 2s 3ms/step - loss: 0.6796 - accuracy: 0.0397
Epoch 2/8
800/800 [==============================] - 2s 2ms/step - loss: 0.6783 - accuracy: 0.0456
Epoch 3/8
800/800 [==============================] - 2s 2ms/step - loss: 0.6781 - accuracy: 0.0440
Epoch 4/8
800/800 [==============================] - 2s 2ms/step - loss: 0.6770 - accuracy: 0.0466
Epoch 5/8
800/800 [==============================] - 2s 2ms/step - loss: 0.6767 - accuracy: 0.0456
Epoch 6/8
800/800 [==============================] - 2s 2ms/step - loss: 0.6764 - accuracy: 0.0461
Epoch 7/8
800/800 [==============================] - 2s 2ms/step - loss: 0.6756 - accuracy: 0.0496
Epoch 8/8
800/800 [==============================] - 2s 2ms/step - loss: 0.6754 - accuracy: 0.0491
<keras.callbacks.History at 0x7fa529bb4700>
```

Εικόνα 35: Διαδικασία εκμάθησης τεχνητού νευρωνικού δικτύου

```
63/63 [==============================] - 0s 2ms/step - loss: 0.6786 - accuracy: 0.0645
```

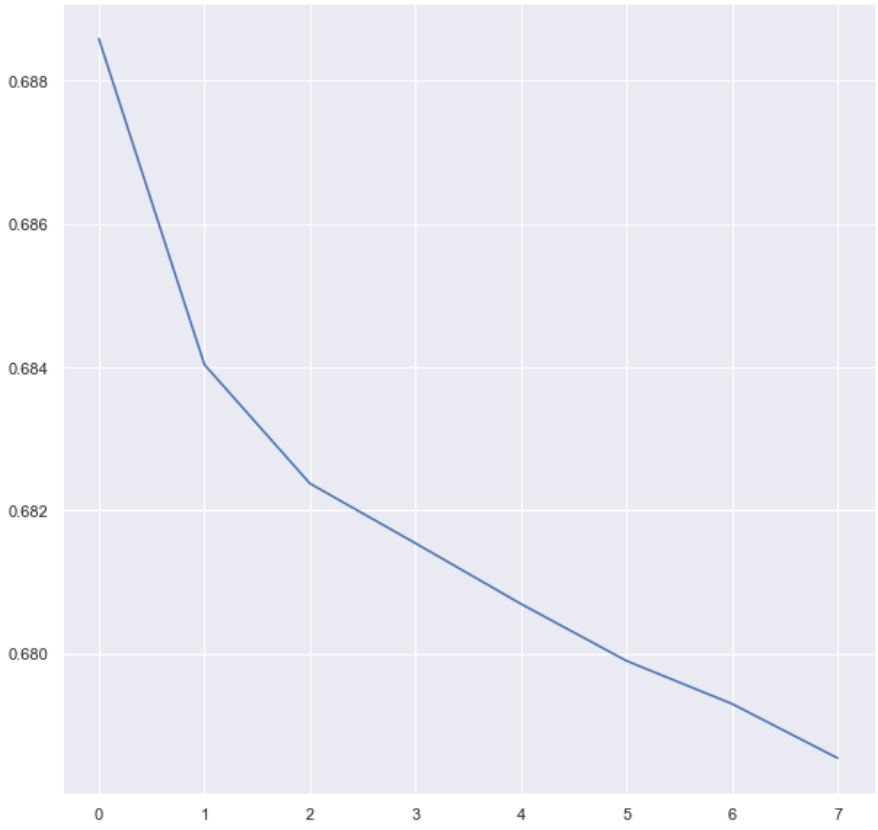
Εικόνα 36: Διαδικασία εκμάθησης τεχνητού νευρωνικού δικτύου

9.3 Προβλέψεις και Οπτικοποίηση

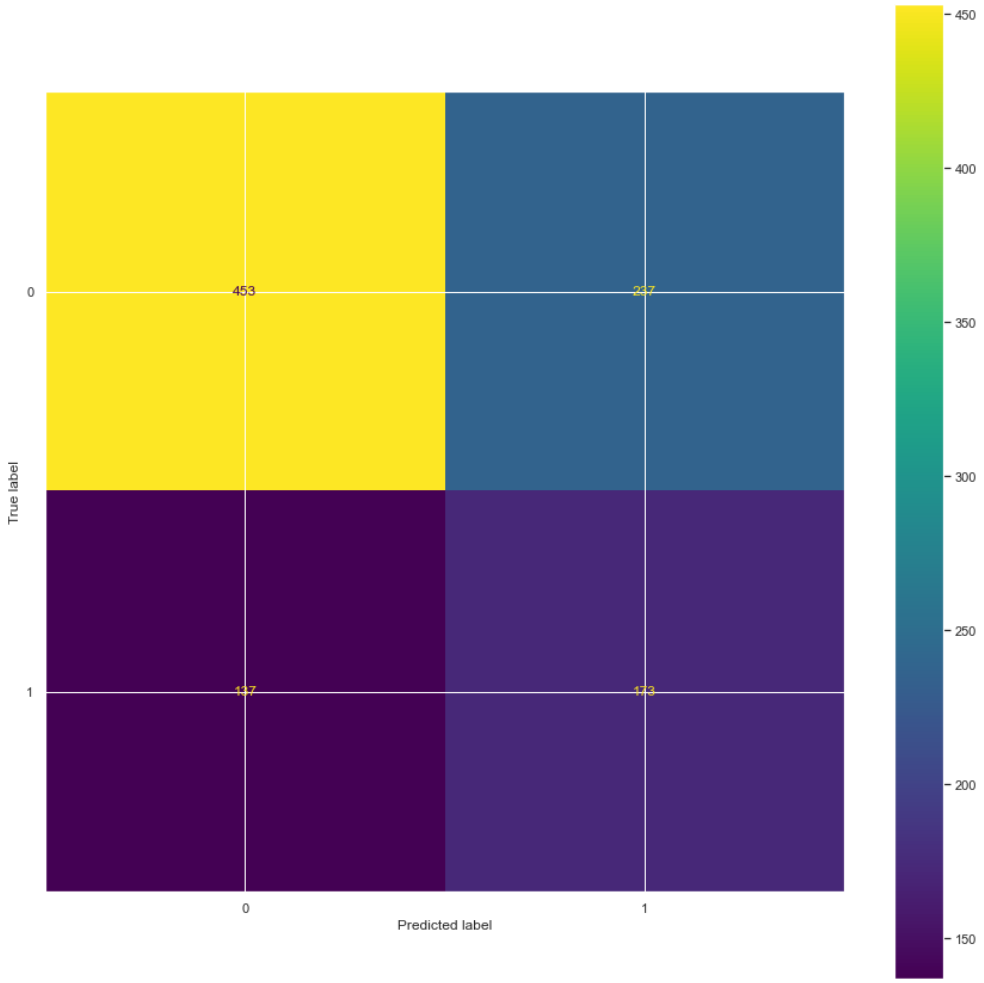
Μετά την επιτυχή ολοκλήρωση της εκπαίδευσης, το μοντέλο είναι πια ικανό να κάνει προβλέψεις πάνω στα χαρακτηριστικά του συνόλου ελέγχου. Τα στατιστικά της διαδικασίας εκμάθησης, όπως επίσης ο πίνακας σύγχυσης φαίνονται στις Εικόνες 37, 38 και 39.

	loss	accuracy	epoch
3	0.683132	0.13550	3
4	0.681936	0.13575	4
5	0.681526	0.13325	5
6	0.680619	0.13275	6
7	0.679497	0.13575	7

Εικόνα 37: Στατιστικά διαδικασίας εκμάθησης



Εικόνα 38: Στατιστικά διαδικασίας εκμάθησης



Εικόνα 39: Πίνακας σύγχυσης

10. Παλινδρόμηση Ελάχιστων Τετραγώνων

Η παλινδρόμηση ελάχιστων τετραγώνων είναι μια μέθοδος γραμμικής παλινδρόμησης, κατά την οποία τα σημεία δεδομένων απέχουν όσο το δυνατόν μικρότερη τετραγωνική απόσταση από τη γραμμή. Η μέθοδος ονομάζεται «Ελάχιστα Τετράγωνα», διότι καλύτερη γραμμή θεωρείται αυτή που ελαχιστοποιεί το άθροισμα τετραγωνικού σφάλματος. Η μέθοδος των ελάχιστων τετραγώνων θεωρείται η πιο βέλτιστη σε προβλήματα που έχουν γραμμική διαχωρισιμότητα χάρη στην απλότητα της. Παρόλα αυτά, σε πολλές περιπτώσεις ενώ υπάρχει η γνώση πως οι κλάσεις δεν είναι γραμμικά διαχωρίσιμες, οι ερευνητές επιλέγουν την υιοθέτηση ενός γραμμικού ταξινομητή, παρά το γεγονός ότι αυτό θα οδηγήσει σε υποβέλτιστες αποδόσεις.

10.1 Μαθηματική Εφαρμογή

Στην παλινδρόμηση ελάχιστων τετραγώνων θα γίνει εκτίμηση της σοβαρότητας ενός ατυχήματος (εξαρτημένη μεταβλητή), βάσει της θερμοκρασίας (ανεξάρτητη μεταβλητή). Αρχικά για να εφαρμόσουμε τον αλγόριθμο Least Squares είναι απαραίτητο να φέρουμε τα δεδομένα στην κατάλληλη μορφή. Η διαδικασία που θα ακολουθηθεί είναι, αυστηρά, μαθηματική και ανάγεται στον παρακάτω τύπο:

$$(X^T X)\hat{w} = X^T y \Rightarrow$$

$$\hat{w} = (X^T X)^{-1} X^T y$$

Πρώτο βήμα που πρέπει να εκτελεστεί είναι ο διαχωρισμός των ετικετών (labels) από τα χαρακτηριστικά (features). Στη συνέχεια, γίνονται οι απαραίτητες ενέργειες ώστε τα δεδομένα να έρθουν σε μορφή, τέτοια ώστε να μπορούν να εισαχθούν στον μαθηματικό τύπο (αντιστροφή μητρών, υπολογισμός bias). Επιπλέον, κρίθηκε απαραίτητη η μετατροπή τους σε constants ώστε να εξασφαλιστεί η σταθερότητά τους.

Πλέον με μια απλή αντικατάσταση μπορεί να υπολογισθεί το $\hat{w} = (X^T X)^{-1} X^T y$ και κατ' επέκταση η μήτρα βαρών A.

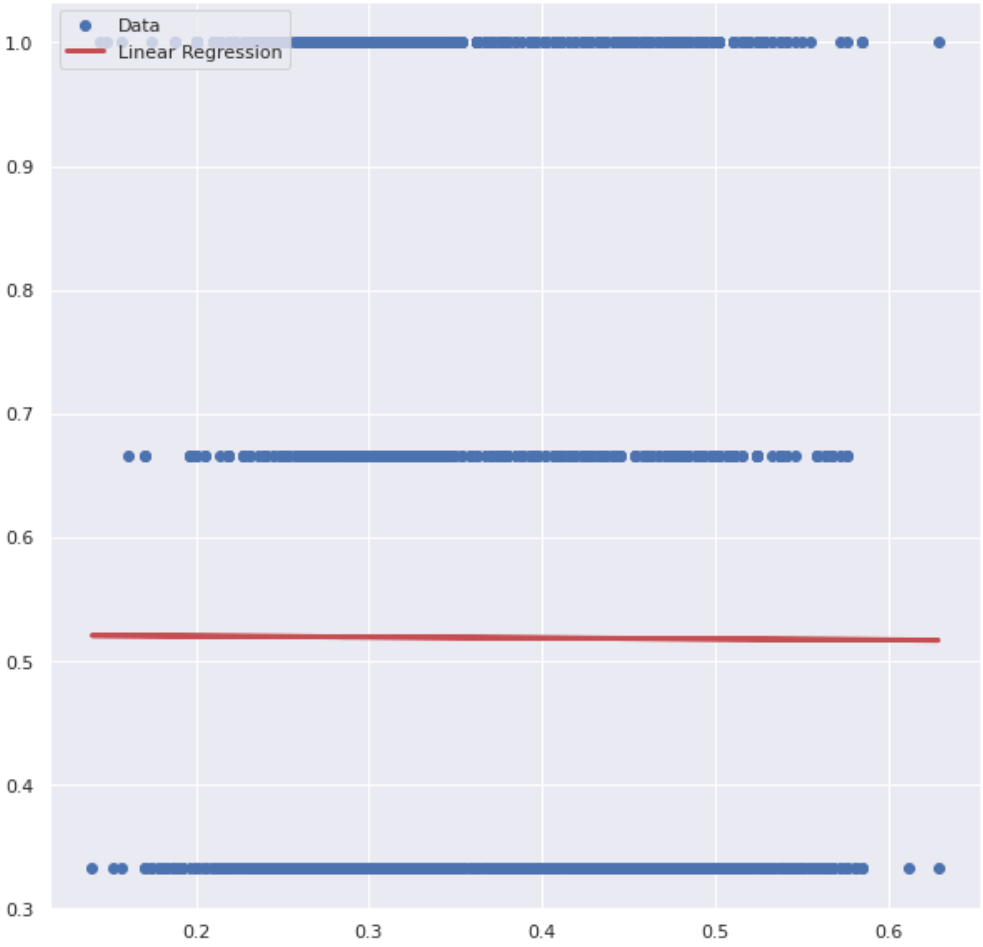
10.2 Οπτικοποίηση

Μετά την εφαρμογή του τύπου υπολογίζεται η κλίση και η απόκλιση της ευθείας, όπως φαίνεται και στην εικόνα που ακολουθεί (Εικόνα 39).

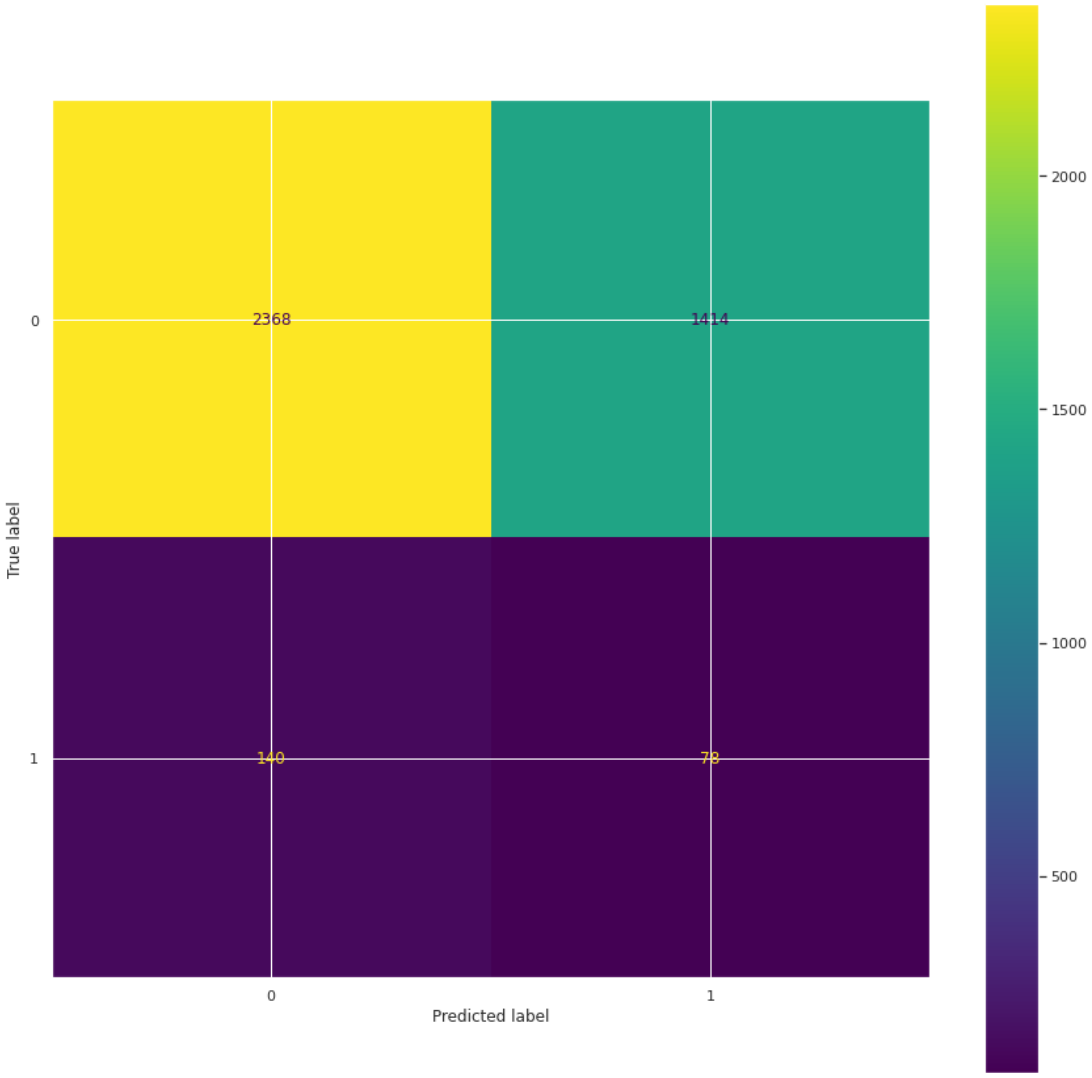
```
slope (m): tf.Tensor(-0.00791816385631638, shape=(), dtype=float64)
intercept (b): tf.Tensor(0.5217788759929788, shape=(), dtype=float64)
```

Εικόνα 39: Κλίση και απόκλιση ευθείας ελάχιστων τετραγώνων

Πλέον, έχουν γίνει όλες οι απαραίτητες μαθηματικές πράξεις και μπορεί να αναπαρασταθεί η γραμμική παλινδρόμηση ελαχίστων τετραγώνων με εξαρτημένη μεταβλητή την στήλη “Severity” και ανεξάρτητη μεταβλητή την στήλη “Temperature(F)”, όπως επίσης και ο αντίστοιχος πίνακας σύγχυσης (Εικόνες 40 & 41).



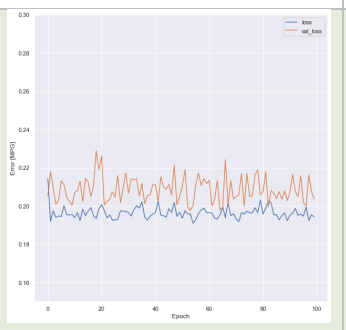
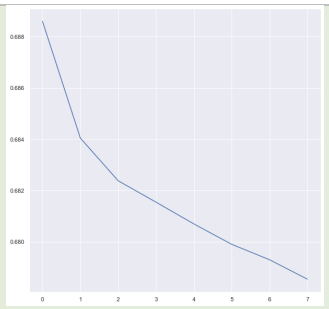
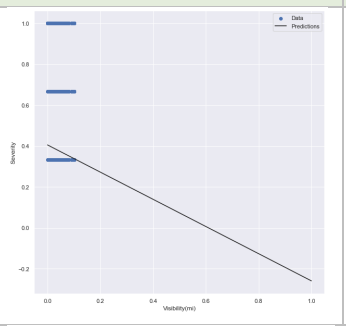
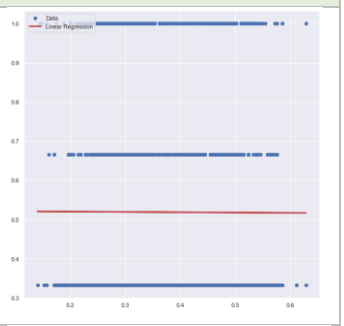
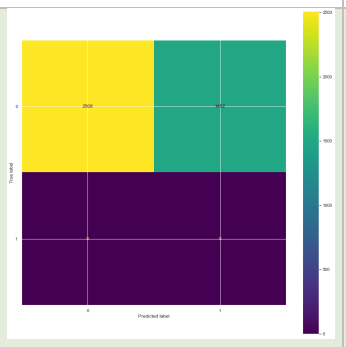
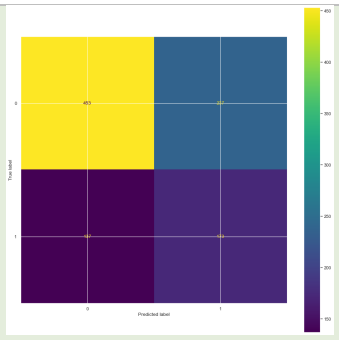
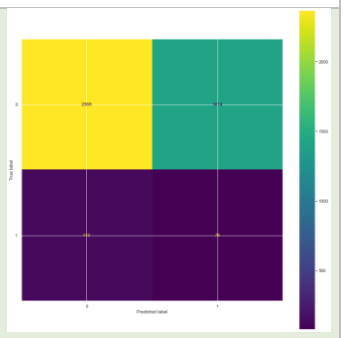
Εικόνα 40: Γραμμική παλινδρόμηση ελάχιστων τετραγώνων



Εικόνα 41: Πίνακας σύγχυσης ελάχιστων τετραγώνων

11. Σύγκριση Μοντέλων Μηχανικής Μάθησης

Παρόλο που, για τεχνικούς λόγους, έπρεπε να περιοριστούν οι διαστάσεις του συνόλου δεδομένων, η εξαγωγή συμπερασμάτων ήταν ακόμα εφικτή. Στον πίνακα που ακολουθεί, αναγράφονται κάποιες διαφοροποιήσεις που παρατηρήθηκαν κατά την εφαρμογή των αλγορίθμων μηχανικής και αποσκοπούν στην σύγκρισή τους.

Basic Linear Regression	Neural Network	Least Squares LR
		
		
		
• Σκορ: 16%	• Σκορ: 13%	• Σκορ: (σχεδόν) 0%

Με βάση τον παραπάνω πίνακα, το καλύτερο μοντέλο και το πιο αξιόπιστο ήταν αυτό της απλής γραμμικής παλινδρόμησης.