Efstratios Karkanis

Απόσπασμα

[Τραβήξτε την προσοχή του αναγνώστη σας με ένα ελκυστικό απόσπασμα. Συνήθως πρόκειται για μια σύντομη περίληψη του εγγράφου.   
Όταν είστε έτοιμοι να προσθέσετε το περιεχόμενό σας, απλώς κάντε κλικ εδώ και αρχίστε να πληκτρολογείτε.]

ΑΝΑΛΥΤΙΚΗ ΔΕΔΟΜΕΝΩΝ

Ομαδική Απαλλακτική Εργασία

**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 αφορούν στην μελέτη μας. Για παράδειγμα, στήλες όπως η χώρα (country), ο ταχυδρομικός κώδικας (zipcode), o Airport\_Code κ.α. δεν μας βοηθούν στην λήψη αποφάσεων. Επομένως, οι στήλες αυτές αφαιρέθηκαν από το σύνολο των δεδομένων. Η διαδικασία αυτή έχει περιγραφτεί σε σχετικό κομμάτι του κώδικα.

**β) Καθαρισμός των δεδομένων**

* 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.
* To dataset έχει καθαριστεί από outliers. Αρχικά αποφασίσαμε πως οι μόνες στήλες που χρειαζόταν να ασχοληθούμε με outliers θα ήταν οι << Temperature(F), Distance(mi), Humidity(%), Visibility(mi) >> καθώς είναι οι μόνες που έχουν αριθμητικά δεδομένα(numerical data) και συνεπώς μόνο εκεί μπορούν να βρεθούν αποτελεσματικά οι outliers και να αντιμετωπιστούν. Για την εύρεση χρησιμοποιήσαμε την μέθοδο IQR.(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) Μετασχηματισμό των δεδομένων:**

* Στις τελευταίες 4 στήλες του dataset (Sunrise\_Sunset, Civil\_Twilight, Nautical\_Twilight, 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(που εμπεριέχεται από ακεραίους είτε 0 είτε 1).
* Θεωρήθηκε επίσης σκόπιμο να στρογγυλοποιήσουμε τους δεκαδικούς αριθμούς της στήλης Temperature(F) καθώς στην μελέτη μας δεν είναι απαραίτητη η τόσο μεγάλη ακρίβεια στην αναπαράσταση των θερμοκρασιών και θεωρήσαμε ευκολότερο να δουλέψουμε 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,Very Bad) έτσι ώστε όχι μόνο να διευκολύνει την επεξεργασία της στήλης αλλά και να δώσει 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) καθώς μία τέτοια μετατροπή θα έκανε ευκολότερη την μετέπειτα χρησιμοποίηση τους στους υπολογισμούς .Παρ’ όλα αυτά προτιμήσαμε να κρατήσουμε ως <<extra>> στήλη και την αρχική State με τους χαρακτήρες κειμένου ώστε να ξέρουμε και το ποιος αριθμός αντιστοιχεί σε ποια στήλη.

**δ) Οπτικοποίηση 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. Μία αρκετά ενδιαφέρουσα και λογική παρατήρηση είναι η αρνητική συσχέτιση που έχει η στήλη Temperature(F) με την στήλη Humidity(%), λογικό αποτέλεσμα αφού η θερμοκρασία είναι παράγοντας της υγρασίας και είναι λογικό να υπάρχει αρνητική συσχέτιση αφού όσο αυξάνεται η θερμοκρασία μειώνεται η υγρασία π.χ. το καλοκαίρι τα επίπεδα υγρασίας είναι όντως πολύ χαμηλά. Οι παρατηρήσεις αυτές μας βοήθησαν πολύ να καταλάβουμε τη μορφή θα προτιμούσαμε να έχουν οι ιστορίες μας.
* Μία ακόμα οπτικοποίηση που θεωρήσαμε σκόπιμη για την μελέτη μας ήταν και η χρήση Ιστογραμμάτων για της στήλες των δεδομένων μας. Τα οποία θα μας επέτρεπαν, για παράδειγμα, να δούμε το ποσοστό της εμφάνισης κάθε εγγραφής μέσα σε κάποια στήλη κ.ο.κ.

**Ερώτημα 2: Ομαδοποίηση**

**α) Clustering**

Σε αυτό το κομμάτι τις εργασίας αποφασίσαμε να δημιουργήσουμε κάποια Clusters. Μετά από αρκετές δοκιμές αποφασίσαμε να υλοποιήσουμε 4 <<ζευγάρια>> στηλών των οποίων τα αποτελέσματα θα μπορούσαν να μας αποκαλύψουν επιπλέον πληροφορία για τα δεδομένα μας. Συγκεκριμένα χρησιμοποιήσαμε τα ζευγάρια:

1. Severity AND Weather Condition
2. Weather\_Condition AND Distance
3. Severity AND State
4. Day\_Night AND Visibility(mi)
5. Side and 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).

(Για όλα τα παραπάνω 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) <Feature(s)>, <Label(s)> (+ Train and test Split)**

Σε αυτό το κομμάτι του δεύτερου ερωτήματος δώσαμε ιδιαίτερη βάση στην σωστή επιλογή των Features και Labels 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), Day\_Night
* Labels = 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 λόγο μεθόδους. Αναλυτικότερα αποφασίσαμε το test να αποτελεί το 20% των δεδομένων μας. Επίσης προβήκαμε και σε μία κανονικοποίηση ώστε να φέρουμε και τα δύο μεταξύ του διαστήματος [0,1] για δικιά μας διευκόλυνση στην ερμηνεία των αποτελεσμάτων.

**γ) Τεχνικές ταξινόμησης**

Στο συγκεκριμένο κομμάτι του ερωτήματος θα χρησιμοποιήσουμε 3 τεχνικές ταξινόμησης όπως μας υποδεικνύεται και από την εκφώνηση:

* Gaussian (Bayes) Classifier
* Random Forest Classifier
* Tree Classifier

Από όλες τις παραπάνω τεχνικές πήραμε τα ανάλογα αποτελέσματα και μπορέσαμε να κάνουμε συγκρίσεις μέσω της ακρίβειας αυτών, δημιουργώντας Confusion Matrixes, αλλά και μέσω του χρόνου του οποίου απαιτούταν για την διεξαγωγή τους. Συγκεκριμένα ο γρηγορότερος όλων αποδείχθηκε ο Gaussian ταξινομητής με χρόνους μικρότερους του ενός λεπτού για όλο το σύνολο των δεδομένων ενώ η ακρίβεια ήταν αρκετά ικανοποιητική για όλες τις τεχνικές αγγίζοντας το 0,88-,090 % βάση των προβλέψεων των οποίων έκαναν. Τέλος για να μπορέσουμε να έχουμε μεγαλύτερη ακρίβεια στην σύγκριση αυτή τρέξαμε τον κώδικα σε 3 διαφορετικούς υπολογιστές με διαφορετική << ιπποδύναμη >>.