ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΣΤΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

Διπλωματική Εργασία

Ανοικτά Κυβερνητικά Δεδομένα Κυκλοφορίας Δρόμων: Διερεύνηση Δεδομένων και Δημιουργία Μοντέλου Πρόβλεψης

της

ΜΟΥΡΑΤΙΔΟΥ ΜΑΡΙΑ

Επιβλέπων καθηγητής: Κ. ΤΑΡΑΜΠΑΝΗΣ

Υποβλήθηκε ως απαιτούμενο για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στα Πληροφοριακά Συστήματα

Μάιος 2023

Αφιερώσεις / Ευχαριστίες

Για την ανάθεση και εκπόνηση της παρούσας εργασίας θέλω να ευχαριστήσω τον επιβλέποντα καθηγητή Κωνσταντίνο Ταραμπάνη, τον επίκουρο καθηγητή Ευάγγελο Καλαμπόκη για τις συμβουλές και τις κατευθυντήριες γραμμές του καθώς και την εργαστηριακή βοηθό Αρετή Καραμάνου.

Τέλος, ευχαριστώ τους γονείς μου και την οικογένεια μου για την στήριξή τους με κάθε τρόπο κατά την διάρκεια φοίτησης για την απόκτηση του μεταπτυχιακού τίτλου σπουδών στα Πληροφοριακά Συστήματα.

Περίληψη

Ο σκοπός της παρούσας διπλωματικής εργασίας, η οποία εκπονήθηκε στα πλαίσια της απόκτησης του μεταπτυχιακού διπλώματος στα Πληροφοριακά Συστήματα, είναι αρχικά η εξερεύνηση των δεδομένων της κυκλοφορίας των δρόμων στην Αττική με άντληση πληροφοριών από την πύλη ανοιχτών κυβερνητικών δεδομένων **data.gov.gr**, η απεικόνιση αυτών των δεδομένων με την χρήση της πλατφόρμας **Tableau** για την δημιουργία δυναμικών και μη γραφημάτων προκειμένου να κατανοήσουμε τα δεδομένα αυτά και το πώς μπορεί να συσχετίζονται μεταξύ τους. Στη συνέχεια αναλύεται το πώς μπορούμε να χρησιμοποιήσουμε το **Climate data store API** του **Copernicus** για να πάρουμε τα δεδομένα καιρού για τα συγκεκριμένα γεωγραφικά μήκη(longitude) και πλάτη(latitude) που μας ενδιαφέρουν, στην προκειμένη για τις συντεταγμένες της Αττικής όπου βρίσκονται οι αισθητήρες των δρόμων που θα μελετήσουμε, και να τα χρησιμοποιήσουμε για να εξετάσουμε τον τρόπο με τον οποίο η κυκλοφορία των δρόμων (αυξημένη κίνηση, μειωμένη ταχύτητα διέλευσης κλπ) επηρεάζεται από τις καιρικές συνθήκες με τελικό στόχο της διπλωματικής εργασίας να δημιουργήσουμε ένα μοντέλο πρόβλεψης της μελλοντικής κυκλοφορίας των δρόμων της Αττικής που επιλέξαμε με βάση τα ιστορικά δεδομένα και τα δεδομένα καιρού με την χρήση της προγραμματιστικής γλώσσας **Python** και αντίστοιχες βιβλιοθήκες.

Η μεθοδολογία που χρησιμοποιήθηκε ακολουθεί την προσέγγιση που διδαχτήκαμε στο μάθημα «Συστήματα Υποστήριξης Αποφάσεων» κατά το 2^{ο} εξάμηνο φοίτησης και τα βήματα της προσέγγισης είναι πρώτα ο προσδιορισμός του προβλήματος, έπειτα η συλλογή δεδομένων, η προετοιμασία, ο καθαρισμός και η εξερεύνησή τους με στόχο την κατανόησή αυτών, η οποία ακολουθείται από τη δημιουργία ενός μοντέλου πρόβλεψης, του **XGBoost** που είναι βασισμένο στην θεωρία των Δέντρων Αποφάσεων (Decision Trees). Μετά την πρόβλεψη γίνεται η αξιολόγηση του μοντέλου και η επεξήγηση των αποτελεσμάτων.

Τελειώνοντας την εργασία καταλήγουμε σε κάποια συμπεράσματα και προτάσεις με βάση τα αποτελέσματα και τις παρατηρήσεις που προέκυψαν από τα γραφήματα και το μοντέλο πρόβλεψης.

Λέξεις κλειδιά: data.gov.gr, tableau, Copernicus data store api, python, κυκλοφορία δρόμων στην Αττική, επιβλεπόμενη μηχανική μάθηση, XGBoost, Regression

Abstract

The purpose of this thesis, which was prepared in the context of acquiring a Master's Degree in Information Systems, is at first the exploration of the road data of Attica which were extracted from the open government data portal **data.gov.gr**, and then the depiction of these data using the platform **Tableau** for the creation of dynamic, or not, graphs and charts in order to better understand the data and how they may be related to each other. We later analyze how we can use the **Climate data store API** of **Copernicus** in order to attain the weather data for the specific longitude and latitude that we desire, in our case the coordinates of Attica where the road censors are placed, and use them to examine the way that the road traffic (increased traffic, lower speed etc) is impacted by the weather with our final goal of this thesis being the creation of a prediction model for the future road traffic in Attica (for the roads that we have selected) based on the historical data and the past weather data using the programming language **Python** and the corresponding libraries.

The methodology which was used follows the approach that we were taught in the course «Decision Support Systems» during the 2nd semester of the Master's program and the steps of this approach is that firstly we define the problem, then we collect the data, we prepare them, we clean them and we explore them in order to understand them, which is then followed by the creation / implementation of a prediction model called **XGBoost** which is based on the theory of the **Decision Trees**. After the prediction an evaluation of the model and the explanation of the results take place.

In the end of this thesis we come up with some conclusions and some suggestions based on the results and the observations that derived from the study, the graphs and the prediction model.

Keywords: data.gov.gr, tableau, Copernicus data store api, road traffic in Attica, supervised machine learning, XGBoost

Πίνακας Περιεχομένων

Πίνακας Εικόνων

1. Εισαγωγή

Τα τελευταία χρόνια έχει αρχίσει να γίνεται όλο και πιο ευρέως γνωστό πως τα δεδομένα που συλλέγονται από δημόσιους και ιδιωτικούς φορείς αποτελούν σημαντική πηγή πληροφοριών καθώς με την κατάλληλη επεξεργασία μπορούμε να μετατρέψουμε καταγεγραμμένα στοιχεία και μετρήσεις σε συγκεκριμένες, κατηγοριοποιημένες, υπολογισμένες και συμπυκνωμένες πληροφορίες που μας ενδιαφέρουν και στη συνέχεια μέσω της κατανόησης αυτών, της εμπειρίας, της διορατικότητας και της τοποθέτησής τους σε εννοιολογικό πλαίσιο να περάσουμε στο επόμενο επίπεδο, της γνώσης. Με αυτόν τον τρόπο μπορούμε να χρησιμοποιήσουμε παρελθοντικά / ιστορικά δεδομένα για να δημιουργήσουμε γραφήματα και μοντέλα πρόβλεψης και με βάση αυτά να είμαστε σε θέση να πάρουμε πιο εμπεριστατωμένες αποφάσεις με μειωμένο ρίσκο.

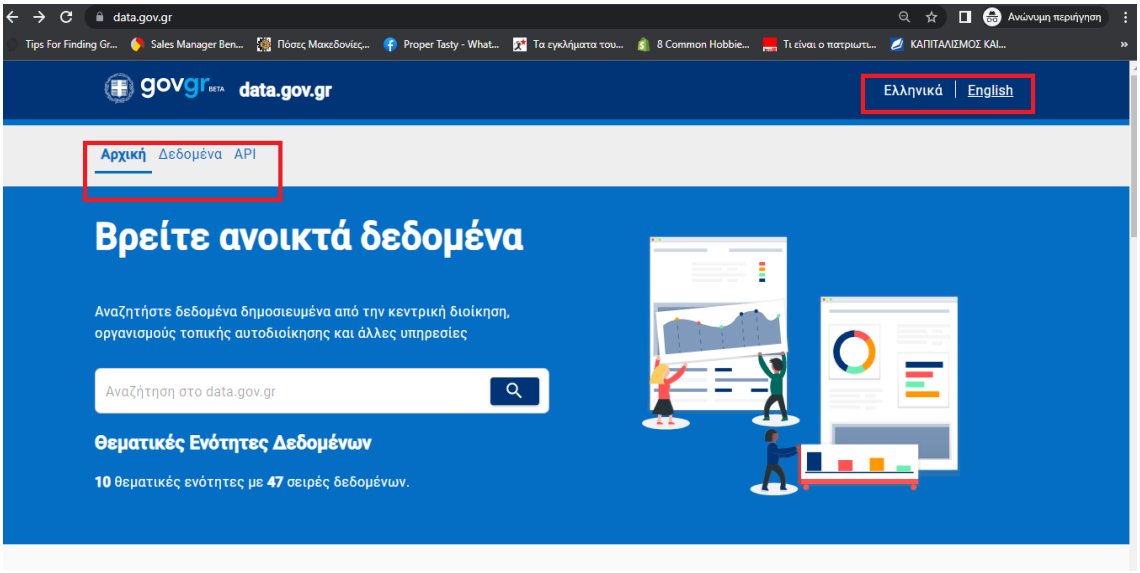
Το κεντρικό θέμα αυτής της διπλωματικής εργασίας αποτελεί το σύνολο δεδομένων «Κυκλοφορία δρόμων στην Αττική» που αντλήθηκε από την πύλη ανοιχτών κυβερνητικών δεδομένων data.gov.gr και περιέχει πληροφορίες όπως ο αριθμός των διελεύσεων αυτοκινήτων ανά ώρα στους δρόμους όπου υπάρχουν οι αισθητήρες και η μέση ταχύτητα διελεύσεων ανά ώρα. Επιλέχτηκε το εργαλείο Google Colaboratory που είναι ένα δωρεάν online εργαλείο παρόμοιο με το Jupyter Notebook που ενδείκνυται για προβλήματα μηχανικής μάθησης, ανάλυση δεδομένων και εκπαιδευτικούς σκοπούς.

Η εργασία εστιάζει στον τρόπο που μπορούμε να πάρουμε δεδομένα από έγκυρες πηγές, την μορφή των δεδομένων (αρχεία csv, διεπαφή api), την προετοιμασία και την μετατροπή τους, την εισαγωγή των δεδομένων στο εργαλείο οπτικοποίησης tableau για την δημιουργία γραφημάτων και απεικονίσεων για την καλύτερη κατανόηση του συνόλου μας, την χρήση της γλώσσας προγραμματισμού Python και των βιβλιοθηκών όπως pandas, math, holidays κλπ και του μοντέλου επιβλεπόμενης μηχανικής μάθησης XGBoost για την πρόβλεψη της μελλοντικής κυκλοφορίας στους δρόμους της Αττικής με βάση τα ιστορικά δεδομένα κυκλοφορίας, τα παρελθοντικά δεδομένα καιρού (όπως η θερμοκρασία και η ταχύτητα του ανέμου) και αν συνυπάρχει κάποια αργία ή γιορτή.

Στο τέλος της εργασίας ακολουθεί η ερμηνεία των ευρημάτων, η αξιολόγηση του μοντέλου μας, τα συμπεράσματα και παρατίθενται αναλυτικά οι αναφορές και η βιβλιογραφία που χρησιμοποιήθηκε.

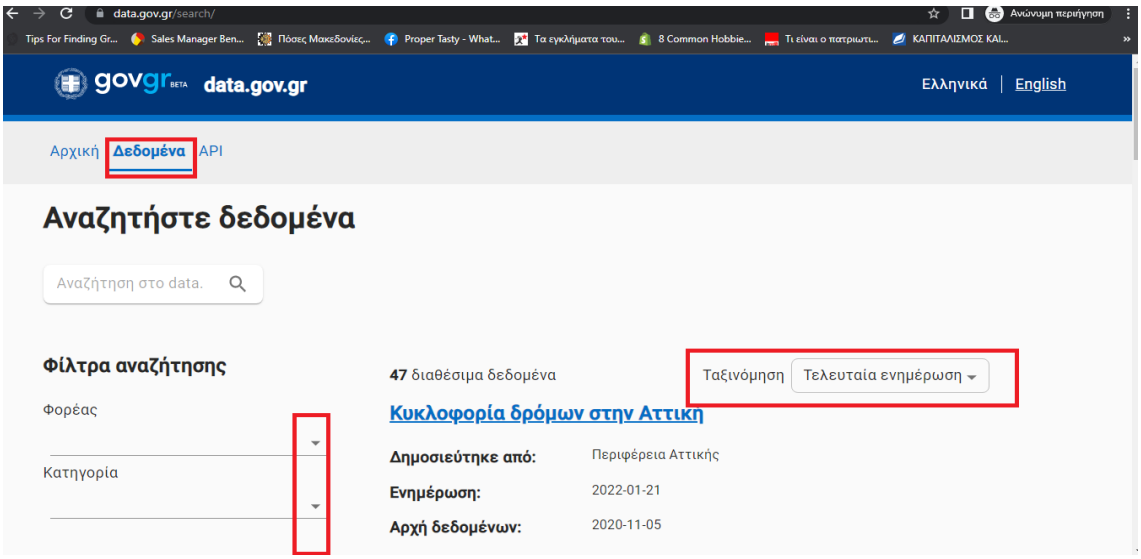
1.1 Πύλη ανοικτών κυβερνητικών δεδομένων data.gov.gr

Τον Δεκέμβριο του 2020 τέθηκε σε λειτουργία το data.gov.gr που είναι μια πύλη με **ανοικτά δεδομένα** και αποτελεί ουσιαστικά, όπως αναγράφεται στο αρχείο της ιστοσελίδας, «έναν κεντρικό κατάλογο δημόσιων δεδομένων που παρέχει πρόσβαση σε βάσεις δεδομένων των φορέων της Ελληνικής Κυβέρνησης. Ο σκοπός του data.gov.gr είναι να αυξηθεί η πρόσβαση σε υψηλής αξίας, μηχανικά αναγνώσιμα σύνολα δεδομένων (datasets) με την παροχή ενιαίων υπηρεσιών καταλογογράφησης, ευρετηρίασης, αποθήκευσης, αναζήτησης και διαθεσιμότητας των δεδομένων και των πληροφοριών δημόσιου τομέα, καθώς και διαδικτυακές υπηρεσίες προς τους πολίτες και τρίτα συστήματα πληροφοριών.». (archive.data.gov.gr, Σχετικά)



Εικόνα 1 Αρχική σελίδα Πύλης Ανοιχτών Κυβερνητικών Δεδομένων

Η πλατφόρμα διαθέτει μέχρι στιγμής **10 θεματικές ενότητες** (Περιβάλλον, Υγεία, Τεχνολογία κ.α.) με **47 σειρές δεδομένων** (Στατιστικά εμβολιασμού Covid-19, Επιβατικό κοινό στον ΟΑΣΑ, Τροχαία ατυχήματα κ.α.). Υπάρχει η δυνατότητα χρήσης **φίλτρων αναζήτησης** με βάση τον φορέα (π.χ. ΔΕΔΗΕ, ΕΦΕΤ, Κτηματολόγιο) ή την κατηγορία των δεδομένων(π.χ. εκπαίδευση, μετακινήσεις, περιβάλλον) και έχει **3 επιλογές ταξινόμησης** με βάση α. την τελευταία ενημέρωση, β. τα πιο σχετικά και γ. τα πιο δημοφιλή. Υπάρχει η **δυνατότητα εναλλαγής της γλώσσας** στα αγγλικά και στα ελληνικά και τα δεδομένα ανανεώνονται **καθημερινά**.



Εικόνα 2 Σελίδα «Δεδομένα» Πύλης Ανοιχτών Κυβερνητικών Δεδομένων

Η διεπαφή είναι εύκολη στην χρήση καθώς κάποιος μπαίνοντας στην ιστοσελίδα μπορεί να πλοηγηθεί στην Αρχική, στα Δεδομένα και στο API. Για πλήρη πρόσβαση και χρήση των δεδομένων απαιτείται η δημιουργία ενός api token από την σελίδα API του μενού.

Όσον αφορά τις άδειες χρήσης της πλατφόρμας πηγαίνοντας στην αρχική σελίδα κι επιλέγοντας το κουμπί πάνω αριστερά που γράφει gov.gr BETA μεταφερόμαστε στο www.gov.gr και κάτω στο footer υπάρχουν διάφορες πληροφορίες όπως το «Όροι Χρήσης». Επιλέγοντας το μας μεταφέρει στην αντίστοιχη σελίδα όπου υπάρχουν αναλυτικά οι «Γενικοί Όροι & Πολιτικές Χρήσης» και η «Πολιτική Προστασίας Προσωπικών Δεδομένων». Επίσης, πηγαίνοντας στο αρχείο του data.gov.gr μπορούμε να βρούμε το θεσμικό πλαίσιο μόνο για την συγκεκριμένη πλατφόρμα το οποίο παρατίθεται αναλυτικά στο τέλος της εργασίας στα προσαρτήματα.

1.1.1 Σύνολο δεδομένων: Κυκλοφορία δρόμων στην Αττική

Το κύριο σύνολο δεδομένων που χρησιμοποιήθηκε για την εργασία ονομάζεται «Κυκλοφορία δρόμων στην Αττική» κι επιλέχθηκε καθώς είναι μεγάλο, περιέχει πολλά δεδομένα και ξέρουμε τι σημαίνουν οι μεταβλητές. Η σύντομη περιγραφή των δεδομένων, όπως αναγράφεται στο data.gov.gr/datasets/road_traffic_attica, είναι «Μετρήσεις αριθμού διελεύσεων και μέσης ταχύτητας ανά σταθμό μέτρησης του

δικτύου παρακολούθησης της κυκλοφορίας στην Αττική» και μας δίνει μια πρώτη ιδέα για τις πληροφορίες που περιέχονται.

Ακολουθώντας τα βήματα που περιγράφονται στην επόμενη ενότητα κι επαναλαμβάνοντας την διαδικασία μεταφόρτωσης των αρχείων 8 φορές έχουμε πλέον 7 αρχεία csv από τις 4.6.2021 μέχρι τις 31.12.2021. Αυτό είχε σαν αποτέλεσμα να απαιτείται μια προεργασία των δεδομένων πριν μπορέσουμε να περάσουμε στο στάδιο της εξερεύνησης.

Αρχικά, πρέπει να ενώσουμε τα πολλαπλά αρχεία csv σε ένα ενιαίο αρχείο csv. Αυτό μπορούμε να το επιτύχουμε με διάφορους τρόπους, μερικοί από τους οποίους είναι:

i. μέσω του command panel
ii. μέσω του excel από την καρτέλα Δεδομένα
iii. μέσω του google sheets online κάνοντας εισαγωγή των αρχείων είτε από τον drive είτε μεταφορτώνοντας αρχεία από τον υπολογιστή κι έπειτα από την καρτέλα «Εισαγωγή Αρχείου» επιλέγοντας από το drop down menu Εισαγωγή Τοποθεσίας την «Προσάρτηση στο τρέχων φύλλο»
iv. μέσω της συγχώνευση αρχείων csv με online εργαλεία (π.χ. merge-csv.com ή extendclass.com)
v. μέσω εντολών με την κάποια γλώσσα προγραμματισμού (π.χ. python)

Στην προκειμένη αρχικά χρησιμοποιήθηκε η τελευταία τεχνική συγχώνευσης των αρχείων μέσω του περιβάλλοντος ανάπτυξης ολοκληρωμένης πλατφόρμας ανοιχτού κώδικα Spyder για επιστημονικό προγραμματισμό στην γλώσσα Python(Spyder, Wikipedia) καθώς λόγω του μεγάλου όγκου δεδομένων ήταν ο πιο αποτελεσματικός και γρήγορος τρόπος. Αφότου κατέβασα τα αρχεία csv τα έβαλα μέσα σε έναν φάκελο που τον ονόμασα CSV και κάνοντας δεξί κλικ στον φάκελο αυτό πήγα στις ιδιότητες κι αντέγραψα την θέση του αρχείου. Έπειτα, άνοιξα το λογισμικό Spyder κι έβαλα τον παρακάτω κώδικα:

```
import os
import glob
import pandas as pd
os.chdir("C:/Users/MAPIA/Downloads/Διπλωματική αρχεία/CSV")
```

```
extension = 'csv'
all_filenames = [i for i in glob.glob('*.{}'.format(extension))]

#combine all files in the list
combined_csv = pd.concat([pd.read_csv(f) for f in all_filenames ])
#export to csv
combined_csv.to_csv( "combined_csv.csv", index=False, encoding='utf-8-sig')
```

Αφού «έτρεξε» ο κώδικας πηγαίνοντας στον φάκελο CSV είχε δημιουργηθεί ένα συνολικό αρχείο με το όνομα «combined_csv.csv». Το αρχείο περιέχει περισσότερες από 1.048.576 γραμμές ή 16.384 στήλες, που είναι το μέγιστο όριο που μπορεί να εμφανιστεί σε ένα ενιαίο φύλλο του Excel, οπότε μας εμφανίζει προειδοποιητικό μήνυμα πως δεν μπορεί να τα εμφανίσει όλα.

Για να διαπιστώσουμε αν έγινε σωστά η συγχώνευση δοκιμάστηκε ένας ακόμη τρόπος. Περάσαμε τα 7 διαφορετικά αρχεία στο online εργαλείο Google Colaboratory το οποίο θα χρησιμοποιήσουμε και στην συνέχεια της διπλωματικής και θα το δούμε πιο αναλυτικά στο κεφάλαιο 2.2.2. και 2.2.3. Εκεί αφότου φορτώθηκε ξεχωριστά το κάθε csv και αποθηκεύτηκε εκ νέου σαν pandas dataframe με το όνομα της επιλογής μας προχωρήσαμε στην συνένωση και των 7 αρχείων μέσω της εντολής pd.concat() όπως φαίνεται παρακάτω

```
frames = [june_road_data, july_road_data, august_road_data, september_road_data, oct
ober_road_data, november_road_data, december_road_data]

result = pd.concat(frames)

print(result)
```

Προσθέτοντας τις εγγραφές του κάθε συνόλου μπορούμε να επιβεβαιώσουμε πως η συνένωση έγινε σωστά βάση του αποτελέσματος.

Με μια πρώτη ματιά μπορούμε εύκολα να δούμε πως υπάρχουν 1.680.421 σειρές και 6 στήλες σε αυτό το σύνολό δεδομένων, εκ των οποίων η πρώτη στήλη είναι το «deviceid» κι αντιπροσωπεύει το πλήθος των αισθητήρων που καταγράφουν την διέλευση των οχημάτων ανά ώρα και την μέση ταχύτητα διέλευσης των οχημάτων. Αυτές οι συσκευές είναι 425 συνολικά, έχουν τοποθετηθεί σε 93 δρόμους και καταγράφουν τα παραπάνω στοιχεία ανά ώρα.

Η δεύτερη στήλη ονομάζεται «countedcars» και είναι το πλήθος των οχημάτων που διέσχισαν έναν συγκεκριμένο δρόμο στη διάρκεια μίας ώρας.

Η τρίτη στήλη ονομάζεται «appprocesstime» και μας ενημερώνει για την ημερομηνία και την ώρα της κάθε καταγραφής.

Η τέταρτη στήλη ονομάζεται «road_name» και περιέχει το όνομα του κάθε δρόμου στον οποίο υπάρχει ο αντίστοιχος αισθητήρας και ακριβώς από δίπλα υπάρχει η πέμπτη στήλη που ονομάζεται «road_info» και μας παρέχει περαιτέρω πληροφορίες σχετικά με τον κάθε κύριο δρόμο. Για να γίνει πιο κατανοητή η πέμπτη στήλη των δεδομένων παρατίθεται σαν παράδειγμα ο δρόμος Λ. Κηφισού που είναι ένας πολύ μεγάλος κεντρικός δρόμος της Αττικής και κατ' επέκταση για να είναι ακριβείς οι μετρήσεις έχει τοποθετηθεί ένα πλήθος αισθητήρων σε συγκεκριμένα σημεία του δρόμου τα οποία περιγράφονται στην στήλη «road_info». Ειδικότερα, στην Λ. Κηφισού υπάρχουν 24 αισθητήρες τοποθετημένοι σε σημεία που περιγράφονται όπως «Κύριος δρόμος με κατεύθυνση Πειραιά μετά τη ράμπα εξόδου της Λ. Κηφισού προς Αγ. Ιω. Ρέντη» ή «Ράμπα εξόδου προς οδό Πειραιώς και γέφυρα αναστροφής του κλάδου της Λ. Κηφισού με κατεύθυνση Πειραιά».

Η έκτη στήλη ονομάζεται «average_speed» και περιέχει την μέση ταχύτητα διέλευσης των οχημάτων σε χιλιόμετρα ανά ώρα με βάση την μέτρηση κάθε αισθητήρα.

Εφόσον έχουμε πλέον τα δεδομένα που θέλουμε να χρησιμοποιήσουμε σε μορφή csv σε ένα ενοποιημένο αρχείο μπορούμε να τα εισάγουμε στο λογισμικό που επιθυμούμε και να ξεκινήσουμε την διερεύνηση των δεδομένων.

1.2 Copernicus Climate Data Store API

Το Copernicus Climate Data Store API (CDS API) ανήκει στο C3S (Copernicus Climate Change Service) της Ευρωπαϊκής Ένωσης και σύμφωνα με την επίσημη ιστοσελίδα προσφέρει «πληροφορίες για το κλίμα που αφορούν το παρελθόν, το παρόν και το μέλλον. Παρέχει εύκολη πρόσβαση σε μια ευρεία γκάμα κλιματικών δεδομένων μέσω ενός καταλόγου αναζήτησης. Μια online εργαλειοθήκη είναι διαθέσιμη κι επιτρέπει στους χρήστες να χτίσουν ροές εργασιών και εφαρμογές προσαρμοσμένες στις ανάγκες τους».

Η βασική λογική του CDS είναι πως παρέχει πρόσβαση μέσω ενός σημείου σε μια ευρεία γκάμα κλιματικών συνόλων δεδομένων υψηλής ποιότητας που υπάρχουν κατανεμημένα στο «σύννεφο» (cloud). Ανάμεσα σε αυτά τα σύνολα περιέχονται παρατηρήσεις, καταγραφές ιστορικών κλιματικών δεδομένων, υπολογισμοί βασικών κλιματικών μεταβλητών που προέρχονται από παρατηρήσεις της γης, παγκόσμιες και τοπικές αναλύσεις παρελθοντικών κλιματικών παρατηρήσεων, εποχικές προβλέψεις και κλιματικές προβολές. Η πρόσβαση στα δεδομένα είναι ανοιχτή, δωρεάν και χωρίς περιορισμούς.

Έχει σχεδιαστεί για να υποστηρίζει χρήστες με διαφορετικές ανάγκες βοηθώντας στην επεξεργασία μεγάλων όγκων δεδομένων και στην δημιουργία απλών οπτικοποιήσεων με δεδομένα προερχόμενα από πολλαπλές πηγές. Τα δεδομένα και τα εργαλεία του CDS αποτελούν την βάση του Κλαδικού Πληροφοριακού Συστήματος (C3S Sectoral Information System) το οποίο παρέχει εργαλεία και εφαρμογές για την αντιμετώπιση της κλιματικής επίδρασης σε διαφορετικούς βιομηχανικούς τομείς συμπεριλαμβανομένης της διαχείρισης του νερού, της ενέργειας και της γεωργίας.

Εδώ αξίζει να αναφέρουμε την λογική της ενασχόλησης με το CDS στην παρούσα εργασία καθώς μπορούμε να πάρουμε δεδομένα καιρού για συγκεκριμένες γεωγραφικές συντεταγμένες (γεωγραφικό μήκος και πλάτος).

Παρακάτω καθώς και στο 2^{ο} κεφάλαιο βλέπουμε πιο αναλυτικά πως μπορούμε να αποκτήσουμε πρόσβαση, την δημιουργία λογαριασμού, την αναζήτηση κι επιλογή του συνόλου δεδομένων που επιθυμούμε, στην προκειμένη το σύνολο «ERA5-Land hourly data from1950 to present» και την άντληση δεδομένων μέσω του API.

1.2.1 Σύνολο δεδομένων: ERA5-Land hourly data from 1950 to present

Το CDS API δίνει πρόσβαση στην «οικογένεια» συνόλου δεδομένων «ERA5» που περιλαμβάνει το σύνολο «ERA5» και το σύνολο «ERA5-land», το οποίο είναι το σύνολο για την στεριά μόνο.

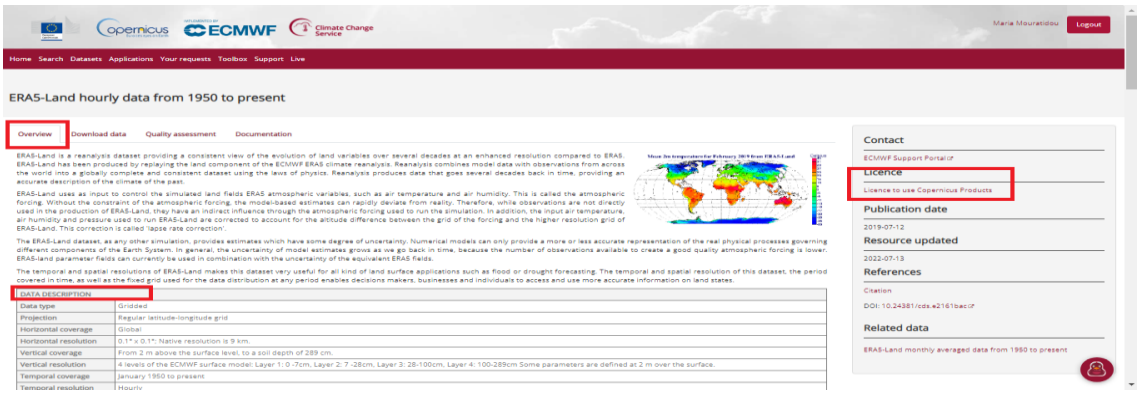
Το «ERA5» είναι η ατμοσφαιρική επανανάλυση 5^{ης} γενιάς του παγκόσμιου κλίματος που καλύπτει την περίοδο από τον Ιανουάριο του 1950 μέχρι σήμερα. Παράγεται από το Copernicus Climate Change Service (C3S) στο ECMWF, ή Ευρωπαϊκό Κέντρο Προγνώσεων Καιρού Μεσαίου Εύρους που είναι ένας ανεξάρτητος διακυβερνητικός οργανισμός που υποστηρίζεται από τα περισσότερα έθνη της Ευρώπης.

Στο «ERA5» υπάρχουν 4 κύρια υποσύνολα στο «ERA5-land» υπάρχουν 2 κύρια υποσύνολα. Για την παρούσα εργασία επιλέχθηκε το υποσύνολο «ERA5-land hourly data» καθώς περιέχει τα μετεωρολογικά στοιχεία που χρειαζόμαστε σε καθημερινή και ωριαία βάση ώστε να μπορέσουμε να εξετάσουμε το αν υπάρχει κάποια συσχέτιση μεταξύ της κυκλοφορίας στους δρόμους της Αττικής και του καιρικών φαινομένων, όπως π.χ. άνεμος, βροχή, πολύ υψηλή / χαμηλή θερμοκρασία, χιονόπτωση.

Για αυτό το σύνολο δεδομένων και τα υποσύνολά του υπάρχουν οι μηνιαίες ενημερώσεις με εξασφάλιση ποιότητας που δημοσιεύονται εντός 3 μηνών σε πραγματικό χρόνο. Επίσης, παρέχονται και προκαταρκτικές ημερήσιες ενημερώσεις του συνόλου δεδομένων που είναι διαθέσιμες στους χρήστες εντός 5 ημερών σε πραγματικό χρόνο.

Στην ιστοσελίδα του CDS API αφότου επιλέξουμε το σύνολο δεδομένων που επιθυμούμε να εξερευνήσουμε παρέχονται πολύ αναλυτικές πληροφορίες όπως φαίνεται και στην κάτω εικόνα.

Αρχικά, βλέπουμε πως στην καρτέλα «Overview» υπάρχει μια σύνοψη του συνόλου δεδομένων, όπως π.χ. ο τρόπος που παράγονται τα δεδομένα που συνδυάζει μοντελοποιημένα δεδομένα και παρατηρήσεις από όλο τον κόσμο σε ένα συνεπές σύνολο δεδομένων χρησιμοποιώντας τους νόμους της φυσικής και άλλα τεχνικά χαρακτηριστικά.



Εικόνα 3 Σελίδα «Overview» του Copernicus Climate Data Store Api

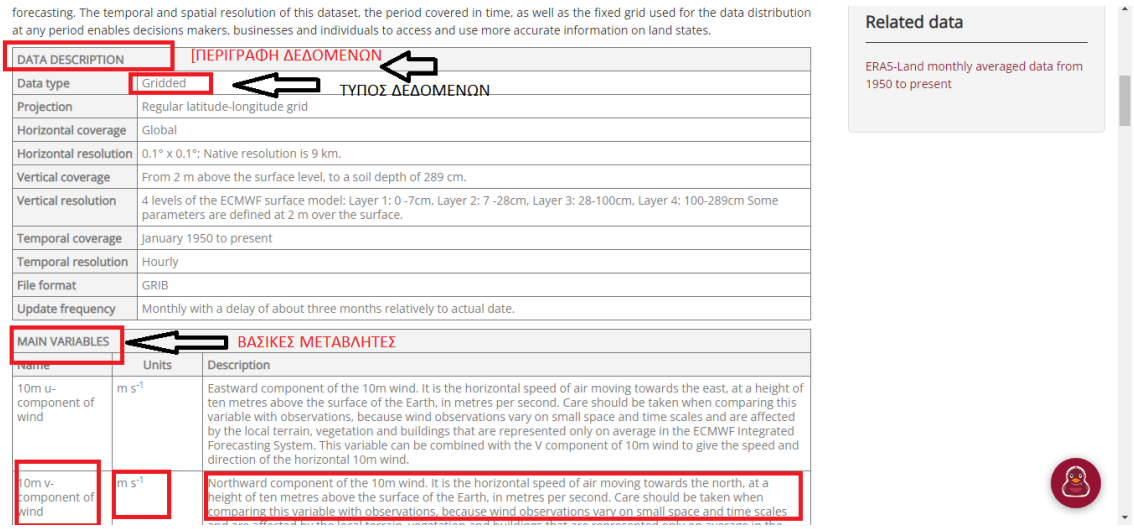
Στο δεξί μέρος της σελίδας υπάρχουν διαθέσιμοι σε pdf η άδεια χρήσης των προϊόντων του Copernicus η οποία είναι διαθέσιμη στο παράρτημα.

Όπως αναγράφεται στην ιστοσελίδα, «το σύνολο δεδομένων ERA5-Land, όπως κάθε άλλη προσομοίωση, παρέχει εκτιμήσεις που έχουν κάποιο βαθμό αβεβαιότητας. Τα αριθμητικά μοντέλα μπορούν μόνο να παρέχουν μια περισσότερο ή λιγότερο ακριβή

αναπαράσταση των πραγματικών φυσικών διεργασιών που διέπουν τα διάφορα στοιχεία του Συστήματος της Γης. Γενικά, η αβεβαιότητα των εκτιμήσεων των μοντέλων μεγαλώνει όσο πηγαίνουμε πίσω στο χρόνο, επειδή ο αριθμός των διαθέσιμων παρατηρήσεων για τη δημιουργία καλής ποιότητας ατμοσφαιρικής πίεσης είναι μικρότερος.

Η χρονική και χωρική ανάλυση αυτού του συνόλου δεδομένων, η χρονική περίοδος που καλύπτεται, καθώς και το σταθερό πλέγμα που χρησιμοποιείται για τη διανομή δεδομένων σε οποιαδήποτε περίοδο επιτρέπει στους υπεύθυνους λήψης αποφάσεων, στις επιχειρήσεις και στα άτομα να έχουν πρόσβαση και να χρησιμοποιούν πιο ακριβείς πληροφορίες για τις πολιτείες γης.»

Επίσης υπάρχει περιγραφή των δεδομένων, του τύπου των δεδομένων μας και των βασικών μεταβλητών μαζί με την επεξήγησή τους και τις μονάδες μέτρησης.



Εικόνα 4 Περιγραφή δεδομένων του συνόλου δεδομένων που έχουμε επιλέξει

Πιο συγκεκριμένα, ο τύπος δεδομένων είναι «gridded» δηλαδή επιτρέπει την αποθήκευση των δεδομένων με τρόπο που μοιάζει με πλέγμα και δίνει στους χρήστες τη δυνατότητα πρόσβασης, τροποποίησης και μεταφοράς εξαιρετικά μεγάλων ποσοτήτων δεδομένων με γεωγραφική κατανομή για σκοπούς έρευνας.

Το αρχείο που μπορούμε να κατεβάσουμε ονομάζεται «grib» και είναι μια μορφή που χρησιμοποιείται για την αποθήκευση και τη μεταφορά μετεωρολογικών δεδομένων. Αφότου υποστούν προεπεξεργασία μπορούν να χρησιμοποιηθούν για οπτικοποιήσεις ή ως «input» για αριθμητικές εφαρμογές πρόβλεψης καιρού ή γενικά σε μοντέλα πρόβλεψης.

Για την δική μας μελέτη περίπτωσης τα δεδομένα που μας ενδιαφέρουν είναι τα εξής και θα τα δούμε πιο αναλυτικά στο 2ο κεφάλαιο:

I. 2m temperature,
II. snowfall,
III. 10m u-component of wind,
IV. 10m v-component of wind,
V. total precipitation.

Οι μονάδες μέτρησης για την θερμοκρασία είναι σε Κελβιν, για την χιονόπτωση και την συνολική βροχόπτωση είναι τα χιλιοστά (mm) και για τους πλευρικούς ανέμους είναι τα m/s ^ -1.

1.3 Πλατφόρμα οπτικοποίησης δεδομένων Tableau

Το Tableau είναι ένα διαδραστικό λογισμικό οπτικοποίησης δεδομένων που επικεντρώνεται στην Επιχειρηματική Ευφυΐα (Business Intelligence) και δημιουργήθηκε το 2003 από την αμερικάνικη εταιρεία Mountain View στην Καλιφόρνια. Το 2019 εξαγοράστηκε από την εταιρεία Salesforce έναντι του ποσού των 15.7 δισεκατομμυρίων δολαρίων. (wiki)

Ένα σύνολο δεδομένων, στο γενικό πλαίσιο του Tableau, περιέχει δεδομένα που χρησιμοποιούνται για να δημιουργηθούν οπτικοποιήσεις. Μπορεί να υποστηρίξει την σύνδεση με σχεσιακές βάσεις δεδομένων, με την λογική της OLAP ανάλυσης (διαστάσεις και μετρούμενα μεγέθη), με βάσεις δεδομένων στο Cloud, με υπολογιστικά φύλλα (excel, google sheets), με αρχεία json και διάφορους ακόμη τύπους και μορφές δεδομένων που αναφέρονται αναλυτικά στα προσαρτήματα.

Υπάρχουν διάφορες εκδόσεις και προϊόντα εκ των οποίων τα πιο ευρέως διαδεδομένα είναι:

- **Tableau Desktop**

  η βασική, πλήρης έκδοση του Tableau όπου επιτρέπεται η δημιουργία διαδραστικών φύλλων εργασίας (worksheets) και ταμπλό (dashboards)

- **Tableau Public**

η δωρεάν έκδοση του Tableau Desktop όπου δεν επιτρέπεται η τοπική αποθήκευση των βιβλίων εργασίας

- **Tableau Prep**

διευκολύνει τον καθαρισμό και την μετατροπή των δεδομένων

- **Tableau Online**

η Cloud εκδοχή του Tableau Desktop με κάποιους περιορισμούς

- **Tableau Server**

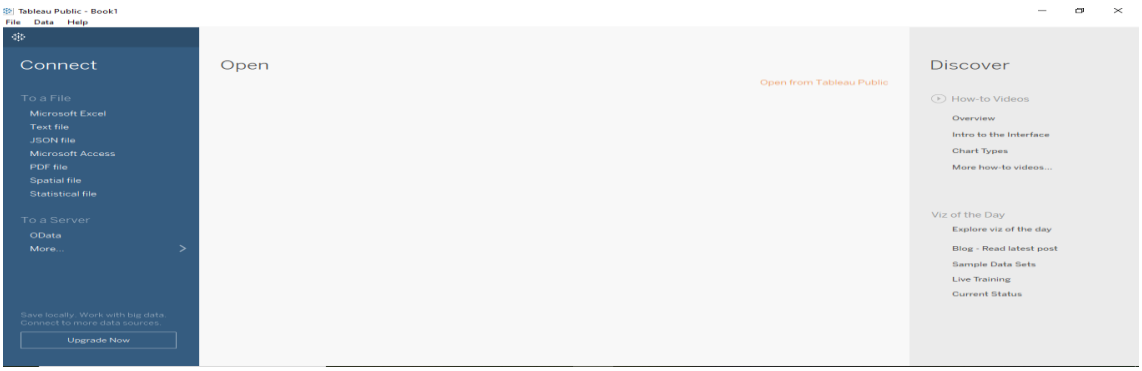
επιτρέπεται η δημοσίευση, ο διαμοιρασμός και η διαχείριση του περιεχομένου του Tableau Desktop

- **Tableau Reader**

δωρεάν εφαρμογή με την οποία μπορείς να ανοίξεις απεικονίσεις δεδομένων που δημιουργήθηκαν στο Tableau Public και να αλληλεπιδράσεις με αυτές μέσω φίλτρων, ιεράρχησης τύπου drill down κλπ

Στην παρούσα διπλωματική εργασία χρησιμοποιήθηκε η έκδοση Tableau Public 2021.4 που προσφέρεται δωρεάν στην ιστοσελίδα της εταιρείας. Μπαίνοντας στην αρχική συμπληρώνουμε το email μας και πατάμε το κουμπί «download the app». Το αρχείο κατεβαίνει σε .exe, ολοκληρώνουμε την εγκατάσταση και είναι έτοιμο για χρήση.

Όταν ανοίξουμε την εφαρμογή για πρώτη φορά αυτό είναι το παράθυρο που εμφανίζεται:



Εικόνα 5 Αρχική σελίδας της πλατφόρμας οπτικοποίησης δεδομένων Tableau

Στο 3ο κεφάλαιο θα αναλυθεί περαιτέρω η διεπαφή του λογισμικού, η εισαγωγή των δεδομένων, η δημιουργία sheets και dashboards, ο τρόπος αποθήκευσης αυτών, η μετατροπή της στατικής πληροφορίας σε δυναμική, το ποιος είναι ο κατάλληλος τρόπος απεικόνισης και η λογική των μετρούμενων μεγεθών και διαστάσεων.

1.4 Χρήση της γλώσσας Python για δημιουργία μοντέλου πρόβλεψης

Αφότου πάρουμε τα δεδομένα που επιθυμούμε από τις πηγές που έχουμε επιλέξει και τα έχουμε πλέον στην κατάλληλη μορφή για επεξεργασία, κι έπειτα από την εξερεύνηση και την κατανόηση αυτών μέσω των οπτικοποιήσεων και των γραφημάτων, γίνεται μια προσπάθεια δημιουργίας ενός μοντέλου πρόβλεψης για την μελλοντική κυκλοφορία συγκεκριμένων δρόμων στην Αττική με την χρήση της γλώσσας προγραμματισμού Python.

Ο κώδικας που θα χρησιμοποιηθεί για την δημιουργία του μοντέλου είναι βασισμένος στην εργασία «Προβλέποντας και επεξηγώντας τις αφίξεις στην Εντατική ενός Δημοσίου Νοσοκομείου» του Σπυρίδων Πετσή, της Αρετής Καραμάνου, του Ευάγγελου Καλαμπόκη και του Κωνσταντίνου Ταραμπάνη, η οποία δημοσιεύτηκε από την Springer Nature το 2021 στα πλαίσια του εργαστηρίου Πληροφοριακών Συστημάτων του Πανεπιστημίου Μακεδονίας.

Θα εξετάσουμε επίσης τον τρόπο με τον οποίο τα έντονα καιρικά φαινόμενα όπως η βροχόπτωση, η χιονόπτωση, οι δυνατοί άνεμοι και οι πολύ υψηλές θερμοκρασίες επηρεάζουν την ταχύτητα με την οποία κινούνται τα αυτοκίνητα στους κεντρικούς δρόμους της Αττικής. Επίσης, θα εξετάσουμε την πιθανότητα ύπαρξης κάποιας συσχέτισης με την μειωμένη ή αυξημένη κυκλοφορία ανάλογα με την ώρα της ημέρας ή την ημέρα της εβδομάδας και τον ρόλο των εθνικών εορτών ή / και των αργιών.

2. Ανασκόπηση Βιβλιογραφίας

Προκειμένου να αντιληφθούμε την λογική που ακολουθείτε στην εκπόνηση της διπλωματικής γίνεται μια σύντομη αναφορά στο θεωρητικό υπόβαθρο της εργασίας, τι περιλαμβάνεται σε αυτήν και στην προσέγγιση που επιλέχθηκε.

2.1 Ανοιχτά κυβερνητικά δεδομένα

Για την παρούσα εργασία αρχικά χρησιμοποιήθηκε το σύνολο δεδομένων «Κυκλοφορία δρόμων στην Αττική» από την πλατφόρμα data.gov.gr που φιλοξενεί σύνολα από ανοιχτά κυβερνητικά δεδομένα.

Εδώ θα δώσουμε έναν ορισμό για τα Ανοικτά Δεδομένα (Open Data) , όπου σύμφωνα με την WorldBank «ένα κομμάτι δεδομένων ή περιεχομένου είναι ανοικτό αν μπορεί να χρησιμοποιηθεί ελεύθερα από τον καθένα, να επαναχρησιμοποιηθεί και να αναδιανεμηθεί , και να υπόκειται μόνο, το πολύ, στην απαίτηση να αποδίδονται ή να παρέχονται οι σωστές παραπομπές ή/ και να αναδιανεμηθεί υπό τους ίδιους όρους και προϋποθέσεις». (“Open Data for Economic Growth”, The WorldBank, June25, 2014.)

Όσον αφορά το data.gov.gr προηγήθηκε ολικός ανασχεδιασμός της πλατφόρμας προκειμένου τα δεδομένα που παρέχονται να είναι άμεσα αξιοποιήσιμα για ερευνητικούς και επιχειρηματικούς σκοπούς. (Ναυτεμπορική, 22 Δεκεμβρίου 2020).

2.2 Δεδομένα καιρού από το Copernicus

Για να μπορέσουμε να εξετάσουμε την υπόθεσή μας ως προς το εάν ο καιρός παίζει ρόλο στην κυκλοφοριακή συμφόρηση πήραμε το σύνολο δεδομένων «ERA5-Land hourly data from 1950 to present» από το Copernicus επιλέγοντας την γεωγραφική περιοχή και την χρονική περίοδο που θέλουμε να εξετάσουμε.

Ακολουθεί μια σύντομη παράθεση για την θεωρητική υπόστασης όπου σύμφωνα με την επίσημη ιστοσελίδα «το Copernicus είναι το πρόγραμμα γεωσκόπησης της Ευρωπαϊκής Ένωσης, το οποίο παρατηρεί το περιβάλλον και τον πλανήτη μας προς όφελος όλων των ευρωπαίων πολιτών. Παρέχει υπηρεσίες πληροφόρησης με βάση δορυφορικά δεδομένα γεωσκόπησης και επίγεια (μη διαστημικά) δεδομένα.

Τον συντονισμό και τη διαχείριση του προγράμματος έχει αναλάβει η Ευρωπαϊκή Επιτροπή. Το πρόγραμμα υλοποιείται σε συνεργασία με τα κράτη μέλη, τον Ευρωπαϊκό Οργανισμό Διαστήματος (ESA), τον Ευρωπαϊκό Οργανισμό Εκμετάλλευσης Μετεωρολογικών Δορυφόρων (EUMETSAT), το Ευρωπαϊκό Κέντρο Μεσοπρόθεσμων Μετεωρολογικών Προβλέψεων (ECMWF), οργανισμούς της ΕΕ και την εταιρεία Mercator Océan.

Το πρόγραμμα χρησιμοποιεί μεγάλες ποσότητες παγκόσμιων δεδομένων προερχόμενων από δορυφορικά και από επίγεια, αερομεταφερόμενα και θαλάσσια συστήματα μέτρησης, για την παροχή πληροφοριών που βοηθούν τους παρόχους υπηρεσιών, τις δημόσιες αρχές και άλλους διεθνείς οργανισμούς να βελτιώσουν την ποιότητα ζωής των πολιτών της Ευρώπης. Οι χρήστες του προγράμματος έχουν ελεύθερη και απρόσκοπτη πρόσβαση στις παρεχόμενες υπηρεσίες πληροφόρησης.» («Λίγα λόγια για το Copernicus», Copernicus.eu, available at https://www.copernicus.eu/el/liga-logia-gia-copernicus)

2.3 Μηχανική Μάθηση

Η μελέτη περίπτωσης που επιλέχθηκε για την συγκεκριμένη διπλωματική εμπίπτει στο πεδίο της Μηχανικής Μάθησης που αποτελεί μια υποκατηγορία της Τεχνητής Νοημοσύνης. Σύμφωνα με την Sap η Μηχανική Μάθηση «επικεντρώνεται στη διδασκαλία υπολογιστών για να μαθαίνουν από τα δεδομένα και να βελτιώνονται με την εμπειρία – αντί να είναι ρητά προγραμματισμένοι να το κάνουν. Στη μηχανική μάθηση, οι αλγόριθμοι εκπαιδεύονται για να βρίσκουν μοτίβα και συσχετίσεις σε μεγάλα σύνολα δεδομένων και να λαμβάνουν τις καλύτερες αποφάσεις και προβλέψεις βάσει αυτής της ανάλυσης. Οι εφαρμογές μηχανικής εκπαίδευσης βελτιώνονται με τη χρήση και γίνονται πιο ακριβείς όσο περισσότερα δεδομένα έχουν πρόσβαση.». («Τεχνητή Νοημοσύνη, Τι είναι η Μηχανική Μάθηση», sap.com, available at https://www.sap.com/greece/products/artificial-intelligence/what-is-machine-learning.html)

Πιο συγκεκριμένα το δικό μας πρόβλημα εμπίπτει στην Επιβλεπόμενη Μηχανική Μάθηση όπου σύμφωνα με την Wikipedia «το υπολογιστικό πρόγραμμα δέχεται τις παραδειγματικές εισόδους καθώς και τα επιθυμητά αποτελέσματα από έναν «δάσκαλο», και ο στόχος είναι να μάθει έναν γενικό κανόνα προκειμένου να αντιστοιχίσει τις εισόδους με τα αποτελέσματα.». Στο 4^{ο} κεφάλαιο αναλύεται λεπτομερώς το θεωρητικό υπόβαθρο που μας οδήγησε στην επιλογή του αλγορίθμου XGBoost για την εκπόνηση της μελέτης.

2.4 Μεθοδολογία

Η λογική της μεθοδολογίας μας είναι η εύρεση των πηγών, η άντληση των συνόλων δεδομένων που μας ενδιαφέρουν και η προετοιμασία αυτών προκειμένου να είναι στην

κατάλληλη μορφή για την περεταίρω χρήση στο μοντέλο μας, ακολουθούν κάποιες οπτικοποιήσεις στο Tableau και στην γλώσσα Python για να διερευνήσουμε καλύτερα τα δεδομένα μας και στο τέλος προχωράμε στην δημιουργία του μοντέλου πρόβλεψης.

2.4.1 Άντληση Δεδομένων

Το πρώτο βήμα της μεθοδολογίας μας είναι η άντληση των δεδομένων κυκλοφορίας από το data.gov.gr και η άντληση των δεδομένων καιρού από το Copernicus. Μετά την άντληση τα αποθηκεύουμε εκ νέου και προχωράμε στην επεξεργασία και στην προετοιμασία των δεδομένων, στον καθαρισμό αυτών και στην μετατροπή τους στην εκάστοτε επιθυμητή μορφή.

2.4.2 Οπτικοποιήσεις Tableau

Εφόσον έχουμε πλέον τα σύνολα που θέλουμε να εξερευνήσουμε στην μορφή και στον βαθμό λεπτομέρειας που επιθυμούμε προχωράμε σε κάποιες οπτικοποιήσεις μέσω του εργαλείου Tableau για την καλύτερη κατανόηση των δεδομένων μας, την διερεύνηση πιθανών προβλημάτων, όπως ακραίων τιμών, που διακρίνονται πιο ξεκάθαρα με τις οπτικοποιήσεις και στην διερεύνηση πιθανών σχέσεων ανάμεσα στις μεταβλητές μας.

2.4.3 Δημιουργία μοντέλου πρόβλεψης

Έχοντας πλέον μια καλύτερη κατανόηση του συνόλου μας μέσω της προετοιμασίας των δεδομένων μας, του καθαρισμού και των οπτικοποιήσεων προχωράμε στην δημιουργία του μοντέλου πρόβλεψης. Έχουμε πλέον ένα ενοποιημένο σύνολο δεδομένων με τις στήλες που έχουμε επιλέξει να κρατήσουμε και από τα δύο σύνολα καθώς κρίναμε πως αυτές θα χρειαστούμε. Προχωράμε στον διαχωρισμό ενός test και ενός train κομματιού των δεδομένων και χρησιμοποιούμε διάφορες τεχνικές για να βρούμε τις καλύτερες παραμέτρους και υπερπαραμέτρους του αλγορίθμου.

2.5 Αποτελέσματα μοντέλου πρόβλεψης

Σε αυτό το σημείο αφότου έχουμε τρέξει πολλές φορές το μοντέλο πρόβλεψης με διάφορες τιμές και τροποποιήσεις προκειμένου να πάρουμε τα καλύτερα δυνατά αποτελέσματα και να αποφύγουμε όσο γίνεται το πρόβλημα του overfitting και του underfitting. Προχωρήσαμε στην δημιουργία ενός πίνακα αποτελεσμάτων για να αποκτήσουμε μια πληρέστερη εικόνα ως προς το πώς επηρεάζεται ο αλγόριθμος XGBoost από τις αλλαγές στις τιμές των παραμέτρων και των υπερπαραμέτρων. Σταματήσαμε τις επαναλήψεις όταν τα αποτελέσματα ήταν κοντά σε θεμιτές τιμές.

2.6 Αξιολόγηση μοντέλου πρόβλεψης

Τέλος, περάσαμε στην αξιολόγηση του μοντέλου πρόβλεψης μέσω των λεγόμενων error metrics, δηλαδή μετρήσεων σφάλματος, εφόσον το πρόβλημά μας ανήκει στην κατηγορία Παλινδρόμησης (Regression) καθώς πρόκειται για πρόβλεψη τιμής. Ουσιαστικά αξιολογούμε την ικανότητα πρόβλεψης του μοντέλου μας και την απόκλιση που έχουν οι τιμές που προέκυψαν από αυτό σε σχέση με τις πραγματικές τιμές των ιστορικών δεδομένων στις οποίες εκπαιδεύτηκε ο αλγόριθμος.

Οι μέθοδοι μέτρησης αξιολόγησης που επιλέξαμε είναι το Μέσο Απόλυτο Σφάλμα (Mean Absolute Error - MAE), η Μέση Τετραγωνική Ρίζα Σφάλματος (Root Mean Squared Error - RMSE) και το Μέσο Απόλυτο Ποσοστιαίο Σφάλμα (Mean Absolute Percentage Error - MAPE). Η εργασία κλείνει με την ερμηνεία των ευρημάτων, τα συμπεράσματα, τους περιορισμούς της μελέτης και κάποιες μελλοντικές προτάσεις προς διερεύνηση.

3. Μεθοδολογία

Η μεθοδολογία που επιλέχθηκε για την εκπόνηση της διπλωματικής εργασίας βασίζεται στην προσέγγιση του μαθήματος «Συστήματα Υποστήριξης Αποφάσεως» του μεταπτυχιακού στα Πληροφοριακά Συστήματα, στο βιβλίο «Business Intelligence, Analytics, and Data Science: A managerial perspective» του Ramesh Sharda, Dursun Deller και Efraim Turban και σε διάφορες έγκυρες διαδικτυακές πηγές που αναφέρονται στην βιβλιογραφία στο τέλος της εργασίας.

Σύμφωνα με τα παραπάνω, η λογική που ακολουθούμε είναι οι 4 φάσεις σε ένα πρόβλημα επιστήμης δεδομένων. Αρχικά ξεκινάμε με την πρώτη φάση που είναι

εύρεση έγκυρων πηγών προκειμένου να προχωρήσουμε στη συλλογή και στην αποθήκευσης των συνόλων δεδομένων που μας ενδιαφέρουν.

Στην δεύτερη φάση συνεχίζουμε με την προετοιμασία των δεδομένων, όπου αυτό απαιτείται, και είναι συνήθως η πιο χρονοβόρα διαδικασία αλλά ταυτόχρονα και η πιο σημαντική καθώς τα αποτελέσματα του μοντέλου μας εξαρτώνται πολύ από την ποιότητα των δεδομένων που τροφοδοτούνται σε αυτό. Σε αυτήν την φάση ¨καθαρίζουμε¨ τα δεδομένα μας από συχνά προβλήματα όπως ελλείπουσες τιμές, εύρεση ακραίων τιμών, διπλές καταχωρήσεις κλπ και ουσιαστικά τα μετατρέπουμε σε πιο οργανωμένη μορφή.

Ακολουθεί η τρίτη φάση, όπου ανάλογα με την προσέγγιση μπορεί να προηγείται της δεύτερης ή και να εναλλάσσονται, όπου γίνεται η αρχική διερεύνηση των καθαρισμένων δεδομένων μέσω της περιγραφικής στατιστικής και της οπτικοποίησης για να γίνουν αντιληπτά και να ελεγχθούν τα δεδομένα (σελ. 65, 5ο μάθημα, ΣΥΑ 2020) π.χ. μέσω των εντολών df.head(), df.describe, df.mean(), df.shape και διαφόρων βιβλιοθηκών για οπτικοποιήσεις στην Python όπως η matplotlib που χρησιμοποιεί διάφορες ενσωματωμένες συναρτήσεις για δημιουργία γραφημάτων όπως το pyplot κλπ, καθώς και με την χρήση του λογισμικού Tableau Public.

Η τελευταία φάση που ακολουθεί είναι η τέταρτη όπου αφού έχουμε κατανοήσει το πρόβλημα και με βάση τα δεδομένα μας προχωράμε στην δημιουργία του μοντέλου πρόβλεψης και την εκτέλεση πειραμάτων. Σε αυτό το σημείο αφού έχουμε επιλέξει τις πιο σημαντικές μεταβλητές που χρειαζόμαστε για τα αποτελέσματα που περιμένουμε από το μοντέλο μας είναι απαραίτητο να μετατρέψουμε τις κατηγορικές μεταβλητές σε αριθμητικές καθώς οι αλγόριθμοι μηχανικής μάθησης απαιτούν αριθμητικούς τύπους δεδομένων.

Σε αυτή την ενότητα βλέπουμε κυρίως βλέπουμε τις τρεις πρώτες φάσεις της προσέγγισης του προβλήματος επιστήμης δεδομένων και στην επόμενη ενότητα θα περάσουμε στην τέταρτη φάση της προσέγγισής μας. Πιο συγκεκριμένα ακολουθούν η φάση της συλλογής και αποθήκευσης εκ νέου, ο αρχικός καθαρισμός, η συνένωση και η διερεύνηση των δεδομένων.

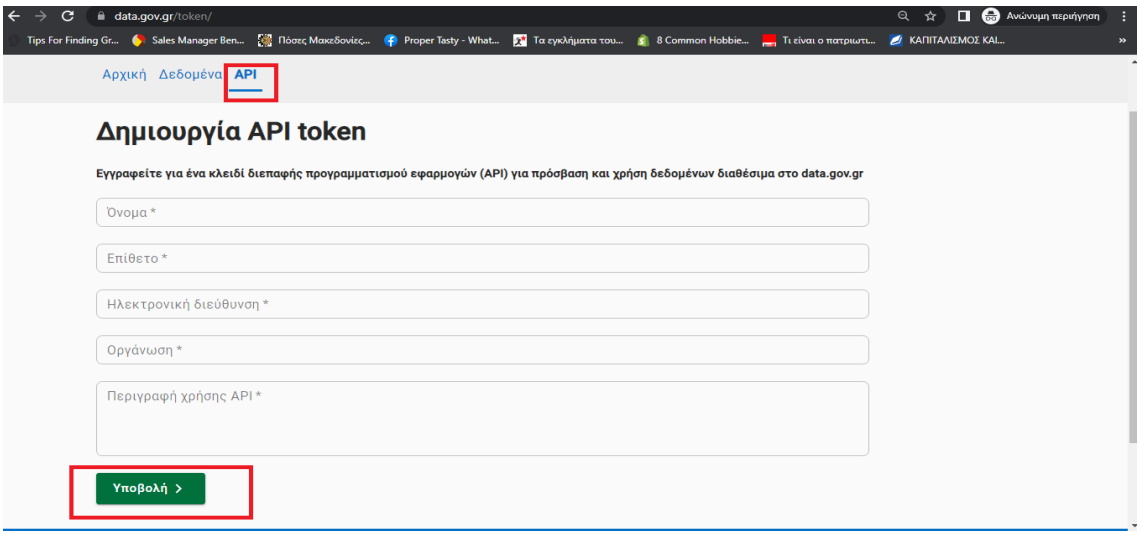
3.1 Άντληση δεδομένων (συλλογή και αποθήκευση)

Σε αυτή την ενότητα θα δούμε αναλυτικά τον τρόπο που προχωρήσαμε στην άντληση των δεδομένων μας από τις ιστοσελίδες των οργανισμών που έχουμε επιλέξει καθώς θεωρούνται αξιόπιστες πηγές δεδομένων, τις δυσκολίες που αντιμετωπίστηκαν στην διαδικασία συλλογής και αποθήκευσης. Αυτή η φάση είναι ιδιαιτέρως σημαντική και χρονοβόρα αλλά και απαραίτητη ώστε να μπορέσουμε να συγκεντρώσουμε τις πληροφορίες που χρειαζόμαστε για να τροφοδοτήσουμε το μοντέλο πρόβλεψης και να περάσουμε στην ανάλυση των δεδομένων μας, στις οπτικοποιήσεις, στην ερμηνεία των ευρημάτων μας και τέλος στα συμπεράσματα που προκύπτουν μέσα από την παρατήρηση και τις επιστημονικές μεθόδους που ακολουθούνται στην παρούσα εργασία.

3.1.1 Άντληση δεδομένων από data.gov.gr

Προκειμένου να μπορέσουμε να αντλήσουμε τα δεδομένα που θέλουμε από την πλατφόρμα data.gov.gr πρέπει να έχουμε ένα προσωπικό api key οπότε αναλύεται ο τρόπος απόκτησης του api key και ο τρόπος χρήσης του.

Πηγαίνοντας στην σελίδα API εμφανίζεται μια φόρμα στην οποία συμπληρώνουμε τα στοιχεία που ζητούνται στα αντίστοιχα 5 πεδία.



Εικόνα 6 Σελίδα Api Ανοιχτών Κυβερνητικών Δεδομένων

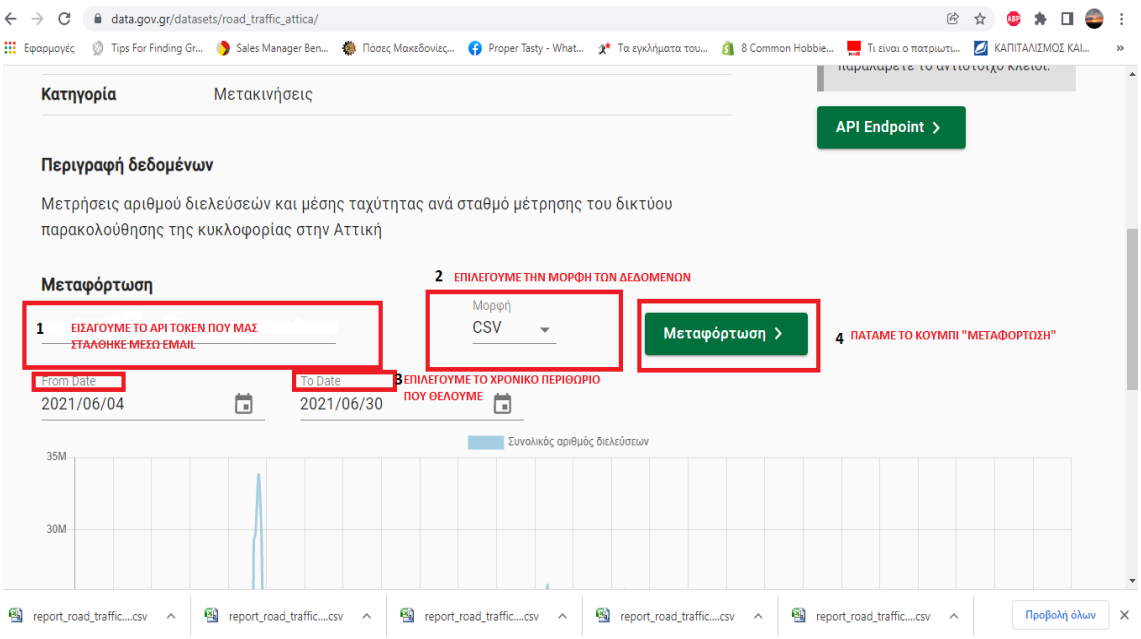
Αυτά είναι το όνομα, το επίθετο, η ηλεκτρονική διεύθυνση, η οργάνωση στην οποία ανήκουμε (π.χ. εταιρεία ή πανεπιστήμιο) και η περιγραφή χρήσης του API (π.χ. χρήση δεδομένων για εκπόνηση διπλωματικής εργασίας). Μόλις συμπληρώσουμε την φόρμα πατάμε το κουμπί «Υποβολή», εμφανίζεται το μήνυμα «Επιτυχής υποβολή, ελέγξτε τα

μηνύματά σας!» και αποστέλλεται στην ηλεκτρονική διεύθυνση που καταχωρήσαμε προηγουμένως ένα email με το Api Token.

Αυτό το Api Token είναι προσωπικό για τον καθένα και είναι ουσιαστικά σαν ένα επαναχρησιμοποιούμενο κλειδί, καθώς δεν έχει συγκεκριμένη ημερομηνία λήξης, το οποίο μας δίνει την δυνατότητα να κατεβάσουμε τοπικά το σύνολο δεδομένων που μας ενδιαφέρει για οποιαδήποτε χρήση.

Το σύνολο δεδομένων που θα ασχοληθούμε, όπως αναφέρθηκε και προηγουμένως, είναι η κυκλοφορία δρόμων στην Αττική οπότε είτε μπορούμε να το ψάξουμε στην Αρχική είτε στα Δεδομένα. Πηγαίνοντας στο landing page «Δεδομένα» μπορούμε να χρησιμοποιήσουμε το φίλτρο αναζήτησης, στην προκειμένη χρησιμοποιήθηκε το φίλτρο «Μετακινήσεις», κι εμφανίζει τα αντίστοιχα σύνολα δεδομένων. Πατώντας πάνω στο σύνολο που μας ενδιαφέρει εμφανίζεται ο τίτλος του συνόλου, από ποια Περιφέρεια δημοσιεύθηκε, πότε ήταν η τελευταία ενημέρωση του συνόλου, την αρχή των δεδομένων (ημερομηνία) και την κατηγορία στην οποία ανήκει.

Από κάτω υπάρχει μια σύντομη περιγραφή των δεδομένων και ακολουθούν οι επιλογές μεταφόρτωσης.



Εικόνα 7 Περιγραφή δεδομένων κυκλοφορίας και διαδικασία άντλησης

Στην Μεταφόρτωση έχει ένα κενό πεδίο στο οποίο πρέπει να βάλουμε το Api Token που μας έχει σταλεί μέσω email, έπειτα επιλέγουμε την μορφή των δεδομένων που θέλουμε να πάρουμε, η οποία είναι είτε csv file είτε json file, και από κάτω ακριβώς

επιλέγουμε την ημερομηνία, δηλαδή από πότε θέλουμε να ξεκινήσουμε να παίρνουμε δεδομένα και πότε θέλουμε να είναι η τελευταία ημερομηνία των δεδομένων που θα πάρουμε και στο τέλος πατάμε το κουμπί «Μεταφόρτωση».

Αν το σύνολο δεδομένων είναι πολύ μεγάλο όταν πατήσουμε «Μεταφόρτωση» εμφανίζεται το παρακάτω προειδοποιητικό μήνυμα «Το χρονικό διάστημα για αυτό το σύνολο δεδομένων σε κάθε αίτημα πρέπει να είναι μικρότερο από αυτό που ζητήθηκε. Προσπαθήστε να εκτελέσετε πολλαπλά αιτήματα για μικρότερα και διαδοχικά χρονικά διαστήματα». Σε άλλα σύνολα δεδομένων της πλατφόρμας δεν εμφανίστηκε αυτό το μήνυμα (π.χ. ενεργειακό ισοζύγιο) αλλά στο συγκεκριμένο εμφανίστηκε καθώς περιέχει πάρα πολλά δεδομένα κυκλοφορίας από τις 5.11.2020, οπότε για να μπορέσουμε να κατεβάσουμε τα δεδομένα μας βρήκα μετά από πειραματισμούς πως το βέλτιστο χρονικό περιθώριο είναι η διάρκεια ενός μήνα επαναλαμβάνοντας την διαδικασία μεταφόρτωσης των αρχείων όσες φορές επιθυμούμε για να πάρουμε τα δεδομένα για τα αντίστοιχα χρονικά διαστήματα που θέλουμε.

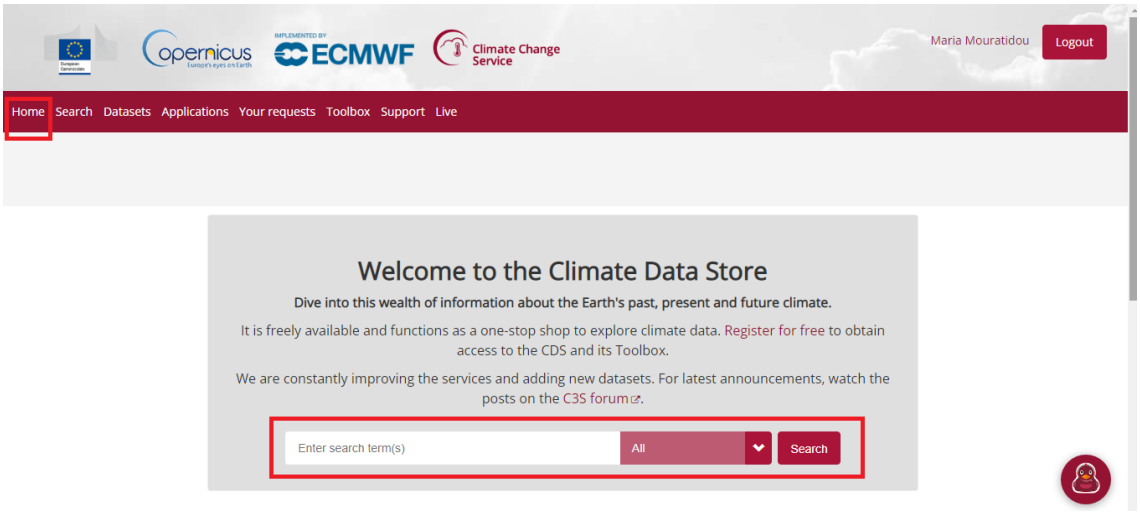
Πέρα από την επιλογή μεταφόρτωσης των δεδομένων σε αρχεία csv ή json μας δίνεται η δυνατότητα να πάρουμε τα δεδομένα που θέλουμε μέσω του api(application programming interface). Στο δεξί μέρος της ίδιας σελίδας υπάρχει η «Πρόσβαση Api» και από κάτω το κουμπί «Api Endpoint». Μόλις πατήσουμε πάνω στο «Api Endpoint» μας εμφανίζει στα αγγλικά τις οδηγίες πρόσβασης στο Api και περιέχει δύο παραδείγματα, ένα στην γλώσσα Javascript όπου μπορούμε να κάνουμε ένα απλό ajax request μέσω της jquery και ένα στην γλώσσα Python.

Στην παρούσα εργασία, τα δεδομένα αντλήθηκαν μέσω του 1^{ου} τρόπου που περιγράφεται και όχι μέσω του Api Endpoint.

3.1.2 Άντληση δεδομένων από Copernicus Climate data store api

Για να μπορέσουμε να αποκτήσουμε πρόσβαση στα δεδομένα μέσω του API του CDS πρέπει να έχουμε λογαριασμό χρήστη οπότε απαιτείται η εγγραφή στην πλατφόρμα. Πηγαίνοντας στην Αρχική σελίδα του Copernicus Climate Data Store πάνω δεξιά υπάρχει ένα κουμπί «Σύνδεση / Εγγραφή» που όταν το πατήσεις σε ανακατευθύνει σε άλλη σελίδα και έχει τρεις επιλογές, την σύνδεση, την δημιουργία λογαριασμού και την επαναφορά κωδικού. Επιλέγουμε την «Δημιουργία νέου λογαριασμού» και συμπληρώνουμε τα στοιχεία που ζητούνται στη φόρμα τα οποία είναι το email, το όνομα, το επώνυμο, η χώρα, ο τομέας (π.χ. Ακαδημαϊκός / Ερευνητικός), τον

Οργανισμό και συμπληρώνουμε το CAPCHA πριν υποβάλλουμε την φόρμα. Επίσης, πρέπει να αποδεχτούμε τους Όρους Χρήσης και την Δήλωση Προστασίας Δεδομένων και Απορρήτου και μπορούμε να επιλέξουμε να εγγραφούμε εκ μέρους κάποιου οργανισμού. Μόλις πατήσουμε το κουμπί υποβολής «Δημιουργία νέου λογαριασμού» εμφανίζεται το μήνυμα «Ένα μήνυμα καλωσορίσματος με περεταίρω οδηγίες έχει σταλεί στην ηλεκτρονική σας διεύθυνση» και μας ανακατευθύνει στην Αρχική σελίδα.



Εικόνα 8 Αρχική σελίδα αναζήτησης δεδομένων του Copernicus Data Climate Store Api

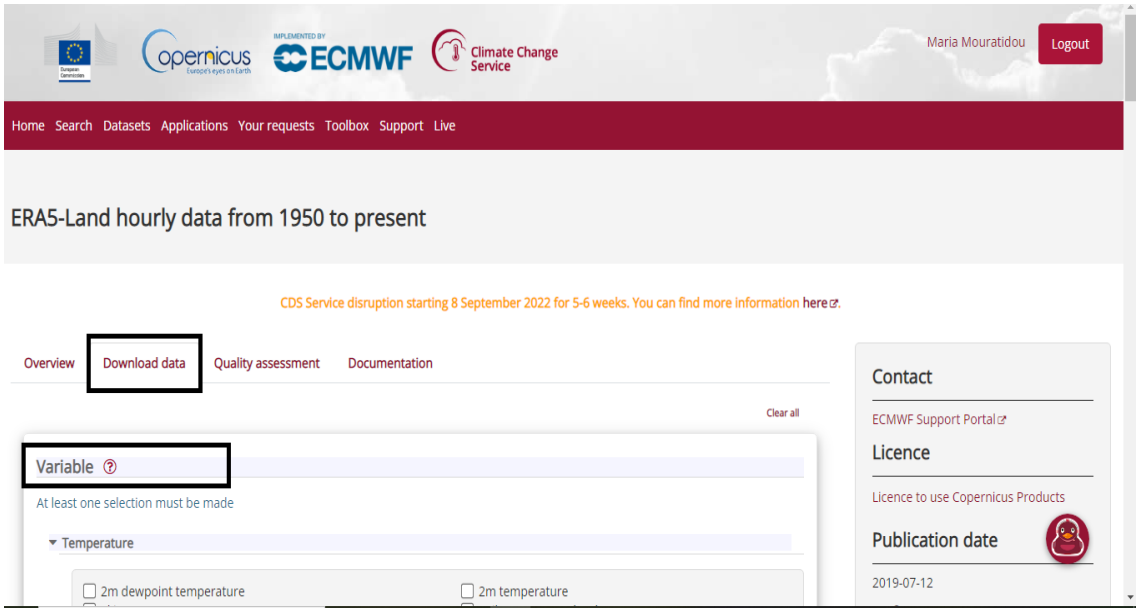
Ανοίγοντας το email που μας έχει σταλεί υπάρχει ένας σύνδεσμος ο οποίος λειτουργεί μόνο την πρώτη φορά που θα πραγματοποιήσουμε την σύνδεση μας και μας οδηγεί σε μια σελίδα όπου θα ορίσουμε τον προσωπικό κωδικό πρόσβασης. Ορίζουμε τον κωδικό ο οποίος προτείνεται να περιέχει τουλάχιστον 12 χαρακτήρες, πεζά, κεφαλαία και αριθμούς και επιλέγουμε το κουμπί «Αποθήκευση και είσοδος ως Όνομα Χρήση».

Στο site υπάρχουν αναλυτικές οδηγίες εγκατάστασης και χρήσης του API ανάλογα με το λειτουργικό σύστημα (Windows, Mac, Linux). Στην περίπτωσή μας ακολουθούμε τις οδηγίες εγκατάστασης για τα Windows και τα προαπαιτούμενα είναι:

- Ένας λογαριασμός CDS
- Η γλώσσα Python εγκατεστημένη
- Η προσθήκη της εγκατάστασης και ο ορισμός ενός PATH μέσω του Command Prompt
- Η προσθήκη της εγκατάστασης και ο ορισμός ενός PATH μέσω του λογισμικού Anaconda

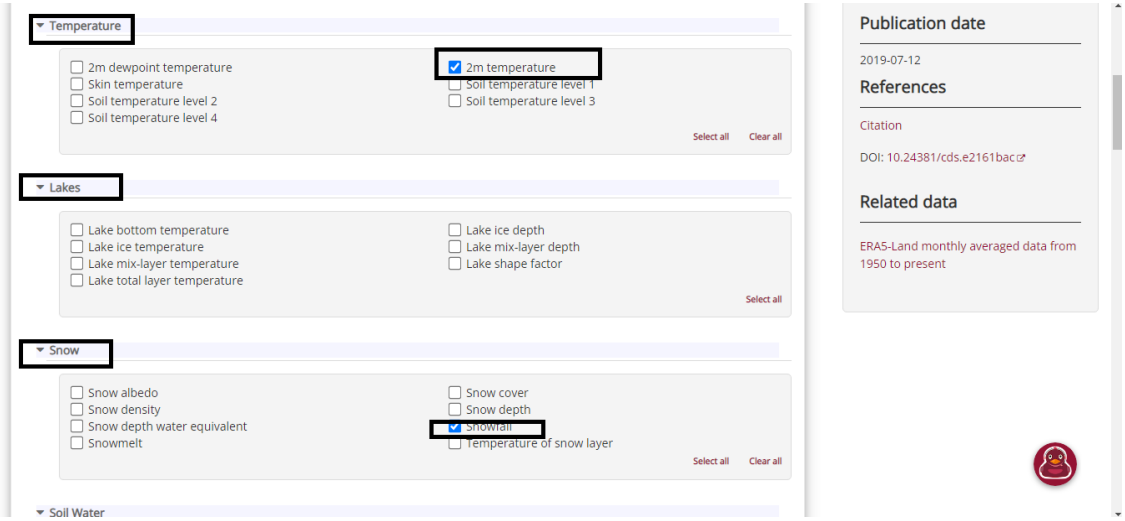
Για να βρούμε το προσωπικό API key αφότου πραγματοποιήσουμε την σύνδεσή μας στην πλατφόρμας πατάμε επάνω δεξιά στο Όνομα Χρήστη και μας ανακατευθύνει στο User Profile όπου κάτω υπάρχει το UID και το API key.

Αφότου βρούμε το προσωπικό μας API Key μπορούμε πλέον να κατεβάσουμε τα δεδομένα που θέλουμε. Όπως αναφέρεται στο 1ο κεφάλαιο για την εργασία μας επιλέχθηκε το σύνολο δεδομένων “Era5 Land Hourly Data from 1950 to present”. Πηγαίνουμε στην καρτέλα “Datasets” και έχουμε την επιλογή να ψάξουμε χειροκίνητα τα διαθέσιμα σύνολα δεδομένων ή να χρησιμοποιήσουμε την μπάρα αναζήτησης. Επιλέγουμε το σύνολο μας και μεταφερόμαστε στην κεντρική σελίδα “Overview” με τις πληροφορίες για το dataset, που καλύφθηκε αναλυτικά στο 1ο κεφάλαιο και πατώντας στην καρτέλα “Download Data” μας εμφανίζει τις μεταβλητές που μπορούμε να επιλέξουμε και αναγράφεται πως πρέπει να επιλέξουμε τουλάχιστον μία μεταβλητή, την χρονολογία από το 1950 μέχρι σήμερα, τον μήνα, τις μέρες, την ακριβή ώρα, την γεωγραφική περιοχή και το format που θα έχουνε τα δεδομένα μας. Ακολουθούν μερικά στιγμιότυπα οθόνης από το Cds.



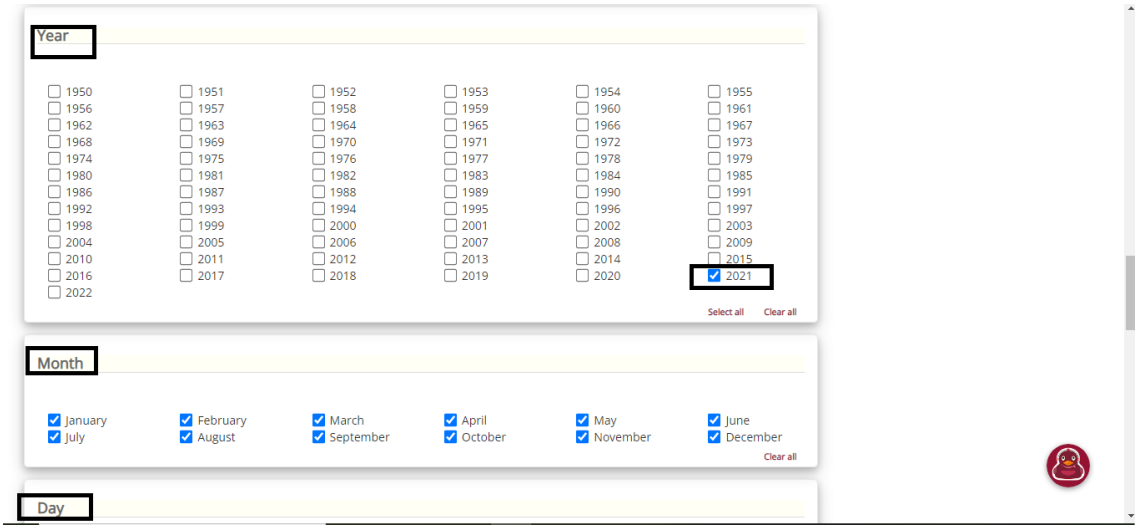
Εικόνα 9 Άντληση δεδομένων καιρού και επιλογή μεταβλητών που επιθυμούμε

Επιλέγουμε τις μεταβλητές που μας ενδιαφέρουν από την κάθε κατηγορία και αυτές είναι skin 2m temperature, snowfall, 10m u-component of wind, 10m v-component of wind, total precipitation.



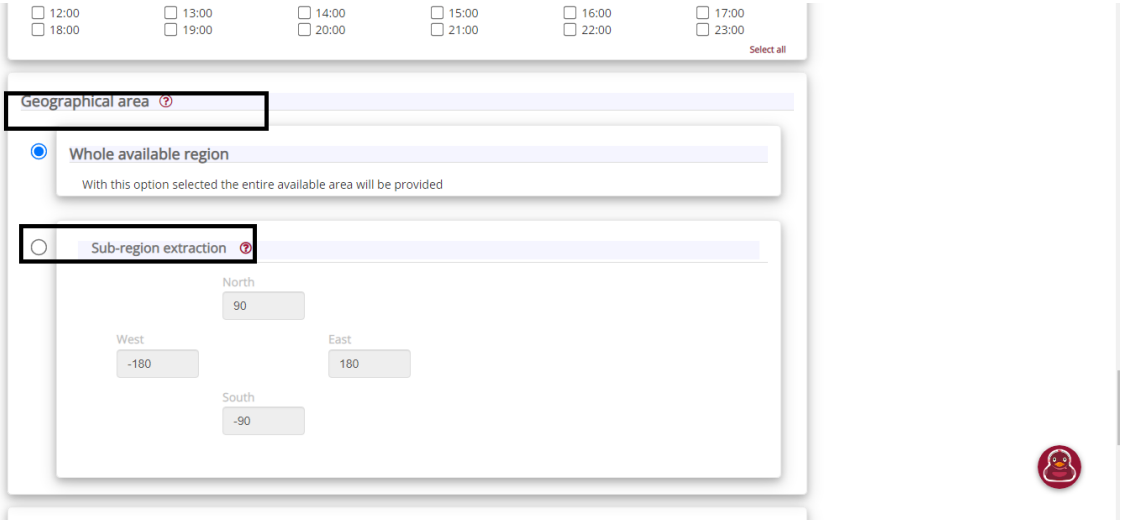
Εικόνα 10 Άντληση δεδομένων καιρού και επιλογή μεταβλητών που επιθυμούμε

Συνεχίζουμε και επιλέγουμε την χρονιά που μας ενδιαφέρει, δηλαδή το 2021, τους μήνες που αν τους θέλουμε όλους μπορούμε να πατήσουμε το “Select All” , στην δική μας περίπτωση επιλέχθηκαν οι μήνες από τον Ιούνιο του 2021 μέχρι τον Δεκέμβριο του 2021. Συνεχίζουμε επιλέγοντας όλες τις μέρες του μήνα και έπειτα όλες τις ώρες της μέρας σε 24ωρη μορφή.



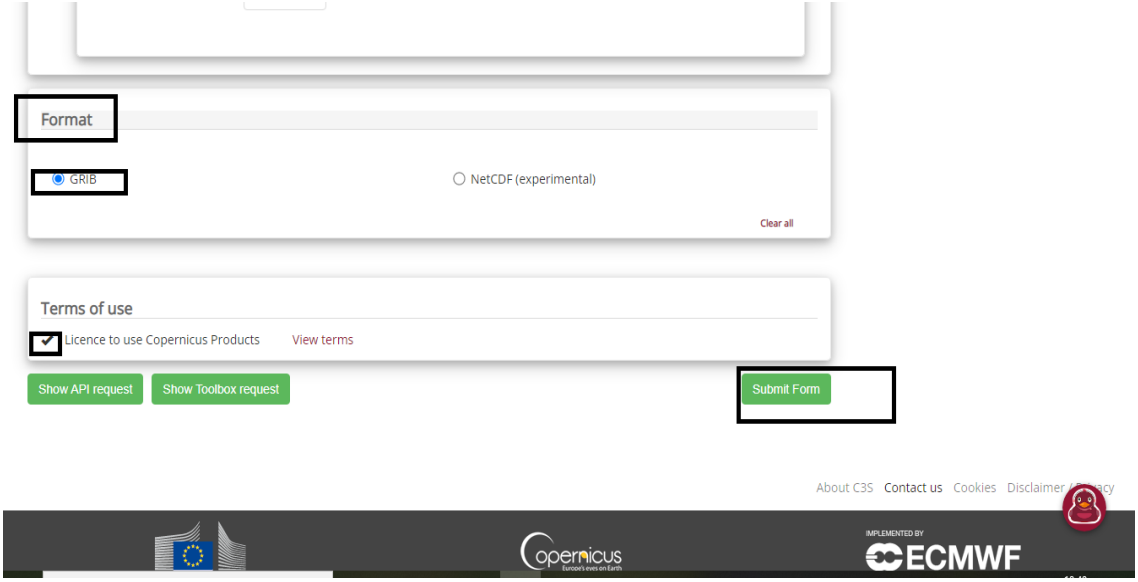
Εικόνα 11 Άντληση δεδομένων καιρού και επιλογή μεταβλητών που επιθυμούμε

Μετά μπορούμε να επιλέξουμε αν μας ενδιαφέρει ολόκληρη η διατιθέμενη γεωγραφική περιοχή ή μια πιο συγκεκριμένη sub region συμπληρώνοντας τις γεωγραφικές μας συντεταγμένες.



Εικόνα 12 Άντληση δεδομένων καιρού και επιλογή μεταβλητών που επιθυμούμε

Οι γεωγραφικές συντεταγμένες της Αθήνας είναι: Γεωγραφικό πλάτος(Βοράς): **37.983917º** και Γεωγραφικό(Ανατολή) μήκος: **23.72936º.** Βάζουμε αυτές τις συντεταγμένες στα πεδία North και East αντίστοιχα. Συνεχίζουμε επιλέγοντας την μορφή GRIB και αποδεχόμαστε τους όρους χρήσης πριν προχωρήσουμε στην υποβολή του αιτήματος μας.



Εικόνα 13 Άντληση δεδομένων καιρού και επιλογή μεταβλητών που επιθυμούμε

Στο επόμενο βήμα αφότου κάνουμε την υποβολή μπορούμε να πάμε στην καρτέλα “Your requests” από το κεντρικό μενού και να δούμε σε τι κατάσταση βρίσκεται το αίτημα μας και αν έχει ολοκληρωθεί.

Όπως φαίνεται από πάνω πέρα από την επιλογή «Submit Form» μπορούμε να πάρουμε τα δεδομένα μας επίσης με 2 ακόμη τρόπους μέσω κώδικα. Ο 2ος τρόπος είναι να πατήσουμε στο "Show api request" και μας εμφανίζει τον κώδικα για να κατεβάσουμε το δεδομένα που επιλέξαμε μέσω του Api endpoint και ο 3ος τρόπος είναι αν πατήσουμε στο «Show Toolbox request» μας δείχνει τον κώδικα που μπορούμε να χρησιμοποιήσουμε στον Online Editor που έχει το Toolbox του Cds Api προκειμένου να πάρουμε τα δεδομένα μας,

Στην προκειμένη εργασία χρησιμοποιήθηκε ο 1ος τρόπος ανάκτησης δεδομένων που αναφέραμε. Το αίτημα μας μπορεί να πάρει αρκετές ώρες ανάλογα με το μέγεθος των δεδομένων που θέλουμε. Πηγαίνοντας από το κεντρικό μενού στην καρτέλα "Your requests" και στην καρτέλα "All" βλέπουμε τα αιτήματα που έχουμε κάνει, αν είναι σε εξέλιξη και πόση ώρα κατεβαίνουν, αν έχουν μπει στην ουρά λόγω πολλών αιτημάτων, αν έχουν αποτύχει από κάποια δυσλειτουργία ή αν έχουνε ολοκληρωθεί. Στην τελευταία περίπτωση όταν έχει ολοκληρωθεί το αίτημα λήψης για τα αρχεία μας μπορούμε να δούμε τον χρόνο που διήρκησε η διαδικασία, το μέγεθος του αρχείου μας και πλέον μπορούμε να το κατεβάσουμε τοπικά πατώντας το "Download" και να αρχίσουμε την επεξεργασία μας.

3.2 Προετοιμασία & Καθαρισμός δεδομένων

Εφόσον έχουμε πλέον κατεβάσει τα δύο σύνολα δεδομένων που χρειαζόμαστε για το 2ο μέρος της εργασίας που είναι το μοντέλο πρόβλεψης είναι απαραίτητο τα δεδομένα μας να έρθουν στην κατάλληλη μορφή, δηλαδή να προετοιμαστούν και να καθαριστούν για να μπορέσουμε να τροφοδοτήσουμε το μοντέλο μας.

Για αυτήν την εργασία χρησιμοποιήθηκε το online εργαλείο Google Collab που σύμφωνα με την επίσημη ιστοσελίδα είναι ένα προϊόν της Google Research το οποίο χρησιμοποιείται για την σύνταξη και εκτέλεση της γλώσσας Python μέσα από κάποιο πρόγραμμα περιήγησης με μηδενική διαμόρφωση, με πρόσβαση χωρίς χρέωση σε GPU και εύκολη κοινοποίηση σε άλλους χρήστες.

Πριν προχωρήσουμε στην δημιουργία του notebook γίνεται μια σύντομη αναφορά στα αρχεία Grib που είναι η μορφή που έχουν κατέβει τα δεδομένα καιρού και την

μετατροπή τους σε κατάλληλη μορφή προκειμένου να τα χρησιμοποιήσουμε στο μοντέλο πρόβλεψης.

3.2.1 Μετατροπή δεδομένων καιρού από Grib file σε csv/ txt file

Σύμφωνα με την Wikipedia το όνομα Grib προκύπτει από τα αρχικά του «GRIdded Binary» ή «General Regularly distributed Information in Binary form» και είναι μια συνοπτική μορφή δεδομένων που χρησιμοποιείται συνήθως στην μετεωρολογία για την αποθήκευση ιστορικών δεδομένων καιρού και προγνωστικών δεδομένων καιρού, όπως και στη δική μας περίπτωση. Είναι ουσιαστικά μια συλλογή αυτοτελών εγγραφών δισδιάστατων δεδομένων και οι εγγραφές αυτές στέκονται μόνες τους χωρίς αναφορές σε άλλες εγγραφές ή σε ένα συνολικό σχήμα. Κάθε εγγραφή Grib αποτελείται από δύο στοιχεία, την κεφαλίδα που περιγράφει την εγγραφή και τα ίδια τα δυαδικά δεδομένα.

Το σύνολο δεδομένων για την κυκλοφορία των δρόμων στην Αττική είναι ήδη σε μορφή csv οπότε μπορούμε να το εισάγουμε στο εργαλείο της επιλογής μας, ενώ το σύνολο δεδομένων καιρού που έχουμε κατεβάσει από το Copernicus Cds Api είναι αρχείο Grib οπότε χρειάζεται μια προεργασία προκειμένου να το φέρουμε σε μορφή csv ή txt.

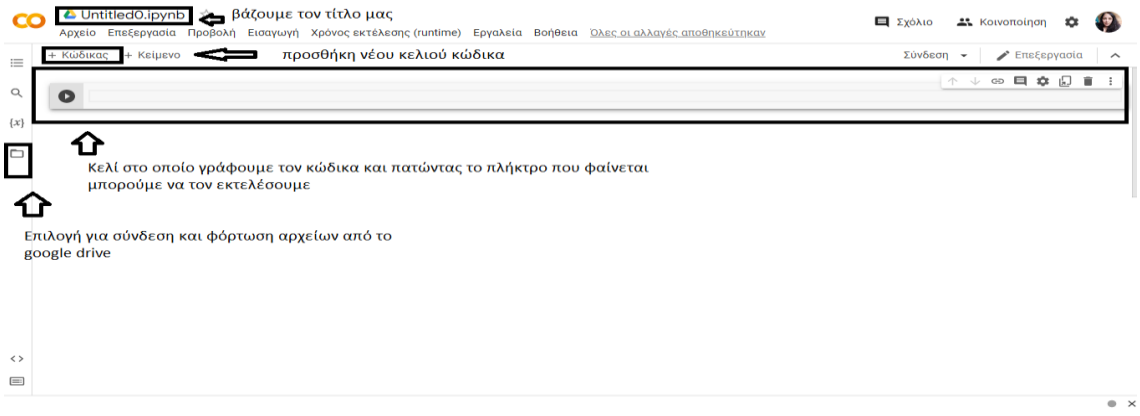
Για να γίνει αυτό υπάρχουν διάφοροι τρόποι αλλά στην προκειμένη δημιουργήθηκε ένα νέο περιβάλλον εργασίας μέσα από το τερματικό του υπολογιστή με την βοήθεια του Conda. Σύμφωνα με την επίσημη ιστοσελίδα το Conda είναι ένα ανοιχτού κώδικα σύστημα διαχείρισης πακέτων και σύστημα διαχείρισης περιβάλλοντος συμβατό με Windows, macOS, Linux και z/OS. Μέσω του Conda μπορούμε να εγκαταστήσουμε γρήγορα πακέτα και τις εξαρτήσεις τους, να τα τρέξουμε και να κάνουμε ενημερώσεις. Ο λόγος που το επιλέξαμε είναι γιατί δημιουργήθηκε για προγράμματα Python και δίνει την δυνατότητα να εναλλάσσεις γρήγορα περιβάλλοντα τοπικά χωρίς να χρειάζεται να εγκαθιστάς συνεχώς νέες εκδόσεις των προγραμμάτων σε όλο το σύστημα.

Χρειάζεται να εγκαταστήσουμε επίσης το Pygrib που είναι μια υψηλού επιπέδου διεπαφή Python για την βιβλιοθήκη ECCODES για τα αρχεία grib. Σύμφωνα με τον ECMWF που είναι το Ευρωπαϊκό Κέντρο Μεσοπρόθεσμων Μετεωρολογικών Προβλέψεων και είναι υπεύθυνοι για την εφαρμογή του Copernicus το ECCODES είναι ουσιαστικά το βασικό πακέτο κωδικοποίησης και αποκωδικοποίησης για τα αρχεία grib. Ο πιο εύκολος τρόπος να εγκαταστήσουμε τα πακέτα που χρειαζόμαστε και τις εξαρτήσεις τους είναι με το pip install pygrib (που είναι πρόγραμμα εγκατάστασης πακέτων για την Python) μέσα από το περιβάλλον που δημιουργήσαμε στο Conda αφού το ενεργοποιήσουμε με την εντολή « conda activate C:\conda_env\environment_name» . Οπότε συνεχίζοντας διαβάζουμε το αρχείο grib και

το κάνουμε export σε csv και μετά μετατρέπουμε το αρχείο csv σε txt για να είναι πιο διαχειρίσιμο λόγω του μεγάλου όγκου δεδομένων.

3.2.2 Δημιουργία σημειωματάριου Google Collab

Για να δημιουργήσουμε το Google Collab Notebook το μόνο που χρειαζόμαστε είναι ένας λογαριασμός email της Google. Αφότου συνδεθούμε στον προσωπικό μας λογαριασμό πηγαίνουμε στον Drive που είναι συνδεδεμένος με αυτόν και από εκεί μπορούμε με δεξί κλικ να επιλέξουμε το «Περισσότερα» και από εκεί το «Google Collaboratory» και να δημιουργήσουμε το σημειωματάριο που θα χρησιμοποιήσουμε.



Εικόνα 14 Δημιουργία σημειωματάριου Google Colab και επεξήγηση διεπαφής

Στην παραπάνω εικόνα φαίνεται η διεπαφή του όταν το δημιουργούμε και οι επιλογές που υπάρχουν. Πάνω αριστερά βάζουμε τον τίτλο που θέλουμε, από κάτω μας παρέχει πολλές επιλογές από την βασική γραμμή εργαλείων και όπως φαίνεται πατώντας το +Κώδικας μπορούμε να προσθέσουμε επιπλέον κελιά. Στο κάθε κελί πατώντας το πλήκτρο αριστερά μπορούμε να εκτελέσουμε τον κώδικα και από τις επιλογές που υπάρχουν δεξιά πάνω από το κελί μας παρέχει την δυνατότητα να το μετακινήσουμε πάνω ή κάτω ανάλογα με την σειρά που θέλουμε να εκτελεστεί ο κώδικάς μας, να βάλουμε έναν σύνδεσμο προς το κελί, να προσθέσουμε ένα σχόλιο, να προβούμε σε κάποιες επιπλέον ρυθμίσεις, να αντιγράψουμε ένα κελί ή να το διαγράψουμε.

Επίσης, στην αριστερή στήλη στο πλάι υπάρχουν πάλι κάποιες επιλογές όπως η προσθήκη πίνακα περιεχομένων, η εύρεση και αντικατάσταση, η εισαγωγή μεταβλητών και η σύνδεση αρχείων από τον drive μας. Ουσιαστικά τα σημειωματάρια Colab είναι σημειωματάρια Jupyter που φιλοξενούνται στο Colab.

3.2.3 Εισαγωγή Datasets στο Google Colab

Προκειμένου να μπορέσουμε να δουλέψουμε με τα δεδομένα μας πρέπει πρώτα να τα ανεβάσουμε στον google drive και μετά να τα εισάγουμε στο Google Colab οπότε δημιουργήσαμε πρώτα μέσα στον Drive έναν φάκελο που ονομάσαμε Colab Notebooks για να έχουμε όλα τα αρχεία που θα χρειαστούμε στον ίδιο φάκελο. Για να μπορέσουμε να συνδέσουμε τα δεδομένα μας στο Google Colab και να τα μετατρέψουμε σε pandas dataframes πρέπει πρώτα να κάνουμε authenticate user και να δώσουμε πρόσβαση στο Google Colab στα αρχεία του Google Drive. Υπάρχουν διάφοροι τρόποι για να γίνει αυτό αλλά στην προκειμένη επιλέξαμε να το κάνουμε με το open by key σύμφωνα με τις οδηγίες του άρθρου Towards Data Science, Different ways to connect google drive to a google colab notebook Part 2 .Το πρώτο κομμάτι του παρακάτω κώδικα είναι για να δώσουμε πρόσβαση στον drive και πιο κάτω βλέπουμε πως μπορούμε να ανοίξουμε ένα αρχεία σαν gspread με την χρήση ενός κλειδιού, να το κάνουμε dataframe και έπειτα να δούμε τι περιέχει. Από εδώ και πέρα δεν θα χρειαστεί να κάνουμε ξανά αυθεντικοποίηση χρήση και μπορούμε να διαβάσουμε κανονικά τα αρχεία που έχουμε ανεβάσει στον google drive είτε είναι σε μορφή txt είτε είναι σε μορφή csv.

```
# Authorizing google colab
from google.colab import auth
auth.authenticate_user()

# Credentials for google sheets
import gspread
from google.auth import default
creds, _ = default()

# Authotizing the connection
gc = gspread.authorize(creds)

# Connecting
worksheet = gc.open_by_key('1arIkiOuggnXz3O75eXd4l9kSsOr5E1MrkT_ZSTyXgrc'
).sheet1

# Exporting data to get_all_values gives a list of rows.
rows = worksheet.get_all_values()
```

```
# Using pandas to convert to a DataFrame and render.
import pandas as pd
dftest = pd.DataFrame.from_records(rows)

#creating columns name
dftest.columns = dftest.iloc[0]
dftest = dftest.iloc[1:]
dftest
```

Ένας πιο απλός τρόπος για να αποκτήσει το Google Colab πρόσβαση στα αρχεία του drive είναι να μοντάρουμε τον drive μέσω της παρακάτω εντολής:

```
from google.colab import drive
drive.mount('/content/drive')
```

και μετά μπορούμε μέσω της εντολής pd.read_csv(file_path) και της διαδρομής του αρχείου μας μέσα στον δίσκο να ανοίξουμε / συνδέσουμε τα αρχεία που χρειαζόμαστε.

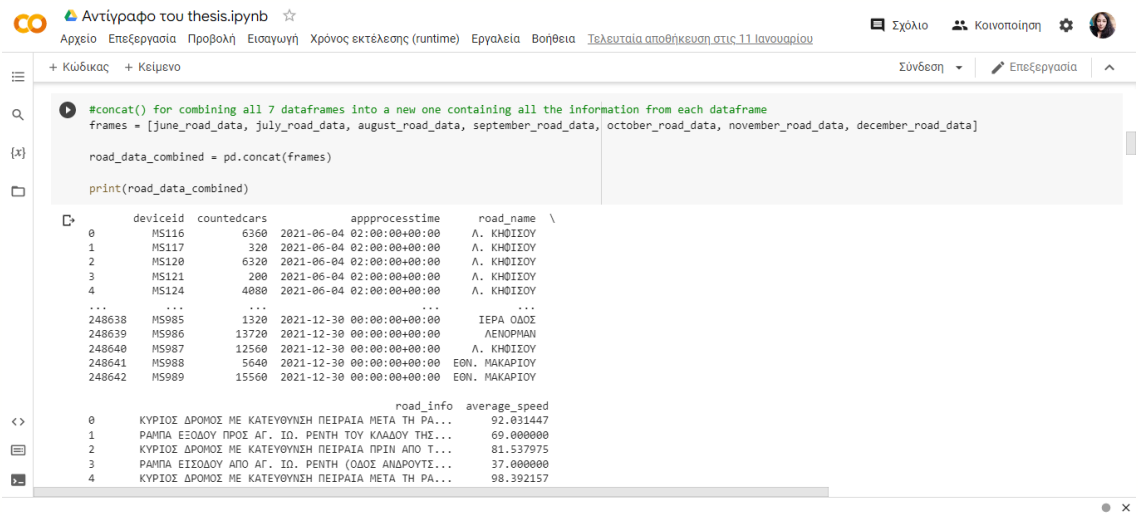
Συνεχίζοντας προτού περάσουμε στην φάση της προετοιμασίας και του καθαρισμού πρώτα πρέπει να «διαβάσουμε» το αρχείο που περιέχονται τα δεδομένα καιρού με την εντολή

```
final_data = pd.read_csv ('/content/drive/MyDrive/Colab Notebooks/final_data.csv')
```

και έπειτα πρέπει να κάνουμε το ίδιο και για τα δεδομένα κυκλοφορίας με την διαφορά πως θα επαναλάβουμε την διαδικασία 7 φορές γιατί έχουμε 7 αρχεία. Παρακάτω φαίνεται ο κώδικας για τα δεδομένα κυκλοφορίας του Ιουνίου.

```
june_road_data = pd.read_csv ('/content/drive/MyDrive/Colab Notebooks/june_data.csv
')
```

Αφότου επαναλάβουμε λοιπόν την διαδικασία για τους μήνες που έχουμε επιλέξει προχωράμε στην συνένωση αυτών με την εντολή που αναγράφεται στο 1.1.1 και έχουμε πλέον ένα νέο dataframe που το έχουμε αποθηκεύσει προς το παρόν με το όνομα road_data_combined. Στην συνέχεια της εργασίας θα το μετονομάσουμε σε ένα πιο κατάλληλο όνομα dataframe όπως π.χ. road_data προκειμένου να είναι κατανοητό πως περιέχει τα δεδομένα κυκλοφορίας και να είναι σύντομο.



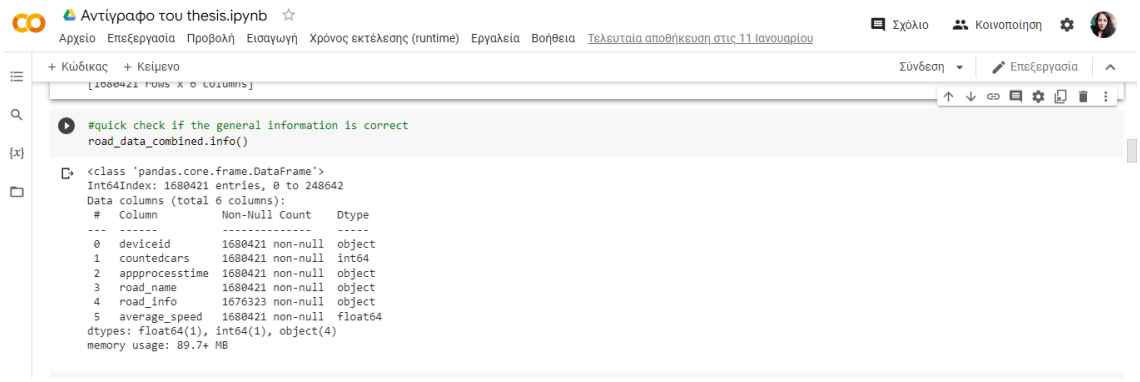
Εικόνα 15 Συνένωση δεδομένων κυκλοφορίας σε ένα κοινό σύνολο δεδομένων

3.2.4 Προετοιμασία, Καθαρισμός & Διερεύνηση δεδομένων Κυκλοφορίας

Σε αυτό το σημείο έχουμε ήδη εισάγει τις βιβλιοθήκες και τα εργαλεία που θα χρειαστούμε στην συνέχεια και έχουμε φορτώσει τα σύνολα δεδομένων που έχουμε επιλέξει για την κυκλοφορία των δρόμων στην Αττική. Πριν ξεκινήσουμε την ανάλυση πρέπει να ελέγξουμε τους τύπους των δεδομένων μας και την κατανομή των δεδομένων μας προκειμένου να εντοπίσουμε πιθανά προβλήματα και να τα αντιμετωπίσουμε.

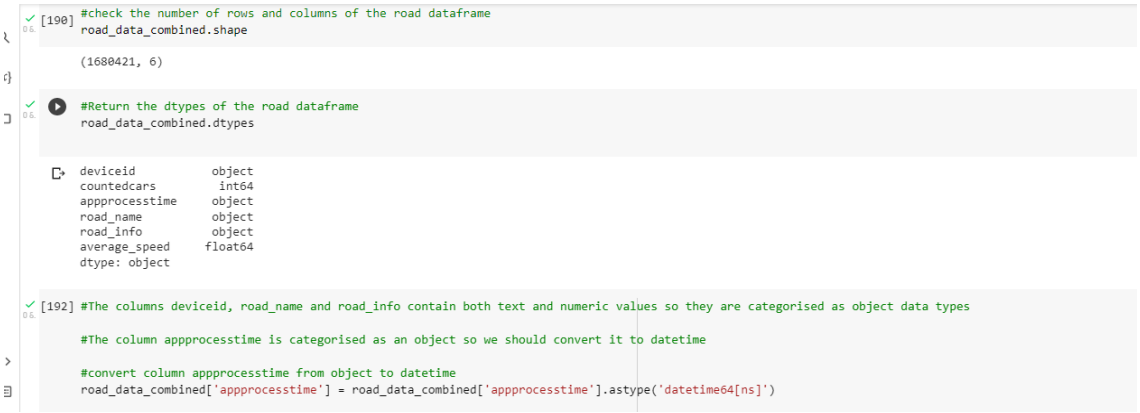
Οι βασικές βιβλιοθήκες που θα χρησιμοποιήσουμε είναι η βιβλιοθήκη Pandas και η βιβλιοθήκη Numpy. Ξεκινάμε λοιπόν με απλές εντολές για να δούμε τι περιέχουν τα σύνολα μας, να δούμε αν έχουν τον σωστό τύπο δεδομένων, αν υπάρχουν ελλείπουσες τιμές, διπλότυπα, τυπογραφικά λάθη, έκτροπες τιμές, μονάδες μέτρησης κάθε στήλης κλπ .

Χρησιμοποιούμε την εντολή info() για να δούμε κάποιες πληροφορίες συνοπτικά όπως οι στήλες, ο αριθμός των εγγραφών της κάθε στήλης, ο τύπος δεδομένων και πόσο χώρο πιάνει στη μνήμη.



Εικόνα 16 Εντολή info() για το σύνολο δεδομένων κυκλοφορίας

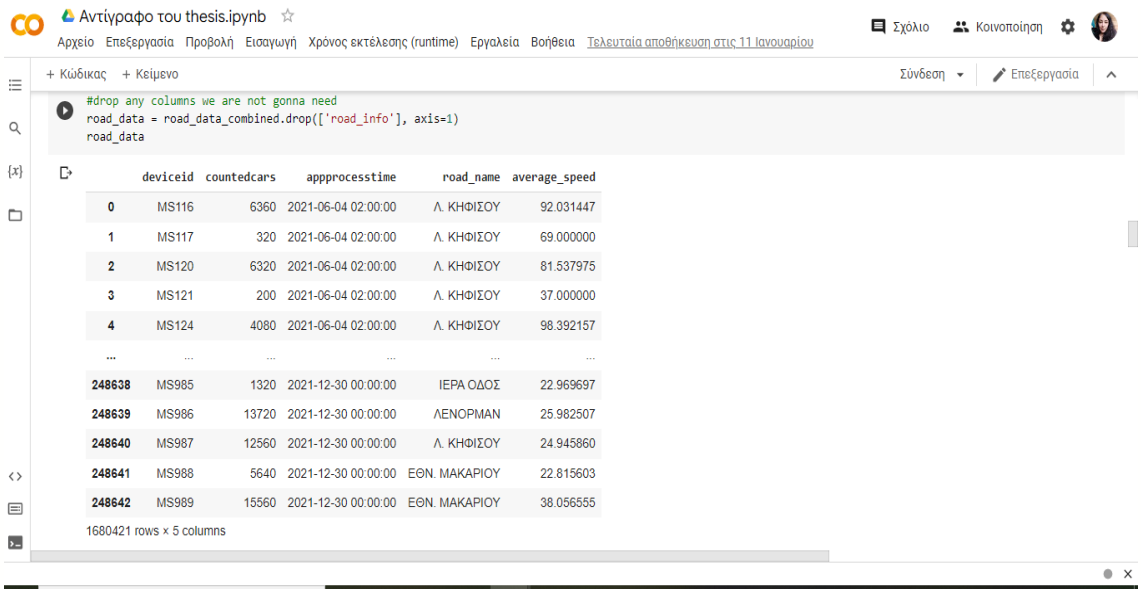
Με το road_data_combined.shape βλέπουμε πως το σύνολο μας έχει 1680421 σειρές και 6 στήλες, με την εντολή road_data_combined.dtypes βλέπουμε τους τύπους δεδομένων που έχουν αποδοθεί στις μεταβλητές μας και αν χρειάζεται να τις αλλάξουμε, όπως π.χ. την στήλη appprocesstime που έχει καταταχθεί σαν object και με την εντολή astype('datetime64[ns]')την μετατρέπουμε σε τύπο datetime64[ns] που είναι ο επιθυμητός τύπος για την στήλη της ημερομηνίας και ώρας.



Εικόνα 17 Εντολή df.dtypes & μετατροπή τύπου object σε datetime64[ns]

Με την ίδια λογική συνεχίζουμε την διερεύνηση με εντολές όπως η describe, min(), max(), head(), tail() κλπ καθώς και την προετοιμασία και τον καθαρισμό μέχρι τα δεδομένα μας να φτάσουν στην κατάλληλη μορφή πχ βρίσκοντας και αφαιρώντας στήλες που περιέχουν ελλείπουσες τιμές. Με το nununique() στις στήλες deviceid και road_name βλέπουμε πως υπάρχουν 425 αισθητήρες τοποθετημένοι σε 93 δρόμους.

Αν κρίνουμε πως δεν χρειαζόμαστε όλες τις στήλες, όπως την στήλη road_info που είδαμε πως περιέχει 4098 Nan εγγραφές και είναι βοηθητική στήλη ως προς την στήλη road_name, χρησιμοποιούμε το drop() καθώς έχουμε στη διάθεσή μας την στήλη deviceid που περιέχει τις πληροφορίες που θα χρειαστούμε για τους αισθητήρες.



	deviceid	countedcars	appprocesstime	road_name	average_speed
0	MS116	6360	2021-06-04 02:00:00	Λ. ΚΗΦΙΣΟΥ	92.031447
1	MS117	320	2021-06-04 02:00:00	Λ. ΚΗΦΙΣΟΥ	69.000000
2	MS120	6320	2021-06-04 02:00:00	Λ. ΚΗΦΙΣΟΥ	81.537975
3	MS121	200	2021-06-04 02:00:00	Λ. ΚΗΦΙΣΟΥ	37.000000
4	MS124	4080	2021-06-04 02:00:00	Λ. ΚΗΦΙΣΟΥ	98.392157
...	...	...	...	...	...
248638	MS985	1320	2021-12-30 00:00:00	ΙΕΡΑ ΟΔΟΣ	22.969697
248639	MS986	13720	2021-12-30 00:00:00	ΛΕΝΟΡΜΑΝ	25.982507
248640	MS987	12560	2021-12-30 00:00:00	Λ. ΚΗΦΙΣΟΥ	24.945860
248641	MS988	5640	2021-12-30 00:00:00	ΕΘΝ. ΜΑΚΑΡΙΟΥ	22.815603
248642	MS989	15560	2021-12-30 00:00:00	ΕΘΝ. ΜΑΚΑΡΙΟΥ	38.056555

1680421 rows × 5 columns

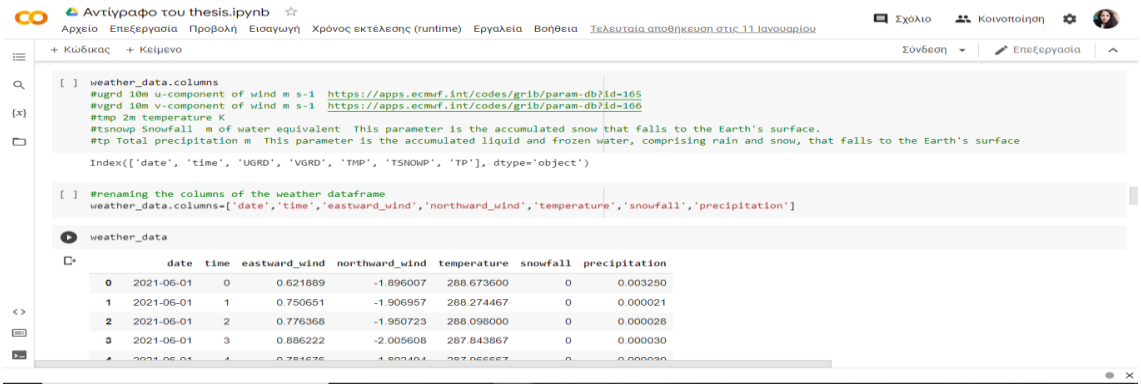
Εικόνα 18 Εντολή drop() για την αφαίρεση της βοηθητικής στήλης road_info

Σε αυτό το σημείο είναι σημαντικό να αναφέρουμε πως η πλατφόρμα data.gov.gr από όπου αντλήσαμε τα δεδομένα κυκλοφορίας για τους μήνες επιλογής μας βρίσκεται ακόμη σε αρχικό στάδιο και περιέχει συχνά ελλειπή δεδομένα ή λανθασμένα δεδομένα με πολύ μεγάλες τιμές που δεν ανταποκρίνονται στις πραγματικές. Για παράδειγμα, στην αρχική διερεύνηση για τα δεδομένα του Αυγούστου 2021 ενώ επιλέξαμε από το portal να πάρουμε τα δεδομένα για την περίοδο από 1.8.2021 μέχρι 31.8.2021 σε αρχείο csv όταν περάσαμε στην διερεύνηση με την εντολή august_road_data.head(10) βλέπουμε πως τα δεδομένα μας ξεκινάνε από 4.8.2021 και με την εντολή .tail(10) βλέπουμε πως τα δεδομένα μας τελειώνουν στις 9.8.2021. Αυτό μας οδηγεί στο συμπέρασμα πως τον Αύγουστο του 2021 οι αισθητήρες ή τα δίκτυα λογικά παρουσίασαν κάποια δυσλειτουργία στην καταγραφή, μεταφορά ή αποθήκευση των δεδομένων οπότε για όλο τον μήνα έχουμε μόνο 5 καταγεγραμμένες ημέρες. Αντίστοιχα για τον Νοέμβριο με τις ίδιες εντολές βλέπουμε πως η καταγραφή των δεδομένων ξεκινάει από 3.11.2021 και φτάνει μέχρι 29.11.2021.

3.2.5 Προετοιμασία, Καθαρισμός & Διερεύνηση δεδομένων Καιρού

Με την ίδια λογική χρησιμοποιούμε τις παραπάνω εντολές για να δούμε τους τύπους δεδομένων του συνόλου με τα δεδομένα καιρού και με την εντολή df.columns μετονομάζουμε τις στήλες μας όπως φαίνεται παρακάτω για να είναι πιο κατανοητές.

weather_data.columns=['date','time','eastward_wind','northward_wind','temperature','snowfall','precipitation']

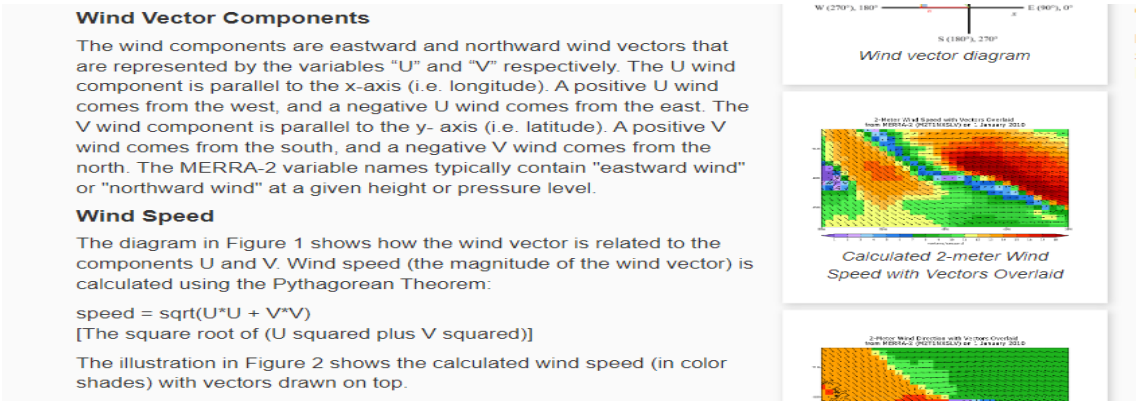


Εικόνα 19 Μετονομασία ονομάτων στηλών του συνόλου δεδομένων καιρού

Γνωρίζοντας ήδη από την ιστοσελίδα και το documentation του dataset τις μονάδες μέτρησης της κάθε στήλης μπορούμε να κάνουμε μετατροπές όπου κρίνουμε πως χρειάζεται, όπως για παράδειγμα στην στήλη temperature όπου έχουμε την θερμοκρασία σε Kelvin και την μετατρέπουμε σε Celsius και με την εντολή max ελέγχουμε την μέγιστη καταγεγραμμένη θερμοκρασία για να βεβαιωθούμε πως η μετατροπή μας έγινε σωστά. Μετατρέπουμε πάλι τον τύπο δεδομένων της στήλης date από object σε datetime64[ns].

Εικόνα 20 Μετατροπή μονάδων μέτρησης θερμοκρασίας από Kelvin σε Celsius

Ακολουθώντας τις οδηγίες του άρθρου https://daac.gsfc.nasa.gov/information/data-in-action?title=Derive%20Wind%20Speed%20and%20Direction%20With%20MERRA-2%20Wind%20Components όπως φαίνεται και στην παρακάτω φωτογραφία μπορούμε γνωρίζοντας τους πλευρικούς ανέμους eastward wind και northward wind να υπολογίσουμε την ταχύτητα του ανέμου με την συνάρτηση της τετραγωνικής ρίζας.



Εικόνα 21 Θεωρία για την δημιουργία της στήλης wind speed με βάση τα u wind component & v wind component

Μπορούμε επίσης να χρησιμοποιήσουμε την συνάρτηση της υποτείνουσας που είναι ισοδύναμε με αυτή. Έτσι προκύπτει μια νέα στήλη που την ονομάσαμε «average wind speed» και περιέχει τα δεδομένα της ταχύτητας του ανέμου σε ωριαίο επίπεδο σε m/s. Εφόσον έχουμε πλέον την νέα στήλη ονομάσαμε «average wind speed» μπορούμε να αφαιρέσουμε τις ξεχωριστές στήλες eastward_wind και northward_wind.

Με την εντολή weather_data['snowfall'].max() βλέπουμε πως το αποτέλεσμα είναι 0 οπότε για την περίοδο που εξετάζουμε δεν υπάρχει καταγεγραμμένη χιονόπτωση και μπορούμε να κάνουμε drop την στήλη snowfall.

Επίσης, ενώσαμε τις στήλες date και time σε μια ενιαία στήλη datetime προκειμένου τα δύο σύνολα δεδομένων που χρησιμοποιούμε να έχουν ίδιο format ημερομηνίας και ώρας datetime64[ns] ώστε αργότερα να μπορέσουμε να συνενώσουμε τα δύο σύνολα δεδομένων μόνο με τις στήλες που μας ενδιαφέρουν για το μοντέλο πρόβλεψης.

Η κανονικοποίηση των δεδομένων μας είναι σημαντική καθώς διαφορετικές στήλες αριθμητικών δεδομένων μπορεί να έχουν διαφορετικό εύρος και η απευθείας σύγκριση τους δεν οδηγεί σε ωφέλιμα συμπεράσματα. Η κανονικοποίηση μας βοηθάει να φέρουμε όλα τα δεδομένα μας σε παρόμοιο εύρος προκειμένου να έχει νόημα η όποια σύγκριση.

Παρακάτω πριν ξεκινήσουμε να τροφοδοτούμε το μοντέλο μας πρέπει να μετατρέψουμε τα κατηγορικά δεδομένα σε αριθμητικά όπου χρειάζεται καθώς τα περισσότερα στατιστικά μοντέλα δεν δέχονται τύπους object και strings σαν input και για την εκπαίδευση του μοντέλου μας θα χρειαστούμε αριθμούς σας input.

3.3 Οπτικοποιήσεις μέσω Python

Σε αυτό το σημείο έχουμε ήδη δει περιληπτικά κάποια βασικά χαρακτηριστικά των δεδομένων μας και συνεχίζουμε την διερεύνηση μέσω οπτικοποιήσεων προκειμένου να κατανοήσουμε καλύτερα τα σύνολά μας και να ανακαλύψουμε τυχόν σχέσεις ανάμεσα στις μεταβλητές μας ώστε να μπορέσουμε να βρούμε τις πιο σημαντικές.

Εφόσον τα δεδομένα έχουν υποστεί μια προετοιμασία και είναι πλέον σε πιο κατάλληλη μορφή και βαθμό λεπτομέρειας θα χρησιμοποιήσουμε δύο ακόμη βιβλιοθήκες της Python για οπτικοποιήσεις, την Matplotlib και την Seaborn. Υπάρχουν πολλοί τύποι Διαγραμμάτων και Γραφημάτων που μπορούμε να χρησιμοποιήσουμε ανάλογα με το είδος της πληροφορίας που θέλουμε να πάρουμε και ανάλογα με τον τύπο των δεδομένων μας, δηλαδή αν είναι αριθμητικά δεδομένα ή κατηγορικά. Πριν προχωρήσουμε στη δημιουργία κάποιων γραφημάτων ακολουθεί μια σύντομη αναφορά στο θεωρητικό υπόβαθρο των πιο συνηθισμένων τύπων οπτικοποιήσεων.

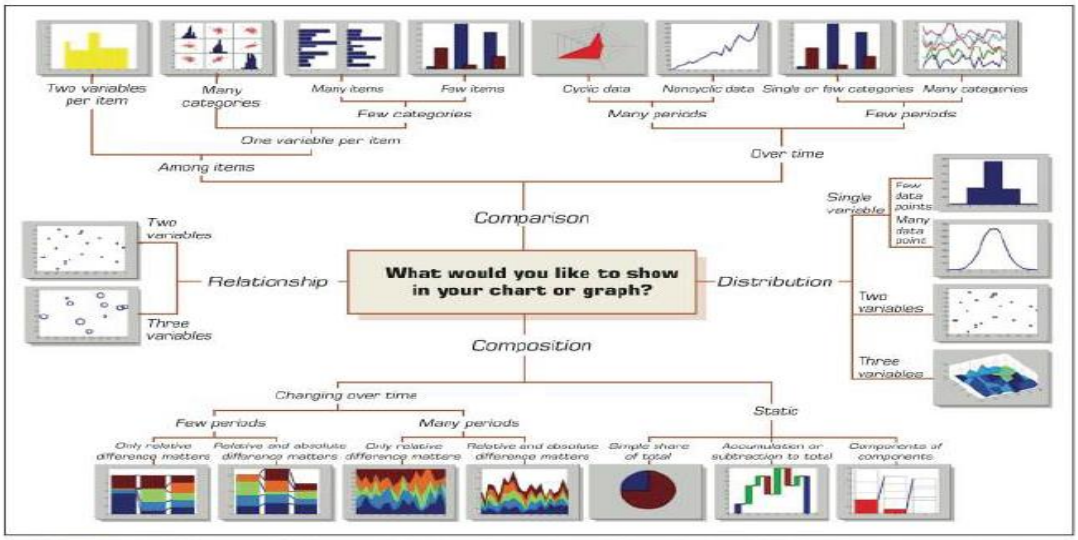


FIGURE 2.21 A Taxonomy of Charts and Graphs. *Source:* Adapted from Abela, A. (2008). Advanced presentations by design: Creating communication that drives action. New York: Wiley.

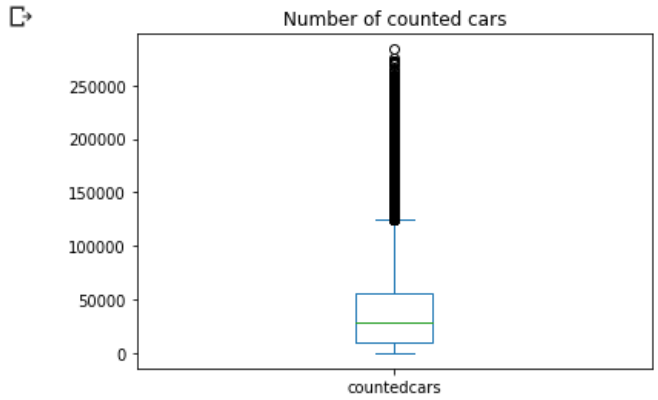
Εικόνα 22 Ταξινομία Διαγραμμάτων και Γραφημάτων (Πηγή: Sharda, Business Intelligence, Analytics and Data Science – A Managerial Perspective, σελ. 109, 4th edition)

Κάποιοι από τους πιο βασικούς τύπους Διαγραμμάτων και Γραφημάτων σύμφωνα με τον Sharda είναι το Διάγραμμα Γραμμής (Line Chart) για δεδομένα χρονοσειρών όπου μας βοηθούν να εξετάσουμε την σχέση ανάμεσα σε δύο μεταβλητές προκειμένου να ανακαλύψουμε αλλαγές ή τάσεις σε βάθος χρόνου, το Ραβδόγραμμα (Bar Chart) όταν έχουμε ονομαστικά ή αριθμητικά δεδομένα που χωρίζονται ομοιόμορφα σε διαφορετικές κατηγορίες ώστε να δούμε εύκολα συγκριτικές αποτελέσματα και τάσεις, το Διάγραμμα Διασποράς (Scatter Plot) για να ανακαλύψουμε σχέσεις ανάμεσα σε 2 ή 3 μεταβλητές και να εξερευνήσουμε την ύπαρξη τάσεων, συγκεντρώσεων ή έκτροπων τιμών (outliers). Επίσης, σημαντικό είναι και το Ιστόγραμμα (Histogram) που μοιάζει

με το Ραβδόγραμμα αλλά χρησιμοποιείται για την συχνότητα κατανομής μιας ή παραπάνω μεταβλητών και η χρήση του Χάρτη Θερμότητας (Heat map) για την σύγκριση συνεχόμενων τιμών και την εξερεύνηση συσχέτισης μεταξύ αριθμητικών μεταβλητών.

Αρχικά ξεκινάμε την εξερεύνηση των δεδομένων κυκλοφορίας με ένα Box Plot για τον αριθμό των μετρούμενων οχημάτων (countedcars) για να δούμε ποια είναι η μικρότερη αξία με εξαίρεση τα outliers, η μεγαλύτερη αξία με εξαίρεση τα outliers, το άνω τεταρτημόριο (που σημαίνει πως το 25% των δεδομένων είναι μεγαλύτερο από αυτή την αξία), το κάτω τεταρτημόριο (που σημαίνει πως το 25% των δεδομένων είναι μικρότερο από αυτή την αξία), ο διάμεσος των δεδομένων και οι έκτροπες τιμές που βρίσκονται πάνω και κάτω από τις αντίστοιχες μεγαλύτερες και μικρότερες αξίες των δεδομένων.

```
#create a boxplot for counted cars to find out the min and max values and the outliers
road_data['countedcars'].plot(kind='box', title='Number of counted cars')
plt.show()
```



Εικόνα 23 Κώδικας και Γράφημα Box Plot για την μεταβλητή countedcars

Με το παραπάνω γράφημα είναι εύκολο να διακρίνουμε πως σε αυτήν την στήλη έχουμε πολλές τιμές που είναι έκτροπες πράγμα που μας δείχνει πως στα δεδομένα κυκλοφορίας περιέχονται πολλά outliers. Ακολουθεί μια επεξηγηματική εικόνα για διευκόλυνση της κατανόησης από το βιβλίο του Sharda.

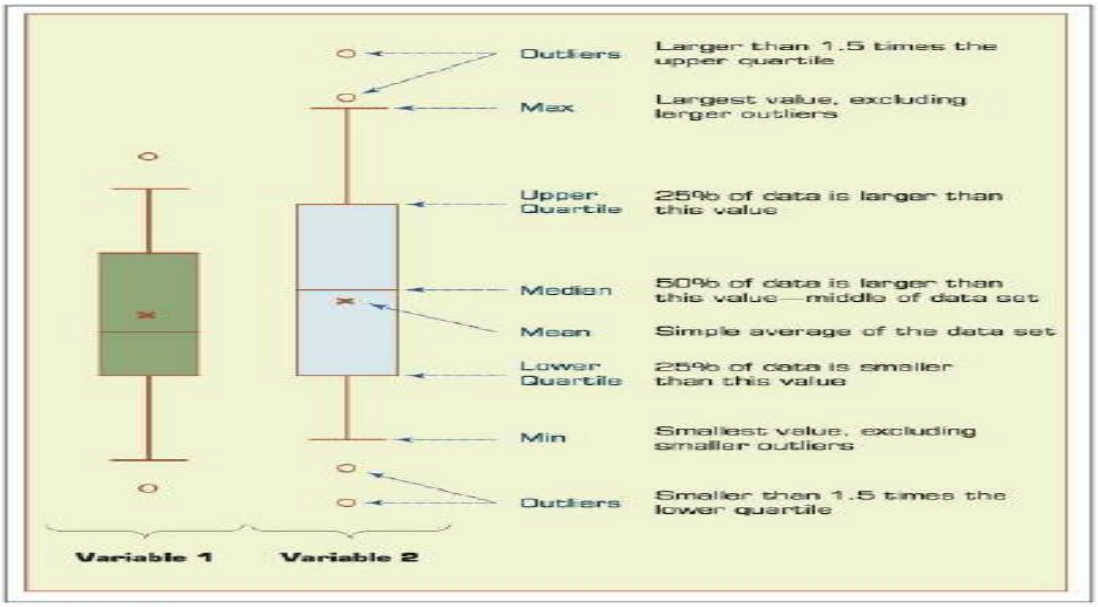
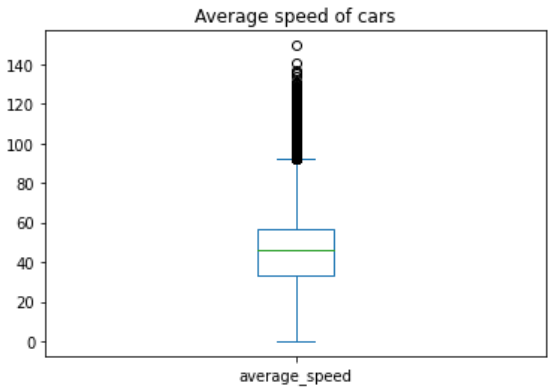


FIGURE 2.8 Understanding the Specifics about Box-and-Whiskers Plots.

Εικόνα 24 Κατανοώντας τα ιδιαίτερα χαρακτηριστικά των Box και Whiskers Γραφημάτων(Πηγή: Sharda, Business Intelligence, Analytics and Data Science – A Managerial Perspective, σελ. 79, 4th edition)

Έπειτα κάνουμε πάλι ένα Box Plot για την στήλη average_speed για να δούμε ποιες είναι οι πιο συνηθισμένες ταχύτητες των οχημάτων.

```
#create a boxplot for average speed to find out the min and max values and the outliers
road_data['average_speed'].plot(kind='box', title='Average speed of cars')
plt.show()
```



Εικόνα 25 Κώδικας και Γράφημα Box Plot για την μεταβλητή average_speed

Βλέπουμε πως το κατώτατο όριο της καταγεγραμμένης ταχύτητας είναι μηδέν και το ανώτερο είναι λίγο κάτω από 100. Από εκεί και πάνω είναι έκτροπες τιμές και η μέση ταχύτητα είναι 50 χιλιόμετρα την ώρα.

Ακολουθεί ένα Ραβδόγραμμα με τους 10 δρόμους που εμφανίζονται πιο συχνά στο δεδομένα κυκλοφορίας μας. Οι δρόμοι που εμφανίζονται είναι κεντρικοί δρόμοι και λεωφόροι της Αττικής.

```
#use the counter method from the collections modules to generate a dictionary of count values for each category in the road name categorical column.
from collections import Counter

print(Counter(road_data['road_name']))
```

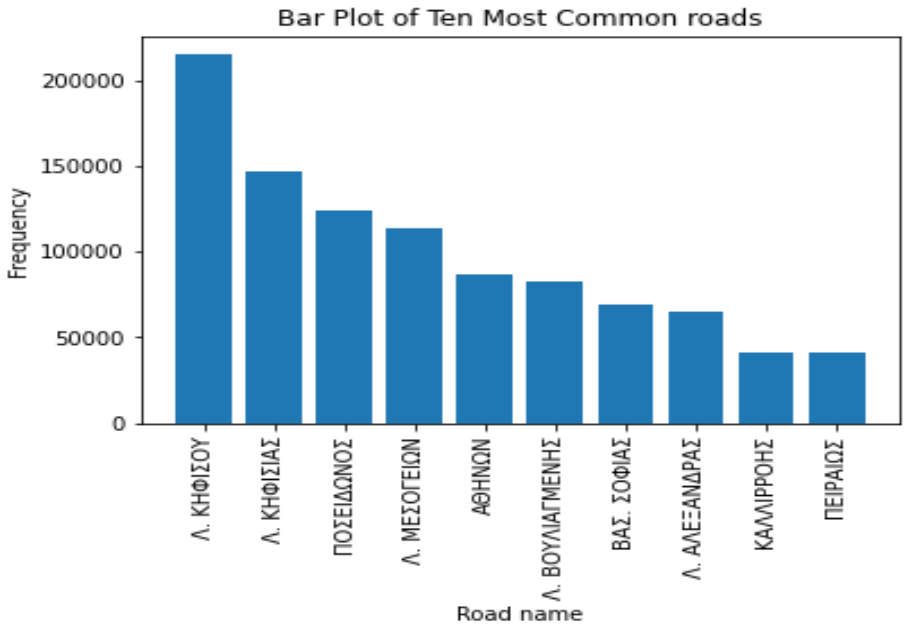
```
Counter({'Λ. ΚΗΦΙΣΟΥ': 215100, 'Λ. ΚΗΦΙΣΙΑΣ': 147378, 'ΠΟΣΕΙΔΩΝΟΣ': 124313, 'Λ. ΜΕΣΟΓΕΙΩΝ': 113632, 'ΑΘΗΝΩΝ': 86758, 'Λ. ΒΟΥΛΙΑΓΜΕΝΗΣ': 82847, 'ΒΑΣ. ΣΟΦΙΑΣ': 69365, 'Λ. ΑΛΕΞΑΝΔΡΑΣ':
```

```
[60] #filter this dictionary using the most_common method to find out which are the most common roads appearing in the road dataset
print(dict(Counter(road_data['road_name']).most_common(10)))
```

```
{'Λ. ΚΗΦΙΣΟΥ': 215100, 'Λ. ΚΗΦΙΣΙΑΣ': 147378, 'ΠΟΣΕΙΔΩΝΟΣ': 124313, 'Λ. ΜΕΣΟΓΕΙΩΝ': 113632, 'ΑΘΗΝΩΝ': 86758, 'Λ. ΒΟΥΛΙΑΓΜΕΝΗΣ': 82847, 'ΒΑΣ. ΣΟΦΙΑΣ': 69365, 'Λ. ΑΛΕΞΑΝΔΡΑΣ': 65490,
```

```
#generate a bar plot of the 10 most common roads appearing in the road dataset
road_name_dict = dict(Counter(road_data['road_name']).most_common(10))

plt.bar(road_name_dict.keys(), road_name_dict.values())

plt.xlabel('Road name')

plt.ylabel('Frequency')

plt.title('Bar Plot of Ten Most Common roads')

plt.xticks(rotation=90)

plt.show()
```

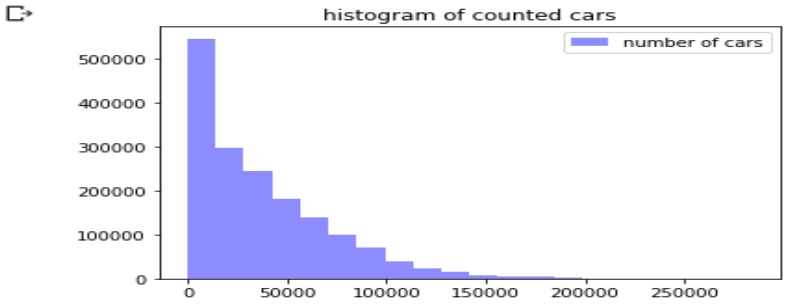
Εικόνα 26 Κώδικας για Ραβδόγραμμα με τους 10 πιο συχνούς δρόμους



Εικόνα 27 Ραβδόγραμμα με τους 10 πιο συχνούς δρόμους

Συνεχίζουμε με ένα Ιστόγραμμα για τον αριθμό των μετρούμενων οχημάτων και την συχνότητα κατανομής.

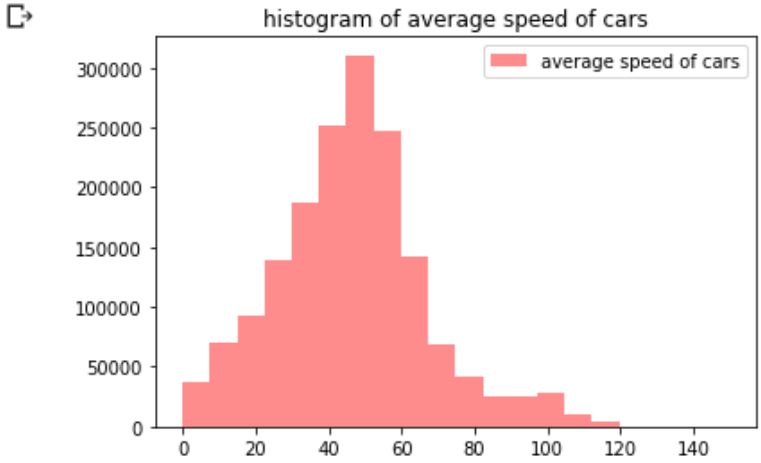
```
plt.hist(road_data['countedcars'], bins=20, alpha=0.45, color='blue')
plt.title("histogram of counted cars")
plt.legend(['number of cars'])
plt.show()
```



Εικόνα 28 Κώδικας και Ιστόγραμμα για συχνότητα κατανομής μετρούμενων οχημάτων

Επαναλαμβάνουμε την διαδικασία για να δημιουργήσουμε ένα Ιστόγραμμα για την μεταβλητή average_speed. Όπως φαίνεται παρακάτω όντως η πιο συχνή ταχύτητα είναι τα 50 χιλιόμετρα την ώρα.

```
plt.hist(road_data['average_speed'], bins=20, alpha=0.45, color='red')
plt.title("histogram of average speed of cars")
plt.legend(['average speed of cars'])
plt.show()
```



Εικόνα 29 Κώδικας και Ιστόγραμμα για συχνότητα κατανομής μέσης ταχύτητας οχημάτων

Μέσα από τα παραπάνω γραφήματα μπορούμε να κατανοήσουμε καλύτερα τα δεδομένα μας και την κατανομή τους πριν προχωρήσουμε στην δημιουργία του μοντέλου μας. Στην επόμενη ενότητα της διπλωματικής συνεχίζουμε την εξερεύνηση των δεδομένων μας μέσω οπτικοποιήσεων στο λογισμικό Tableau Public.

4. Οπτικοποιήσεις στο Tableau

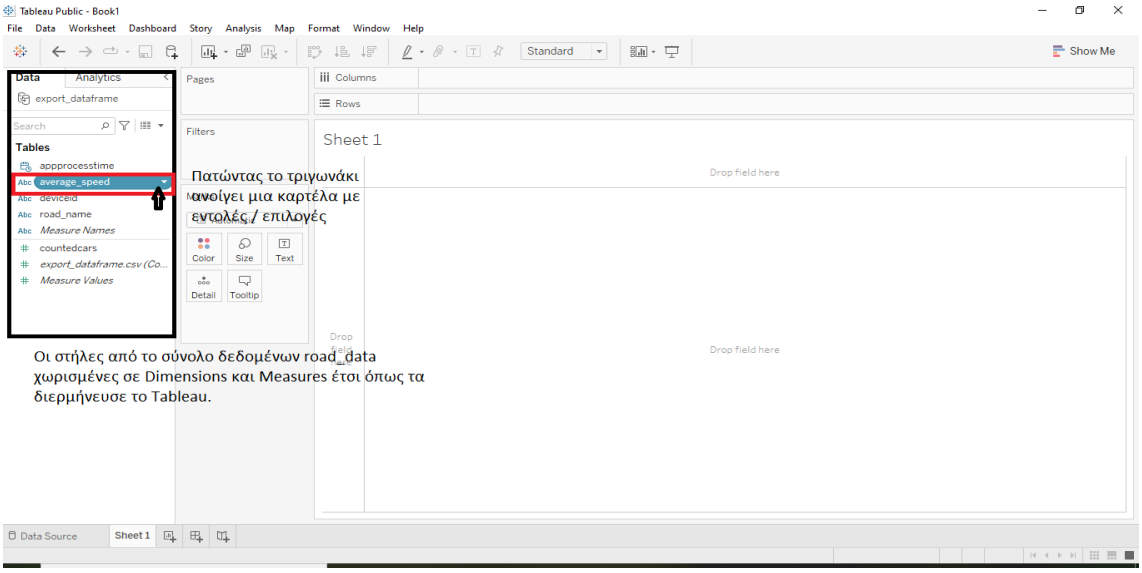
Σύμφωνα με τον Sharda (Business Intelligence, Analytics and Data Science – A Managerial Perspective, σελ. 203, 4th edition) «η οπτικοποίηση των δεδομένων μπορεί να χρησιμοποιηθεί σε συνδυασμό με άλλες τεχνικές εξόρυξης δεδομένων για να αποκτήσουμε καλύτερη κατανόηση των υποκείμενων σχέσεων. Καθώς η σημασία της οπτικοποίησης έχει αυξηθεί τα τελευταία χρόνια έχει προκύψει ένας νέος όρος, Οπτική Αναλυτική (Visual Analytics).»

Καθώς λοιπόν τα διαγράμματα και τα γραφήματα μπορούν να μας βοηθήσουν να κατανοήσουμε καλύτερα τα δεδομένα προς εξέταση σε αυτό το κεφάλαιο γίνεται χρήση του λογισμικού Tableau. Όπως αναγράφεται στην Wikipedia το Tableau είναι «μια αμερικάνικη εταιρεία λογισμικού διαδραστικής απεικόνισης δεδομένων που επικεντρώνεται στην επιχειρηματική ευφυΐα.». Σύμφωνα με την επίσημη ιστοσελίδα δημιουργήθηκε το 2003 σαν επιστημονικό έργο στο Πανεπιστήμιο Stanford και είχε σκοπό την βελτίωση της ροής της ανάλυσης καθώς και να κάνει τα δεδομένα πιο προσβάσιμα μέσω της οπτικοποίησης. Το 2019 εξαγοράστηκε από την εταιρεία Salesforce.

4.1 Εισαγωγή δεδομένων στο Tableau

Έχοντας στην κατοχή μας τα σύνολα δεδομένων προς ανάλυση στην κατάλληλη μορφή μπορούμε να ξεκινήσουμε εισάγοντάς τα στην πλατφόρμα οπτικοποίησης Tableau. Αρχικά, όταν ανοίγουμε την εφαρμογή στην Αρχική σελίδα επιλέγουμε την «Σύνδεση» της με το αρχείο που επιθυμούμε ή με τον Server που επιθυμούμε. Εμείς θα εξετάσουμε μόνο την σύνδεση με της εφαρμογής με αρχεία. Οι επιλογές που μας δίνει είναι αρχείο excel, text file, json file, Microsoft access, pdf file, spatial file και statistical file.

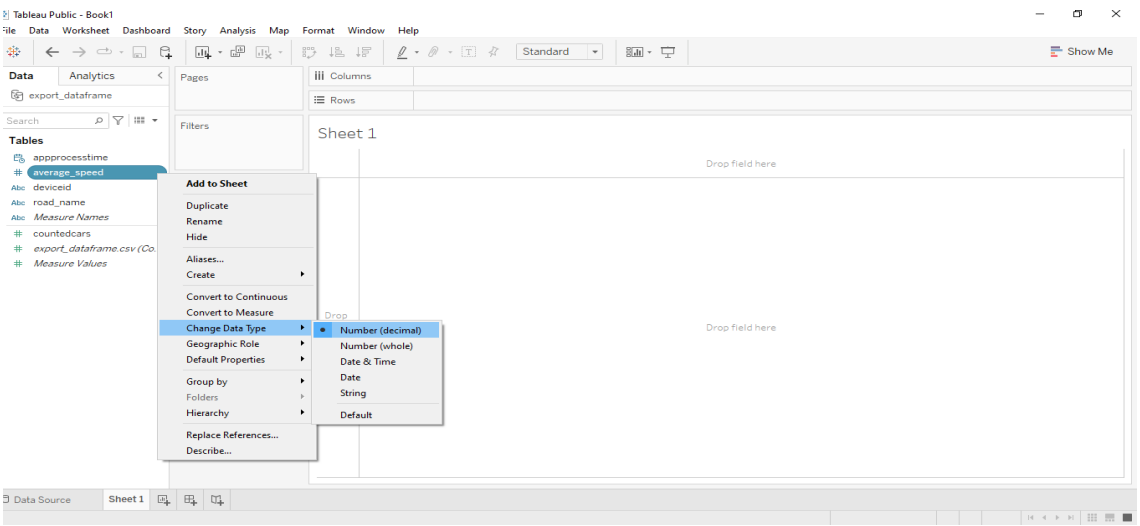
Στην δική μας περίπτωση έχουμε τα δεδομένα μας σε αρχείο csv κι επιλέγουμε σύνδεση με έγγραφο κειμένου (text file). Τότε ανοίγει ένα παράθυρο για να επιλέξουμε το σύνολο που θέλουμε να συνδέσουμε από τα τοπικά μας αρχεία κι επιλέγουμε το «csv_combined». Το σύστημα μας ενημερώνει πόση ώρα κάνει να φορτώσει τα δεδομένα κι όταν είναι έτοιμα προς χρήση μας ανακατευθύνει στο παρακάτω παράθυρο.



Εικόνα 30 Διεπαφή Tableau μετά την εισαγωγή του συνόλου δεδομένων κίνησης

Εδώ μπορούμε να δούμε το αρχείο που έχει συνδεθεί και πριν προχωρήσουμε μας εμφανίζονται κάποιες βασικές πληροφορίες, όπως τα ονόματα των στηλών και τον τύπο των δεδομένων. Μπορούμε είτε να χρησιμοποιήσουμε την επιλογή «Use Data Interpreter» στο Αρχείο, αριστερά κάτω από το Συνδέσεις για να «καθαρίσει» το text file βιβλίο εργασίας, διαφορετικά μπορούμε να ελέγξουμε τον τύπο δεδομένων κάθε στήλης και να τον αλλάξουμε χειροκίνητα, όπως φαίνεται στην πάνω εικόνα.

Η στήλη «average_speed» βλέπουμε πως έχει λάθος τύπο δεδομένων, της μορφής string ενώ πρόκειται για δεκαδικό αριθμό κι εμφανίζεται ως «Abc» πάνω από το όνομα της στήλης, οπότε πηγαίνουμε στο βελάκι δίπλα από το «Abc» κι όταν το πατήσουμε εμφανίζεται ένα drop down menu με κάποιες επιλογές, πηγαίνουμε στο «Change data type» κι επιλέγουμε «Number(decimal)».

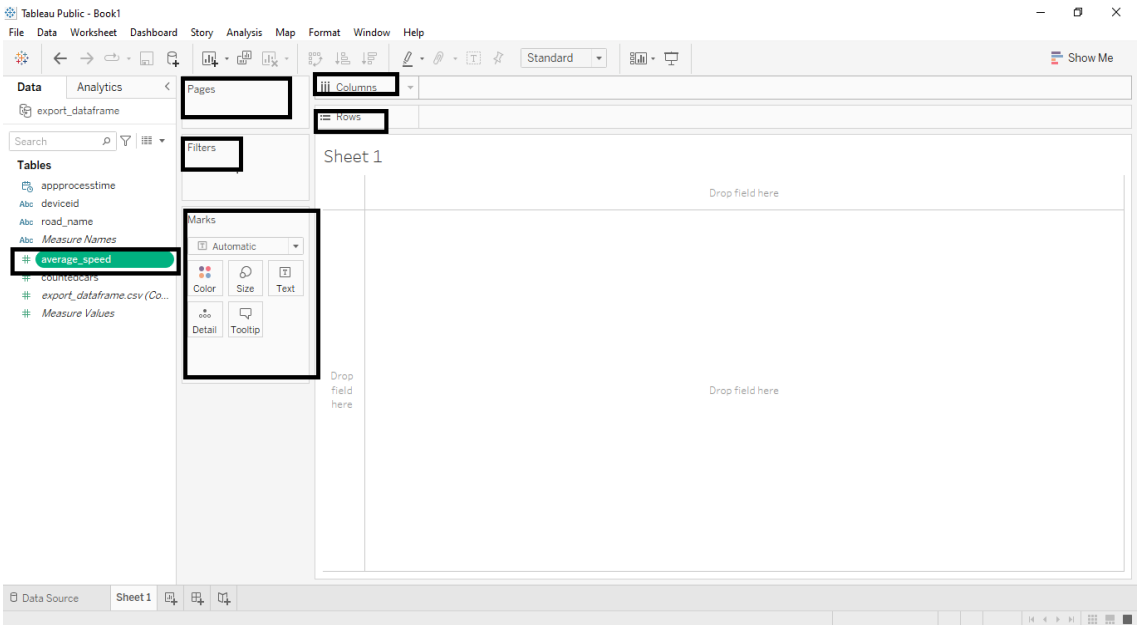


Εικόνα 31 Αλλαγή τύπου δεδομένων της στήλης average speed χειροκίνητα

Επίσης, η στήλη «appprocesstime» έχει τύπο δεδομένων string ενώ στην πραγματικότητα είναι ημερομηνία και ώρα οπότε αλλάζουμε με τον ίδιο τρόπο τον τύπο δεδομένων αυτής της στήλης και τον μετατρέπουμε σε «Date and Time». Στην περίπτωση που δεν αντιληφθούμε από την αρχή πως ένας τύπος δεδομένων είναι λανθασμένος υπάρχει η δυνατότητα τον αλλάξουμε όταν είμαστε μέσα στο Sheet πριν ξεκινήσουμε την ανάλυση.

4.2 Γνωριμία με το Tableau και δυνατότητες

Τώρα που τα δεδομένα μας είναι στην σωστή μορφή είμαστε έτοιμοι να περάσουμε στο κομμάτι της ανάλυσης μέσω διερευνητικών οπτικοποιήσεων.



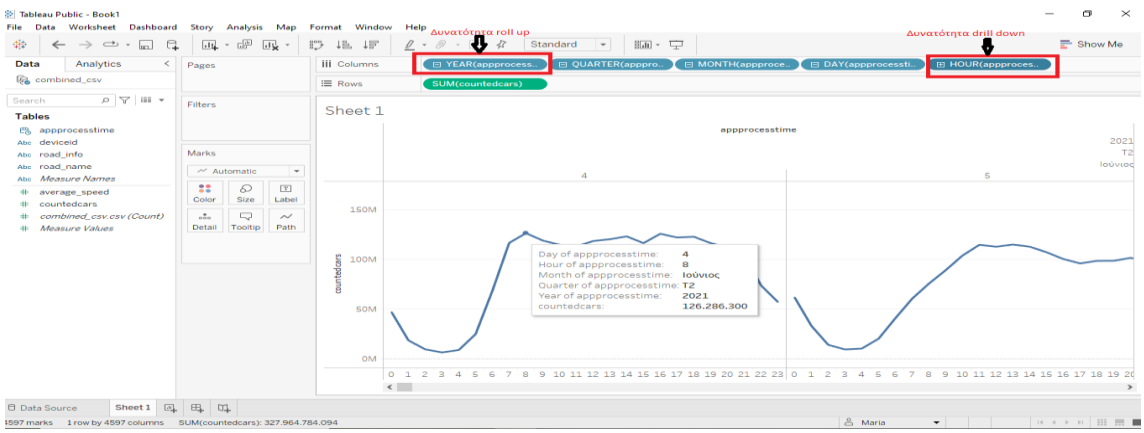
Εικόνα 32 Πεδία (shelves) του εργαλείου Tableau. Pages, Columns, Rows, Marks Cards which contain other shelves like color, size, text, detail, tooltip

Εδώ, βλέπουμε στα αριστερά τα δεδομένα(στήλες) με τα οποία θα εργαστούμε κι ακριβώς από δίπλα βλέπουμε τις επιλογές που υπάρχουν για να δημιουργήσουμε τις απεικονίσεις που επιθυμούμε.

«Εξ' ορισμού, τα πεδία που περιέχουν κείμενο, ημερομηνίες ή δυαδικές τιμές είναι διαστάσεις (dimensions), ενώ τα πεδία που περιέχουν αριθμητικές τιμές είναι μετρήσεις (measures).» (Σημειώσεις Εργαστηρίου Πληροφοριακών Συστημάτων, Tableau 2021, σελίδα 21). Γενικά, στο Tableau οι διαστάσεις είναι συνήθως κατηγορικά δεδομένα και οι μετρήσεις είναι συνήθως αριθμητικά δεδομένα. Οι στήλες με πράσινο χρόνο είναι συνεχής και οι στήλες με μπλε χρώμα είναι διακριτές.

Όταν δημιουργούμε μια απεικόνιση (ή όψη) αναλύουμε τις τιμές μιας ή παραπάνω μέτρησης σύμφωνα με μια ή παραπάνω κατηγορία διάστασης. Γενικά, οι οπτικοποιήσεις δημιουργούνται χρησιμοποιώντας Σειρές (Rows) και Στήλες (Columns), καθώς επίσης και Σελίδες (Pages), Φίλτρα (Filters) και Σημάδια (Marks). Ο τρόπος για να χτίσουμε μια οπτικοποίηση είναι με την τοποθέτηση ενός πεδίου δεδομένων στις παραπάνω περιοχές του χώρου εργασίας.

Ανάλογα με την δομή των δεδομένων μας το Tableau μας δίνει την δυνατότητα να κάνουμε drill down ή roll up σε μεγέθη, όπως για παράδειγμα σε μια στήλη με ημερομηνίες, όπου μπορούμε να κάνουμε drill down, δηλαδή πιο λεπτομερή απεικόνιση, ανάλογα με τον χρόνο, με το τρίμηνο, τον μήνα, την εβδομάδα, την ημέρα και την ώρα, φυσικά εφόσον αυτές οι πληροφορίες περιέχονται στα δεδομένα μας. Αντίστοιχα, αν θέλουμε πιο συγκεντρωτικά μεγέθη μπορούμε να κάνουμε roll up από την ώρα πχ στο έτος. Αυτό μπορούμε να το επιτύχουμε με πολύ απλό τρόπο απλά πατώντας το + ή το - δίπλα από το όνομα της στήλης. Στην προκειμένη αν σύρουμε π.χ. την στήλη «appprocesstime» στο «Columns» μας την εμφανίζει με την μορφή Year, πατώντας το + μας εμφανίζει τον μήνα και ούτω καθ' εξής μέχρι να φτάσει στην ώρα και αντίστοιχα μετά πατάμε το – αν θέλουμε να γυρίσουμε πχ στον μήνα από την ώρα ή την μέρα.

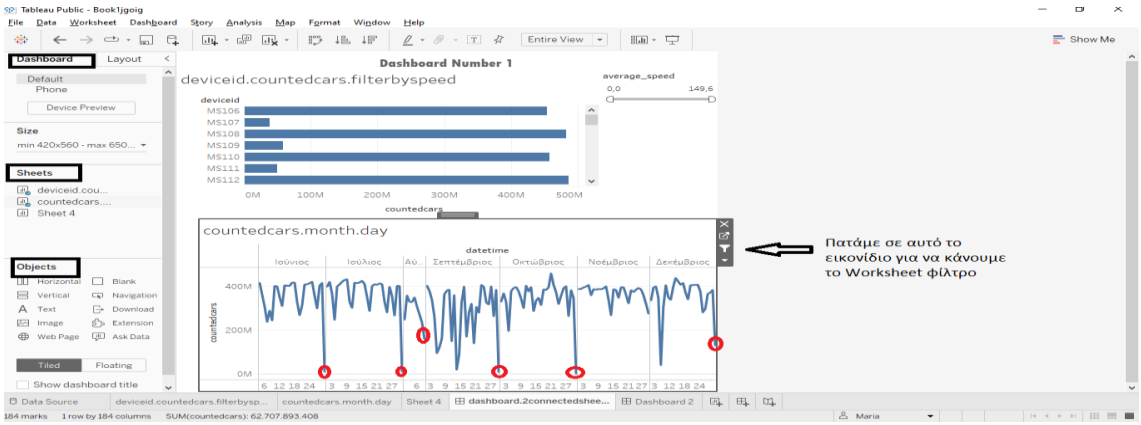


Εικόνα 33 Δυνατότητα drill down ή roll up στην πληροφορία στο Tableau

Καθώς επιλέγουμε τις στήλες που θέλουμε να οπτικοποιήσουμε με την λογική του drag and drop το Tableau δημιουργεί αυτόματα τον τύπο γραφήματος που θεωρεί κατάλληλο για την κάθε οπτικοποίηση. Αν πατήσουμε πάνω δεξιά στο «Show me» μας εμφανίζει όλους τους τύπους γραφημάτων και μπορούμε να επιλέξουμε κάποιον άλλο τύπο.

Υπάρχει η δυνατότητα να δημιουργήσουμε Dashboards με την ίδια λογική που δημιουργούμε τα Sheets πατώντας στην κάτω γραμμή εργαλείων το 2ο σήμα που

εμφανίζεται. Αφότου πατήσουμε το κουμπί λοιπόν «New Dashboard» μας εμφανίζει στο δεξί μέρος της διεπαφής τα διαθέσιμα Sheets που έχουμε δημιουργήσει και να επιλέξουμε ένα ή παραπάνω Sheets για να δημιουργήσουμε το Dashboard που επιθυμούμε. Πατώντας ταυτόχρονα το «shift» όταν σέρνουμε ένα φύλλο εργασίας μπορούμε να αλλάξουμε την θέση του στο Dashboard και το μέγεθός του και να αλλάξουμε γενικά την διάταξη. Μπορούμε να επιλέξουμε το μέγεθος από την επιλογή Size π.χ. για browser laptop, να επιλέξουμε πως θα φαίνεται η προεπισκόπηση για συσκευές και από το πεδίο που περιέχει τα Objects μπορούμε να βάλουμε κείμενο, εικόνες κλπ.



Εικόνα 34 Διεπαφή Dashboard Tableau, Χαρακτηριστικά και Δυνατότητες

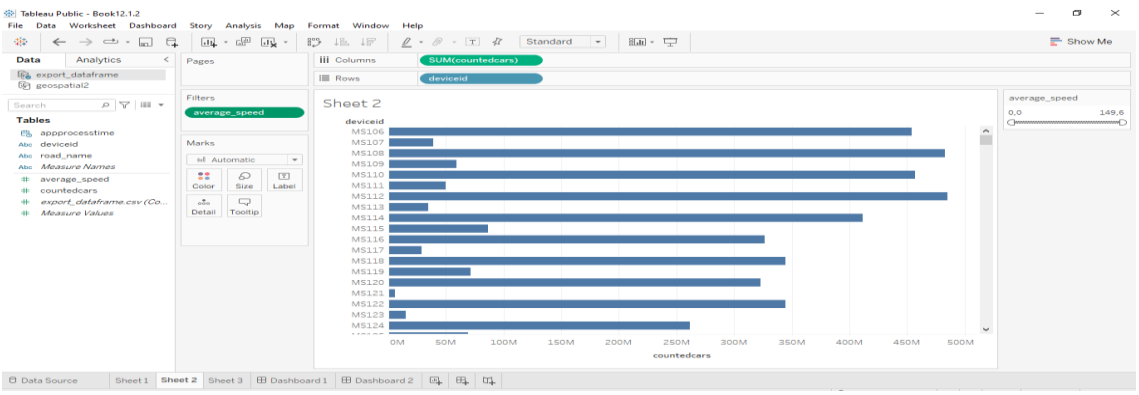
Μια πολύ χρήσιμη ιδιότητα του Tableau είναι πως μπορούμε να πάμε σε όποιο φύλλο εργασίας επιθυμούμε και πατώντας το βελάκι δίπλα στο φίλτρο που έχουμε βάλει να εφαρμόσουμε το φίλτρο σε όποια worksheets θέλουμε ή και σε όλα τα worksheets που χρησιμοποιούν την ίδια πηγή δεδομένων. Με αυτό τον τρόπο το dashboard γίνεται διαδραστικό και κάθε φορά που πατάμε το φίλτρο π.χ. την ταχύτητα να μας δείχνει αντίστοιχα πως αλλάζουνε και τα συνδεδεμένα worksheets με βάση το αντίστοιχο φίλτρο. Επίσης, όσο είμαστε στο Dashboard πατώντας πάνω στο Sheet που θέλουμε πάνω δεξιά στο πλάι υπάρχει μια επιλογή που έχει το σχήμα της χοάνης και αν την επιλέξουμε λειτουργεί όλο το sheet σαν φίλτρο και για τα υπόλοιπα Sheets που βρίσκονται στο ίδιο Dashboard. Έτσι π.χ. μπορούμε να δούμε πιο συγκεκριμένα χαρακτηριστικά. Ενδιαφέρον παρουσιάζει στο παραπάνω διάγραμμα με τον αριθμό των καταγεγραμμένων οχημάτων ως προς τον μήνα και την μέρα πως την τελευταία μέρα του κάθε μήνα έχουμε τις λιγότερες καταγραφές οχημάτων που πιθανότητα οφείλεται σε κάποια δυσλειτουργία των αισθητήρων ή σε σφάλμα του δικτύου.

4.3 Δημιουργία γραφημάτων στο Tableau

Κάθε οπτικοποίηση ή όψη που θέλουμε να δημιουργήσουμε στο Tableau είναι καλό να πηγάζει από κάποιο ερώτημα στο οποίο ψάχνουμε να βρούμε κάποια απάντηση. Κάθε φορά που «σέρνουμε» ένα πεδίο σε κάποιο «shelf» , π.χ. στις στήλες, κάνουμε μια ερώτηση για τα δεδομένα μας και αυτή η ερώτηση διαφέρει ανάλογα με το που θα σύρουμε τα διάφορα πεδία, τους τύπους των πεδίων και την σειρά που θα σύρουμε τα πεδία στο γράφημα ή στην όψη. Ο τρόπος που δημιουργούμε κάθε γράφημά εξαρτάται αρκετά από τον τύπο των δεδομένων που έχουμε και από την ερώτηση που θέλουμε να απαντήσουμε.

Υπάρχουν κάποια χαρακτηριστικά στο Tableau που μπορούν να εφαρμοστούν σε όλους τους τύπους γραφημάτων. Αρχικά χρησιμοποιούμε το ιστόγραμμα (bar chart) για να συγκρίνουμε δεδομένα ανάμεσα σε διαφορετικές κατηγορίες πληροφοριών όπως φαίνεται παρακάτω.

Στην παρακάτω γραφική αναπαράσταση μπορούμε να δούμε τα μετρούμενα οχήματα (countedcars) από τον κάθε αισθητήρα (deviceid).

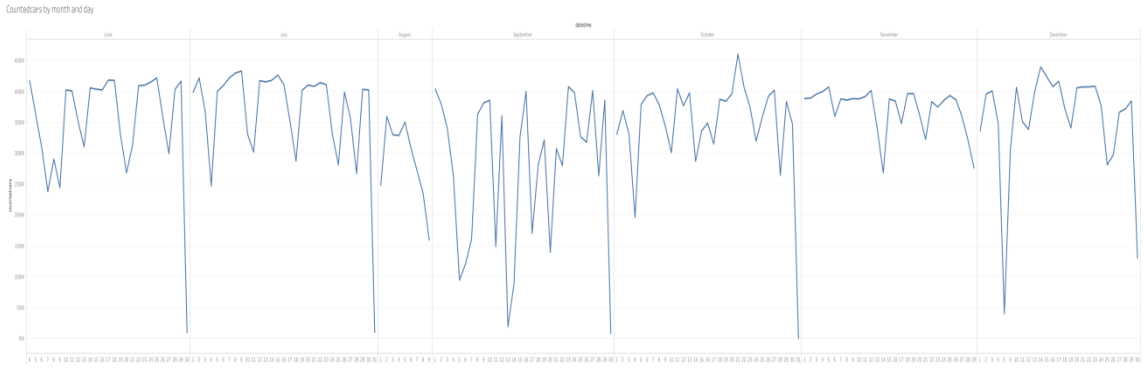


Εικόνα 35 Οριζόντιο Ιστόγραμμα με το αναγνωριστικό του κάθε αισθητήρα στις σειρές και τα καταγεγραμμένα οχήματα σαν στήλες και φίλτρο την μέση ταχύτητα οχημάτων

Έχουμε βάλει σαν φίλτρο την μέση ταχύτητα των οχημάτων (average speed) και αν πατήσουμε στην μπάρα πάνω δεξιά που γράφει «average speed» και είναι από 0 μέχρι 149 βλέπουμε πως μεταβάλλεται το διάγραμμα και πόσα αμάξια κατέγραψε ο κάθε αισθητήρας ανάλογα με την ταχύτητα που επιλέξαμε.

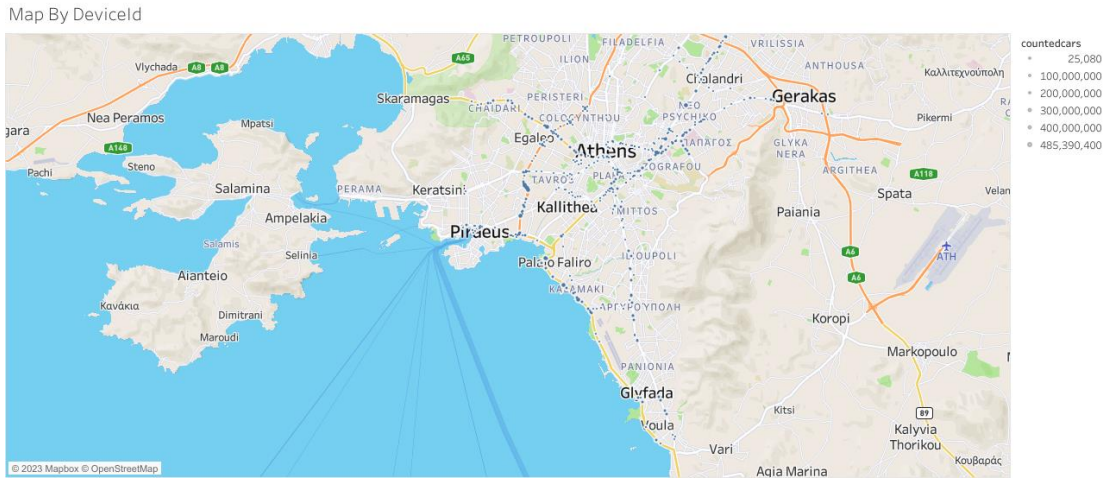
Στη συνέχεια χρησιμοποιούμε ένα γράφημα γραμμής (line chart) για να ενώσουμε μεμονωμένα σημεία δεδομένων προκειμένου να οπτικοποιήσουμε μια ακολουθία

καταγεγραμμένων οχημάτων ανά μήνα και ανά ημέρα ώστε να δούμε την μεταβολή αυτών με τον χρόνο.



Εικόνα 36 Line Chart για καταγεγραμμένα οχήματα ανά μήνα και ανά ημέρα από Ιούνιο μέχρι Δεκέμβριο

Όπως αναφέρθηκε και προηγουμένως μπορούμε να διακρίνουμε πως την τελευταία μέρα του κάθε μήνα πέφτει κατακόρυφα ο αριθμός των καταγεγραμμένων οχημάτων που οφείλεται πιθανώς σε κάποια δυσλειτουργία των αισθητήρων ή σε κάποιο σφάλμα στο σύστημα. Τον Αύγουστο έχουμε πολύ λίγα δεδομένα όπως είδαμε και στην αρχική διερεύνηση των δεδομένων μας και πατώντας πάνω στα σημεία που παρουσιάζουν άκρα προς τα πάνω ή προς τα κάτω μπορούμε να δούμε πως έχουμε πχ στις 28.10 μειωμένη διέλευση οχημάτων ή πχ στις 25.12 έχουμε πάλι μειωμένη διέλευση οχημάτων καθώς είναι και οι δύο μέρες αργίας. Για αυτό τον λόγο θα χρησιμοποιήσουμε μια βιβλιοθήκη για τις αργίες και τις γιορτές στην Ελλάδα στο μοντέλο πρόβλεψης για να δούμε πως επηρεάζουν την κυκλοφορία των δρόμων.

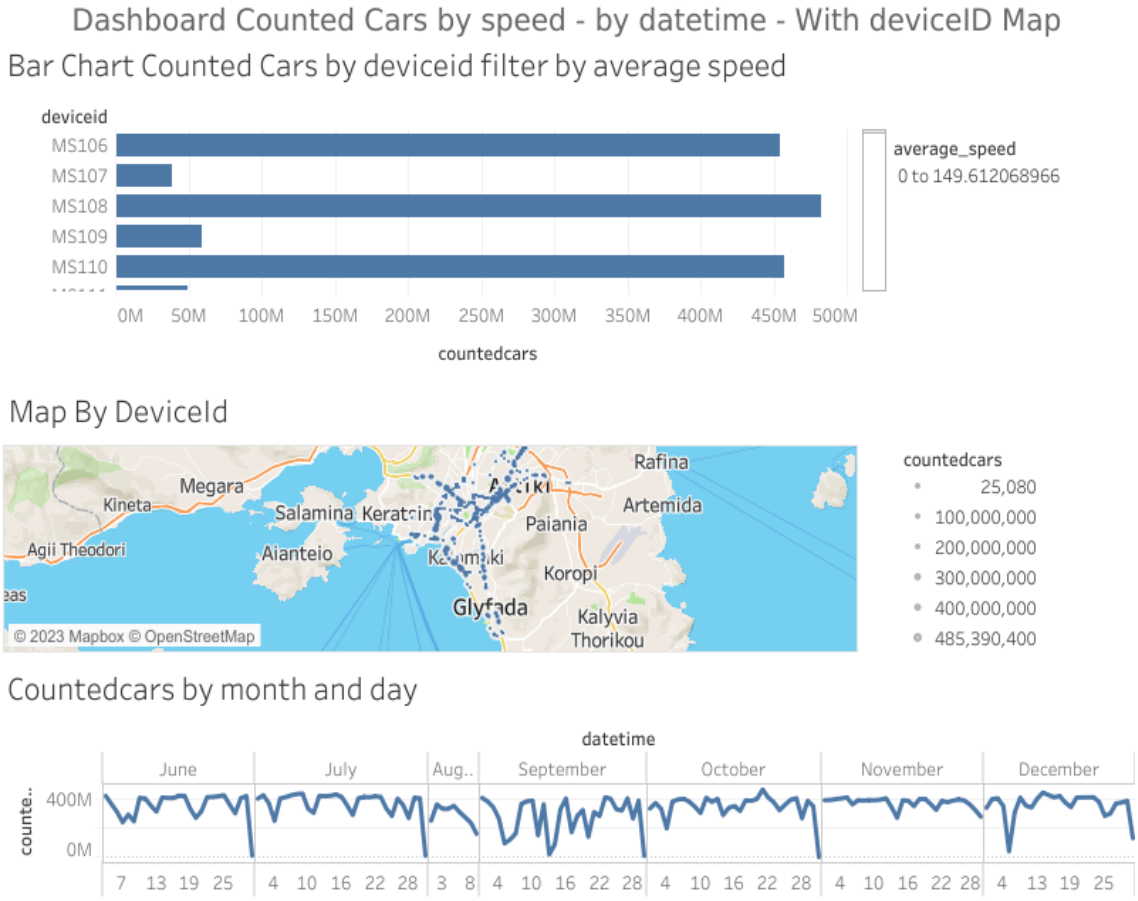


Εικόνα 37 Χάρτης Αττικής στο Tableau με βάση τις γεωχωρικές συντεταγμένες των αισθητήρων του data.gov.gr και με mark το μέγεθος των καταγεγραμμένων οχημάτων

Μια επιπλέον λειτουργικότητα του Tableau που παρουσιάζει ενδιαφέρον είναι η δυνατότητα χαρτογράφησης (mapping functionality) καθώς μπορεί να δημιουργήσει

γραφικές αναπαραστάσεις βάσει των συντεταγμένων γεωγραφικού πλάτους και μήκους και να συνδεθεί με χωροταξικά (spatial) αρχεία. Η ενσωματωμένη γεωκωδικοποίηση επιτρέπει να χαρτογραφηθούν αυτόματα μεταξύ άλλων η χώρα, η περιφέρεια και ο ταχυδρομικός κώδικας. (Tableau, Wikipedia). Χρησιμοποιώντας λοιπόν τα χωροταξικά δεδομένα των αισθητήρων στους δρόμους στην Αττική που έχουμε στην κατοχή μας από την εργασία «**Ανακαλύπτοντας την ποιότητα των Δυναμικών Ανοιχτών Κυβερνητικών Δεδομένων με την χρήση στατιστικών μεθόδων και μεθόδων μηχανικής μάθησης**» των Καραμάνου, Μπρίμος, Καλαμπόκης, Ταραμπάνης μπορούμε να δημιουργήσουμε ένα χάρτη στο Tableau.

Στη συνέχεια αφότου έχουμε δημιουργήσει τα παραπάνω γραφήματα προχωράμε στην δημιουργία ενός ταμπλό (dashboard) που περιλαμβάνει όλα τα παραπάνω και έχοντας βάλει σαν φίλτρο και στα τρία την μέση ταχύτητα μπορούμε να έχουμε πληροφορία που αλλάζει δυναμικά και στα τρία με βάση την ταχύτητα.



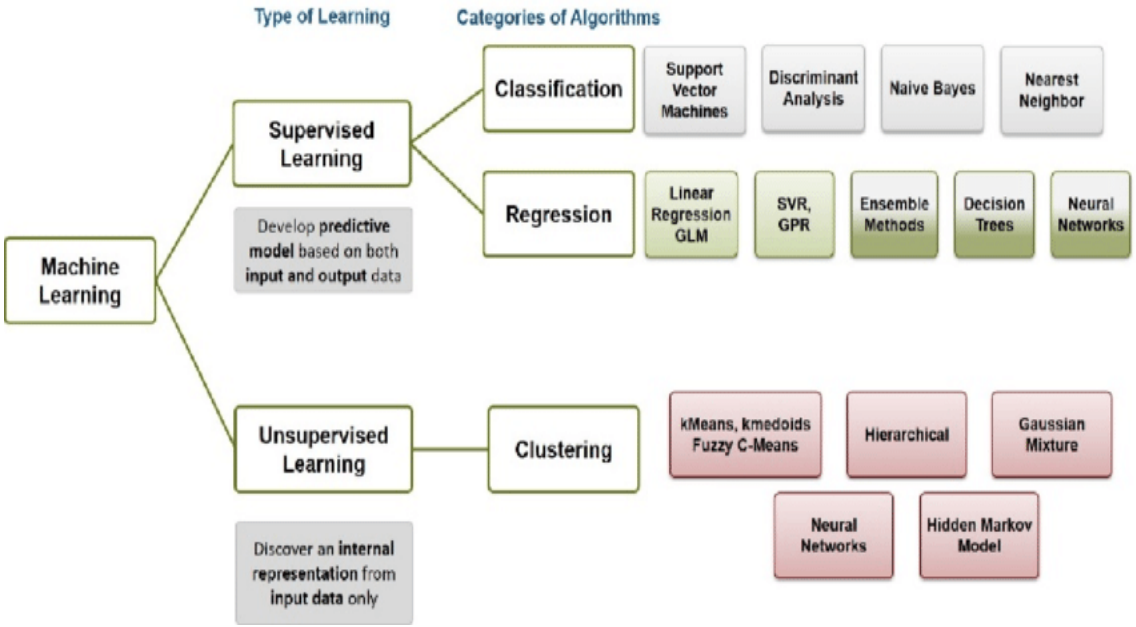
Εικόνα 38 Dashboard στο Tableau με καταγεγραμμένα οχήματα και μέση ταχύτητα, με βάση τον χρόνο και τα καταγεγραμμένα οχήματα και χάρτης Αττικής με αισθητήρες

Αυτό μας βοηθάει να αποκτήσουμε μια πιο σφαιρική εικόνα και να απαντήσουμε πιο εύκολα σε πιθανά ερωτήματα για το σε ποιους αισθητήρες παρατηρούνται

μεγαλύτερες ταχύτητες, που είναι τοποθετημένοι οι αισθητήρες π.χ. μακριά ή κοντά στο κέντρο της πόλης και πως αλλάζει ο αριθμός των καταγεγραμμένων οχημάτων με βάση τον μήνα, την μέρα και την ταχύτητα.

5. Πειραματισμός και προβλέψεις

Ουσιαστικά το ζητούμενο της εργασίας μας είναι μέσω της ανάλυσης δεδομένων και με την χρήση διαφόρων τεχνικών και αλγορίθμων να μπορέσουμε να κάνουμε μια πρόβλεψη σχετικά με τα μετρούμενα οχήματα στη μελλοντική κυκλοφορία της Αττικής. Σύμφωνα με τον Sharda (2017, σελ.196) «η εξόρυξη δεδομένων είναι μια διαδικασία που χρησιμοποιεί στατιστικές τεχνικές, μαθηματικές τεχνικές και τεχνικές τεχνητής νοημοσύνης για να εξάγει και να αναγνωρίσει χρήσιμες πληροφορίες και επακόλουθες γνώσεις (ή μοτίβα) από μεγάλα σύνολα δεδομένων. Αυτά τα μοτίβα μπορούν να έχουν την μορφή επιχειρηματικών κανόνων, σχέσεων, συσχετίσεων, τάσεων ή μοντέλων πρόβλεψης.» Στην δική μας περίπτωση βρισκόμαστε στα μοντέλα πρόβλεψης. Συνεχίζοντας, (2017, σελ. 200) «οι εργασίες ανάλυσης δεδομένων μπορούν να χωριστούν σε τρεις κατηγορίες, την πρόβλεψη, την συσχέτιση και την κατηγοριοποίηση. Ανάλογα με τον τρόπο που εξάγονται τα μοτίβα από τα ιστορικά δεδομένα, οι αλγόριθμοι μπορούν να κατηγοριοποιηθούν σαν επιβλεπόμενοι ή μη επιβλεπόμενοι.»



Εικόνα 39 Διάγραμμα ταξινόμησης Επιβλεπόμενης και Μη επιβλεπόμενης Μηχανικής Μάθησης & Κατηγορίες Αλγορίθμων

Στο παραπάνω διάγραμμα (Πηγή: Researchgate, Abdullah, & Hasan, Mohammad. (2017). An application of pre-trained CNN for image classification. 1-6. 10.1109/ICCITECHN.2017.8281779.) αναγράφονται οι βασικοί τύποι μηχανικής μάθησης με τις κύριες κατηγοριοποιήσεις προβλημάτων (Classification, Regression, Clustering) και τους πιο συνηθισμένους αλγορίθμους μηχανικής μάθησης. Προτού προχωρήσουμε λοιπόν ακολουθεί μια σύντομη αναφορά στο θεωρητικό υπόβαθρο σχετικά με την τεχνητή νοημοσύνη και την μηχανική μάθηση, τα είδη της και σε ποια κατηγορία εμπίπτει το δικό μας πρόβλημα.

5.1 Ταξινόμηση μοντέλων μηχανικής μάθησης

Γενικά, η κατηγοριοποίηση των μοντέλων μηχανικής μάθησης μπορεί να παρουσιάσει κάποιες διαφοροποιήσεις στην βιβλιογραφία ανάλογα με την φύση των χαρακτηριστικών της ανάλυσης. Ακολουθούν επιγραμματικά οι διάφορες κατηγοριοποιήσεις που είναι πιθανό να συναντήσουμε.

1. **Ταξινόμηση με βάση την φύση των δεδομένων** που τροφοδοτούμε για να εκπαιδεύσουμε το μοντέλο μας:
    - Επιβλεπόμενη Μηχανική Μάθηση (Supervised Machine Learning)
    - Μη Επιβλεπόμενη Μηχανική Μάθηση (Unsupervised Machine Learning)
    - Ημι-επιβλεπόμενη Μηχανική Μάθηση (Semi-supervised Machine Learning)
    - Ενισχυτική Μάθηση (Reinforcement Learning)

2. **Ταξινόμηση με βάση τη φύση του προβλήματος** που προσπαθούμε να επιλύσουμε, η οποία είναι η ταξινόμηση που έχουμε επιλέξει για την εργασία μας
    - Πρόβλημα Ταξινόμησης (Classification Problem)
    - Πρόβλημα Παλινδρόμησης (Regression Problem)
    - Πρόβλημα Ομαδοποίησης (Clustering Problem)

3. **Ταξινόμηση με βάση την φύση του αλγορίθμου** που χρησιμοποιούμε
    - Κλασσική Μηχανική Μάθηση (Classical Machine Learning)

- Νευρωνικά Δίκτυα (Neural Networks)
- Βαθιά Μάθηση (Deep Learning)

4. **Ταξινόμηση με βάση τη φύση της επίλυσης**
    - Παραμετρικά Μοντέλα (Parametric Models)
    - Μη Παραμετρικά Μοντέλα (Non-Parametric Models)

5. **Ταξινόμηση με βάση την φύση του αποτελέσματος**
    - Πιθανολογικά Μοντέλα (Probabilistic Models)
    - Μη Πιθανολογικά Μοντέλα (Non-Probabilistic Models)

Στην δική μας περίπτωση, όπως αναφέρεται και πιο πάνω, επιλέξαμε την ταξινόμηση με βάση την φύση του προβλήματος καθώς το ζητούμενο μας είναι η πρόβλεψη τιμής.

5.2 Επιβλεπόμενη Μηχανική Μάθηση

Η Μηχανική Μάθηση είναι από τις πιο συχνές εφαρμογές της Τεχνητής Νοημοσύνης κατά την οποία μια μηχανή «μαθαίνει» να εκτελεί εργασίες από τα δεδομένα που της τροφοδοτούμε. Καθώς μαθαίνει από αυτά τα δεδομένα βελτιώνεται και η απόδοση του μοντέλου σε μια συγκεκριμένη εργασία.

Η Επιβλεπόμενη Μηχανική Μάθηση είναι ουσιαστικά μια διαδικασία ανάλυσης δεδομένων που χρησιμοποιεί σύγχρονες μεθόδους όπως π.χ. τα δέντρα απόφασης (decision trees), τα τυχαία δάση (random forests) και τις μηχανές ενίσχυσης κλίσης (gradient boosting machines). Οι αλγόριθμοι προσπαθούν να μοντελοποιήσουν σχέσεις και εξαρτήσεις μεταξύ των χαρακτηριστικών εισόδου (input features) και του αποτελέσματος της στοχευόμενης πρόβλεψης (target prediction output) ώστε να μπορέσουμε να προβλέψουμε τις τιμές του αποτελέσματος για νέα δεδομένα με βάση αυτές τις σχέσεις που οι αλγόριθμοι έχουν εκπαιδευτεί στα προηγούμενα σύνολα δεδομένων.

Για να γίνει αυτό χρησιμοποιούμε labeled σύνολα δεδομένων (που έχουμε συγκεντρώσει μέσω της άντλησης δεδομένων από έγκυρες πηγές) και άλλες διαδικασίες σαν εισροή (inputs) για να εκπαιδεύσουμε το μοντέλο μας ώστε να καταλήξουμε σε ένα σωστό αποτέλεσμα(output), είτε πρόκειται για κατηγοριοποίηση δεδομένων είτε

πρόκειται για πρόβλεψη τιμής. Κατά την φάση της εκπαίδευσης καθώς τροφοδοτούμε το μοντέλο με δεδομένα ο αλγόριθμος επιβλέπει συνεχώς την ακρίβεια του μοντέλου και προσαρμόζεται μέχρι να έχουμε το μικρότερο ποσοστό λάθους και το μοντέλο να είναι σωστό (fitted) καταλλήλως.

Η Επιβλεπόμενη Μηχανική Μάθηση χωρίζεται σε δύο τύπους προβλημάτων όπως αναφέρθηκε και προηγουμένως, στην Ταξινόμηση (Classification) και στην Παλινδρόμηση (Regression). Στην Ταξινόμηση χρησιμοποιείται ένας αλγόριθμος προκειμένου να αναθέσει με ακρίβεια τα test data σε συγκεκριμένες κατηγορίες , π.χ. το πρόβλημα customer churn ή αν μια παραγγελία θα γίνει εκ νέου ή όχι. Η Παλινδρόμηση (Regression) χρησιμοποιείται προκειμένου να γίνει αντιληπτή η σχέση ανάμεσα σε μια εξαρτημένη μεταβλητή και μη εξαρτημένες μεταβλητές. Χρησιμοποιείται συνήθως για προβλέψεις τιμών, όπως ο μελλοντικός αριθμός πωλήσεων ή η τιμή της θερμοκρασίας μια συγκεκριμένη μέρα. Η βασική διαφορά ανάμεσα στους δύο τύπους προβλημάτων είναι πως για την Παλινδρόμηση η εξαρτημένη μεταβλητή είναι αριθμητική ενώ για την Ταξινόμηση είναι κατηγορική.

Οι πιο συχνά χρησιμοποιούμενοι αλγόριθμοι σε αυτή την κατηγορία μηχανικής μάθησης είναι οι εξής:

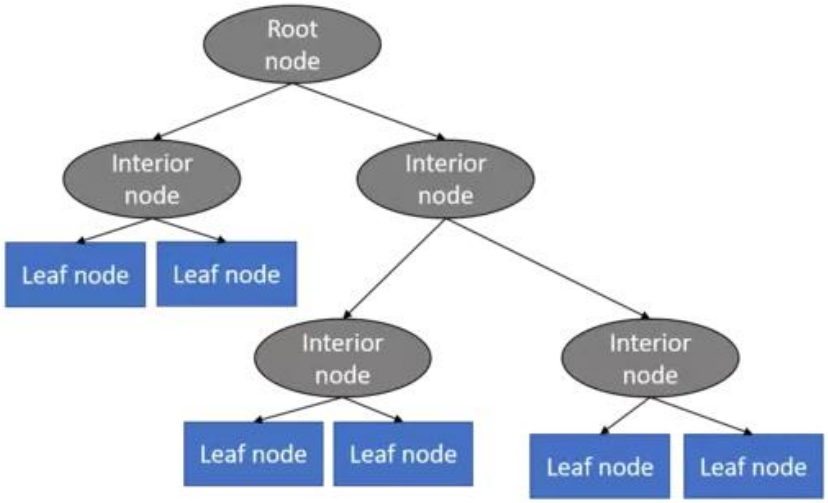
- Neural Networks
- Naïve bayes
- Linear Regression
- Logistic Regression
- Support Vector Machines (SVM)
- K-Nearest Neighbor
- Random Forest
- Decision Tree

Το πρόβλημα της παρούσας εργασίας εμπίπτει στην κατηγορία της Επιβλεπόμενης Μηχανικής Μάθησης και πιο συγκεκριμένα στην Παλινδρόμηση (Regression) καθώς το ζητούμενό μας είναι η πρόβλεψη τιμής, δηλαδή ο αριθμός των καταγεγραμμένων οχημάτων σε έναν συγκεκριμένο αισθητήρα. Με βάση τα παραπάνω είμαστε πλέον σε θέση να επιλέξουμε το κατάλληλο μοντέλο για το πρόβλημα μας καθώς και τις παραμέτρους που θα χρησιμοποιήσουμε.

5.3 Δέντρο απόφασης (Decision Tree) και μοντέλο XGBoost

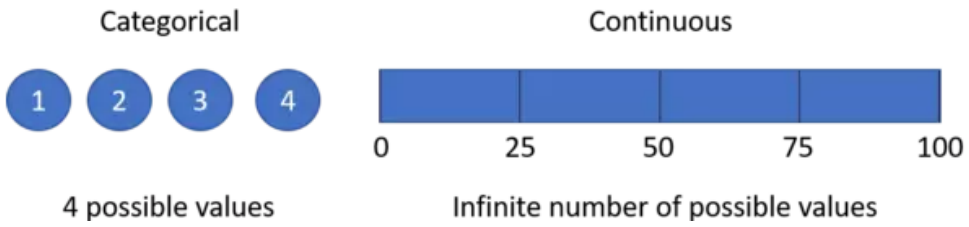
Σύμφωνα με την Wikipedia «το **μοντέλο δέντρου απόφασης** είναι το μοντέλο υπολογισμού στο οποίο ένας αλγόριθμος αντιμετωπίζεται ως ένα δέντρο αποφάσεων, δηλαδή μια ακολουθία ερωτημάτων ή δοκιμών που γίνονται προσαρμοστικά, οπότε το αποτέλεσμα των προηγούμενων δοκιμών / ελέγχων μπορεί να επηρεάσει τη δοκιμή που πρόκειται να εκτελεστεί στη συνέχεια.» .

Ένα δέντρο απόφασης είναι μια δομή που μοιάζει με Διάγραμμα Ροής όπου ο αλγόριθμος χρησιμοποιεί ένα Δέντρο απόφασης σαν μοντέλο πρόβλεψης για να περάσει από παρατηρήσεις για ένα πράγμα(που αντιπροσωπεύεται στα κλαδιά) σε συμπεράσματα για την στοχευόμενη αξία του πράγματος (που αντιπροσωπεύεται στα φύλλα). Είναι μία από τις προσεγγίσεις της προβλεπτικής μοντελοποίησης που χρησιμοποιείται στην Στατιστική, στην Εξόρυξη Δεδομένων και στην Μηχανική Μάθηση.



Εικόνα 40 Διάγραμμα Decision Tree and Nodes

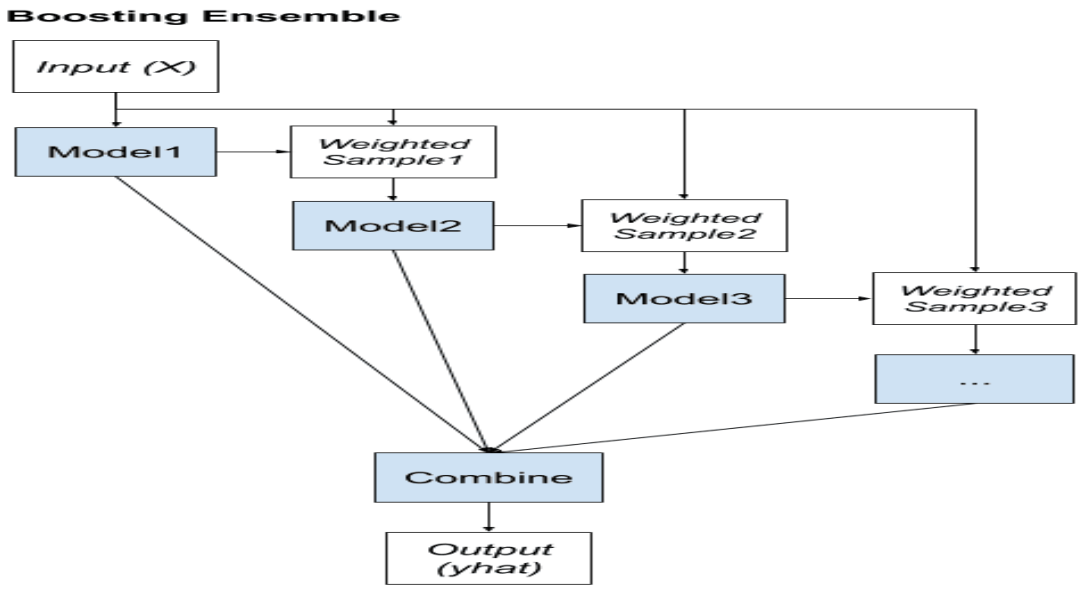
Τα μοντέλα Δέντρου Απόφασης όπου η target μεταβλητή μπορεί να πάρει ένα σύνολο διακριτών (discrete) τιμών ονομάζονται Δέντρα Ταξινόμησης. Τα Δέντρα απόφασης όπου η target μεταβλητή μπορεί να πάρει συνεχή (continuous) τιμή ονομάζονται Δέντρα Παλινδρόμησης και σε αυτές τις δομές δέντρων τα φύλλα (leaves) αντιπροσωπεύουν τις στοχευόμενες μεταβλητές (target variable) και τα κλαδιά (branches) αντιπροσωπεύουν τα χαρακτηριστικά (features).



Εικόνα 41 Πιθανές τιμές Μοντέλου Δέντρου Απόφασης ανάλογα με το αν είναι πρόβλημα Ταξινόμησης ή Παλινδρόμησης

Παρόλο που τα μοντέλα Δέντρου Απόφασης χρησιμοποιούνται συχνά στην Μηχανική Μάθηση είναι επιρρεπή σε προβλήματα όπως η μεροληψία (bias) και η «υπερβολική προσαρμογή» (overfitting). Αυτό μπορεί να διορθωθεί όταν πολλά δέντρα απόφασης δημιουργούν ένα σύνολο (ensemble) όπως π.χ. με τον αλγόριθμο random forest ή με τον αλγόριθμο XGBoost ώστε τα αποτελέσματα της πρόβλεψης έχουν μεγαλύτερη ακρίβεια.

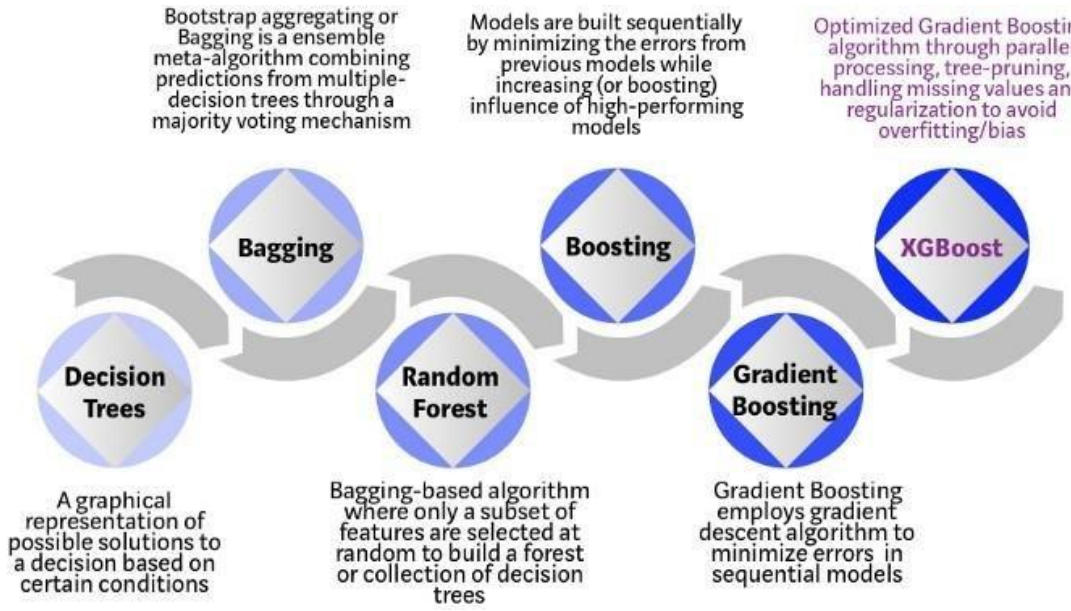
Οι μέθοδοι εκμάθησης συνόλου (ensemble learning methods), όπως ο αλγόριθμος XGBoost που επιλέχθηκε για την συγκεκριμένη μελέτη περίπτωσης, αποτελούνται από ένα σύνολο ταξινομητών, όπως τα Δέντρα Απόφασης, και οι προβλέψεις τους παρουσιάζουν καλύτερη απόδοση συνδυάζοντας τις προβλέψεις από πολλά μοντέλα αντί για ένα. Οι τρεις βασικές κατηγορίες χωρίζονται σε bagging, stacking και boosting. Η ενίσχυση (boosting), που είναι η κατηγορία που μας αφορά με βάση την επιλογή του αλγορίθμου μας, περιλαμβάνει την διαδοχική προσθήκη μελών (δέντρων) του συνόλου που διορθώνουν τις προβλέψεις που έγιναν από τα προηγούμενα μοντέλα και βγάζει σαν αποτέλεσμα έναν σταθμισμένο μέσο όρο προβλέψεων.



Εικόνα 42 Διάγραμμα Ενίσχυσης Συνόλου (Boosting Ensemble)

Η Extreme Gradient Boosting (ή XGBoost) είναι μια αποτελεσματική εφαρμογή ανοιχτού κώδικα (open source) του αλγορίθμου ενίσχυσης κλίσης (gradient boosting algorithm). Πέρα από αλγόριθμος είναι επίσης και βιβλιοθήκη στην Python. Ο όρος «gradient boosting» προέρχεται από την λογική της ενίσχυσης ή βελτίωσης ενός αδύναμου μοντέλου όπου συνδυάζοντας το με έναν αριθμό άλλων αδύναμων μοντέλων παράγεται ένα συλλογικά δυνατό μοντέλο.

Ουσιαστικά στην ενίσχυση κλίσης (gradient boosting) τα Δέντρα απόφασης χτίζονται το ένα μετά το άλλο και κάθε Δέντρο που προστίθεται δημιουργείται για να βελτιώσει τις ελλείψεις των προηγούμενων δέντρων και ελαχιστοποιεί την κλίση της συνάρτησης απώλειας καθώς ο αλγόριθμος δημιουργεί κάθε δέντρο. Για να αποφύγουμε το overfitting είναι σημαντικό να ρυθμίσουμε προσεκτικά τις παραμέτρους.



Εικόνα 43 Εξέλιξη του αλγορίθμου XGBoost από τα Δέντρα Απόφασης & Βασικά χαρακτηριστικά κάθε αναγραφόμενου αλγορίθμου

Στην παραπάνω εικόνα (Πηγή: Krishnan, Sundar & Neyaz, Ashar & Liu, Qingzhong. (2021). IoT Network Attack Detection using Supervised Machine Learning. 10. 18-32.) βλέπουμε την εξέλιξη του αλγορίθμου XGBoost που ξεκίνησε από τα Δέντρα Απόφασης και τα βασικά χαρακτηριστικά των αναγραφόμενων μοντέλων. Είναι ένας βελτιστοποιημένος αλγόριθμος ενίσχυσης κλίσης που λειτουργεί με παράλληλη επεξεργασία, «κλάδεμα» δέντρων, διαχείριση ελλειπουσών τιμών και κανονικοποίησης για να αποφευχθεί το bias και το overfitting.

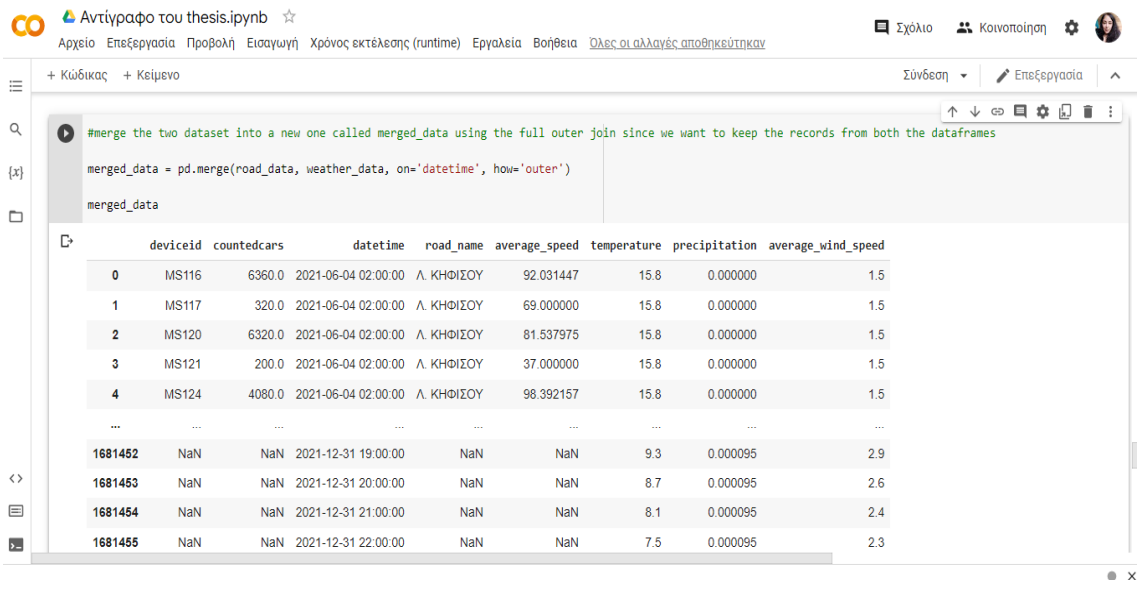
Ο αλγόριθμος XGBoost είναι μια επεκτάσιμη (scalable) και υψηλής ακρίβειας εφαρμογή της ενίσχυσης κλίσης (gradient boosting) που ανεβάζει την υπολογιστική ισχύ των ενισχυμένων αλγορίθμων Δέντρων Απόφασης και δημιουργήθηκε για να ενεργοποιεί την απόδοση του μοντέλου μηχανικής μάθησης και την υπολογιστική ταχύτητα.

Τα τελευταία χρόνια έχει γίνει πολύ δημοφιλής καθώς έχει κερδίσει πολλούς διαγωνισμούς μηχανικής μάθησης που οργανώνονται από εταιρείες και ερευνητές μέσα στην πλατφόρμα Kaggle όπου ο στόχος είναι να δημιουργηθεί το καλύτερο μοντέλο για προβλέψεις και περιγραφές δεδομένων. Η επιλογή αυτού του αλγορίθμου έγινε καθώς μπορεί να χρησιμοποιηθεί απευθείας σε προβλεπτικά μοντέλα παλινδρόμησης και επίσης λόγω της υψηλής ταχύτητας εκτέλεσης, της βελτιωμένης απόδοσης του μοντέλου και των μειωμένων σφαλμάτων.

5.4 Εφαρμογή μοντέλου XGBoost

Εφόσον τα δεδομένα κυκλοφορίας και τα δεδομένα καιρού είναι πλέον στην κατάλληλη επιθυμητή μορφή μπορούμε να προχωρήσουμε στην συνένωση αυτών έχοντας πλέον κρατήσει τις στήλες (μεταβλητές) από το κάθε σύνολο δεδομένων που θα μας χρειαστούν για να εκπαιδεύσουμε το μοντέλο μας. Αυτό σημαίνει πως δεν έχουμε ελλείπουσες τιμές, πως έχουμε τις μονάδες μέτρησης που προτιμάμε, τους κατάλληλους τύπους δεδομένων και πως έχουμε αφαιρέσει τις στήλες που δεν θα χρειαστούμε για το μοντέλο πρόβλεψης, όπως τις στήλες road_info και road_name από το σύνολο δεδομένων road_data και τις στήλες snowfall, eastward_wind και northward_wind από το σύνολο δεδομένων weather_data.

Καθώς θέλουμε να συμπεριλάβουμε όλες τις στήλες που έχουμε κρατήσει μέχρι στιγμής προχωράμε στην συνένωση των δύο συνόλων με την εντολή pd.merge με την συνένωση outer join προκειμένου το νέο σύνολο να περιλαμβάνει όλα τα δεδομένα και από τα δύο σύνολα. Η συνένωση γίνεται στην στήλη της ημερομηνίας και ώρας datetime που είναι κοινή και στα δύο σύνολα και απλά μετονομάζουμε την στήλη «appprocesstime» του συνόλου road_data σε «datetime» όπως φαίνεται παρακάτω.



	deviceid	countedcars	datetime	road_name	average_speed	temperature	precipitation	average_wind_speed
0	MS116	6360.0	2021-06-04 02:00:00	Λ. ΚΗΦΙΣΟΥ	92.031447	15.8	0.000000	1.5
1	MS117	320.0	2021-06-04 02:00:00	Λ. ΚΗΦΙΣΟΥ	69.000000	15.8	0.000000	1.5
2	MS120	6320.0	2021-06-04 02:00:00	Λ. ΚΗΦΙΣΟΥ	81.537975	15.8	0.000000	1.5
3	MS121	200.0	2021-06-04 02:00:00	Λ. ΚΗΦΙΣΟΥ	37.000000	15.8	0.000000	1.5
4	MS124	4080.0	2021-06-04 02:00:00	Λ. ΚΗΦΙΣΟΥ	98.392157	15.8	0.000000	1.5
...	...	...	...	...	...	...	...	...
1681452	NaN	NaN	2021-12-31 19:00:00	NaN	NaN	9.3	0.000095	2.9
1681453	NaN	NaN	2021-12-31 20:00:00	NaN	NaN	8.7	0.000095	2.6
1681454	NaN	NaN	2021-12-31 21:00:00	NaN	NaN	8.1	0.000095	2.4
1681455	NaN	NaN	2021-12-31 22:00:00	NaN	NaN	7.5	0.000095	2.3

Εικόνα 44 Εντολή merge για συνένωση των δύο συνόλων δεδομένων

Στην παραπάνω εικόνα βλέπουμε πως στην ημερομηνία 31.12.2021 περιέχονται NaN στις στήλες που έχουμε κρατήσει από το σύνολο δεδομένων κυκλοφορίας. Ελέγχοντας ξανά το road_data σύνολο βλέπουμε πως η τελευταία ημερομηνία που έχουμε καταγεγραμμένες πληροφορίες είναι 30.12.2021 και αυτό λογικά οφείλεται σε κάποια δυσλειτουργία της πλατφόρμας καθώς είναι ακόμη στα πρώτα στάδια της ανάπτυξής της είτε σε κάποια δυσλειτουργία των αισθητήρων. Πρέπει να προχωρήσουμε στην αφαίρεση των εγγραφών με NaN τιμές στο σύνολο merged_data πριν προχωρήσουμε στην τροφοδότηση του μοντέλου μας με τα δεδομένα που έχουμε επιλέξει πως είναι σημαντικά για την σωστή πρόβλεψη. Με τις εντολές .isnull() και dropna() βλέπουμε πως έχουμε 1036 κενές τιμές και στις 3 στήλες των δεδομένων κυκλοφορίας και αφαιρούμε τις σειρές με κενές εγγραφές από το σύνολό μας αντίστοιχα. Αποθηκεύουμε εκ νέου το καθαρισμένο σύνολο με το όνομα clean_data.

Το μοντέλο XGBoost στην συνέχεια θα εκπαιδευτεί στα δεδομένα κυκλοφορίας και στα δεδομένα καιρού σε ωριαία βάση με σκοπό να προβλέψει την τιμή της μεταβλητής countedcars που λαμβάνει αριθμητικές τιμές. Για να γίνει αυτό προχωρήσαμε στην επιλογή ενός αισθητήρα (στήλη deviceid) ως μελέτη περίπτωσης όπου εφαρμόζουμε τον αλγόριθμο πρόβλεψης της κίνησης.

Για να μπορέσουμε να επιλέξουμε τον αισθητήρα που γίνεται η εφαρμογή του μοντέλου μας θα πρέπει να γίνει μια αναφορά στην εργασία «**Ανακαλύπτοντας την ποιότητα των Δυναμικών Ανοιχτών Κυβερνητικών Δεδομένων με την χρήση στατιστικών μεθόδων και μεθόδων μηχανικής μάθησης**» των Καραμάνου, Μπρίμος, Καλαμπόκης, Ταραμπάνης όπου εξετάζεται η αξιοπιστία της πλατφόρμας data.gov.gr . Όπως φαίνεται

στην προαναφερόμενη εργασία μέσα από διάφορες τεχνικές που εφαρμόστηκαν, τα δεδομένα που ήταν διαθέσιμα στην ιστοσελίδα για μεγάλο χρονικό διάστημα ήταν λανθασμένα με πολύ μεγάλες τιμές, ελλείπουσες τιμές και ανωμαλίες κυρίως λόγω δυσλειτουργιών των αισθητήρων και προβλημάτων στα δίκτυα.

Μέσα από την αξιολόγηση του συνόλου δεδομένων «road traffic in Attica» φαίνεται πως για το διάστημα που εξετάστηκε η πλατφόρμα υπήρχαν 20.16% ελλείπουσες τιμές και οι μισοί αισθητήρες παρουσίασαν ελλείπουσες τιμές σε 15% με 33.43% ποσοστό. Ο μέσος όρος των ανωμαλιών ανά αισθητήρα ήταν στο 71.1% με μόνο λίγους αισθητήρες να παρουσιάζουν μικρότερο ποσοστό ανωμαλιών από 10%. Η καλύτερη περίοδος για τον εντοπισμό των ανωμαλιών κρίθηκε πως είναι από 2.1.2022 μέχρι 23.6.2022 όπου μέσα από την Ανάλυση Συσχέτισης Ροής-Ταχύτητας (flow-speed correlation analysis) φαίνεται πως μόνο 9 αισθητήρες είχαν ποσοστό ανωμαλιών μικρότερο από 10% και για αυτό τον λόγο θεωρούνται οι «καλύτεροι» για το συγκεκριμένο χρονικό περιθώριο.

Table 3. Traffic sensors with less than 10% detected anomalies using the flow-speed correlation analysis.

Sensor ID	Hours of Anomaly	Percentage of Anomalies (%)
MS346	378	9.1
MS121	321	7.73
MS941	308	7.41
MS309	295	7.10
MS944	178	4.28
MS145	48	1.15
MS134	18	0.43
MS502	14	0.33
MS734	8	0.19

Εικόνα 45 Πίνακας με τους αισθητήρες που παρουσίασαν ποσοστό ανωμαλιών μικρότερο από 10%, Πηγή: Karamanou, A., Brimos, P., Kalampokis, E., & Tarabanis, K. (2022, December 10). Exploring the quality of dynamic open government data using statistical and machine le

Όπως φαίνεται στον παραπάνω πίνακα λοιπόν ο αισθητήρας με το μικρότερο ποσοστό ανωμαλιών (0.19%) είναι ο MS734 και για αυτό τον λόγο είναι ο αισθητήρας που επιλέξαμε να χρησιμοποιήσουμε στο δικό μας μοντέλο πρόβλεψης.

Αφού έχουμε πλέον το σύνολο δεδομένων μας στην κατάλληλη μορφή και έχοντας καταλήξει στην επιλογή του αισθητήρα MS734 πρέπει να αφαιρέσουμε τους υπόλοιπους αισθητήρες από την στήλη deviceid πριν το χωρίσουμε σε ένα κομμάτι εκπαίδευσης (training set) και ένα κομμάτι δοκιμής/ελέγχου (test set). Οπότε δημιουργούμε ένα νέο σύνολο δεδομένων prepred_data και αφαιρούμε όλους τους αισθητήρες εκτός από τον MS734 με την παρακάτω εντολή.

#drop all rows in column deviceid except from where the deviceid column is equal to ' MS734'

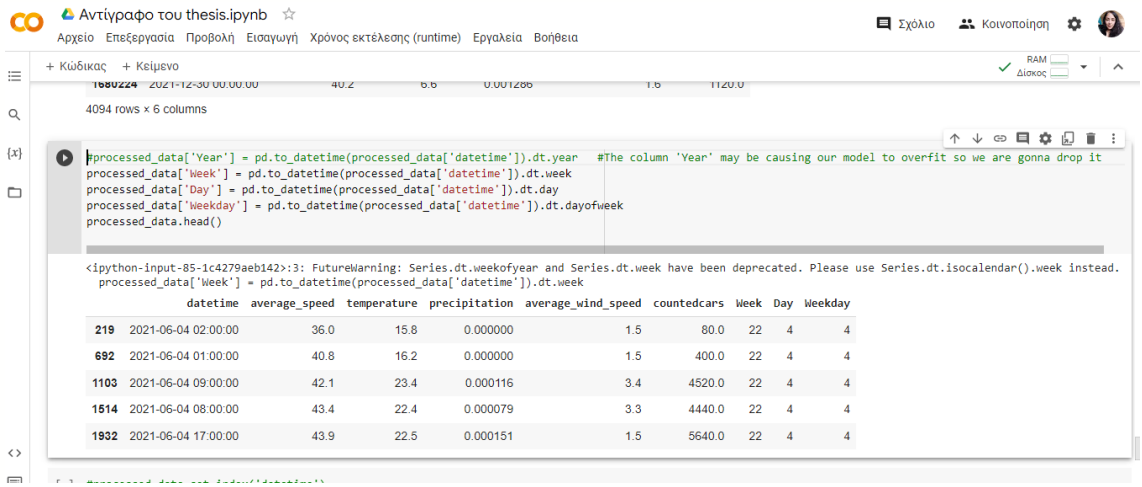
prepred_data = clean_data.query("deviceid == 'MS734'")

Μετά με την εντολή prepred_data.dtypes βλέπουμε πως οι 2 μεταβλητές (στήλες) από τις 7 που έχουμε επιλέξει για να τροφοδοτήσουμε το μοντέλο μας δεν έχουν αριθμητικό τύπο δεδομένων οπότε πρέπει να το διαχειριστούμε πριν προχωρήσουμε. Η στήλη deviceid έχει τύπο object οπότε με την εντολή από pandas pd.get_dummies() μετατρέπουμε την κατηγορική μεταβλητή σε εικονική μεταβλητή για να τρέξει ο αλγόριθμος και αποθηκεύουμε εκ νέου το σύνολο δεδομένων μας με το όνομα prefinal_data όπως φαίνεται παρακάτω.

prefinal_data = pd.get_dummies(prepred_data,prepred_data.columns[prepred_data.dtypes == 'object']).

Ακολούθως χρησιμοποιούμε την εντολή processed_data=prefinal_data.iloc[:, [0,2,3,4,5,6,1] για να μεταφέρουμε την target variable (countedcars) μας στην τελευταία στήλη και να αποθηκεύσουμε εκ νέου το σύνολο δεδομένων μας με το όνομα processed_data.

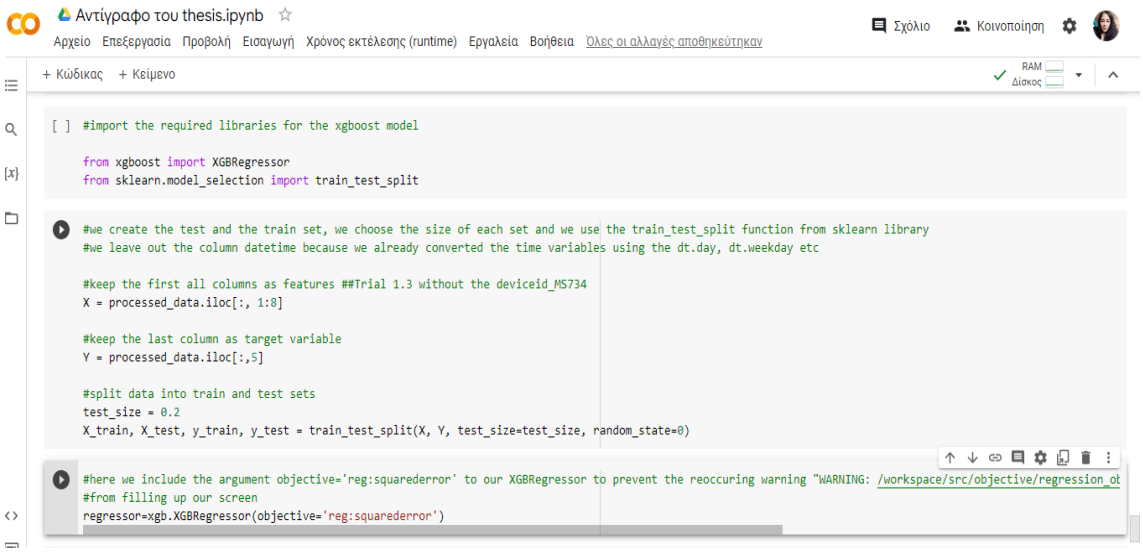
Έπειτα για να μπορέσουμε να συμπεριλάβουμε και τις χρονικές μεταβλητές που περιέχονται στην στήλη datetime χρησιμοποιούμε την συνάρτηση pd.to_datetime και τις μεθόδους dt.year, dt.day, dt.week, dt.dayofweek όπως φαίνεται στην παρακάτω εικόνα.



Εικόνα 46 Αρχείο κώδικα με μετατροπή της στήλης datetime σε κατάλληλη μορφή για τροφοδότηση του μοντέλου XGBoost.

Συνεχίζουμε κάνοντας import τον XGBoostRegressor από την βιβλιοθήκη xgboost και την συνάρτηση train_test_split από την βιβλιοθήκη sklearn προκειμένου να διαχωρίσουμε το σύνολο δεδομένων μας σε train set, όπου θα εκπαιδευτεί το μοντέλο μας, και σε test set, όπου θα ελεγχθεί το κατά πόσο το μοντέλο μας μπορεί να κάνει προβλέψεις σε δεδομένα που δεν έχει εκπαιδευτεί.

Ο λόγος που γίνεται αυτό είναι για να δούμε αν το μοντέλο μας μπορεί να εφαρμοσθεί και σε άλλα σύνολα δεδομένων για πραγματικές μελλοντικές προβλέψεις. Ένας συνηθισμένος διαχωρισμός είναι της τάξης 70%-80% των δεδομένων του συνόλου μας να χρησιμοποιείται για την φάση εκπαίδευσης του μοντέλου και το υπόλοιπο 20%-30% των δεδομένων του συνόλου μας να «αποκρύπτεται» από το μοντέλο κατά την φάση της εκπαίδευσης και να χρησιμοποιείται μόνο κατά την φάση του ελέγχου.



Εικόνα 47 Αρχείο κώδικα με την εντολή train_test_split με test_size=0.2 και random_state=0.

Ο XGBoost είναι ο αλγόριθμος που έχουμε επιλέξει για την δική μας μελέτη περίπτωσης οπότε πριν τον «τρέξουμε» πρέπει να ορίσουμε τις παραμέτρους του(hyperparameters). Σύμφωνα με τις επίσημες οδηγίες (documentation) πρέπει λοιπόν να ορίσουμε:

- Τις γενικές παραμέτρους (General Parameters)
- Τις παραμέτρους ενίσχυσης (Boosting Parameters)
- Τις παραμέτρους μάθησης εργασίας (Learning Task Parameters)

Οι γενικές παράμετροι σχετίζονται με τον ενισχυτή (booster) που χρησιμοποιούμε και η default επιλογή είναι η gbtree, οι παράμετροι ενίσχυσης εξαρτώνται από τον ενισχυτή

που επιλέξαμε, στην δική μας περίπτωση είναι Tree Booster και οι Learning Task Parameters έχουν να κάνουν με το σενάριο μάθησης, δηλαδή πχ μια εργασία παλινδρόμησης πιθανό να χρησιμοποιεί διαφορετικές παραμέτρους από μια εργασία ταξινόμησης.

Οι hyperparameters που μας ενδιαφέρουν σχετίζονται με τον Tree Booster και ο στόχος (objective) μας είναι το «reg:squarederror», δηλαδή παλινδρόμηση με απώλεια τετραγώνου. Με βάση τα παραπάνω επιλέγουμε στην επόμενη υποενότητα και τα eval_metric, δηλαδή τις μετρήσεις αξιολόγησης που θα χρησιμοποιήσουμε για να αξιολογήσουμε την απόδοση του μοντέλου.

Ακολουθεί μια επιγραμματική αναφορά σχετικά με τις παραμέτρους που θεωρούμε πιο σημαντικές και επιλέξαμε για τον αλγόριθμο μαζί με μια σύντομη επεξήγηση αυτών:

- **Eta** [**default=0.3, alias: learning_rate**]
  Είναι η συρρίκνωση του μεγέθους του βήματος που χρησιμοποιείται κατά την ενημέρωση για να αποφευχθεί η υπερβολική προσαρμογή (overfitting). Μετά από κάθε βήμα ενίσχυσης μπορούμε να πάρουμε άμεσα τα βάρη των νέων features και η παράμετρος eta (ή learning rate) τα συρρικνώνει για να κάνει την διαδικασία ενίσχυσης πιο συντηρητική.
- **N_estimators [default=100, range: 0-,∞]**
  Προσδιορίζει τον αριθμό των Δέντρων Απόφασης που θα ενισχυθούν. Εάν η παράμετρος είναι ίση με 1 σημαίνει πως κατασκευάζεται μόνο ένα δέντρο επομένως δεν γίνεται κάποια ενίσχυση. Η default τιμή ισούται με 100 αλλά μπορούμε να πειραματιστούμε για να βρούμε την βέλτιστη απόδοση.
- **Gamma [default=0, alias: min_split_loss, range: 0-,∞]**
  Είναι η ελάχιστη μείωση απώλειας που απαιτείται για να δημιουργηθεί μια περεταίρω κατάτμηση σε έναν κόμβο φύλλου του δέντρου. Όσο μεγαλύτερη είναι παράμετρος gamma τόσο πιο συντηρητικός θα είναι ο αλγόριθμος.
- **Max_depth [default=6, range: 0-,∞]**
  Είναι το μέγιστο βάθος δέντρου. Η αύξηση αυτής της τιμής θα κάνει το μοντέλο πιο περίπλοκο και πιο πιθανή της υπερβολική προσαρμογή. Το 0 υποδεικνύει πως δεν υπάρχει όριο στο βάθος. Απαιτείται προσοχή καθώς ο αλγόριθμος XGBoost καταναλώνει μεγάλο κομμάτι μνήμης όταν εκπαιδεύει ένα βαθύ δέντρο. Η ακριβής μέθοδος Δέντρου απαιτεί τιμή διάφορη του μηδέν.
- **Min_child_weight [default=1, range: 0-,∞]**

Είναι το ελάχιστο άθροισμα του βάρους περίπτωσης που απαιτείτε σε ένα παιδί. Εάν το βήμα κατάτμησης του δέντρου έχει ως αποτέλεσμα έναν κόμβο φύλλου με το άθροισμα του βάρους περίπτωσης μικρότερο από το min_child_weight τότε η διαδικασία δημιουργίας θα σταματήσει την περεταίρω κατάτμηση. Στην γραμμική παλινδρόμηση αυτό αντιστοιχεί στον ελάχιστο αριθμό περιπτώσεων που χρειάζονται σε κάθε βήμα. Όσο μεγαλύτερο είναι το min_child_weight τόσο πιο συντηρητικός θα είναι ο αλγόριθμος.

- **Subsample [default=1, range: 0-1]**
  Είναι η subsample αναλογία των training περιπτώσεων. Αν το ορίσουμε ίσο με 0.5 ο αλγόριθμος XGBoost θα έπαιρνε τυχαία δείγματα από τα μισά δεδομένα εκπαίδευσης πριν αναπτυχθούν τα δέντρα και αυτό μπορεί να αποτρέψει την υπερβολική προσαρμογή. Το subsampling πραγματοποιείται μία φορά σε κάθε επανάληψη ενίσχυσης.
- **Colsample_by_tree [default=1, range: 0-1]**
  Μαζί με τις παραμέτρους colsample_by_level και colsample_by_node παραμέτρους ανήκει στην οικογένεια παραμέτρων για να κάνουμε subsample στις στήλες. Είναι η subsample αναλογία των στηλών όταν κατασκευάζεται το κάθε δέντρο και πραγματοποιείται για κάθε δέντρο που κατασκευάζεται.

Οι παραπάνω παράμετροι του **Tree Booster** (ή αλλιώς υπερπαράμερτοι) διαδραματίζουν σημαντικό ρόλο στην απόδοση του μοντέλου μας και χρειάζεται προσοχή κατά τον ορισμό των τιμών τους καθώς είναι συχνό το πρόβλημα του overfitting ή του underfitting όπως θα δούμε και παρακάτω. Μια μικρή μεταβολή της τιμής μπορεί να βελτιώσει ή να μειώσει αισθητά την απόδοση του μοντέλου.

Προκειμένου να βρούμε τους καλύτερους συνδυασμούς των τιμών των υπερπαραμέτρων του αλγορίθμου χρησιμοποιήσαμε την τεχνική **GridSearchCV** της βιβλιοθήκης sklearn που βρίσκει τις βέλτιστες τιμές των παραμέτρων από ένα σετ τιμών που ορίζουμε εμείς ως πλέγμα (grid) και, όπως φαίνεται και από το όνομα της, είναι ουσιαστικά μια τεχνική διασταυρούμενης επικύρωσης (**cross validation**). Αυτή η διαδικασία ονομάζεται **Hyperparameter Tuning** και το αποτέλεσμά της είναι ένα σύνολο τιμών που χρησιμοποιείται για την παραμετροποίηση του μοντέλου μηχανικής μάθησης.

Η GridSearchCV είναι μια τεχνική αναζήτησης πλέγματος που παράγει εκτενώς υποψήφιους συνδυασμούς από ένα πλέγμα τιμών των παραμέτρων που καθορίζονται

μέσα στην παράμετρο που ονομάζεται param_grid. Όταν εκτελούμε την εντολή fit σε ένα σύνολο δεδομένων αξιολογούνται όλοι οι πιθανοί συνδυασμοί των τιμών των παραμέτρων και κρατείται ο καλύτερος συνδυασμός στο τέλος της διαδικασίας. Όταν λοιπόν εκπαιδεύσουμε το μοντέλο στο training set και αφότου το προσαρμόσουμε προχωράμε στην εφαρμογή της πρόβλεψης με την εντολή predict χρησιμοποιώντας το test set με το εκπαιδευμένο μοντέλο στο training set.

Συνοψίζοντας, η λογική που ακολουθούμε εδώ είναι η εισαγωγή των απαραίτητων βιβλιοθηκών (όπως η sklearn και η xgboost) προκειμένου να χρησιμοποιήσουμε τις διαθέσιμες συναρτήσεις/εντολές, επιλέγουμε εκ νέου το διαμορφωμένο σύνολο δεδομένων στην πλέον κατάλληλη μορφή, διαχωρίζουμε τα δεδομένα σε σετ εκπαίδευσης (training set) και σετ ελέγχου (test set), εκπαιδεύουμε το μοντέλο xgboost και τέλος πραγματοποιούμε προβλέψεις στο σετ ελέγχου.

Η διαδικασία της εκπαίδευσης και των προβλέψεων επαναλαμβάνεται όσες φορές επιθυμούμε μέχρι να φτάσουμε στο επιθυμητό αποτέλεσμα ή κοντά σε αυτό. Μπορούμε να αφαιρέσουμε 1 ή παραπάνω στήλες (μεταβλητές) από αυτές που επιλέξαμε στην αρχική εφαρμογή αν κρίνουμε πως κάποια δεν προσφέρει αξία στο μοντέλο ή αν υποψιαζόμαστε πως ίσως να προκαλεί υπερπροσαρμογή (overfitting), μπορούμε να αλλάξουμε το μέγεθος του test size/train size, ή το random state, ή γενικά να πειραματιστούμε στις τιμές των υπερπαραμέτρων όπως το n_estimators, learning_rate ή όποια θεωρούμε σημαντική.

5.5 Αξιολόγηση μοντέλου XGBoost

Σε αυτό το σημείο προχωράμε στην αξιολόγηση του μοντέλου που έχουμε επιλέξει προκειμένου να ελέγξουμε την απόδοσή του μετά την πρόβλεψη. Καθώς πρόκειται για πρόβλημα Παλινδρόμησης και όχι Ταξινόμησης δεν μπορούμε να χρησιμοποιήσουμε την μέθοδο αξιολόγησης της ακρίβειας των προβλέψεων του μοντέλου. Αντιθέτως, πρέπει να χρησιμοποιήσουμε μετρήσεις σφάλματος (error metrics) που έχουν σχεδιαστεί συγκεκριμένα για την αξιολόγηση προβλέψεων Προβλημάτων Παλινδρόμησης.

Οι μέθοδοι μέτρησης για μοντέλα Παλινδρόμησης περιλαμβάνουν τον υπολογισμό μιας βαθμολογίας σφάλματος (error score) που συνοψίζει την προβλεπτική ικανότητα του μοντέλου. Η απόδοση ενός μοντέλου Παλινδρόμησης πρέπει να αναφέρεται ως σφάλμα σε αυτές τις προβλέψεις καθώς δεν μας ενδιαφέρει η ακριβής πρόβλεψη μιας τιμής αλλά

κατά πόσο οι προβλέψεις των τιμών ήταν κοντά στις αναμενόμενες τιμές με βάση τα παρελθοντικά δεδομένα. Στον παρακάτω πίνακα φαίνονται πολλά από τα διαθέσιμα εργαλεία αξιολόγησης για προβλήματα Παλινδρόμησης.

Regression metrics

See the Regression metrics section of the user guide for further details.

metrics.explained_variance_score(y_true, ...)	Explained variance regression score function.
metrics.max_error(y_true, y_pred)	The max_error metric calculates the maximum residual error.
metrics.mean_absolute_error(y_true, y_pred, *)	Mean absolute error regression loss.
metrics.mean_squared_error(y_true, y_pred, *)	Mean squared error regression loss.
metrics.mean_squared_log_error(y_true, y_pred, *)	Mean squared logarithmic error regression loss.
metrics.median_absolute_error(y_true, y_pred, *)	Median absolute error regression loss.
metrics.mean_absolute_percentage_error(...)	Mean absolute percentage error (MAPE) regression loss.
metrics.r2_score(y_true, y_pred, *[, ...])	R^2 (coefficient of determination) regression score function.
metrics.mean_poisson_deviance(y_true, y_pred, *)	Mean Poisson deviance regression loss.
metrics.mean_gamma_deviance(y_true, y_pred, *)	Mean Gamma deviance regression loss.
metrics.mean_tweedie_deviance(y_true, y_pred, *)	Mean Tweedie deviance regression loss.
metrics.d2_tweedie_score(y_true, y_pred, *)	D^2 regression score function, fraction of Tweedie deviance explained.
metrics.mean_pinball_loss(y_true, y_pred, *)	Pinball loss for quantile regression.
metrics.d2_pinball_score(y_true, y_pred, *)	D^2 regression score function, fraction of pinball loss explained.
metrics.d2_absolute_error_score(y_true, ...)	D^2 regression score function, fraction of absolute error explained.

Εικόνα 48 Πίνακας με Regression Metrics Πηγή: API reference. scikit. (n.d.). Retrieved March 6, 2023, from https://scikit-learn.org/stable/modules/classes.html#regression-metrics

Στην δική μας περίπτωση, καθώς ο κώδικας του μοντέλου πρόβλεψης μας έχει σαν βάση την εργασία «Forecasting ED Visits» όπως έχουμε αναφέρει εκτενώς προηγουμένως, επιλέγουμε τις παρακάτω μεθόδους μέτρησης αξιολόγησης (evaluation metrics), δύο εκ των οποίων είναι και οι πιο συνηθισμένες. Αυτές είναι:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Percentage Error (MAPE)

Το **MAE** είναι το μέσο απόλυτο σφάλμα (είναι επίσης γνωστό σαν απώλεια L1) και είναι μία από τις πιο απλές συναρτήσεις απώλειας. Υπολογίζεται παίρνοντας την απόλυτη διαφορά ανάμεσα στις τιμές πρόβλεψης και στις πραγματικές τιμές και βγάζοντας τον μέσο όρο για όλο το σύνολο δεδομένων. Μας δίνει ουσιαστικά το μέγεθος του αριθμητικού μέσου όρου σφαλμάτων**. Όσο πιο μικρό είναι το MAE** τόσο **καλύτερη είναι η απόδοση του μοντέλου** μας. Μαθηματικά μπορεί να εκφραστεί με τον παρακάτω τρόπο όπου where y_i = πραγματική τιμή, y_hat_i = τιμή πρόβλεψης, n = μέγεθος δείγματος.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i |$$

Εικόνα 49 Μαθηματική έκφραση MAE, Πηγή: M, P. (2022) End-to-end introduction to evaluating Regression Models, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/ (Accessed: March 17, 2023).

Το **RMSE** είναι η μέση τετραγωνική ρίζα σφάλματος και υπολογίζεται παίρνοντας την τετραγωνική ρίζα του MSE (είναι επίσης γνωστό σαν Root Mean Square Deviation). Μετράει το μέσο μέγεθος των σφαλμάτων και ασχολείται με την απόκλιση από την πραγματική τιμή. Μια τιμή ίση με μηδέν σημαίνει πως το μοντέλο μας λειτουργεί τέλεια**.** **Όσο χαμηλότερο είναι το RMSE τόσο καλύτερες είναι οι προβλέψεις του μοντέλου.** Ένα μεγαλύτερο RMSE φανερώνει πως υπάρχει μεγάλη απόκλιση. Μας βοηθάει να δούμε αν ένα χαρακτηριστικό (feature) βελτιώνει ή όχι το μοντέλο. Μαθηματικά μπορεί να εκφραστεί με τον παρακάτω τρόπο όπου where y_i = πραγματική τιμή, y_hat_i = τιμή πρόβλεψης, n = μέγεθος δείγματος.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Εικόνα 50 Μαθηματική έκφραση MAE, Πηγή: M, P. (2022) End-to-end introduction to evaluating Regression Models, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/ (Accessed: March 17, 2023).

Το MAPE είναι το μέσο απόλυτο ποσοστιαίο σφάλμα και υπολογίζεται παίρνοντας της διαφορά ανάμεσα στην πραγματική τιμή και στην τιμή πρόβλεψης και διαιρώντας την με την πραγματική τιμή. Ένα απόλυτο ποσοστό εφαρμόζεται σε αυτή την τιμή και βγαίνει ο μέσος όρος σε όλο το σύνολο δεδομένων. Είναι επίσης γνωστό και σαν Mean Absolute Percentage Deviation (MAPD). Αυξάνεται γραμμικά με την αύξηση στο σφάλμα. Όσο μικρότερο είναι το MAPE τόσο καλύτερη είναι η απόδοση του μοντέλου. Μαθηματικά

μπορεί να εκφραστεί με τον παρακάτω τρόπο όπου where y_i = πραγματική τιμή, y_hat_i = τιμή πρόβλεψης, n = μέγεθος δείγματος.

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i} . 100\%$$

Εικόνα 51 Μαθηματική έκφραση MAE, Πηγή: M, P. (2022) End-to-end introduction to evaluating Regression Models, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/ (Accessed: March 17, 2023).

Γενικά, οι συναρτήσεις απώλειας παίρνουν τις τιμές πρόβλεψης του μοντέλου και τις συγκρίνουν με τις πραγματικές τιμές. Έτσι υπολογίζουμε πόσο καλά ή όχι αποδίδει το μοντέλο όσον αφορά την ικανότητα του να χαρτογραφεί σχέσεις ανάμεσα σε ένα χαρακτηριστικό (ή μια ανεξάρτητη μεταβλητή) και στην στοχευόμενη μεταβλητή (ή αλλιώς εξαρτημένη μεταβλητή). Μέσω λοιπόν αυτών των μετρήσεων αξιολόγησης του μοντέλου μπορούμε να υπολογίσουμε πόσο απέχει το μοντέλο από τις πραγματικές τιμές του συνόλου δεδομένων. Με αυτό τον τρόπο μπορούμε να βρούμε την απόκλιση της τιμής πρόβλεψης από την πραγματική τιμή και να εκπαιδεύσουμε αναλόγως το μοντέλο. Η απώλεια (loss) είναι η διαφορά ανάμεσα στην πραγματική τιμή και στην τιμή πρόβλεψης. Όσο μεγαλύτερη είναι η απώλεια / σφάλμα τόσο χαμηλότερη είναι η απόδοση του μοντέλου.

Πριν προχωρήσουμε στην παράθεση των αποτελεσμάτων του μοντέλου μας και στην ερμηνεία των ευρημάτων μας είναι χρήσιμο να αναφερθεί πως ένα συχνό πρόβλημα στην Μηχανική Μάθηση είναι το λεγόμενο Overfitting του μοντέλου όπου ο αλγόριθμος αποδίδει πολύ καλά στα δεδομένα που εκπαιδεύεται (train set) αλλά έχει μειωμένη απόδοση σε δεδομένα που δεν έχει συναντήσει ξανά (test set). Αυτό μπορεί να οφείλεται σε διάφορους παράγοντες όπως η ποιότητα των δεδομένων όταν περιέχονται πολλά outliers ή ελλείπουσες τιμές, από λανθασμένο χειρισμό, επειδή δίνεται μεγάλη βαρύτητα σε κάποια μεταβλητή ή εξαιτίας της τιμής κάποιας παραμέτρου. Άλλες φορές μπορεί να συναντήσουμε το πρόβλημα του underfitting όπου ο αλγόριθμος αδυνατεί να συλλάβει

τις σχέσεις ανάμεσα στις μεταβλητές εισόδου και εξόδου επακριβώς με αποτέλεσμα να έχουν μεγάλα σφάλματα τόσο στο σετ εκπαίδευσης όσο και στο σετ ελέγχου.

6. Συμπεράσματα και προτάσεις

Σε αυτό το σημείο βλέπουμε αναλυτικά τα αποτελέσματα της εφαρμογής του μοντέλου XGBoost από την πρώτη εφαρμογή του μέχρι τα τελικά αποτελέσματα και τις αλλαγές των τιμών στις παραμέτρους του αλγορίθμου και έπειτα προχωράμε στα συμπεράσματα που προκύπτουν από την εφαρμογή και αξιολόγηση του μοντέλου μας, καθώς και τους περιορισμούς της παρούσας εργασίας μαζί με προτάσεις για μελλοντική έρευνα.

6.1 Ερμηνεία ευρημάτων και Συμπεράσματα

Την πρώτη φορά που έτρεξε το μοντέλο συμπεριλάβαμε και την στήλη deviceid_MS734, καθώς και την στήλη Year. Πήραμε από το GridSearchCV τις παρακάτω τιμές για τις παραμέτρους, έγινε η προσαρμογή στα δεδομένα και στην αξιολόγηση των σφαλμάτων πήραμε τα αποτελέσματα του παρακάτω πίνακα. Να σημειωθεί πως συμπεριλαμβανόταν και η τιμή gamma=0 και subsample=1. Ακολουθεί παράθεση του κώδικα την πρώτη φορά που έτρεξε το μοντέλο και μετά ένας αναλυτικός πίνακας με τα αποτελέσματα της αξιολόγησης του μοντέλου μας.

Εικόνα 52 Αρχείο κώδικα με GridSearchCV, XGBRegressor() και fit

Εξετάζοντας τις παρακάτω τιμές είναι ευδιάκριτο πως το μοντέλο μας συνάντησε το πρόβλημα της υπερπροσαρμογής καθώς είχε πολύ καλή απόδοση στα δεδομένα που εκπαιδεύτηκε αλλά καθόλου καλή απόδοση στα δεδομένα του test set.

Model Parameters & Evaluation Metrics	Values
Best Params: colsample_bytree	1
Best Params: learning_rate	0.015
Best Params: max_depth	5
Best Params: min_child_weight	1
Best Params: n_estimators	998
Test size	0.2
Random State	2
Fit.score(X_test, y_test)	-3.024
mae_train	0.00213
mae_test	3.01831
rmse_train	0.09
rmse_test	57.096
mape_train	2.297
mape_test	0.0003

Εικόνα 53 Πίνακας με τις παραμέτρους και τις αξιολογήσεις του μοντέλου στην πρώτη εφαρμογή

Όπως φαίνεται στον παραπάνω πίνακα το μοντέλο αποδίδει εξαιρετικά καλά στο train test και εξαιρετικά χαμηλά στο test set, για παράδειγμα η μέση τετραγωνική ρίζα σφάλματος είναι (rmse_train) είναι ίση με 0.09 και η αντίστοιχη τιμή της rmse_test είναι 57.09. Επομένως επαναλάβαμε την διαδικασία ξανά και αυτή την φορά αφαιρέθηκαν οι στήλες deviceid_MS734 και Year γιατί ήταν πιθανό να προκαλούσαν αυτές το Overfitting αλλά δεν φάνηκε να ισχύει κάτι τέτοιο μιας και τα αποτελέσματα μεταβλήθηκαν ελάχιστα και μόνο σε δεκαδικό επίπεδο. Αφαιρέσαμε έπειτα επίσης τις παραμέτρους gamma και subsample καθώς μετά από πειραματισμό ξεκινώντας από ελάχιστες μέχρι πολύ μεγάλες τιμές φάνηκε πως δεν ασκούν ιδιαίτερη επιρροή στο μοντέλο μας.

Συνεχίζοντας, αλλάξαμε το μέγεθος του test size από 0.15 μέχρι 0.3 (που είναι το προτεινόμενο μέγεθος) και το random state σε διάφορες τιμές αλλά καθώς το μοντέλο μας φάνηκε να αποδίδει καλύτερα με τις τιμές 0 και 2 δοκιμάστηκαν διάφοροι συνδυασμοί αυτών των τιμών και άλλων τιμών των παραμέτρων μας.

Για οικονομία χώρου και χρόνου δημιουργήθηκε ένας ενδεικτικός πίνακας αποτελεσμάτων με διάφορες τιμές και συνδυασμούς τιμών που χρησιμοποιήθηκαν για να βρούμε τις τιμές που κάνουν πιο αποδοτικό το μοντέλο.

Test size	Random State	n-estimators	learning rate	max depth	min child weight	colsample bytree	Fit Score	mae train	mae test	rmse train	rmse test	mape train	mape test	Extra:
0.2	2	998	0.015	5	1	1	-3.024	0.00213	3.01831	0.09	57.096	2.297	0.0003	First Run including deviceid_MS734 and Year, plus gamma=0, subsample=1
0.2	2	998	0.015	5	1	1	-3.05	0.0018	3.043	0.089	57.64	1.811	0.0003	Second Run without the columns deviceid_MS734 and Year
0.2	2	998	0.01	5	5	1	0.99	0.57	2.5	14.94	54.41	0.0001	0.0003	
0.2	2	998	0.01	5	5	0.9	0.99	15.6	26.4	26.7	74.08	47304.....	802839....	
0.2	2	998	0.01	5	6	1	0.99	0.62	2.69	16	56	0.0001	0.0003	
0.2	2	1500	0.01	5	5	1	0.99	0.49	2.67	11.74	55.52	0.0001	0.0003	
0.2	2	1500	0.01	2	5	1	0.99	1.19	3.46	12.82	55.95	165017....	82403....	
0.2	2	998	0.01	5	5	1	0.99	0.92	3.21	22.4	60.5	31628....	10997..	
0.2	2	200	0.1	5	5	1	0.99	2.4	5.2	20.7	57.6	33987...	16496....	
0.3	2	998	0.01	5	5	1	-1.976	0.66	1.88	15.89	44.70	0.0001	0.0003	
0.15	2	998	0.01	5	5	1	0.99	4.56	7	18	63	504800....	60366....	
0.2	0	998	0.015	5	5	1	0.99	0..47	3.12	11.504	56.223	0.0001	0.0003	
0.2	0	998	0.015	5	6	1	0.99	0.61	3.36	12.73	58.47	27.502.898.487.758.700	4.139.520.684.713.310	
0.2	0	200	0.015	5	6	1	0.99	107.92	109.46	141.86	161.88	544557390057623.3	725854884997442.6	
0.2	0	-	0.015	5	6	1	0.86	482.85	481.84	619.81	636.44	1270633910134454.5	1693661398327366.0	
0.2	0	50	0.015	5	5	1	0.40	1021.82	1016.45	1304.72	1314.36	1225254127629652.8	1633173491244246.0	
0.2	0	1000	0.09	5	5	1	0.99	0.20	3.68	0.81	58.04	27502898487758700	0.0005	
0.2	0	900	0.015	2	1	1	0.99	0.17	2.22	0.62	33.59	90759565009603.89	120975814166240.44	
0.2	0	998	0.015	5	15	1	0.99	4.03	9.65	43.7	105	48.405	8.908	
0.2	0	1000	0.01	5	5	1	0.99	0.54	3.08	14.54	55.58	0.0001	0.0003	
0.2	0	1000	0.01	4	5	1	0.99	0.57	3.08	14.66	55.46	90759565009603.89	60487907083120.22	
0.2	0	997	0.015	10	0.2	1	0.99	0.0632	1.238	1.47	26.19	7.90	0.0001	gamma=1, subsample=0.5
0.3	0	998	0.01	5	5	1	0.99	1.22	3.058	24.72	53.21	33010677896956.516	40308865664015.83	gamma=1, subsample=0.5
0.3	2	998	0.01	5	5	1	0.99	1.0677	2.167	24.38	49.42	22007118597971.01	7328884666184.696	gamma=1, subsample=0.5
0.2	0	1500	0.01	4	1	1	-2.65	0.0021	2.64	0.09	51.82	2.13	0.0002	
0.3	2	998	0.01	4	5	1	0.99	0.66	1.88	15.89	44.70	0.0001	0.0002	booster='gbtree'
0.2	2	998	0.01	2	3	0.77	0.99	43.17	46.53	62.06	79.6	1224538956272815.5	776861774615577.9	booster='gbtree'
0.3	2	998	0.01	2	3	0.77	0.99	43.17	46.53	62.06	79.6	1224538956272815.5	776861774615577.9	booster='gbtree'
0.3	2	998	0.01	5	5	0.77	0.99	31.13	52.7	43.29	93.58	1512203435089150.8	1769925646883604.2	booster='gbtree'
0.3	2	1500	0.01	5	5	1	0.99	0.606	2.039	12.37	45.75	0.0001	0.0002	booster='gbtree'

Εικόνα 54 Πίνακας με τιμές παραμέτρων και τιμές σφαλμάτων από την εφαρμογή του μοντέλου XGBRegressor()

Από τον παραπάνω πίνακα μπορούμε να δούμε πως αλλάζουν οι τιμές των μετρούμενων σφαλμάτων ανάλογα με τις τιμές ή τους συνδυασμούς των τιμών των παραμέτρων του αλγορίθμου, τις τιμές του test size και του random state. Οι περισσότερες από τις τιμές αυτές δοκιμάστηκαν με απευθείας εφαρμογή του παρακάτω κώδικα (αλλάζοντας τις τιμές)

```
regressor=xgb.XGBRegressor(n_estimators=1500, learning_rate=0.01, max_depth=5,
min_child_weight=5, colsample_bytree=1, booster='gbtree')

fit = regressor.fit(X_train, y_train)
```

και όχι με την τεχνική GridSearchCV καθώς είναι αρκετά χρονοβόρα και για να βρει πχ τον καλύτερο συνδυασμό για 5 παραμέτρους με 3 πιθανές τιμές η καθεμία η διάρκεια της αναζήτησης ήταν περίπου 3 ώρες την κάθε φορά. Σε αυτό φυσικά παίζει ρόλο η επεξεργαστική ισχύς τόσο του υπολογιστή που χρησιμοποιήθηκε για την εργασία όσο και του online εργαλείου Google Colab.

Προκειμένου να βελτιστοποιήσουμε το μοντέλο χρησιμοποιήθηκε το επίσημο site για τον xgboost, το documentation των παραμέτρων του αλγορίθμου, τεχνικές αξιολόγησης αυτού, καθώς και συμβουλές για να αποφευχθεί το overfitting σε συνδυασμό φυσικά και με άλλες πηγές όπως άρθρα από το Towards Data Science, Medium, simplilearn για τα οποία υπάρχουν αναφορές στη βιβλιογραφία στο τέλος της παρούσας εργασίας.

Μετά λοιπόν από πολλές δοκιμές και επαναλήψεις καταλήξαμε πως οι πιο κατάλληλες τιμές για τις παραμέτρους του μοντέλου μας με βάση τα αποτελέσματα σφαλμάτων είναι test size=0.3, random state=2, n_estimators=998, learning rate=0.01, max depth=5, min child weight=5, colsample bytree=1.

Το fit score που πήραμε είναι ίσο με -1.97 με βάση το scoring= 'neg_mean_absolute_error' που περάσαμε στην συνάρτηση search την τελευταία φορά που τρέξαμε το μοντέλο, το οποίο είναι το μέσο απόλυτο σφάλμα. Ο λόγος που επιλέξαμε αυτές τις τιμές των παραμέτρων για το τελικό μοντέλο είναι επειδή η διαφορά ανάμεσα στη τιμή του κάθε τύπου σφάλματος στο training set και στο test set είναι μικρότερη σε σχέση με τις άλλες δοκιμές.

Test size	Random State	n-estimators	learning rate	max depth	min child weight	colsample bytree	Fit Score	mae train	mae test	rmse train	rmse test	mape train	mape test
0.3	2	998	0.01	5	5	1	-1.976	0.66	1.88	15.89	44.70	0.0001	0.0003

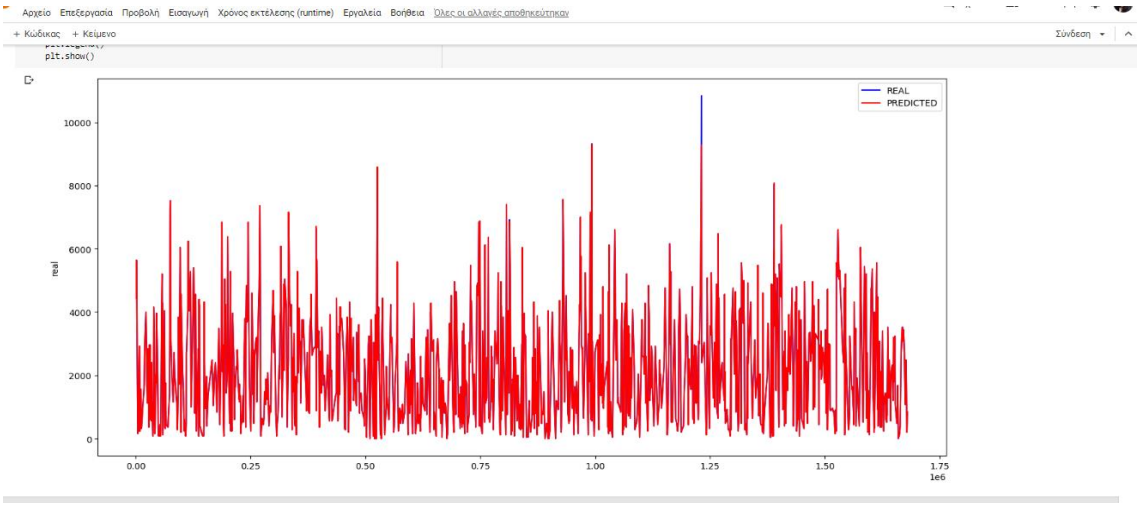
Εικόνα 55 Πίνακας τελικών τιμών παραμέτρων μοντέλου XGBRegressor() και μεγέθη σφαλμάτων

Όπως φαίνεται στον πάνω πίνακα με τις τελικές τιμές η τιμή του mae train=0.66 και η τιμή του mae test=1.88, διαφορά πολύ μικρότερη από την πρώτη φορά που έτρεξε ο αλγόριθμος όπου η τιμή του mae train=0.002 και η τιμή του mae test=3.01.

Αντίστοιχα, η τιμή του rmse train=15.89 και η τιμή του rmse test=44.7, δηλαδή η μέση τετραγωνική ρίζα σφάλματος στο σετ εκπαίδευσης είναι μικρότερη από αυτήν του σετ ελέγχου αλλά σε λογική βαθμό. Η τιμή αυτή είναι επίσης βελτιστοποιημένη σε σχέση με την πρώτη φορά που τρέξαμε τον αλγόριθμο όπου η τιμή του rsme train=0.09 και του rsme test=57.09,όπου η διαφορά στο σετ εκπαίδευσης και στο σετ ελέγχου ήταν υπερβολικά μεγάλη, δηλαδή το μοντέλο μάθαινε υπερβολικά καλά τα δεδομένα εκπαίδευσης αλλά δεν είχε καλή εφαρμογή σε δεδομένα που δεν είχε ξαναδεί. Αυτό σημαίνει πως πλέον το μοντέλο μας μπορεί να εφαρμοστεί και σε δεδομένα που δεν έχει εκπαιδευτεί.

Τέλος, η τιμή του mape train=0.0001 και η τιμή του mape test=0.0003, τιμές θεμιτές καθώς πρώτον όσο πιο μικρή είναι η τιμή του μέσου απόλυτου ποσοστιαίου σφάλματος τόσο καλύτερη είναι η απόδοση του μοντέλου και δεύτερον η διαφορά στο train set και στο test set είναι πολύ μικρή.

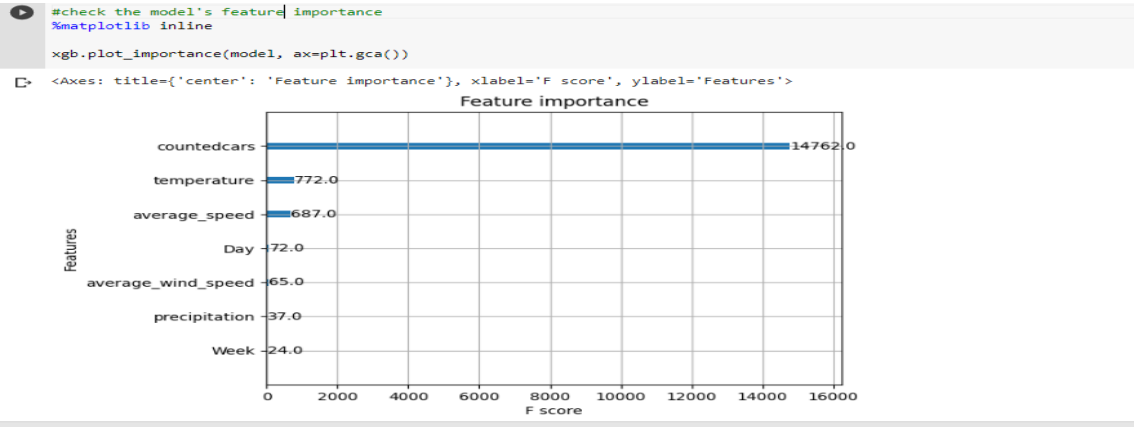
Έπειτα συνεχίζουμε στο κομμάτι της οπτικοποίησης για καλύτερη κατανόηση των αποτελεσμάτων. Χρησιμοποιήθηκε η βιβλιοθήκη Seaborn (alias sns) για να δημιουργήσουμε ένα Γράφημα Γραμμής (lineplot) όπου συγκρίνονται οι πραγματικές τιμές και οι προβλεπόμενες τιμές του κυκλοφοριακού φόρτου. Το μπλε χρώμα αντιπροσωπεύει τις πραγματικές τιμές και το κόκκινο τις τιμές πρόβλεψης που προέκυψαν από την εφαρμογή του μοντέλου.



Εικόνα 56 Γράφημα Lineplot – Σύγκριση πραγματικών δεδομένων και δεδομένων πρόβλεψης.

Γίνεται αντιληπτό πως πάλι αντιμετωπίζουμε το πρόβλημα του Overfitting αλλά σε μικρότερο αποδεκτό βαθμό. Και στα δύο σύνολα δεδομένων έχουμε πάλι κάποιες τιμές που είναι πολύ υψηλότερες από τις συνηθισμένες οι οποίες μπορούν να αποδοθούν σε κάποιο σφάλμα του αισθητήρα ή σε κάποια δυσλειτουργία καθώς το μοντέλο μας φαίνεται να «ακολουθεί» τα πραγματικά δεδομένα.

Χρησιμοποιήθηκε επίσης η ενσωματωμένη συνάρτηση plot_importance() της βιβλιοθήκης XGBoost για να απεικονίσουμε γραφικά τα χαρακτηριστικά (features/μεταβλητές εισόδου) ταξινομημένα με βάση το πόσο σημαντικά τα θεώρησε το μοντέλο μας.



Εικόνα 57 Γράφημα για feature importance του αλγορίθμου XGBoost

Όπως φαίνεται λοιπόν στο πάνω Γράφημα το μοντέλο μας θεωρεί πολύ σημαντική την μεταβλητή countedcars πράγμα λογικό καθώς είναι η target μεταβλητή μας, δηλαδή η μεταβλητή για την οποία θέλουμε να γίνει η πρόβλεψη τιμής.

Ακολουθεί η μεταβλητή temperature και μετά η μεταβλητή average speed. Αυτό σημαίνει πως το μοντέλο μας από τα δεδομένα καιρού θεωρεί πιο σημαντική την θερμοκρασία ως προς την πρόβλεψη του κυκλοφοριακού φόρτου και μάλιστα περισσότερο από την μεταβλητή average_speed που αντιπροσωπεύει την μέση ταχύτητα των οχημάτων.

Αυτό επιβεβαιώνει ως ένα βαθμό την υπόθεσή μας πως τα καιρικά φαινόμενα επηρεάζουν την κυκλοφορία των δρόμων αλλά όχι όλα καθώς η μέση ταχύτητα του ανέμου και η βροχή φαίνεται να μην παίζουν κομβικό ρόλο. Βέβαια αυτό μπορεί να οφείλεται και στην περίοδο που εξετάζει η συγκεκριμένη διπλωματική εργασία που είναι από τον Ιούνιο 2021 μέχρι τον Δεκέμβριο του 2021 όπου ο καιρός στην Αττική είναι ήπιος ως προς τις βροχοπτώσεις και τους δυνατούς ανέμους συνήθως.

6.2 Περιορισμοί μελέτης – Μελλοντικές προτάσεις

Η συγκεκριμένη μελέτη περίπτωσης πραγματοποιήθηκε στα πλαίσια του μεταπτυχιακού προγράμματος στα Πληροφοριακά Συστήματα και το κύριο ζητούμενο ήταν η διερεύνηση πιθανών σχέσεων μεταξύ παρελθοντικών δεδομένων κυκλοφορίας και ιστορικών δεδομένων καιρού οπότε το εύρος της είναι αρκετά περιορισμένο σε σύγκριση με τον πραγματικό κόσμο όπου υπάρχουν πολλοί αστάθμητοι παράγοντες και συσχετίσεις που θα ήταν εξαιρετικά δύσκολο να αποτυπωθούν σε μια έρευνα με περιορισμένο χρόνο και άλλες τεχνικές αντιξοότητες.

Ένας βασικός περιορισμός που παρουσιάστηκε κατά την εκπόνηση της εργασίας αυτής είναι η ποιότητα των Ανοιχτών Κυβερνητικών Δεδομένων από την πύλη του data.gov.gr που περιέχουν ακόμη πολλά σφάλματα στην καταγραφή των δεδομένων καθώς η πλατφόρμα είναι σχετικά καινούργια, αλλά και οι ίδιοι οι αισθητήρες που για διάφορους λόγους έχουν τεχνικές δυσλειτουργίες και σφάλματα κατά την καταγραφή των πληροφοριών.

Οι εφαρμογές της Μηχανικής Μάθησης κερδίζουν έδαφος τα τελευταία χρόνια και είναι πολλά υποσχόμενες. Η ποιότητα των αποτελεσμάτων εξαρτάται σε μεγάλο βαθμό από την εμπειρία αυτών που τις αναπτύσσουν οπότε μια ακόμη δυσκολία κατά την ανάπτυξη ήταν η περιορισμένη εμπειρία. Η βελτιστοποίηση των παραμέτρων του αλγορίθμου είναι μια πολύ ιδιαίτερη και χρονοβόρα διαδικασία και καθώς πρόκειται για πρόβλημα Παλινδρόμησης είναι πιο δύσκολο να βρεθούν οι κατάλληλες ισορροπίες εφόσον δεν υπάρχει κάποιο ακριβές πρότυπο για τις καλύτερες πιθανές τιμές αλλά αποτελεί ιδιαιτερότητα του κάθε προβλήματος.

Οι χρονικές μεταβλητές δεν φαίνεται να παίζουν ιδιαίτερο ρόλο στην κυκλοφορία των οχημάτων στην δική μας περίπτωση αν και αυτό είναι πιθανό να οφείλεται στο γεγονός πως τελικά δεν χρησιμοποιήσαμε τις εθνικές εορτές, τις αργίες κλπ στο μοντέλο μας λόγω τεχνικών δυσκολιών και παραμένει προς εξερεύνηση στο μέλλον για περαιτέρω έρευνα.

Ο τομέας των μεταφορών παίζει κομβικό ρόλο στην καθημερινότητα μας και η πρόβλεψη του κυκλοφοριακού φόρτου, ειδικά στις μεγάλες πόλεις, μπορεί να βοηθήσει στην λήψη αποφάσεων με σκοπό την αποφυγή της υπερφόρτωσης των κυκλοφοριακών

αρτηριών και έτσι να βελτιώσει την εξοικονόμηση χρόνου τόσο σε ατομικό επίπεδο όσο και σε επίπεδο τροφοδοσίας εμπορευμάτων κάθε φύσης με ό,τι αυτό συνεπάγεται.

Καθώς επίσης είναι συχνό φαινόμενο πολλά αυτοκινητιστικά ατυχήματα να λαμβάνουν χώρα σε ώρες αιχμής της κυκλοφορίας η δυνατότητα της πρόβλεψης αυτής μπορεί να αποκτήσει και προληπτικό χαρακτήρα εφόσον υπάρχει η δυνατότητα για ζωντανή παρακολούθηση και αναμετάδοση της κίνησης μέσω εφαρμογών.

Σε αυτό το σημείο αξίζει να αναφερθεί πως τα τελευταία χρόνια έχει αυξηθεί κατά πολύ η αξιοπιστία των εφαρμογών όταν υπάρχει η δυνατότητα σχολίων ή/και αξιολογήσεων από άλλους χρήστες οπότε ίσως να ήταν βοηθητικό να δημιουργηθεί μια επιπλέον εφαρμογή που να επιτρέπει σε διαπιστευμένους χρήστες π.χ. μέσω διασύνδεσης των στοιχείων τους με το taxisnet (για αποφυγή κακόβουλων σχολίων) να αφήνουν σχόλια σε ζωντανό χρόνο σε περίπτωση που κάποιος κεντρικός δρόμος είναι κλειστός λόγω έργων, λόγω ατυχημάτων, διαδηλώσεων, τοπικών εορτασμών κλπ προκειμένου να ενημερώνεται το σύστημα ώστε οι υπόλοιποι χρήστες να ακολουθήσουν διαφορετικές διαδρομές και να εξομαλυνθεί η κυκλοφοριακή συμφόρηση πιο άμεσα αν και είναι εκτός του εύρους της παρούσας μελέτης.

Τέλος, θα παρουσίαζε ιδιαίτερο ενδιαφέρον σε μελλοντικές μελέτες να διερευνηθεί τόσο χωροταξικά όσο και μορφολογικά η κυκλοφορία των δρόμων στην Αττική (και όχι μόνο) καθώς υπάρχουν πολλοί παράγοντες που επηρεάζουν την κίνηση όπως πχ ακόμη και η κλίση των δρόμων, δηλαδή αν πρόκειται για ανηφόρα ή κατηφόρα, ή για στενούς δρόμους με λιγότερες λωρίδες ή για πιο μεγάλες λεωφόρους με πολλές λωρίδες, ή ακόμη αν πρόκειται για ευθύγραμμους δρόμους ή για δρόμους με πολλές συνεχόμενες στροφές.

Κατάλογος Αναφορών / Βιβλιογραφία

[1] Sharda, R., Delen, D., & Turban, E. (2018). Business Intelligence, analytics, and Data Science: A Managerial Perspective (4th ed.). Pearson.

[2] Petsis, S., Karamanou, A., Kalampokis, E., & Tarabanis, K. (2022). Forecasting and explaining emergency department visits in a public hospital. Journal of Intelligent Information Systems, 59(2), 479–500. https://doi.org/10.1007/s10844-022-00716-6

[3] Karamanou, A., Brimos, P., Kalampokis, E., & Tarabanis, K. (2022, December 10). Exploring the quality of dynamic open government data using statistical and machine learning methods. MDPI. Retrieved from https://www.mdpi.com/1424-8220/22/24/9684

[4] «data.gov.gr: Στο φως, για κάθε χρήση, όλα τα δεδομένα του Δημοσίου», Ναυτεμπορική, 22 Δεκεμβρίου 2020.

Διαθέσιμο: https://www.naftemporiki.gr/story/1673307/datagovgr-sto-fos-gia-kathe-xrisi-ola-ta-dedomena-tou-dimosiou

[5] Luke Gloege, Ph.D, «towardsdatascience.com: Read ERA5 Directly into Memory with Python», Towards Data, 21 Μαρτίου 2021.

Διαθέσιμο: https://towardsdatascience.com/read-era5-directly-into-memory-with-python-511a2740bba

[6] Σχετικά. data.gov.gr. (n.d.). Retrieved from http://archive.data.gov.gr/about

[7] Sebastian Carmona A, «towardsdatascience.com: Different Ways to Connect Google Drive to a Google Colab Notebook! (Part 2)», Towards Data, 6 Οκτωβρίου 2022.

Διαθέσιμο: https://towardsdatascience.com/different-ways-to-connect-google-drive-to-a-google-colab-notebook-part-2-b867786aed55

[8] Luke Gloege, P. D. (2021, December 8). *Read ERA5 directly into memory with python*. Medium. Retrieved April 4, 2023, from https://towardsdatascience.com/read-era5-directly-into-memory-with-python-511a2740bba0

[9] Era5 | Ecmwf. (n.d.). Retrieved April 4, 2023, from https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5

[10] Setchell, H. (2023, March 7). ECMWF reanalysis V5. ECMWF. Retrieved April 4, 2023, from https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5

[11] Wikimedia Foundation. (2022, December 24). *European Centre for Medium-range weather forecasts*. Wikipedia. Retrieved April 4, 2023, from https://en.wikipedia.org/wiki/European_Centre_for_Medium-Range_Weather_Forecasts

[12] *Copernicus Climate Data Store*. Copernicus Climate Data Store |. (n.d.). Retrieved April 4, 2023, from https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview

[13] *Datatype grid*. [favicon]. (n.d.). Retrieved April 4, 2023, from http://ucomponents.github.io/data-types/datatype-grid/

[14] Wikimedia Foundation. (2023, March 31). *Data Grid*. Wikipedia. Retrieved April 4, 2023, from https://en.wikipedia.org/wiki/Data_grid

[15] Canada, E. (2023, April 4). *What is Grib - Environment Canada*. What is GRIB - Environment Canada. Retrieved April 4, 2023, from https://weather.gc.ca/grib/what_is_GRIB_e.html

[16] Google. (n.d.). Google colab. Retrieved April 4, 2023, from https://colab.research.google.com/

[17] morato, N. (2022, September 17). *Τι είναι το google colab ή το google colaboratory*. Ikkaro. Retrieved April 4, 2023, from https://www.ikkaro.com/el/google-colaboratory-o-google-colab/

[18] *What is google colab?* Education Ecosystem Blog. (2021, November 18). Retrieved April 4, 2023, from https://educationecosystem.com/blog/what-is-google-colab/

[19] Real Python. (2023, February 7). *Combining data in pandas with merge(), .join(), and CONCAT()*. Real Python. Retrieved April 4, 2023, from https://realpython.com/pandas-merge-join-and-concat/

[20] NASA. (n.d.). Derive Wind Speed and Direction With MERRA-2 Wind Components by: Dana Ostrenga - Updated: Dec 4, 2019, Retrieved April 4, 2023, from https://daac.gsfc.nasa.gov/information/data-in-action?title=Derive+Wind+Speed+and+Direction+With+MERRA-2+Wind+Components

[21] *The average wind speed (M S −1) and wind direction (degrees) for the ...* (n.d.). Retrieved April 4, 2023, from https://www.researchgate.net/figure/The-average-wind-speed-m-s-1-and-wind-direction-degrees-for-the-measurement-period_fig2_242607883

[22] *Θεσμικό Πλαίσιο*. data.gov.gr. (n.d.). Retrieved April 4, 2023, from http://archive.data.gov.gr/pages/thesmikoplaisio

[23] Wikimedia Foundation. (June 8 2022). *Tableau Software*. Wikipedia. Retrieved April 4, 2023, from https://en.wikipedia.org/wiki/Tableau_Software

[24] The Climate Data Store. Copernicus. (n.d.). Retrieved April 4, 2023, from https://climate.copernicus.eu/climate-data-store

[25] *Δικαιώματα πνευματικής ιδιοκτησίας και άδειες*. Copernicus. (n.d.). Retrieved April 4, 2023, from https://www.copernicus.eu/el/pos/dikaiomata-pneymatikis-idioktisias-kai-adeies

[26] *How to install and use cds API on windows*. ECMWF Confluence Wiki. (last updated 13 May 2022). Retrieved April 4, 2023, from https://confluence.ecmwf.int/display/CKB/How+to+install+and+use+CDS+API+on+Windows

[27] Nguyen, C. (June 16 2020). *Tableau: Measures vs. dimensions*. Medium. Retrieved April 4, 2023, from https://medium.com/swlh/tableau-measures-vs-dimensions-cb9986fecef9

[28] Santoalla, D. V. (last updated by Shahram Najm on June 01 2022). *Eccodes Home*. ECMWF Confluence Wiki. Retrieved April 4, 2023, from https://confluence.ecmwf.int/display/ECC

[29] *What is XGBoost?* NVIDIA Data Science Glossary. (n.d.). Retrieved April 4, 2023, from https://www.nvidia.com/en-us/glossary/data-science/xgboost/

[30] Wasike, B. (October 25 2021). *Machine learning with XGBoost and Scikit-Learn*. Section. Retrieved April 4, 2023, from https://www.section.io/engineering-education/machine-learning-with-xgboost-and-scikit-learn/

[31] Raj, R. (n.d.). *Different types of machine learning algorithms*. enjoyalgorithms. Retrieved April 4, 2023, from https://www.enjoyalgorithms.com/blog/classification-of-machine-learning-models

[32] Raj, R. (n.d.). *Classification and regression in machine learning*. enjoyalgorithms. Retrieved April 4, 2023, from https://www.enjoyalgorithms.com/blogs/classification-and-regression-in-machine-learning

[33] GeeksforGeeks. (last updated January 10 2023). *Regression and classification: Supervised machine learning*. GeeksforGeeks. Retrieved April 4, 2023, from https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/

[34] Priya Pedamkar *Supervised machine learning algorithms: 2 types of learning algorithm*. EDUCBA. (last updated March 23 2023). Retrieved April 4, 2023, from https://www.educba.com/supervised-machine-learning-algorithms/

[35] Jaiswal, S. (last updated February 8 2023). *Types of supervised learning you must know – emeritus India*. Emeritus. Retrieved April 4, 2023, from https://emeritus.org/in/learn/types-of-supervised-learning/

[36] Fumo, J. (August 17 2017). *Types of machine learning algorithms you should know*. Medium. Retrieved April 4, 2023, from https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861

[37] *What is supervised learning?* IBM. (n.d.). Retrieved April 4, 2023, from https://www.ibm.com/topics/supervised-learning

[38] Liberman, N. (May 21 2020). *Decision trees and random forests*. Medium. Retrieved April 4, 2023, from https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991

[39] Gurucharan. M. (July 18 2020). *Machine learning basics: Decision tree regression*. Medium. Retrieved April 4, 2023, from https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda

[40] Wikimedia Foundation. (last updated March 27 2023). *Machine learning*. Wikipedia. Retrieved April 4, 2023, from https://en.wikipedia.org/wiki/Machine_learning

[41] Wikimedia Foundation. (last updated April 2 2023). *Decision tree*. Wikipedia. Retrieved April 4, 2023, from https://en.wikipedia.org/wiki/Decision_tree

[42] *31. decision trees in python*. 31. Decision Trees in Python | Machine Learning. (n.d.). Retrieved April 4, 2023, from https://python-course.eu/machine-learning/decision-trees-in-python.php

[43] Brownlee, J. (last updated January 18 2021). *How to develop your first XGBoost model in Python*. MachineLearningMastery.com. Retrieved April 4, 2023, from https://machinelearningmastery.com/develop-first-xgboost-model-python-scikit-learn/

[44] Gupta, A. (last updated June 1 2021). *XGBoost versus Random Forest*. Medium. Retrieved April 4, 2023, from https://medium.com/geekculture/xgboost-versus-random-forest-898e42870f30

[45] Lok, L. (last updated January 6 2022). *Decision trees, random forests, and gradient boosting: What's the difference?* Medium. Retrieved April 4, 2023, from https://towardsdatascience.com/decision-trees-random-forests-and-gradient-boosting-whats-the-difference-ae435cbb67ad

[46] Dante Sblendorio (5 August 2022). *Comparing decision tree algorithms: Random Forest vs. XGBoost*. ActiveState. Retrieved April 4, 2023, from https://www.activestate.com/blog/comparing-decision-tree-algorithms-random-forest-vs-xgboost/

[47] Brownlee, J. (26 April 2021). *A gentle introduction to ensemble learning algorithms*. MachineLearningMastery.com. Retrieved April 4, 2023, from https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/

[48] *Gradient boosting, decision trees and XGBoost with Cuda*. NVIDIA Technical Blog. (10 October 2022). Retrieved April 4, 2023, from https://developer.nvidia.com/blog/gradient-boosting-decision-trees-xgboost-cuda/

[49] Morde, V. (8 April 2019). *XGBoost algorithm: Long may she reign!* Medium. Retrieved April 4, 2023, from https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

[50] Παππάς, Α. (22 December 2020). *Data.gov.gr: Η ελλάδα ανοίγει τα ανώνυμα δεδομένα του δημοσίου*. Techblog. Retrieved April 4, 2023, from https://techblog.gr/internet/data-gov-gr-i-ellada-anoigei-ta-anonyma-dedomena-toy-dimosioy/

[51] Ηγουμενίδη, Τ. (22 December 2020). *Επίσημη mobile εφαρμογή gov.gr - σε λειτουργία το data.gov.gr*. InfoCom. Retrieved April 4, 2023, from https://www.infocom.gr/2020/12/22/episimi-mobile-efarmogi-gov-gr-leitourgia-data-gov-gr/52900/

[52] Παπαστεφάνου, Β. (22 December 2020). *Παρουσίαση της εφαρμογής govapp και της πύλης data.gov.gr - πρόσβαση σε 38 σειρές δεδομένων του ελληνικού δημοσίου*. ertnews.gr. Retrieved April 4, 2023, from https://www.ertnews.gr/eidiseis/ellada/politiki/paroysiasi-tis-efarmogis-govapp-kai-tis-pylis-data-gov-gr-prosvasi-se-38-seires-dedomenon-toy-ellinikoy-dimosioy/

[53] *What are Grib files and how can I read them*. ECMWF Confluence Wiki. (last modified 13 May 2022). Retrieved April 4, 2023, from https://confluence.ecmwf.int/display/CKB/What+are+GRIB+files+and+how+can+I+read+them

[54] *XGBoost parameters:* XGBoost Parameters - xgboost 1.7.4 documentation. (n.d.). Retrieved March 2, 2023, from https://xgboost.readthedocs.io/en/stable/parameter.html#parameters-for-tree-booster

[55] Alam, M. (23 August 2021). *A guide to xgboost hyperparameters*. Medium. Retrieved April 4, 2023, from https://towardsdatascience.com/a-guide-to-xgboost-hyperparameters-87980c7f44a9

[56] *XGBoost parameters*. XGBoost Parameters - xgboost 1.7.5 documentation. (n.d.). Retrieved April 4, 2023, from https://xgboost.readthedocs.io/en/stable/parameter.html

[57] *Notes on parameter tuning*. Notes on Parameter Tuning - xgboost 1.7.5 documentation. (n.d.). Retrieved April 4, 2023, from https://xgboost.readthedocs.io/en/stable/tutorials/param_tuning.html

[58] Brownlee, J. (15 February 2021). *Regression metrics for machine learning*. MachineLearningMastery.com. Retrieved April 4, 2023, from https://machinelearningmastery.com/regression-metrics-for-machine-learning/

[59] *API reference*. scikit. (n.d.). Retrieved April 4, 2023, from https://scikit-learn.org/stable/modules/classes.html#regression-metrics

[60] *3.3. metrics and scoring: Quantifying the quality of predictions*. scikit. (n.d.). Retrieved April 4, 2023, from https://scikit-learn.org/stable/modules/model_evaluation.html

[61] Okamura, S. (30 December 2020). *GRIDSEARCHCV for beginners*. Medium. Retrieved April 4, 2023, from https://towardsdatascience.com/gridsearchcv-for-beginners-db48a90114ee

[61] TrainDataHub. (24 May 2022). *Interpretation of evaluation metrics for regression analysis (MAE, MSE, RMSE, MAPE, R-squared, and...* Medium. Retrieved April 4, 2023, from https://medium.com/@ooemma83/interpretation-of-evaluation-metrics-for-regression-analysis-mae-mse-rmse-mape-r-squared-and-5693b61a9833

[62] *3.34 Regression metrics* scikit. (n.d.). Retrieved April 4, 2023, from https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics

[63] Padhma M. (published October 28, 2021 and last modified on September 1, 2022). *End-to-end introduction to evaluating Regression Models*. Analytics Vidhya. Retrieved April 4, 2023, from https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/

[64] *Loss functions*¶. Loss Functions - machine learning note documentation. (n.d.). Retrieved April 4, 2023, from https://machine-learning-note.readthedocs.io/en/latest/basic/loss_functions.html#regression-loss-functions

[65] *Build a simple map*. Online Help. (n.d.). Retrieved April 4, 2023, from https://help.tableau.com/current/pro/desktop/en-us/maps_howto_simple.htm

[66] Evolytics-p3. (28 June 2022). *Tableau Fundamentals: Dimension vs. measure*. Evolytics. Retrieved April 4, 2023, from https://evolytics.com/blog/tableau-fundamentals-dimension-vs-measure/

[67] *How to choose the Right Chart for Your Data*. Infogram. (n.d.). Retrieved April 4, 2023, from https://infogram.com/page/choose-the-right-chart-data-visualization

[68] Engineer, W. byE. W. T. S. (last updated 8 March 2023). *How to get started with grib2 weather data and python - spire tutorials*. Spire. Retrieved April 4, 2023, from https://spire.com/tutorial/spire-weather-tutorial-intro-to-processing-grib2-data-with-python/

[69] *API*¶. API - pygrib documentation. (n.d.). Retrieved April 4, 2023, from https://jswhit.github.io/pygrib/api.html#example-usage

[70] Dr Christian Mayer, *Python convert CSV to text file (*.csv to .txt) Retrieved April 4, 2023, from https://blog.finxter.com/python-convert-csv-to-text-file-csv-to-txt/

[71] *How to plot grib files with python and matplotlib*. ECMWF Confluence Wiki. (n.d.). Retrieved April 4, 2023, from

https://confluence.ecmwf.int/display/CKB/How+to+plot+GRIB+files+with+Python+and+matplotlib

[72] *How to convert Grib to CSV*. ECMWF Confluence Wiki. (n.d.). Retrieved April 4, 2023, from https://confluence.ecmwf.int/display/CKB/How+to+convert+GRIB+to+CSV

[72] Wikimedia Foundation. (last updated 30 August 2022). *Grib*. Wikipedia. Retrieved April 4, 2023, from https://en.wikipedia.org/wiki/GRIB

[73] Conda-Forge. (n.d.). *Conda-Forge/pygrib-feedstock: A Conda-Smithy Repository for pygrib*. GitHub. Retrieved April 4, 2023, from https://github.com/conda-forge/pygrib-feedstock

[74] *Installation¶*. Installation - pygrib documentation. (n.d.). Retrieved April 4, 2023, from https://jswhit.github.io/pygrib/installing.html

[75] *Managing environments — conda 0.0.0.dev0+placeholder documentation*. (n.d.). Conda — conda documentation. Retrieved April 4, 2023, from https://conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html

[76] *Οι 4 τύποι του business analytics | stepupadvisor.gr*. (n.d.). stepupadvisor.gr. Retrieved April 4, 2023, from https://stepupadvisor.gr/business-analytics-meros-1/

[77] *Ανοιχτά μετεωρολογικά δεδομένα στην Ελλάδα – Ανοιχτά Δεδομένα*. (n.d.). Ανοιχτά Δεδομένα – Ανοιχτά Δεδομένα. Retrieved April 4, 2023, from https://opendata.ellak.gr/2018/02/06/anichta-meteorologika-dedomena-stin-ellada/

[78] *Fact sheet: Reanalysis*. (n.d.). ECMWF Retrieved April 4, 2023, from https://www.ecmwf.int/en/about/media-centre/focus/2020/fact-sheet-reanalysis

[79] *ECMWF | Parameter database*. (n.d.). Retrieved April 4, 2023, from. https://apps.ecmwf.int/codes/grib/param-db

[80] *(PDF) the ERA5 global reanalysis - researchgate*. (n.d.). Retrieved April 4, 2023, from https://www.researchgate.net/publication/341448930_The_ERA5_global_reanalysis

[81] Hersbach H, Bell B, Berrisford P, et al. The ERA5 global reanalysis. QJR Meteorol Soc. 2020;146:1999–2049. https://doi.org/10.1002/qj.3803

[82] Ellis, C. (24 September 2022). *XGBoost overfitting*. Crunching the Data. Retrieved April 4, 2023, from https://crunchingthedata.com/xgboost-overfitting/

[83] Proz, H. H., & Chowdhury, M. F. (n.d.). *Free machine learning tutorial - machine learning & artificial intelligence with python*. Udemy. Retrieved April 4, 2023, from https://www.udemy.com/course/machine-learning-artificial-intelligence-python/

[84] *Evaluation metrics in regression models - regression*. Coursera. (n.d.). Retrieved April 4, 2023, from https://www.coursera.org/lecture/machine-learning-with-python/evaluation-metrics-in-regression-models-5SxtZ

[85] Nair, A. (20 August 2019). *How to judge A machine learning model? (part 1)*. Medium. Retrieved April 4, 2023, from https://medium.com/analytics-vidhya/how-to-judge-a-machine-learning-model-part-1-ca5074a12c69

[86] Zach. (10 May 2021). *What is considered a good RMSE value?* Statology. Retrieved April 4, 2023, from https://www.statology.org/what-is-a-good-rmse/

[87] Brownlee, J. (last updated 14 April 2021). *A gentle introduction to XGBoost loss functions*. MachineLearningMastery.com. Retrieved April 4, 2023, from https://machinelearningmastery.com/xgboost-loss-functions/

[88] GeeksforGeeks. (19 December 2021). *How to create boxplot from pandas DataFrame?* GeeksforGeeks. Retrieved April 4, 2023, from https://www.geeksforgeeks.org/how-to-create-boxplot-from-pandas-dataframe/

[89] GeeksforGeeks. (last updated 8 March 2022). *Box plot in python using Matplotlib*. GeeksforGeeks. Retrieved April 4, 2023, from https://www.geeksforgeeks.org/box-plot-in-python-using-matplotlib/

[90] GeeksforGeeks. (9 November 2022). *Data Visualisation in python using Matplotlib and Seaborn*. GeeksforGeeks. Retrieved April 4, 2023, from https://www.geeksforgeeks.org/data-visualisation-in-python-using-matplotlib-and-seaborn/

[91] Ajitesh KumarI (16 April 2022). *Correlation Concepts, Matrix & Heatmap using Seaborn*. Data Analytics. Retrieved April 4, 2023, from https://vitalflux.com/correlation-heatmap-with-seaborn-pandas/

[92] *How do weather events impact roads?* How Do Weather Events Impact Roads? - FHWA Road Weather Management. (n.d.). Retrieved April 4, 2023, from https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm

[93] *Weather and Climate Data*. xarray. (last updated 22 March 2023). Retrieved April 4, 2023, from https://docs.xarray.dev/en/stable/user-guide/weather-climate.html

Παράρτημα

Data.gov.gr – Θεσμικό πλαίσιο

Το data.gov.gr αποτελεί εργαλείο υλοποίησης της πολιτικής των ανοικτών δεδομένων κατ' εφαρμογή της σχετικής νομοθεσίας, μετά και την ενσωμάτωση της Οδηγίας 2013/37/ΕΕ.

Το θεσμικό πλαίσιο που διέπει τη συγκεκριμένη δράση είναι:

1. Ν. 4305/2014 (ΦΕΚ 237/Α΄) «Ανοικτή διάθεση και περαιτέρω χρήση εγγράφων, πληροφοριών και δεδομένων του δημόσιου τομέα, τροποποίηση του ν. 3448/2006 (Α΄ 57), προσαρμογή της εθνικής νομοθεσίας στις διατάξεις της Οδηγίας 2013/37/ΕΕ του

Ευρωπαϊκού Κοινοβουλίου και του Συμβουλίου, περαιτέρω ενίσχυση της διαφάνειας, ρυθμίσεις θεμάτων Εισαγωγικού Διαγωνισμού Ε.Σ.Δ.Δ.Α. και άλλες διατάξεις».

2. Ν. 3448/2006 (ΦΕΚ 57/Α΄) «Για την περαιτέρω χρήση πληροφοριών του δημόσιου τομέα και τη ρύθμιση θεμάτων αρμοδιότητας Υπουργείου Εσωτερικών, Δημόσιας Διοίκησης και Αποκέντρωσης».

3. Αιτιολογική έκθεση Ν. 4305/2014.

4. Αριθ. ΔΗΔ/Φ.40/407/8.1.2015 εγκύκλιος με θέμα «Εφαρμογή των διατάξεων του Κεφαλαίου Α' του ν. 4305/2014 (ΦΕΚ 237/Α΄) σχετικά με την «ανοικτή διάθεση και περαιτέρω χρήση εγγράφων, πληροφοριών και δεδομένων του δημόσιου τομέα, την τροποποίηση των διατάξεων του πρώτου κεφαλαίου του ν. 3448/2006, προσαρμογή της εθνικής νομοθεσίας στις διατάξεις της οδηγίας 2013/37 του Ευρωπαϊκού Κοινοβουλίου και του Συμβουλίου και την περαιτέρω ενίσχυση της διαφάνειας στο δημόσιο τομέα». (ΑΔΑ: ΩΩΡΜΧ-ΜΒΛ).

5. Αριθ. ΔΗΔ. Φ./ 19710/16.6.2015 εγκύκλιος (ΑΔΑ:7ΧΩΨ465ΦΘΕ-Β2Ι) με θέμα «Ανοικτή διάθεση και περαιτέρω χρήση εγγράφων, πληροφοριών και δεδομένων του δημόσιου τομέα σύμφωνα με το κεφ. Α΄ του Ν. 4305/2014».

6. Αριθμ. Πρωτ.: ΔΗΔ/Φ.40/2369/24.1.2017 εγκύκλιος (ΑΔΑ: 6Γ07465ΧΘΨ-473) με θέμα «Επίσπευση ενεργειών από τους υπόχρεους φορείς για τη διάθεση και περαιτέρω χρήση των ανοικτών δεδομένων σε εφαρμογή του ν. 4305/2014 (ΦΕΚ 237 Α΄)».

7. Αριθμ. Πρωτ.: ΔΗΔ/17544/11.5.2018 εγκύκλιος (ΑΔΑ: 676Ν465ΧΘΨ-ΚΕΘ) με θέμα «Επικαιροποίηση απόφασης του άρθρου 10 του Ν. 4305/2014».

8. Αριθμ. Πρωτ.: ΔΗΔ/3274/22.1.2019 εγκύκλιος με θέμα «Ανοικτή διάθεση των δεδομένων σε εφαρμογή του ν. 4305/2014».

Copernicus – Άδεια χρήσης δεδομένων

I. Licence to Use Copernicus Products

1. Definitions

1.1. 'Licensor' means the European Union, represented by the European Centre for Medium-Range Weather Forecasts (ECMWF).

1.2. ‘Licensee’ means all natural or legal persons who agree to the terms of this Licence.

1.3. ‘Licence’ means this license agreement between the Licensor and the Licensee as amended from time to time.

1.4. ‘Copernicus Services’ means:

1.4.1. the Copernicus Atmosphere Monitoring Service (CAMS), which is to provide information on air quality on a local, national, and European scale, and the chemical composition of the atmosphere on a global scale.

1.4.2. the Copernicus Climate Change Service (C3S), which is to provide information to increase the knowledge base to support policies on adaptation to and mitigation of climate change

1.5. ‘Copernicus Products’ means all products listed in the C3S or CAMS Service Product Specification or any other items available through an ECMWF Copernicus portal, except those items which are labelled/flagged as being subject to their own separate terms of use.

1.6. ‘Intellectual Property Rights’ refers to intellectual property rights of all kinds,

1.6.1. including: all patents; rights to inventions; copyright and related rights; moral rights; trademarks and service marks; trade names and domain names; rights in get-up; rights to goodwill or to sue for passing off or unfair competition; rights in designs; rights in computer software; database rights; rights in confidential information (including know-how and trade secrets); any other rights in the nature of intellectual property rights;

1.6.2. in each case whether registered or unregistered and including all applications (or rights to apply) for, and renewals or extensions of, such rights and all similar or equivalent rights or forms of protection which subsist or will subsist now or in the future in any part of the world together with all rights of action in relation to the infringement of any of the above.

1.7. ‘Copernicus Contractor’ refers to providers of Copernicus related goods and services to ECMWF, including information and data, to the Licensor and/or to the users.

1.8. ‘Copernicus Regulations’ refers to Regulation (EU) No 377/2014 of the European Parliament and of the Council of 3 April 2014 establishing the Copernicus Programme.

1.9. ‘ECMWF Agreement’ refers to the agreement between the European Commission and ECMWF dated 11 November 2014 on the implementation of CAMS and C3S.

2. Introduction

Copernicus is funded under the Copernicus Regulation and operated by ECMWF under the ECMWF Agreement. Access to all Copernicus (previously known as GMES or Global Monitoring for Environment and Security) Information and Data is regulated under Regulation (EU) No 1159/2013 of the European Parliament and of the Council of 12 July 2013 on the European Earth monitoring programme, under the ECMWF Agreement and under the European Commission’s Terms and Conditions. Access to all Copernicus information is regulated under Regulation (EU) No 1159/2013 and under the ECMWF Agreement.

3. Terms of the Licence

This Licence sets out the terms for use of Copernicus Products. By agreeing to these terms, the Licensee agrees to abide by all of the terms and conditions in this Licence for the use of Copernicus Products.

4. Licence Permission

4.1. This Licence is free of charge, worldwide, non-exclusive, royalty free and perpetual.

4.2. Access to Copernicus Products is given for any purpose in so far as it is lawful, whereas use may include, but is not limited to: reproduction; distribution; communication to the public; adaptation, modification and combination with other data and information; or any combination of the foregoing.

5. Attribution

5.1. All users of Copernicus Products must provide clear and visible attribution to the Copernicus programme. The Licensee will communicate to the public the source of the Copernicus Products by crediting the Copernicus Climate Change and Atmosphere Monitoring Services:

5.1.1. Where the Licensee communicates or distributes Copernicus Products to the public, the Licensee shall inform the recipients of the source by using the following or any similar notice:

• 'Generated using Copernicus Climate Change Service information [Year]' and/or
• 'Generated using Copernicus Atmosphere Monitoring Service information [Year]'.

5.1.2. Where the Licensee makes or contributes to a publication or distribution containing adapted or modified Copernicus Products, the Licensee shall provide the following or any similar notice:
• 'Contains modified Copernicus Climate Change Service information [Year]'; and/or
• 'Contains modified Copernicus Atmosphere Monitoring Service information [Year]'

5.1.3. Any such publication or distribution covered by clauses 5.1.1 and 5.1.2 shall state that neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains.

6. Intellectual Property Rights

6.1. All Intellectual Property Rights in the Copernicus Products belong, and will continue to belong, to the European Union.

6.2. All Intellectual Property Rights of new items created as a result of modifying or adapting the Copernicus Products through the applications and workflows accessible on the ECMWF Copernicus portals (e.g. through the CDS Toolbox) will belong to the European Union.

6.3. All other new Intellectual Property Rights created as a result of modifying or adapting the Copernicus information will be owned by the creator.

7. Provision of Third Party Information and Data

This Licence only covers Copernicus Products. Access to third party products, information, and data related to Copernicus information to which the Licensee is directed or which can be directly accessed through any Copernicus portal will be subject to different licence terms.

8. Disclaimers

8.1. Neither the Licensor nor ECMWF warrant that Copernicus Products will be free from errors or omissions or that such errors or omissions can or will be rectified, or that the Licensee will have uninterrupted, continuous, or timely access to Copernicus Products.

8.2. The Licensor, as well as ECMWF, exclude all warranties, conditions, terms, undertakings, obligations whether express or implied by statute including but not

limited to the implied warranties of satisfactory quality and fitness for a particular purpose or otherwise to the fullest extent permitted by law.

9. Liabilities

Neither the Licensor nor ECMWF will accept liability for any damage, loss whether direct, indirect or consequential resulting from the Licensee's use of the Copernicus Products.

10. Termination of and Changes to this Licence

The Licensor may terminate this licence if the Licensee breaches its obligations under these terms. The Licensor may revise this Licence at any time and will notify the Licensee of any revisions.

11. Arbitration Clause and Governing Law

In the event of a dispute arising in connection with this License, the parties shall attempt to settle their differences in an amicable manner. If any dispute cannot be so settled, it shall be settled under the Rules of Arbitration of the International Chamber of Commerce by one arbitrator appointed in accordance with the said rules sitting in London, United Kingdom. The proceedings shall be in the English language. The right of appeal by either party to regular Courts on a question of law arising in the course of any arbitral proceedings or out of an award made in any arbitral proceedings is hereby agreed to be excluded.

It is the intention of the parties that this License shall comprehensively govern the legal relations between the parties to the Licence, without interference or contradiction by any unspecified law. However, where a matter is not specifically covered by these terms or a provision of the Licence terms is ambiguous or unclear, resolution shall be found by reference to the laws of England and Wales, including any relevant law of the European Union.

Nothing stated in this License shall be construed as a waiver of any privileges or immunities of the Licensor or of ECMWF.

Version 1.2 (November 2019)