ΑΝΑΛΥΤΙΚΗ ΔΕΔΟΜΕΝΩΝ

Ομαδική Απαλλακτική Εργασία

Μέλη που Συμμετείχαν:
Χριστοφορίδης Χαράλαμπος – ΑΜ: Π19188
Καρκάνης Ευστράτιος – ΑΜ: Π19064
Ιωάννης Μπρισίμης-- ΑΜ: Π19118

01/07/2022

Πίνακας περιεχομένων

Σημαντική Παρατήρηση: Για να λειτουργήσει ο κώδικας, θα πρέπει το αρχείο με το dataset να μετονομαστεί σε dataset.csv.

Ερώτημα 1: εξοικείωση με τα δεδομένα

α) Επιλογή των δεδομένων

- Στο σύνολο των δεδομένων που μας δόθηκε, παρατηρήσαμε ότι κάποιες από τις στήλες του περιείχαν περιττές πληροφορίες που δεν μας αφορούσαν στην μελέτη μας. Για παράδειγμα, στήλες όπως η χώρα (Country), ο ταχυδρομικός κώδικας (Zipcode), ο Airport_Code κ.α. δεν μας βοηθούν στην λήψη αποφάσεων. Επομένως, οι στήλες αυτές αφαιρέθηκαν από το σύνολο των δεδομένων. Η διαδικασία αυτή έχει περιγραφεί με σχόλια σε σχετικό κομμάτι του κώδικα.

β) Καθαρισμός των δεδομένων[1]

- Το dataset έχει καθαριστεί από null τιμές. Λόγω του γεγονότος ότι οι εγγραφές του dataset είναι πάρα πολλές, αυτό που κάναμε είναι να διαγράψουμε όσες γραμμές έχουν **έστω μία null τιμή**. Αυτή μας η επιλογή δεν επηρεάζει την αξιοπιστία των δεδομένων, καθώς στις 2.8 εκατομμύρια εγγραφές του dataset (περίπου), οι 500.000 μόνο εγγραφές διαγράφηκαν. Επομένως, δεν θεωρούμε ότι η απώλεια αυτή θα μειώσει την ποιότητα της μελέτη μας.

- Το dataset έχει καθαριστεί από outliers. Αρχικά, αποφασίσαμε πως οι μόνες στήλες που χρειάζεται να ασχοληθούμε με outliers είναι οι «Temperature(F)», «Distance(mi)», «Humidity(%)» και «Visibility(mi)», καθώς είναι οι μόνες που έχουν αριθμητικά δεδομένα (numerical data). Συνεπώς, μόνο εκεί μπορούν να βρεθούν αποτελεσματικά οι outliers και να αντιμετωπιστούν. Για την εύρεση χρησιμοποιήσαμε την μέθοδο IQR[2]. Ειδικότερα, για την στήλη Visibility(mi) έγινε μία παραλλαγή, καθώς η πλειονότητα των τιμών μέσα σε αυτήν είχε την τιμή 10 με αποτέλεσμα να κρατάει μόνο τις τιμές με αριθμό 10. Για να το

αποφύγουμε αυτό ρίξαμε λίγο το κάτω όριο της μεθόδου κατά -4 όπως φαίνεται και στο σχετικό κομμάτι του κώδικα με επεξηγηματικό σχόλιο).

γ) Μετασχηματισμός των δεδομένων

- Στις τελευταίες 4 στήλες του dataset («Sunrise_Sunset», «Civil_Twilight», «Nautical_Twilight» και «Astronomical_Twilight») παρατηρήσαμε ότι μας δινόταν η ίδια πληροφορία (αν ήταν μέρα ή νύχτα) από διαφορετική πηγή. Συνεπώς, θεωρήσαμε σκόπιμο να βρούμε την «μέση τιμή» της κάθε γραμμής και να φτιάξουμε μία νέα στήλη που θα αντικαθιστά τις προηγούμενες 4 και θα περιέχει την μέση τιμή τους σε κάθε γραμμή αντίστοιχα. Για να το πετύχουμε αυτό κάναμε δυαδική αναπαράσταση των τιμών ορίζοντας το **Day ως 1** και το **Night ως 0** και εφαρμόσαμε την μέση τιμή δίνοντας μας την νέα στήλη «Day_Night» (αποτελείται από ακεραίους είτε 0 είτε 1).

- Θεωρήθηκε επίσης σκόπιμο να στρογγυλοποιήσουμε τους δεκαδικούς αριθμούς της στήλης «Temperature(F)», αφού στην μελέτη μας δεν είναι απαραίτητη η τόσο μεγάλη ακρίβεια στην αναπαράσταση των θερμοκρασιών. Μάλιστα, είναι ευκολότερο να δουλεύουμε με ακεραίους.

- Παρατηρήσαμε πως στη στήλη «Weather_Condition» υπάρχει πληθώρα περιγραφών καιρικών συνθηκών, πολλές φορές και για το ίδιο φαινόμενο. Στην δικιά μας μελέτη όμως, πολλές καιρικές συνθήκες δίνουν το ίδιο αποτέλεσμα με τις άλλες, για παράδειγμα σε ένα ατύχημα το να έχει λιακάδα με το να είναι μερικώς νεφελώδες δεν αποτελεί διαφορά που θα άλλαζε την εξέλιξη του ατυχήματος. Ως αποτέλεσμα, θεωρήσαμε σωστό να κατηγοριοποιήσουμε τις διάφορες περιγραφές σε μονάχα 4: **Perfect**,**Medium**,**Bad** και **Very Bad**. Έτσι, όχι μόνο διευκολύνεται η ανάλυση και η επεξεργασία της στήλης, αλλά και να δίδεται περισσότερη σημασία στα ίδια τα δεδομένα.

- Θεωρήθηκε επίσης σκόπιμο να μετατρέψουμε τα ονόματα των πολιτειών της στήλης «State» σε αριθμούς (κάνοντας χρήση της ειδικής συνάρτησης LabelEncoder). Μία τέτοια μετατροπή θα έκανε ευκολότερη την μετέπειτα χρησιμοποίηση τους στους υπολογισμούς.

Παρ' όλα αυτά, προτιμήσαμε να κρατήσουμε ως επιπλέον στήλη και την αρχική State με τους χαρακτήρες κειμένου ώστε να ξέρουμε και το ποιος αριθμός αντιστοιχεί σε ποια στήλη.

δ) Οπτικοποίηση των δεδομένων

- Για την οπτικοποίηση των δεδομένων θεωρήσαμε σκόπιμο να κάνουμε χρήση heatmap, έτσι ώστε να μπορέσουμε να δούμε την συσχέτιση (correlation) που έχουν οι στήλες μεταξύ τους. Τελικά, διαπιστώθηκε ότι δεν υπάρχει καθόλου σημαντική συσχέτιση μεταξύ των δεδομένων (ούτε θετική, ούτε αρνητική). Η παρατήρηση αυτή **δεν** μας βοήθησε πολύ να καταλάβουμε το πως θα επιλέξουμε την ιστορία που θα εξάγουμε.

- Μία ακόμα οπτικοποίηση που θεωρήσαμε σκόπιμη για την μελέτη μας ήταν και η χρήση ιστογραμμάτων για της στήλες «Severity» και «Weather_Condition» των δεδομένων μας. Τα ιστογράμματα αυτά μας επέτρεψαν να καταλάβουμε ότι το εν λόγω dataset είναι αρκετά **unbalanced**.

Ερώτημα 2: Ομαδοποίηση

α) Clustering

Σε αυτό το κομμάτι της εργασίας, αποφασίσαμε να δημιουργήσουμε κάποια clusters, να ομαδοποιήσουμε δηλαδή το dataset[3]. Μετά από αρκετές δοκιμές αποφασίσαμε να υλοποιήσουμε 5 «ζευγάρια» δεδομένων (columns), των οποίων τα αποτελέσματα θα μπορούσαν να μας αποκαλύψουν επιπλέον πληροφορία για τα δεδομένα μας. Συγκεκριμένα χρησιμοποιήσαμε τα ακόλουθα ζευγάρια/ συνδυασμούς δεδομένων:

- «Severity» και «Weather_Condition»
- «Weather_Condition» και «Distance»
- «Severity» και «State»

- ❖ «Day_Night» και «Visibility(mi)»
- ❖ «Side» και «Severity»

Σε όλα τα παραπάνω, κάναμε χρήση του αλγόριθμου ομαδοποίησης **K-means**. Για να μπορέσουμε να βρούμε το βέλτιστο K (αριθμό Cluster) για τον αλγόριθμο κάναμε χρήση της μεθόδου του αγκώνα (**Elbow Method**)[3].

Σημείωση: για όλα τα παραπάνω clusterings, υπάρχουν επεξηγηματικά σχόλια στο κώδικα καθώς και η σχετική οπτικοποίηση των αποτελεσμάτων τους, τόσο για τον K-means όσο και για την Elbow Method.

β) Χωρισμός σε Features και Label

Σε αυτό το κομμάτι του δεύτερου ερωτήματος δώσαμε ιδιαίτερη βάση στην σωστή επιλογή των **Features** και **Label**, καθώς μία σωστή επιλογή θα έκανε τις μετέπειτα μεθόδους ταξινόμησης να μας αποκαλύψουν ακόμα περισσότερη πληροφορία για τα δεδομένα μας. Η επιλογή μας ήταν:

- Features: «Weather_Condition», «Visibility(mi)» και «Day_Night»
- Label: «Severity»

Η παραπάνω επιλογή δεν αποτέλεσε τυχαία επιλογή, αλλά μία λογική επιλογή καθώς η κατάσταση του καιρού, η ορατότητα και το αν είναι πρωί ή μέρα αποτελούν σημαντικούς παράγοντες στην σοβαρότητα ενός ατυχήματος.

Εν συνεχεία, και πριν μπορέσουμε να εφαρμόσουμε οποιαδήποτε τεχνική ταξινόμησης, έπρεπε να χωρίσουμε τα δεδομένα μας σε train και test ώστε να τροφοδοτήσουμε κατάλληλα αυτές τις εν λόγω μεθόδους[4]. Αναλυτικότερα αποφασίσαμε το test να αποτελεί το 20% των δεδομένων μας. Επίσης προβήκαμε και μία κανονικοποίηση των δεδομένων εκπαίδευσης, ώστε να φέρουμε και τα δύο μεταξύ του διαστήματος [0,1].

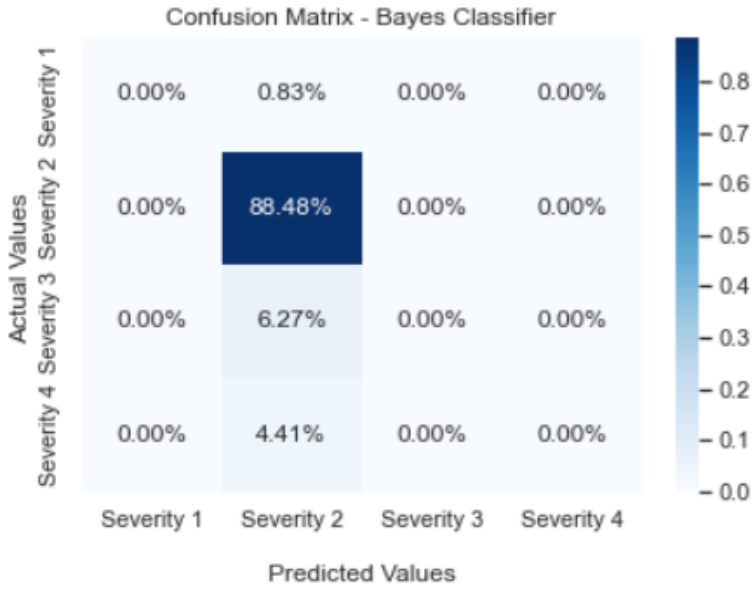
γ) Τεχνικές ταξινόμησης

Στο συγκεκριμένο κομμάτι του ερωτήματος χρησιμοποιήσαμε 3 τεχνικές ταξινόμησης[5], όπως μας υποδεικνύεται και από την εκφώνηση:

1. Gaussian (Bayes) Classifier[6]
2. Random Forest Classifier[7]
3. Decision Tree Classifier

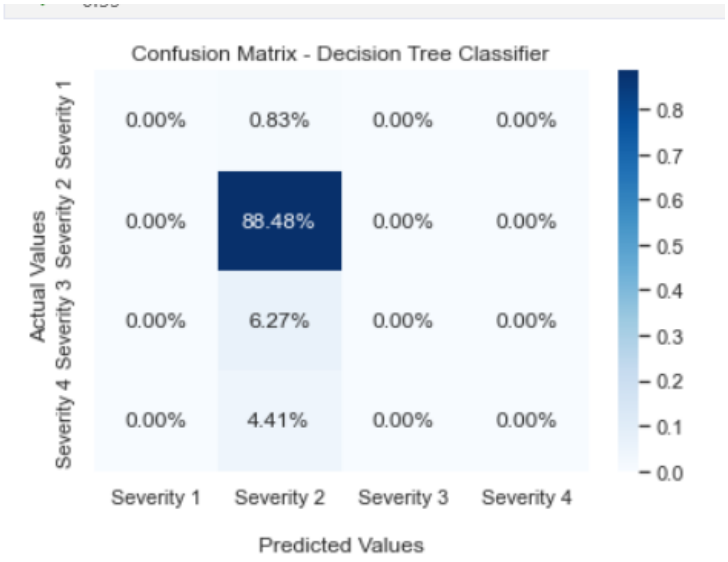
Από όλες τις παραπάνω τεχνικές πήραμε τα ανάλογα αποτελέσματα και μπορέσαμε να κάνουμε συγκρίσεις μέσω της ακρίβειας αυτών, δημιουργώντας Confusion Matrixes[8], αλλά και μέσω του χρόνου του οποίου απαιτούταν για την διεξαγωγή τους. Αναλυτικότερα:

1. Ο Bayes ταξινομητής σημείωσε ένα πολύ υψηλό accuracy της τάξης του 88% και ένα F1 – macro score 23%. Η confusion matrix του ταξινομητή φαίνεται στην επόμενη εικόνα.



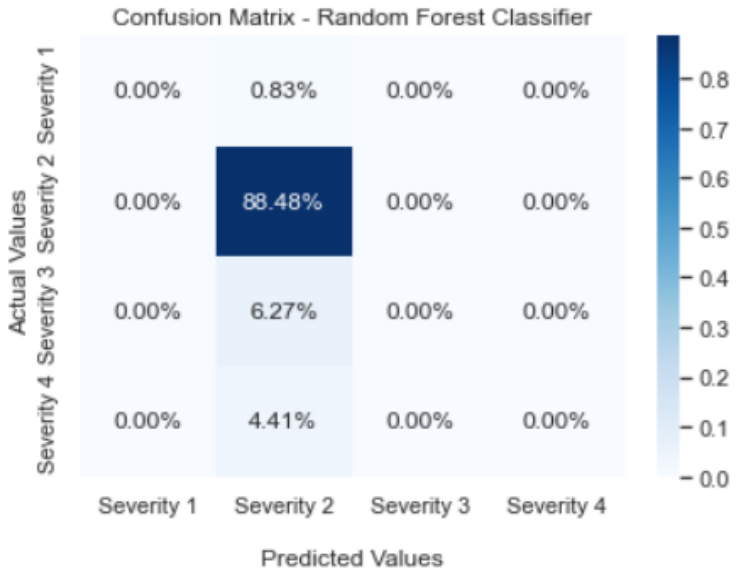
Confusion Matrix - Bayes Classifier

2. Ο Decision Tree ταξινομητής σημείωσε ένα πολύ υψηλό accuracy της τάξης του 88% και ένα F1 – macro score 23%. Η confusion matrix του ταξινομητή φαίνεται στην επόμενη εικόνα.



Confusion Matrix - Decision Tree

3. Ο Random Forest ταξινομητής σημείωσε ένα πολύ υψηλό accuracy της τάξης του 88% και ένα F1 – macro score 23%. Η confusion matrix του ταξινομητή φαίνεται στην επόμενη εικόνα.



Random Forest Classifier - Confusion Matrix

Ο γρηγορότερος όλων αποδείχθηκε ο **Gaussian ταξινομητής**. Επιπλέον, και οι τρεις ταξινομητές σημείωσαν παρόμοια αποτελέσματα. Τέλος για να μπορέσουμε να έχουμε μεγαλύτερη ακρίβεια στην σύγκριση αυτή τρέξαμε τον κώδικα σε 3 διαφορετικούς υπολογιστές με διαφορετική επεξεργαστική ισχύ.

Ωστόσο, το γεγονός ότι οι ταξινομητές μας προέβλεπαν συνεχώς την κλάση «Severity 2» προκύπτει από το γεγονός ότι τα δεδομένα μας είναι αρκετά μη ισορροπημένα (τα περισσότερα δείγματα του dataset ανήκουν στην κλάση «Severity 2»). Επομένως, **η μετρική accuracy δεν είναι πολύ έμπιστη σε δεδομένα που είναι imbalanced.** Με άλλα λόγια, το 88% accuracy δεν σημαίνει ότι ο ταξινομητής μας είναι πολύ καλός. Στην ουσία, το καλύτερο που έχει να κάνει, είναι να ταξινομεί κάθε παρατήρηση στην κλάση με τη μέγιστη πιθανότητα. Για αυτό τον λόγο το accuracy δεν γίνεται να μην είναι υψηλό σε τέτοιες περιπτώσεις.

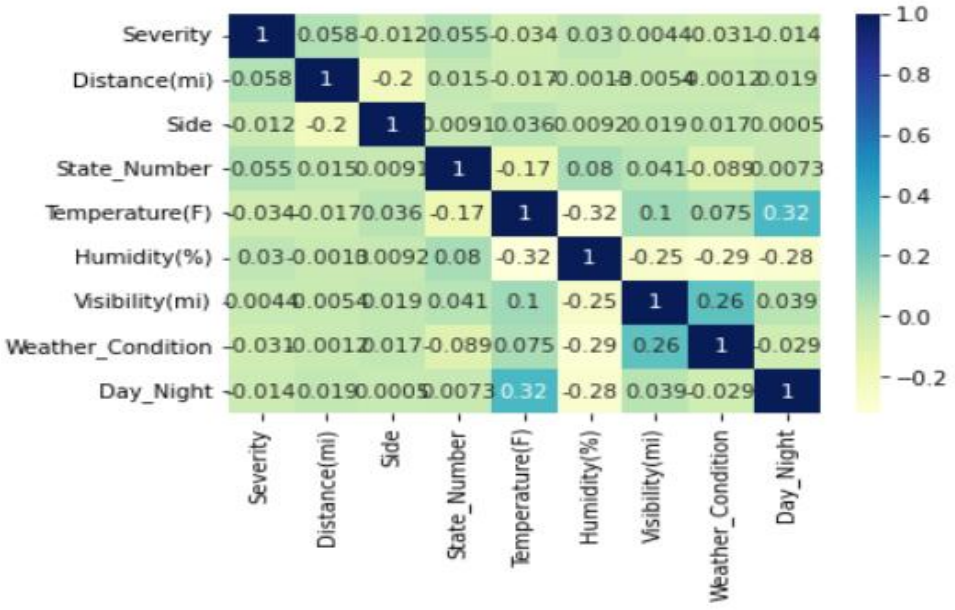
Επομένως, στην εργασία μας εισάγαμε την μετρική του F1 score με average macro[9]. Το F1 score δουλεύει άριστα σε περιπτώσεις που υπάρχουν imbalanced data (βασίζεται σε δύο άλλες μετρικές: precision και recall [10]), δίδοντάς μας μία αντικειμενική προσέγγιση της ποιότητας του ταξινομητή.

Ερώτημα 3: Classification/Regression/Storytelling[11]

α) Storytelling

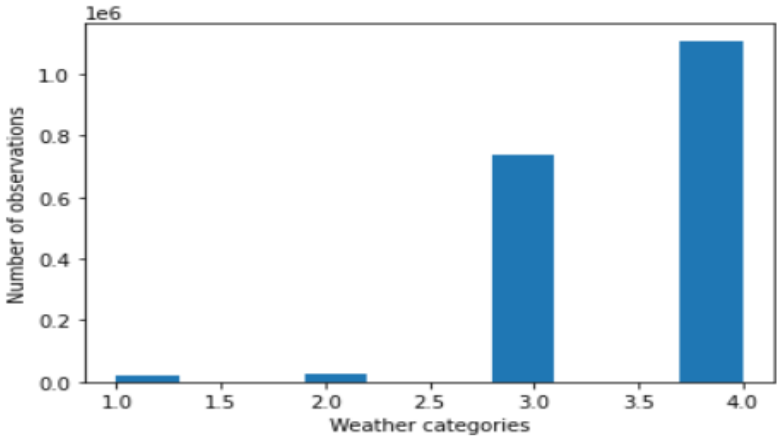
Ως τελική φάση στην εργασίας μας οφείλουμε να παράγουμε μια ιστορία, μέσα από τα αποτελέσματα της επεξεργασίας που κάναμε στα δεδομένα μας. Η αρχική μας σκέψη ήταν να παράγουμε μία πληθώρα ιστοριών από όλο το σύνολο των δεδομένων. Παρ' όλα αυτά, μετά από τα βήματα που μας υποδείχθηκαν από την εκφώνηση, κάτι τέτοιο δεν ήταν εφικτό, καθώς τα ίδια επεξεργασμένα δεδομένα δεν ήταν ικανά να μας δώσουν την απαραίτητη πληροφορία για την δημιουργία πολλαπλών ιστοριών. Συγκεκριμένα:

1. Κατά τη διάρκεια σχηματισμού του χάρτη συσχετίσεων (correlation map) δεν υπήρξε καμία ουσιαστική συσχέτιση μεταξύ των μεταβλητών του συνόλου δεδομένων.

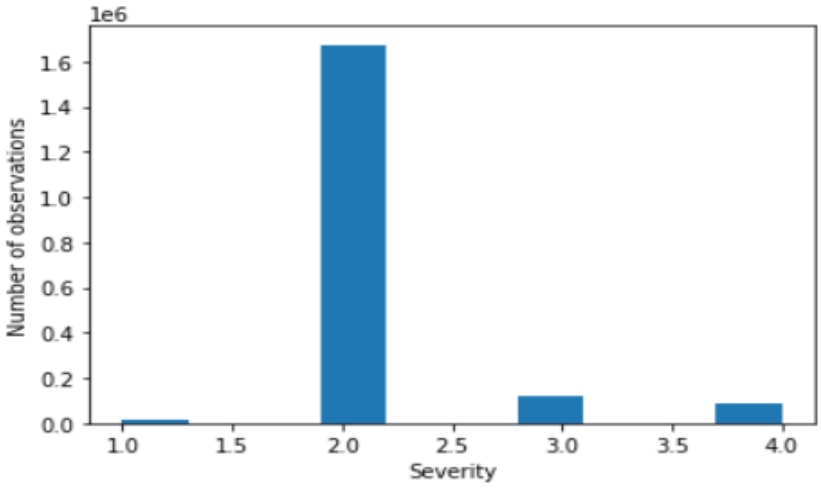


Στην εικόνα αυτή απεικονίζεται η μήτρα συσχετίσεων. Όπως φαίνεται, δεν υπάρχει καμία σημαντική συσχέτιση (θετική ή αρνητική) μεταξύ των μεταβλητών. Αυτή η διαπίστωση δυσχεραίνει στην λήψη αντικειμενικών αποτελεσμάτων από τους αλγορίθμους μηχανικής μάθησης και ταξινόμησης που υιοθετήθηκαν.

2. Διαπιστώσαμε μεγάλη ανισορροπία (imbalance) στα δεδομένα μας, κάτι που υποδεικνύεται και στα παρακάτω γραφήματα:



Το ιστόγραμμα αυτό απεικονίζει ότι η συντριπτική πλειοψηφία δεδομένων ανήκει στις κατηγορίες καιρού 4 (δηλαδή perfect weather) και 3 (δηλαδή μέτριος καιρός). Όλες οι άλλες κατηγορίες καιρού αντιπροσωπεύονται από έναν πολύ χαμηλό (μηδαμινό) αριθμό δειγμάτων.



Το ιστόγραμμα αυτό απεικονίζει ότι η κατηγορία severity 2 αντιπροσωπεύει πάνω από το 90% (περίπου) των δειγμάτων του συνόλου δεδομένων.

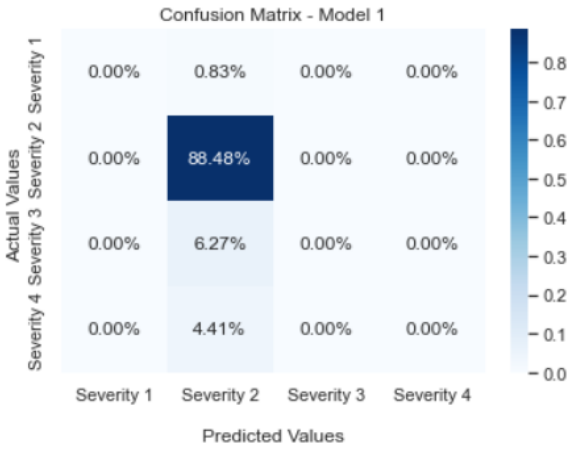
Αυτή η ανισορροπία δυσχεραίνει την παραγωγή της ιστορίας μας, καθώς τα μοντέλα των ταξινομητών και των νευρωνικών δικτύων **προβλέπουν πάντα την κλάση με τα περισσότερα δείγματα**.

Ως αποτέλεσμα των παραπάνω, θεωρήσαμε λογικό και σκόπιμο να περιοριστούμε μονάχα σε μία ιστορία η οποία σε σύγκριση με τις υπόλοιπες

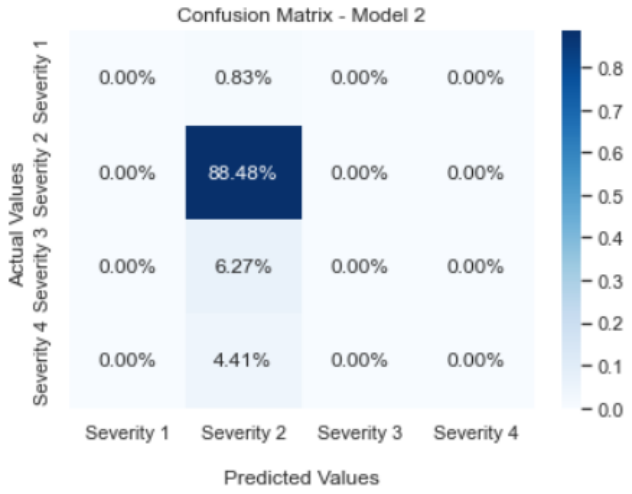
που είχαμε στο μυαλό μας υπερτερούσε ποσοτικά και ποιοτικά. Η ιστορία αυτή αφορά την πρόβλεψη της σοβαρότητας του ατυχήματος («Severity») , βάση των χαρακτηριστικών που αναφέραμε προηγουμένως («Weather_Condition», «Day_Night» και «Visibility(mi)»). Ο χαρακτήρας αυτής της ιστορίας μας φαίνεται και ο πιο λογικός, καθώς η σοβαρότητα ενός ατυχήματος αποτελεί και η σημαντικότερη παράμετρος σε ένα τέτοιο συμβάν.

Έχοντας καταλήξει στη παραπάνω ιστορία ξεκινήσαμε και τις ανάλογες διαδικασίες, ώστε να μας βοηθήσουν στην διατύπωσή της. Συγκεκριμένα, προβήκαμε στη δημιουργία **τριών νευρωνικών δικτύων**, τα οποία και εκπαιδεύσαμε ώστε να μπορέσουν να μας αποκαλύψουν επιπρόσθετη πληροφορία για να ταιριάξει στην ιστορία μας. Όπως αναφέρεται και στην αρχή του 3ου ερωτήματος της εργασίας, τα 3 αυτά μοντέλα διαφέρουν μεταξύ τους ως προς τον αριθμό των κρυφών επιπέδων. Ο στόχος αυτής της αλλαγής μεταξύ των μοντέλων ήταν να αποφύγουμε την χασούρα πληροφορίας, που πιθανόν να υπήρχε από μία μόνο μοναδική αρχιτεκτονική. Γενικά, προσπαθήσαμε ποικιλοτρόπως να αλλάξουμε παραμέτρους στα δίκτυά μας, κάτι το οποίο μας δυσκόλεψε αρκετά, καθώς τα δίκτυα προβλέπανε πάντα την κλάση με τα περισσότερα δείγματα. Αυτό μάλιστα φαίνεται και από τις παρακάτω μήτρες (confusion matrixes):

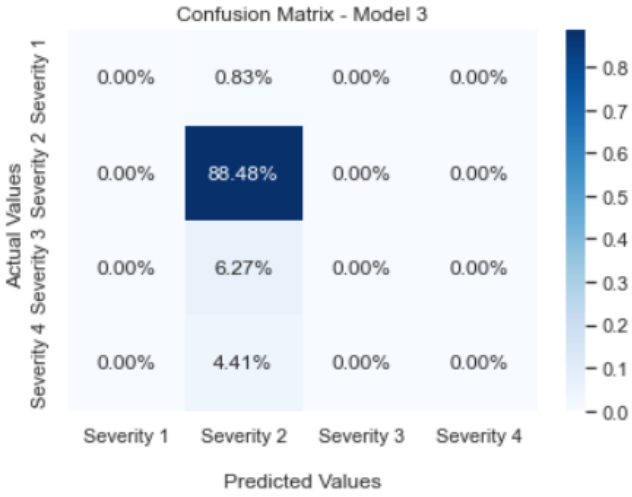
- Νευρωνικό Δίκτυο 1 (πρώτο μοντέλο)



- Νευρωνικό Δίκτυο 2 (δεύτερο μοντέλο)



- Νευρωνικό Δίκτυο 1 (τρίτο μοντέλο)



Κατά συνέπεια, δεν ήταν δυνατή η εξαγωγή κάποιας επιπρόσθετης πληροφορίας, καθώς για άλλη μια φορά στη μελέτη μας η ανισορροπία των δεδομένων δεν επέτρεψε στην επεξεργασία να επιφέρει καρπούς.

β) Γενικές Παρατηρήσεις/Συμπεράσματα

- Στο σύνολο των δεδομένων μας υπήρχε μεγάλη ανισορροπία των τιμών. Η παραπάνω παρατήρηση αναφέρεται σε κάθε σημείο της επεξεργασίας στο οποίο μας πρόσθεσε μια παραπάνω δυσκολία, καθώς τα αποτελέσματα από τις διάφορες ενέργειες ήταν ιδιαίτερα προβληματικά. Αυτή η ανωμαλία λοιπόν, καθιστά τα πορίσματα της εργασίας μας όχι τόσο ποιοτικά όσο θα επιθυμούσαμε εξαρχής.

- Επίσης παρατηρήθηκε, ότι η πολιτεία με τα περισσότερα ατυχήματα ήταν η Καλιφόρνια.

	State	Counts
0	CA	476520
1	FL	282610
2	TX	117003
3	OR	83386
4	VA	75577
5	SC	70262
6	NY	70041
7	NC	65085
8	PA	64114
9	MN	56317

The state with the majority of accidents is: CA

- Τέλος, μια ιδιαίτερα αναπάντεχη παρατήρηση για εμάς, είναι ότι τα περισσότερα ατυχήματα λάβαν χώρα σε καλές καιρικές συνθήκες όπως φαίνεται και στο ιστόγραμμα που αναφέρθηκε παραπάνω.

Βιβλιογραφία

[1] "Data Preprocessing in Data Mining -A Hands On Guide - Analytics Vidhya." https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/ (accessed Jul. 01, 2022).

[2] "(73) Outlier detection and removal using IQR | Feature engineering tutorial python # 4 - YouTube." https://www.youtube.com/watch?v=A3gClkbIXK8&ab_channel=codebasics (accessed Jul. 01, 2022).

[3] "A Guide to Data Clustering Methods in Python | Built In." https://builtin.com/data-science/data-clustering-python (accessed Jun. 15, 2022).

[4] "How To Split Train and Test Data. Concepts explained for beginners! | by Anya | Medium." https://medium.com/@karyaozmen/how-to-split-train-and-test-data-c1381d240fc4 (accessed Jun. 18, 2022).

[5] "Top 6 Machine Learning Algorithms for Classification | by Destin Gong | Towards Data Science." https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501 (accessed Jun. 21, 2022).

[6] "Beginners Guide to Naive Bayes Algorithm in Python." https://www.analyticsvidhya.com/blog/2021/01/a-guide-to-the-naive-bayes-algorithm/ (accessed Jun. 18, 2022).

[7] "Sklearn Random Forest Classifiers in Python Tutorial | DataCamp." https://www.datacamp.com/tutorial/random-forests-classifier-python (accessed Jun. 20, 2022).

[8] "Confusion Matrix for Multi-Class Classification - Analytics Vidhya." https://www.analyticsvidhya.com/blog/2021/06/confusion-matrix-for-multi-class-classification/ (accessed Jun. 18, 2022).

[9] "Micro, Macro & Weighted Averages of F1 Score, Clearly Explained | by Kenneth Leung | Towards Data Science." https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f (accessed Jun. 30, 2022).

[10] “How to Calculate F1 Score in Python (Including Example) - Statology.” https://www.statology.org/f1-score-in-python/ (accessed Jun. 30, 2022).

[11] “GUNet2 eClass - Τμήμα Πληροφορικής | ΑΝΑΛΥΤΙΚΗ ΔΕΔΟΜΕΝΩΝ (6ο εξ.) | Έγγραφα.” https://gunet2.cs.unipi.gr/modules/document/document.php?course=TMD104&openDir=/2010092902619o88n75tp/5eb91cb47YWp (accessed Jul. 01, 2022).