**ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ**

**ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ**

**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

**ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ**

# «Διαχείριση δεδομένων τύπου Big Data για την επίτευξη αειφόρου αστικής ανάπτυξης»

**Λαμία Ιούλιος 2018**

Θοδωρής Ψαλλιδάς
Επιβλέπων Καθηγητής: κος. Σταμούλης Γεώργιος
Συνεπιβλέπων Καθηγητής: κος. Κόκκινος Κωνσταντίνος

«Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις (1), που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί χωρίς να τα περικλείω σε εισαγωγικά και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.

2. Δέχομαι ότι η αυτολεξεί παράθεση χωρίς εισαγωγικά, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.

3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ.), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια.

4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.»

# ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια της απόκτησης του πτυχίου μου από το Τμήμα Πληροφορικής του Πανεπιστημίου Θεσσαλίας στην Λαμία υπό την επίβλεψη του καθηγητή κυρίου Κωνσταντίνου Κόκκινου.

Θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή για την βοήθειά του καθώς και για το θέμα της εργασίας που μου πρότεινε, το οποίο διεύρυνε τις γνώσεις μου γύρω από το θέμα των Big Data και μιας νέας πλατφόρμας.

# ΠΕΡΙΛΗΨΗ

Την σύγχρονη εποχή, την εποχή της πληροφορίας, όπου η τεχνολογία ανθίζει, τα δεδομένα που παράγονται πολλαπλασιάζονται συνεχώς και ο ρυθμός αστικοποίησης μιας πόλης ολοένα και αυξάνεται, δημιουργείται έτσι η επιτακτική ανάγκη για την αειφορία της αστικής βιωσιμότητας.

Στην παρούσα πτυχιακή εργασία θα δούμε πώς τα μεγάλα αστικά δεδομένα που παράγονται από τους αισθητήρες μπορούν να βοηθήσουν στην ευεξία της πόλης και στην αειφορία της βιωσιμότητας. Πιο συγκεκριμένα, γίνεται ανάλυση και επεξεργασία δεδομένων κυκλοφορίας οχημάτων και ατμοσφαιρικής ρύπανσης από μία σύγχρονη και έξυπνη πόλη, όπως αυτή είναι η πόλη Ώρχους της Δανίας, με την χρήση της πλατφόρμας του Apache Spark.

Με τη χρήση κάποιων αλγορίθμων μηχανικής μάθησης εκπαιδεύονται κάποια μοντέλα πρόβλεψης, συγκρίνοντας την ακρίβεια που έχουν αυτά μεταξύ τους. Πέρα από την χρήση της πλατφόρμας και της μηχανικής μάθησης, δημιουργήθηκε και μία διαδικτυακή εφαρμογή, ένας τοπικός server με τη χρήση του bokeh, για την απεικόνιση των δεδομένων και των αισθητήρων.

## ABSTRACT

Nowadays, the information age in which technology is developing, the data produced are constantly growing and so does a city's urbanization rate. Thus, an urgent need of urban sustainability is being created.

In this thesis, we will examine how large urban data produced by sensors can assist a city's health and its urban sustainability. More specifically, analysis and processing of traffic data, as well as air pollution data from a modern and intelligent city, such as the Danish town of Arhus, are carried out using the Apache Spark platform.

By using some machine learning algorithms, we have created some prediction models, comparing the accuracy between them. In addition to using the platform, we have also created a dashboard, an online web interface running locally using bokeh server, to display data and sensors information about traffic and air quality index.

# ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

## ΕΙΚΟΝΕΣ



## ΠΙΝΑΚΕΣ



## ΕΞΙΣΩΣΕΙΣ

# 1.ΕΙΣΑΓΩΓΗ

## 1.1 Μεγάλα Δεδομένα

Η εποχή μετά την τεχνολογική επανάσταση και την εκρηκτική εξάπλωση του διαδικτύου μετονομάστηκε ως εποχή της πληροφορίας και όχι άδικα. Με τη συνεχή εξέλιξη της τεχνολογίας και εξαιτίας του IoT κάθε ηλεκτρική συσκευή αναβαθμίζεται πλέον σε ηλεκτρονική. Σε κάθε γωνία του σπιτιού, της πόλης ή ακόμα και στο αυτοκίνητο τοποθετούνται αισθητήρες που είναι άμεσα συνδεδεμένοι με το διαδίκτυο. Η μάστιγα του IoT έχει ως αποτέλεσμα την αμέριστη δημιουργία δεδομένων. Ο όρος του IoT είναι στενά συνδεδεμένος με τον όρο των δεδομένων, επομένως η ραγδαία αύξηση των συσκευών του IoT φέρει ως αποτέλεσμα την ραγδαία αύξηση των παραγόμενων δεδομένων. Ο όγκος των δεδομένων που δημιουργούνται καθημερινά αυξάνεται με γεωμετρική πρόοδο και εδώ είναι που γεννιέται ο όρος Big Data. Τα Big Data χαρακτηρίζουν οποιαδήποτε μορφή δεδομένων, η οποία παράγεται κατά συρροή από διάφορους αισθητήρες συλλογής δεδομένων και ζητούν ειδική και πολύ γρήγορη μεταχείριση.

Ο όρος IoT, ο οποίος βαφτίζει “έξυπνο” ό,τι αγγίζει, έκανε την εμφάνισή του και στις σύγχρονες πόλεις. Πλέον, κάθε πόλη που θεωρείται έξυπνη εξοπλίζεται με έναν τεράστιο στόλο αισθητήρων, για τη συλλογή διαφόρων δεδομένων. Μεγάλης ποικιλίας αισθητήρες κάνουν την εμφάνισή τους σε πολλές αστικές υποδομές. Σε κάθε σπιθαμή των πόλεων τοποθετούνται διάφοροι αισθητήρες, όπως για παράδειγμα σε σταθμούς φόρτισης ηλεκτρικών οχημάτων, σε σταθμούς ηλεκτρικής ενέργειας, σε σταθμούς ύδρευσης, σε φωτεινούς σηματοδότες, σε μέσα μαζικής μεταφοράς ακόμα και στον δρόμο. Αισθητήρες που μπορεί να μετράνε από τη θερμοκρασία, την ποσότητα κατανάλωσης του νερού ή της ηλεκτρικής ενέργειας μέχρι και τις εκπομπές ρύπων των οχημάτων. Η τοποθέτηση όλων αυτών των αισθητήρων μετατρέπει κάθε πόλη σε μία γεννήτρια μεγάλων δεδομένων.

Τα δεδομένα που ξεχειλίζουν από τους αισθητήρες της πόλης αποτελούν τα μεγάλα αστικά δεδομένα τα οποία συνδέονται με την ευφυία της πόλης. Εξαντλώντας όλη την πληροφορία που μπορούν να εξάγουν τα αστικά δεδομένα και μετατρέποντάς την σε πολύτιμη γνώση, αυξάνεται ο δείκτης νοημοσύνης της πόλης, κάνοντας την πόλη πιο έμπειρη να αντιμετωπίσει τα ήδη υπάρχοντα προβλήματα αλλά και νέα εάν προκύψουν [1]. Είναι, επίσης, ζωτικής σημασίας για τις ζωές των ανθρώπων καθώς τους κερδίζουν χρόνο, κρατώντας τους ενήμερους και διευκολύνοντας τον τρόπο κίνησης τους στην πόλη. Τα μεγάλα δεδομένα θα επιφέρουν ριζικές αλλαγές στον τρόπο διοίκησης των αστικών υποδομών και των υπηρεσιών κάθε πόλης, εξελίσσοντας τις υπηρεσίες και αναπτύσσοντας νέες βιομηχανίες.

## 1.2 Αστική αειφορία

Ο υψηλός ρυθμός της αστικοποίησης, μία νέα σημαντική τάση που επικρατεί στην παγκόσμια ανάπτυξη, είναι ο αστάθμητος παράγοντας που ασκεί πολύ μεγάλη πίεση στους πόρους και στις υπηρεσίες των σύγχρονων αστικών πόλεων, με αποτέλεσμα να υποβαθμίζει αρκετά το βιοτικό επίπεδο. Η υψηλή αστικοποίηση αποτελεί πηγή προβλημάτων για τις σύγχρονες πόλεις. Αυξάνεται η κυκλοφοριακή συμφόρηση και η εκπομπή ρύπων, δυσχεραίνονται οι ανθρώπινες και οι αστικές μετακινήσεις, υποβαθμίζεται η υγεία. Αυτά είναι κάποια από τα προβλήματα που δημιουργεί η ταχεία αστικοποίηση σε μία απροετοίμαστη πόλη. Πρέπει κάθε πόλη να βρει λύσεις για να απαλλαγεί από τα προβλήματα της αστικοποίησης και να ξεκινήσει την αναδιάρθρωση της αστικής ευημερίας.

Η έξυπνη πόλη προτείνεται ως μία πιο αποτελεσματική προσέγγιση για την επίτευξη της αστικής διαχείρισης [2]. Τα μεγάλα αστικά δεδομένα, είναι η μόνη διέξοδος των προβλημάτων και αποτελούν μονόδρομο για να επιτευχθεί η αειφορία της αστικής βιωσιμότητας. Η γνώση που θα προκύψει από την διαχείριση των δεδομένων θα δώσει ένα τέλος στα προβλήματα των σύγχρονων πόλεων, όπως για παράδειγμα στην κυκλοφοριακή συμφόρηση, τη ρύπανση του αέρα, τη σπατάλη ενέργειας, νερού και καυσίμων κ.α [3]. Κάθε έξυπνη πόλη θα έχει τον απόλυτο έλεγχο των πόρων της, θα μπορεί να ενημερώνει και να προτείνει στους πολίτες εναλλακτικές διαδρομές για την αποφυγή των ρύπων. Θα μπορεί να διαχειρίζεται πιο σωστά την κίνηση στους δρόμους, ρυθμίζοντας τους σηματοδότες [4]. Η άμβλυνση της κυκλοφοριακής κίνησης παίζει σημαντικό ρόλο στην μείωση των εκπομπών ρύπων από τα οχήματα, δημιουργώντας έτσι ένα καλύτερο περιβάλλον βιωσιμότητας για τους πολίτες της [5]. Δίνοντας έναν πιο υγιεινό τόπο σε ανθρώπους που είναι πιο ευάλωτοι στις ρυπογόνες ουσίες όπως είναι οι ηλικιωμένοι, οι άρρωστοι και τα παιδιά.

Είναι εμφανές πως η αειφορία της βιωσιμότητας και η ευημερία των αστικών υποδομών πηγάζει από την ευφυία και την εμπειρία κάθε πόλης.

## 1.3 Δομή της εργασίας

Η εργασία συνεχίζει με το 2ο κεφάλαιο να αναφέρεται στο θεωρητικό υπόβαθρο της εργασίας, με εκτενή ανάλυση των όρων IoT, Big Data και Hadoop. Στο 3ο κεφάλαιο γίνεται αναλυτική περιγραφή της πλατφόρμας των πειραμάτων, Apache Spark. Ακολουθεί το 4ο κεφάλαιο όπου αναλύεται η μηχανική μάθηση και πιο συγκεκριμένα το πακέτο της μηχανικής μάθησης που φιλοξενεί η πλατφόρμα του Apache Spark. Στο 5ο κεφάλαιο παρουσιάζονται τα δεδομένα και γίνεται η πειραματική ανάλυση και η χρήση των μοντέλων και στο τελευταίο κεφάλαιο, το

κεφάλαιο 6 συνοψίζονται τα αποτελέσματα και τα συμπεράσματα της εργασίας και προτείνονται μελλοντικές ενέργειες.

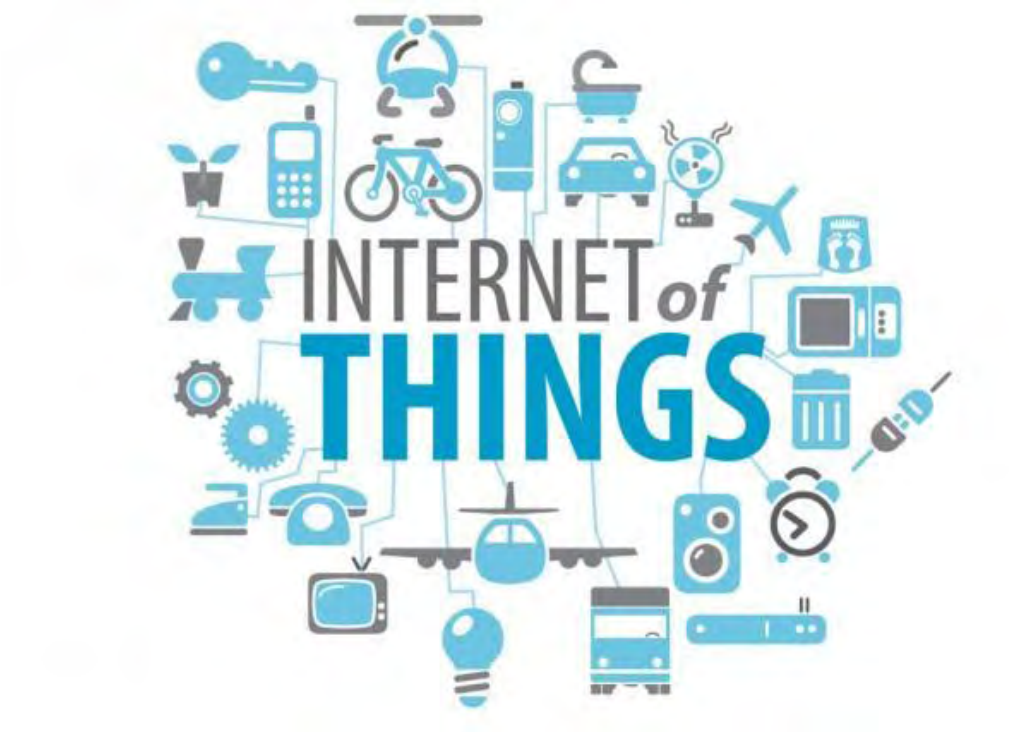
# 2. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

## 2.1 Διαδίκτυο των Πραγμάτων (IoT)

Το IoT (Internet of Things) αποτελεί το δίκτυο επικοινωνίας πληθώρας καθημερινών συσκευών, συσκευών που ενσωματώνουν αισθητήρες για τη συλλογή δεδομένων. Πιο συγκεκριμένα, το διαδίκτυο των πραγμάτων είναι η καθολική σύνδεση όλων των ηλεκτρικών συσκευών μεταξύ τους, κάτι το οποίο υλοποιείται με τη χρήση και σύνδεσή τους με το διαδίκτυο. Αυτός είναι και ο λόγος του Internet στην ονομασία [6].

Η έννοια του things τώρα, δεν περιορίζεται σε κάποια συγκεκριμένη κατηγορία συσκευών αλλά αναφέρεται σε μία ευρεία γκάμα αυτών. Συσκευές ξένες μεταξύ τους, όπως για παράδειγμα έξυπνα κινητά, έξυπνα ρολόγια, κάμερες, έξυπνα αυτοκίνητα και διάφοροι άλλοι ετερογενείς αισθητήρες συνεργάζονται αρμονικά κάτω από ένα στρώμα επικοινωνίας.

Σκοπός λοιπόν του Internet of Things σε αυτές τις συσκευές είναι ο απομακρυσμένος έλεγχος και η διαχείρισή τους μέσω κάποιου υπολογιστή συλλέγοντας πληροφορίες, εξάγοντας πολύτιμη γνώση και απλουστεύοντας έτσι τη ζωή μας.




**Εικόνα 1 Διαδίκτυο των Πραγμάτων**

### 2.1.1 Χρήση του IoT

Με την τεχνολογία να εξελίσσεται συνεχώς, το διαδίκτυο των πραγμάτων στρατολογεί ολοένα και περισσότερες συσκευές/αισθητήρες. Ο ρυθμός αύξησης του στόλου ακολουθεί γεωμετρική πρόοδο. Με το μέγεθος του στόλου να ξεπερνάει τον αριθμό των ανθρώπων, οι συσκευές που απαρτίζουν το στόλο του Internet of Things παράγουν ένα τεράστιο πλήθος δεδομένων από τις πληροφορίες που συλλέγονται από τους ενσωματωμένους αισθητήρες των συσκευών.

Αισθητήρες ξένοι μεταξύ τους, όπως για παράδειγμα συσκευές εντοπισμού, έξυπνα ρολόγια, συσκευές Arduino, αισθητήρες μέτρησης θερμοκρασίας κ.ά. είναι στον κόσμο του IoT. Συνδεδεμένοι με το διαδίκτυο επικοινωνούν με ένα σύστημα και του παρέχουν τις μετρήσεις που συλλέγουν. Ανάλογα την συσκευή και την χρησιμότητα της, το είδος και ο τύπος των δεδομένων που παράγονται διαφέρουν μεταξύ τους, αλλά ο όγκος των δεδομένων μπορεί να είναι εξωφρενικά μεγάλος. Αυτός ο τεράστιος όγκος δεδομένων λοιπόν, μας εισάγει στο φαινόμενο της ψηφιακής εποχής των big data. Τι ακριβώς όμως είναι τα big data;

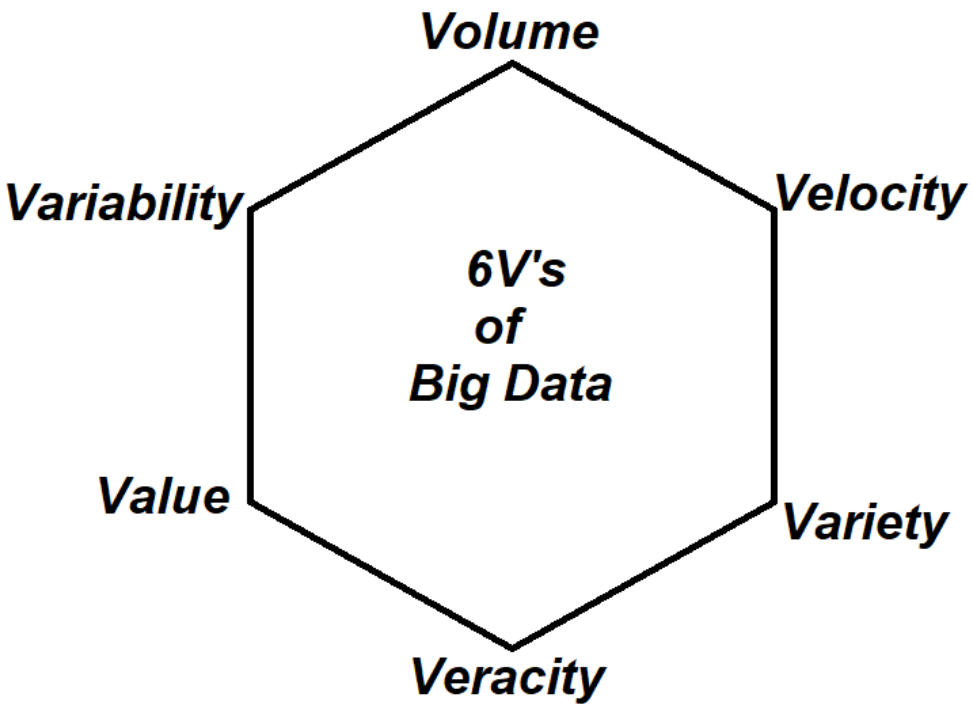
## 2.2 Μεγάλα Δεδομένα (Big Data)

Τα μεγάλα δεδομένα είναι οποιαδήποτε ογκώδη ποσότητα δεδομένων, είτε αυτά είναι δομημένα, ημι-δομημένα ή και αδόμητα. Τα δεδομένα αυτά έχουν τη δυνατότητα να συλλέγονται, είναι η λεγόμενη εξόρυξη δεδομένων για την άντληση χρήσιμων πληροφοριών [7]. Ο όγκος αυτών των δεδομένων τα εμποδίζει να επεξεργαστούν με τις παραδοσιακές μεθόδους, γεγονός που τα χαρακτηρίζει και ως big data και τα ξεχωρίζει από τα υπόλοιπα σύνολα δεδομένων.


**Εικόνα 2 Big Data**

### 2.2.1 Βασικά Χαρακτηριστικά

Η έννοια των Big Data χαρακτηρίζεται από τα εξής 6 V's [7]: τον όγκο (volume), την ταχύτητα (velocity), την ποικιλία (variety), την ακρίβεια (veracity), την αξία (value) και τέλος τη μεταβλητότητα (variability). Λέμε ότι ένα πρόβλημα υπάγεται στην κατηγορία των μεγάλων δεδομένων όταν αυτό παρουσιάζει τα παραπάνω χαρακτηριστικά.




**Εικόνα 3 6 V's**

### 2.2.1.2 Όγκος

Ο Όγκος αναφέρεται στο τεράστιο μέγεθος των δεδομένων. Σαν μονάδα μέτρησης του όγκου των δεδομένων χρησιμοποιούμε τα gigabytes, αν και πλέον το μέγεθος των δεδομένων αγγίζει ακόμα και τα terabytes. Λόγω του Internet of Things, το μέγεθος των δεδομένων γνωρίζει εκθετική αύξηση, γεγονός που δεν αποκλείει να δούμε και δεδομένα που ξεπερνάνε τα zettaBytes.

### 2.2.1.3 Ταχύτητα

Η ταχύτητα είναι αυτή που χαρακτηρίζει τον ρυθμό με τον οποίο λαμβάνουμε τα παραγόμενα δεδομένα προς επεξεργασία και αποθήκευση. Η ταχύτητα ίσως να αποτελεί τη μεγαλύτερη πρόκληση στις μέρες μας καθώς υπάρχει η ανάγκη της ανάλυσης και επεξεργασίας των δεδομένων σε πραγματικό χρόνο.

### 2.2.1.4 Ποικιλία

Η ποικιλία αναφέρεται στον τύπο και στη μορφή που θα έχουν τα δεδομένα. Πιο συγκεκριμένα, συναντάμε τρεις βασικές κατηγορίες τύπων δεδομένων, οι οποίες είναι:

- Δομημένα: όπως για παράδειγμα mysql, access, oracle, rdbs και διάφορες άλλες βάσεις δεδομένων.
- Ημι-δομημένα: όπως τα email, τα tweets, τα log files κ.ά.
- Αδόμητα: είναι η τελευταία κατηγορία και αποτελείται από εικόνες, βίντεο, ήχο, κείμενο ή ακόμα και συνδυασμό αυτών.

### 2.2.1.5 Ακρίβεια ή Εγκυρότητα

Η ποιότητα πληροφορίας των δεδομένων παίζει πολύ σημαντικό ρόλο στην ακρίβεια της ανάλυσης. Αυτός είναι και ο βασικότερος λόγος όπου κάποια δεδομένα χρειάζονται καθάρισμα από τυχόν θορύβους και ανωμαλίες. Ένα απλό παράδειγμα στην φιλαλήθεια των δεδομένων αποτελεί ο έλεγχος της βάσης δεδομένων για τυχόν χρήστες με μη υπαρκτά email.

### 2.2.1.6 Αξία

Η αξία είναι ο πλούτος που κρύβουν τα δεδομένα. Αυτός είναι και ο λόγος που τα αναλύουμε προκειμένου να τον ανακαλύψουμε. Ο κρυφός πλούτος είναι αυτός που καθορίζει και τη σημαντικότητα που έχει η ανάλυση των δεδομένων. Η αξία των δεδομένων συνήθως προκύπτει από κάποια αναγνώριση μοτίβων ή από λήψη αποφάσεων.

### 2.2.1.7 Μεταβλητότητα

Η μεταβλητότητα των δεδομένων αναφέρεται στη διαρκή αλλαγή του νοήματός τους. Τα δεδομένα αλλάζουν σημασία κατά την πάροδο του χρόνου. Επομένως, ένας σημαντικός παράγοντας στη μεταβλητότητα είναι ο χρόνος γιατί ο χρόνος είναι αυτός που καθορίζει τη ζωή των δεδομένων.

## 2.3 Apache Hadoop

Το Hadoop δεν είναι μία ακόμη κλασσική βάση δεδομένων. Το hadoop είναι μία πλατφόρμα ανοιχτού κώδικα που χρησιμοποιεί το πρότυπο nosql. Είναι μία πλατφόρμα γραμμένη σε Java, η οποία βασίζεται σε ένα κατανεμημένο σύστημα αρχείων από την Google, γνωστό και ως Google file system, χρησιμοποιώντας το μοντέλο mapReduce. Είναι ειδικά σχεδιασμένη για να διαχειρίζεται μεγάλες ποσότητες δεδομένων τα οποία λόγω του τεράστιου όγκου τους δεν είναι δυνατόν να τοποθετηθούν στο δίσκο ενός υπολογιστή κι έτσι η επεξεργασία και η αποθήκευση γίνεται κατανεμημένα σε πολλά clusters, ισοσκελίζοντας με αυτό τον τρόπο τον φόρτο εργασίας.

Ένα από τα πιο γνωστά παραδείγματα κατανεμημένου συστήματος αρχείων αποτελεί το Hadoop Distributed File system (HDFS). Ένα σύστημα σχεδιασμένο συγκεκριμένα για την αποθήκευση μεγάλων ποσών δεδομένων μοιρασμένο σε πολλούς διακομιστές. Για το λόγο αυτό, το HDFS δεν είναι κατάλληλο για εργασίες σε μικρές ομάδες δεδομένων [8].


**Εικόνα 4 Hadoop**

# 3.APACHE SPARK

Το κεφάλαιο του Apache Spark αναφέρεται στην πηγή [9].

## 3.1 Τί εστί Spark

Το Apache Spark είναι μία -ανοικτού κώδικα- υπολογιστική πλατφόρμα όπως το Hadoop. Είναι ειδικά σχεδιασμένη να λειτουργεί σε κατανεμημένα υπολογιστικά συστήματα. Επιτρέπει τη δημιουργία κατανεμημένων εφαρμογών για την διαχείριση και επεξεργασία μεγάλου όγκου δεδομένων χρησιμοποιώντας ταυτόχρονα και την παράλληλη επεξεργασία. Πρόκειται, ουσιαστικά, για μία επέκταση του μοντέλου MapReduce που αναφέρθηκε στην πλατφόρμα του Hadoop. Βέβαια, υπάρχει ένα μεγάλο πλήθος εργαλείων που το κάνουν ξεχωριστό και πιο γρήγορο σε σχέση με το Hadoop. Ένα από τα βασικά χαρακτηριστικά του Apache Spark είναι η δυνατότητα του να αποθηκεύει τα δεδομένα στην κύρια μνήμη κάθε υπολογιστή κόμβου στον οποίο δουλεύει, μειώνοντας έτσι δραματικά τον χρόνο αναζήτησης και εγγραφής στο δίσκο. Η διαδικασία αυτή ονομάζεται caching και είναι αυτή που χαρίζει ένα μεγάλο προβάδισμα χρόνου σε σχέση με το μοντέλο του MapReduce.

Η πλατφόρμα του Apache Spark είναι εξ' ολοκλήρου γραμμένη στη γλώσσα προγραμματισμού Scala. Παρέχει ένα υψηλού επιπέδου API στον χρήστη προσφέροντάς του τη δυνατότητα να διαβάζει, να μετασχηματίζει και να εκπαιδεύει στατιστικά μοντέλα με μεγάλη ευκολία, όχι μόνο σε Scala αλλά και σε άλλες γλώσσες προγραμματισμού όπως είναι η Python, η Java και η R. Ταυτόχρονα υποστηρίζει και μία μεγάλη γκάμα εργαλείων, επίσης υψηλού επιπέδου, όπως για παράδειγμα το Spark SQL, η MLlib ή ML, το GraphX και τέλος το Spark Streaming, τα οποία θα δούμε αναλυτικότερα στην συνέχεια.


**Εικόνα 5 Apache Spark**

## 3.2 Η αρχιτεκτονική του Spark

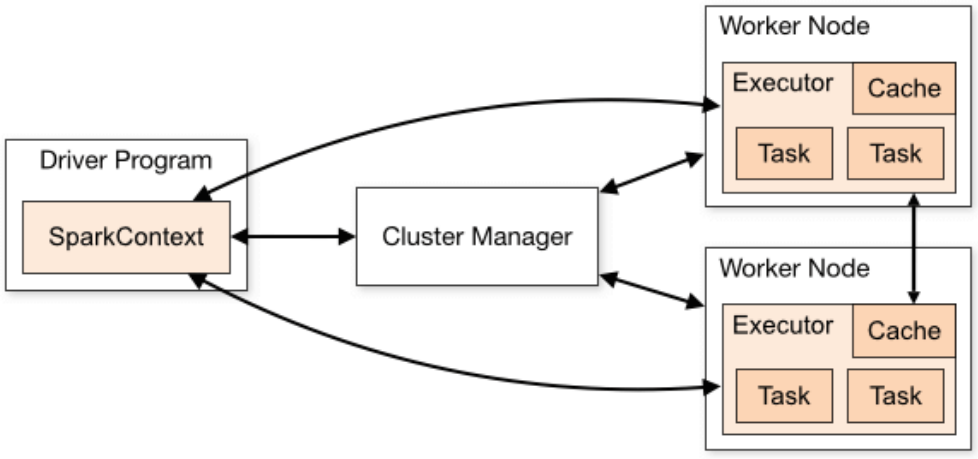
Στα ενδότερα, το Apache Spark ακολουθεί την αρχιτεκτονική του master/slave, όπου στη συγκεκριμένη περίπτωση είναι master και worker. Κατά την εκκίνηση λοιπόν, της πλατφόρμας δημιουργείται μία διεργασία master από τον κόμβο όπου δόθηκε η εντολή δημιουργίας, η δημιουργία του Spark Context δηλαδή, η οποία είναι υπεύθυνη για τον διαμοιρασμό των εργασιών στις διεργασίες workers που εκτελούνται από τους κόμβους slaves.

Παράλληλα με αυτό, η πλατφόρμα του Apache Spark χρειάζεται και έναν manager για τη διαχείριση και τον διαμοιρασμό των πόρων στους workers. Σαν resource manager το Spark παρέχει τη δυνατότητα επιλογής ανάμεσα σε Standalone, Hadoop Yarn και Spark in MapReduce (SIMR) τα οποία εξηγούνται παρακάτω.

Standalone: Η Standalone σημαίνει πως η πλατφόρμα του Apache Spark εγκαθίσταται πάνω από το HDFS (Hadoop Distributed File System) και ο χώρος διατίθεται ρητά για HDFS. Πράγμα που φέρνει τον Spark και το MapReduce να τρέχουν δίπλα-δίπλα για να καλύψουν όλες τις εργασίες spark στον κόμβο. Η Standalone είναι και η υλοποίηση που εφαρμόζεται στη συγκεκριμένη εργασία.

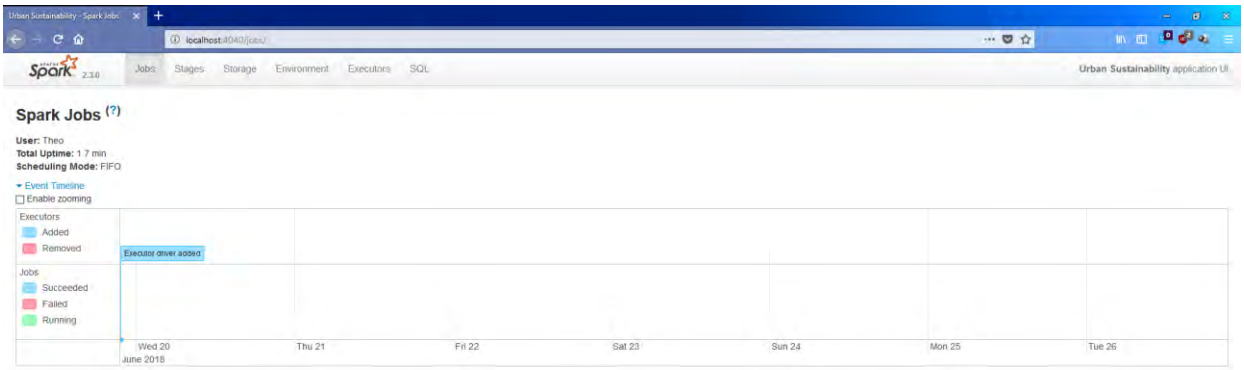
Hadoop Yarn: Hadoop Yarn σημαίνει, πως ο Spark τρέχει στον Yarn Resource Manager χωρίς καμία προεγκατάσταση ή χωρίς να απαιτείται κάποια πρόσβαση διαχειριστή. Βοηθά στην ενσωμάτωση του Spark στον Hadoop, επιτρέποντας έτσι και σε άλλα στοιχεία να τρέχουν πάνω από τη στοίβα.

Spark στο MapReduce (SIMR): Με το SIMR, ο χρήστης μπορεί να ξεκινήσει το Spark και χρησιμοποιεί το κέλυφος του Spark χωρίς καμία πρόσβαση διαχειριστή.




**Εικόνα 6 Εκκίνηση του Spark**

Το Spark παρέχει και τη δυνατότητα παρακολούθησης των πόρων του περιβάλλοντος, των εργασιών καθώς και τα στάδια των εργασιών. Με τη δημιουργία ενός Spark Context εκκινείτε ταυτόχρονα και ένα Web UI τοπικά στον υπολογιστή και είναι διαθέσιμο με έναν περιηγητή στην σελίδα http://localhost:4040. Προκαθορισμένη θύρα είναι η 4040, αν είναι δεσμευμένη συνεχίζει μέχρι να βρει την πρώτη διαθέσιμη (πχ. 4041, 4042, 4043, … ). Το διαδικτυακό περιβάλλον παρακολούθησης είναι προκαθορισμένο από το σύστημα να είναι επισκέψιμο μόνο κατά τη διάρκεια ζωής της εφαρμογής, κάτι το οποίο αν θέλουμε αλλάζει.




**Εικόνα 7 Spark Monitoring**

## 3.3 Τα υποσυστήματα του Spark

Ήδη παραπάνω κάναμε αναφορά σε κάποια από τα εργαλεία υψηλού επιπέδου που μας παρέχει η πλατφόρμα του Apache Spark. Τα εργαλεία αυτά λειτουργούν πάνω από τον πυρήνα της πλατφόρμας και μπορούν να χρησιμοποιηθούν ως βιβλιοθήκες στη γλώσσα προγραμματισμού που θα επιλέξουμε για την υλοποίηση της εφαρμογής. Στη συγκεκριμένη περίπτωση που χρησιμοποιούμε σαν γλώσσα προγραμματισμού την Python, η βιβλιοθήκη που εκπροσωπεί τον Apache Spark ονομάζεται Pyspark. Τα εργαλεία αυτά είναι:

### 3.3.1 Spark Sql

Το εργαλείο του Spark Sql μας παρέχει την υποστήριξη ερωτημάτων τύπου Sql και είναι υπεύθυνο για τον όρο DataFrames. Ο όρος αυτός αποτελεί μία δομή δεδομένων, η οποία αφορά μόνο δομημένα και ημιδομημένα δεδομένα. Τα δεδομένα στην παρούσα εργασία διαβάστηκαν ως Dataframes.

### 3.3.2 MLlib ή Ml

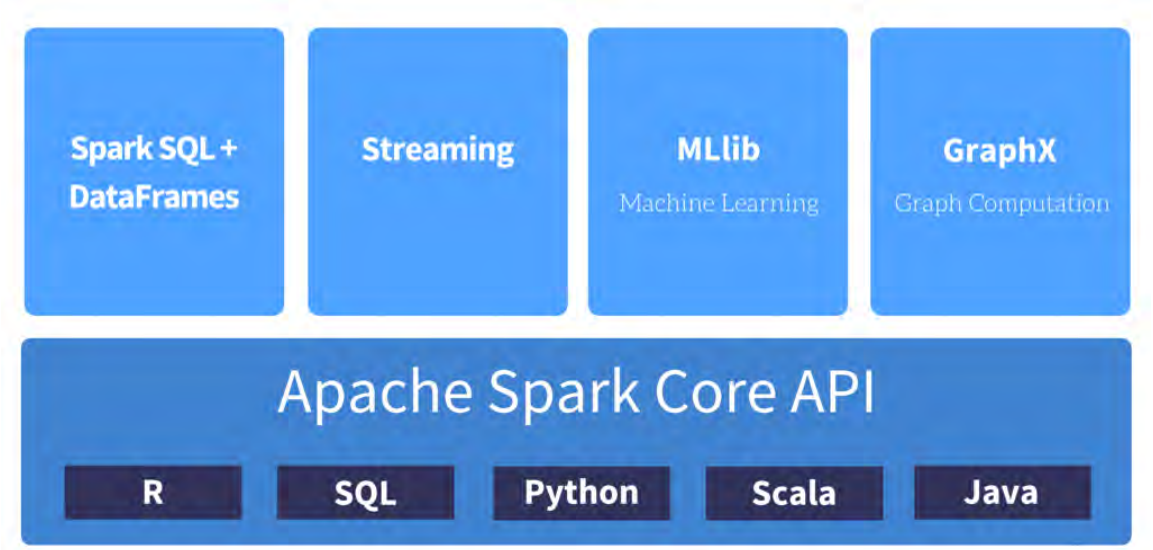
Η βιβλιοθήκη MLlib ή Ml περιέχει μία συλλογή αλγορίθμων μηχανικής μάθησης, ειδικά σχεδιασμένη ώστε να τρέχουν κατανεμημένα και αρκετά γρήγορα. Η διαφορά μεταξύ MLlib και Ml είναι ότι η πρώτη δέχεται σαν είσοδο δεδομένα υπό τη μορφή των RDD (Resilient Distributed Datasets), ενώ η δεύτερη υπό τη μορφή Dataframes.

### 3.3.3 GraphX

Το GraphX αποτελεί ένα εργαλείο για την επεξεργασία γράφων. Αποτελεί μία πολύ καλή υλοποίηση για την εκτέλεση διάφορων υπολογισμών σε γράφους καθώς εκμεταλλεύεται την ταχύτητα και τον παραλληλισμό που του παρέχει η πλατφόρμα του Apache Spark.

### 3.3.4 Spark Streaming

Το υποσύστημα του Spark, Spark Streaming, είναι υπεύθυνο για την ανάλυση και επεξεργασία δεδομένων σε πραγματικό χρόνο. Χρησιμοποιεί στο έπακρο την ταχύτητα πυρήνα του Apache Spark, για την πιο γρήγορη διαχείριση των δεδομένων σε πραγματικό χρόνο.




**Εικόνα 8 Ο πυρήνας του Spark**

# 4.ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Η μηχανική μάθηση αποτελεί τον πιο δημοφιλή τομέα της πληροφορική την εποχή της πληροφορίας. Η μηχανική μάθηση είναι ένας τομέας της τεχνητής νοημοσύνης, η οποία δίνει τη δυνατότητα σε ένα υπολογιστικό σύστημα να “μαθαίνει” μέσα από κάποιους αλγορίθμους και μεθόδους. Η μηχανική μάθηση στηρίζεται καθαρά στην εμπειρία που αποκτά το σύστημα κατά την διάρκεια που ένας αλγόριθμός μηχανικής μάθησης εκπαιδεύεται και προσπαθεί συνεχώς να βελτιώνει τα αποτελέσματα που εξάγει. Μετά την έκρηξη των Big Data, ο όρος μηχανική μάθηση έγινε στενά συνδεδεμένος με τη στατιστική ανάλυση, ένας κλάδος που επικεντρώνεται στην πρόβλεψη, καθώς οι αλγόριθμοι της μηχανικής μάθησης τροφοδοτούνται με τεράστια ποσά δεδομένων. Αυτό έχει ως αποτέλεσμα να μεγαλώνει την ακρίβεια των δεδομένων.

Ένα παράδειγμα μηχανικής μάθησης είναι το φίλτρο ανίχνευσης εάν ένα email είναι spam ή όχι. Το φίλτρο αποτελεί ένα μοντέλο μηχανικής μάθησης, το οποίο έχει εκπαιδευτεί μέσα από μία σειρά email που κάποια χαρακτηρίστηκαν ως spam και κάποια όχι, με σκοπό να αποκτήσει τη γνώση και την εμπειρία ώστε να τα ξεχωρίζει, ενημερώνοντας τον χρήστη για να μην τα ανοίγει. Επομένως, η μηχανική μάθηση επινοεί πολύπλοκα μαθηματικά μοντέλα που οδηγούν σε πρόβλεψη με αποτέλεσμα αυτά να μας οδηγούν στην καλύτερη λήψη αποφάσεων.

Δεν υπάρχει ένας αυστηρός ορισμός για τη μηχανική μάθηση. Ένας ορισμός, όμως, που θα μπορούσαμε να δώσουμε είναι πως μηχανική μάθηση είναι η δημιουργία και συνεχής βελτίωση μαθηματικών μοντέλων και προτύπων μέσα από ένα σύνολο δεδομένων, με σκοπό να εξάγει πολύτιμα αποτελέσματα.

## 4.1 Τα είδη της μηχανικής μάθησης

Η μηχανική μάθηση χωρίζεται σε τρεις μεγάλες κατηγορίες, ανάλογα με τη φύση των δεδομένων.

### 4.1.1 Επιβλεπόμενη Μάθηση

Στην επιβλεπόμενη μάθηση (Supervised Learning), ή αλλιώς μάθηση με παραδείγματα, το υπολογιστικό σύστημα καλείται να μάθει και να δημιουργήσει μία συνάρτηση στην οποία αντιστοιχεί το σύνολο των δεδομένων εισόδου στο υπάρχον γνωστό σύνολο εξόδου. Στη συγκεκριμένη μάθηση έχουμε τα features που αποτελούν την είσοδο και τα labels, τις ετικέτες των χαρακτηριστικών, που αποτελούν τη γνωστή έξοδο των features. Σκοπός του μοντέλου που δημιουργεί είναι μέσα από την εμπειρία του να μπορεί προβλέψει την τιμή εξόδου, το label, σε

μία νέα πιθανή είσοδο. Οι αλγόριθμοι της επιβλεπόμενης μάθησης χωρίζονται σε 2 υποενότητες οι οποίες είναι:

Αλγόριθμοι ταξινόμησης (Classification algorithms), όπου αφορούν σύνολα των οποίων οι τιμές είναι αυστηρά διακριτές τιμές, όπως για παράδειγμα είναι η ηλικία. Η ταξινόμηση με τη σειρά της χωρίζεται και αυτή σε δύο υποενότητες, στην δυαδική ταξινόμηση, όπου οι κλάσεις πρόβλεψης ή πιο απλά τα label παίρνουν μόνο δύο τιμές και στην πολλαπλή ταξινόμηση, όπου οι κλάσεις των δεδομένων είναι περισσότερες από δύο.

Αλγόριθμοι παλινδρόμησης (Regression algorithms) που αφορούν σύνολα δεδομένων που έχουν συνεχείς τιμές και όχι διακριτές, όπως για παράδειγμα η θερμοκρασία. Οι αλγόριθμοι αυτοί δεν χωρίζονται περαιτέρω όπως της ταξινόμησης.

### 4.1.2 Μη-Εποπτευόμενη Μάθηση

Στην κατηγορία της μη-εποπτευόμενης μάθησης (UnSupervised Learning) το σύστημα καλείται μόνο του να ανακαλύψει συσχετίσεις ή ομάδες μέσα από το σύνολο των δεδομένων.

Το σύστημα αυτή τη φορά δεν γνωρίζει τις επιθυμητές εξόδους των δεδομένων. Οι αλγόριθμοι της μη εποπτευόμενης μάθησης χωρίζουν τα δεδομένα με βάση τα χαρακτηριστικά τους, τα features, σε ομάδες.

### 4.1.3 Ενισχυτική μάθηση

Η ενισχυτική μάθηση (Reinforcement Learning), αποτελεί την τελευταία και πιο σύνθετη κατηγορία μηχανικής μάθησης. Οι αλγόριθμοι της κατηγορίας αυτής καλούνται να μάθουν μία στρατηγική μέσα από μία σειρά ενεργειών καθώς αυτοί αλληλοεπιδρούν με το περιβάλλον. Χρησιμοποιείται κυρίως σε προβλήματα σχεδιασμού, όπως είναι για παράδειγμα ο έλεγχος κίνησης ρομπότ.

## 4.2 Μηχανική Μάθηση με Spark

Η μηχανική μάθηση, ένας τόσο δημοφιλής τομέας της πληροφορικής την εποχή της πληροφορίας, δεν θα μπορούσε να απουσιάζει από την πλατφόρμα του Apache Spark.

Το πακέτο μηχανικής μάθησης του Apache Spark εμπεριέχει μία μεγάλη γκάμα αλγορίθμων που καλύπτουν τόσο την εποπτευόμενη μάθηση (Supervised Learning) όσο και τη μη-Εποπτευόμενη μάθηση (Unsupervised Learning) [9].

### 4.2.1 Εποπτευόμενη Μάθηση

Στο πακέτο της Εποπτευόμενης μάθησης του Spark, οι αλγόριθμοι χωρίζονται σε δύο μεγάλες υποενότητες, τους αλγορίθμους ταξινόμησης (classification algorithms) και τους αλγορίθμους παλινδρόμησης (regression algorithms) με οκτώ και εφτά μοντέλα αλγορίθμων αντίστοιχα στην έκδοση του Spark 2.3.1 .

#### 4.2.1.1 Ταξινόμηση Classifier

Στην κατηγορία της ταξινόμησης, οι αλγόριθμοι προσπαθούν να προβλέψουν αποκλειστικά και μόνο διακριτές τιμές. Οι ετικέτες των δεδομένων που αποτελούν και την επιθυμητή έξοδο, έχουν πεπερασμένο πλήθος τάξεων με το οποίο θα τις ταξινομήσει ο αλγόριθμος. Επίσης, η ταξινόμηση χωρίζεται και αυτή με τη σειρά της σε δύο μικρότερες ενότητες, τη δυαδική ταξινόμηση, όπου ανήκουν σύνολα δεδομένων των οποίων οι τάξεις είναι μόνο δύο και την πολλαπλή ταξινόμηση, όπου οι τάξεις των δεδομένων είναι περισσότερες από δύο. Οι περισσότεροι αλγόριθμοι ταξινόμησης δέχονται και τα δύο είδη ταξινόμησης.

##### 4.2.1.1.1 LinearSVC Classifier

Όσο αφορά στον Linear Support Vector Machine, πρόκειται για ένα μοντέλο που υπολογίζει την απώλεια άρθρωσης των παρατηρήσεων με τη χρήση κάποιου αλγορίθμου βελτιστοποίησης. Το συγκεκριμένο μοντέλο του LinearSVC στον Spark υποστηρίζει μόνο δυαδικές τιμές.

##### 4.2.1.1.2 LogisticRegression Classifier

Το μοντέλο LogisticRegression χρησιμοποιεί μία λογική συνάρτηση για τον υπολογισμό της πιθανότητας για κάθε παρατήρηση ώστε να την εντάξει σε μία συγκεκριμένη ομάδα.

##### 4.2.1.1.3 DecisionTree Classifier

Ο αλγόριθμος Decision Tree αποτελεί τον πιο κλασικό αλγόριθμο ταξινόμησης. Ο αλγόριθμος δημιουργεί ένα δέντρο αποφάσεων, προβλέποντας έτσι την κλάση των δεδομένων προς παρατήρηση. Είναι αρκετά εύκολος στη χρήση τόσο σε δυαδικές όσο και σε πολλαπλές κλάσεις εξόδου και δεν απαιτεί τα δεδομένα να είναι ταξινομημένα. Η υλοποίησή του στην πλατφόρμα του Apache Spark κατακερματίζει τα δεδομένα βάσει των γραμμών του συνόλου, επιτρέποντας του να εκπαιδεύει μοντέλα με εκατομμύρια ή και δισεκατομμύρια γραμμές.

#### 4.2.1.1.4 RandomForest Classifier

Ο αλγόριθμος Random Forest πρόκειται για μία επέκταση του αλγορίθμου DecisionTree, καθώς δημιουργεί ένα δάσος από πολλά δέντρα αποφάσεων. Συνδυάζει έτσι τις πληροφορίες από όλα τα δέντρα, εκμηδενίζοντας την πιθανότητα να εξάγει λάθος αποτελέσματα λόγω θορύβων από τα δεδομένα. Ο ταξινομητής RandomForest υποστηρίζει και τις δύο υποενότητες ταξινόμησης. Επίσης, ο αλγόριθμος έχει τα ίδια ορίσματα με τον DecisionTreeClassifier με επιπλέον την παράμετρο που καθορίζει το πλήθος των δέντρων που θα αποτελούν το δάσος.

#### 4.2.1.1.5 GBT Classifier

Ο αλγόριθμος Gradient Boosted Tree αποτελεί και αυτός με τη σειρά του μία επέκταση του αλγορίθμου DecisionTree, καθώς συνδυάζει πολλά αδύναμα μοντέλα δέντρων αποφάσεων με σκοπό να σχηματίσει ένα πιο ισχυρό μοντέλο προβλέψεων. Η υλοποίηση του αλγορίθμου από την πλατφόρμα του Spark υποστηρίζει μόνο δυαδική ταξινόμηση.

#### 4.2.1.1.6 NaiveBayes Classifier

Ο αλγόριθμος του NaiveBayes είναι βασισμένος στο θεώρημα πιθανοτήτων Bayes. Χρησιμοποιεί δηλαδή, το μοντέλο της υποθετικής πιθανότητας για την ταξινόμηση των παρατηρήσεων στις κλάσεις. Ο NaiveBayes στην πλατφόρμα του Spark ανήκει τόσο στη δυαδική ταξινόμηση όσο και στην πολλαπλή ταξινόμηση.

#### 4.2.1.1.7 Multilayer Perceptron Classifier

Ο συγκεκριμένος αλγόριθμος είναι υπεύθυνος για τη δημιουργία νευρωνικών δικτύων με απλή τροφοδότηση, που σημαίνει πως δεν υπάρχουν συνδέσεις μεταξύ νευρώνων ενός επιπέδου και νευρώνων προηγούμενου επιπέδου. Η υλοποίησή του στον Spark δεν έχει κάποιον απαραίτητο περιορισμό. Βασική του προϋπόθεση είναι να έχει τουλάχιστον τρία στενά συνδεδεμένα επίπεδα. Το πρώτο επίπεδο, ή αλλιώς στρώμα εισόδου, περιέχει τις εισόδους του νευρωνικού δικτύου. Οι νευρώνες του πρώτου επιπέδου πρέπει να είναι ίσοι σε αριθμό με το πλήθος των χαρακτηριστικών (features) του συνόλου δεδομένων. Το επόμενο επίπεδο αφορά τα κρυφά στρώματα του νευρωνικού δικτύου, πρέπει να περιέχει τουλάχιστον ένα κρυφό στρώμα και τέλος το επίπεδο εξόδου στο οποίο ο αριθμός των νευρώνων πρέπει να είναι ίσος με τις κλάσεις παρατηρήσεων που θέλουμε να δημιουργήσουμε. Στα δύο πρώτα επίπεδα η συνάρτηση ενεργοποίησης είναι σιγμοειδής, ενώ το στρώμα εξόδου έχει την softmax σαν συνάρτηση ενεργοποίησης.

#### 4.2.1.1.8 OneVsRest Classifier

Ο αλγόριθμος OneVsRest δέχεται δεδομένα με τάξεις περισσότερες από δύο και τις διαιρεί κάθε φορά σε δυαδικές, εκπαιδεύοντας έτσι πολλαπλά μοντέλα LogisticRegression. Το μοντέλο με το υψηλότερο σκορ ακρίβειας είναι και αυτό που κερδίζει και αποτελεί το μοντέλο πρόβλεψης του OneVsRest.

### 4.2.1.2 Παλινδρόμηση

Εν αντιθέσει με την κατηγορία της ταξινόμησης, οι αλγόριθμοι της παλινδρόμησης δημιουργούν μοντέλα που προσπαθούν να προβλέψουν συνεχείς τιμές εξόδου. Εδώ δεν έχουμε μικρότερες κατηγορίες παλινδρόμησης καθώς οι τάξεις των δεδομένων είναι πάντα περισσότερες από δύο λόγω της συνέχειας των τιμών.

#### 4.2.1.2.1 AFTSurvival Regressor

Ο AFT, ή αλλιώς Accelerated Failure Time, είναι ένα παραμετρικό μοντέλο παλινδρόμησης που υποθέτει ότι ένα από τα χαρακτηριστικά του συνόλου επιταχύνει ή επιβραδύνει αρκετά γρήγορα.

#### 4.2.1.2.2 DecisionTree Regressor

Είναι ακριβώς ίδιος αλγόριθμος με τον DecisionTree που αναφέρθηκε στην ενότητα της ταξινόμησης. Πρόκειται για την ίδια υλοποίηση του δέντρου αποφάσεων με τη μόνη διαφορά ότι εδώ, οι τιμές εξόδου του συνόλου δεδομένων είναι συνεχόμενες και όχι διακριτές.

#### 4.2.1.2.3 GBT Regressor

Όμοια με τον αντίστοιχο αλγόριθμο ταξινόμησης μόνο που αλλάζει ο τύπος των ετικετών από τα δεδομένα.

#### 4.2.1.2.4 LinearRegression Regressor

Το πιο απλό μοντέλο της παλινδρόμησης. Προϋποθέτει την ύπαρξη μιας γραμμικής σχέσης μεταξύ των χαρακτηριστικών, ενώ οι ετικέτες του συνόλου πρέπει να είναι συνεχές τιμές.

#### 4.2.1.2.5 GeneralizedLinear Regressor

Γνωστός και ως GLM, αποτελεί μία γενίκευση της γραμμικής παλινδρόμησης (Linear Regression) καθώς επιτρέπει στις μεταβλητές να έχουν και άλλα μοντέλα κατανομής λάθους πέρα από την κανονική κατανομή που ακολουθεί ο Linear Regression.

#### 4.2.1.2.6 Isotonic Regressor

Αφορά έναν τύπο παλινδρόμησης ο οποίος εφαρμόζει μία ελεύθερη και αύξουσα γραμμή στα δεδομένα προς παρατήρηση. Είναι χρήσιμος σαν αλγόριθμος όταν οι παρατηρήσεις των δεδομένων είναι ταξινομημένες.

#### 4.2.1.2.7 RandomForest Regressor

Ο αλγόριθμος Random Forest αποτελεί και αυτός την παρόμοια υλοποίηση με αυτόν της ταξινόμησης. Η μόνη διαφορά εδώ είναι πως το σύνολο των επιθυμητών εξόδων των δεδομένων έχει συνεχείς τιμές, ενώ παράλληλα τα δέντρα που δημιουργεί ο αλγόριθμος ανήκουν στην κατηγορία Decision Tree Regression.

### 4.2.2 Μη-Εποπτευόμενη μάθηση

Στην Μη-Εποπτευόμενη μάθηση, οι αλγόριθμοι που υποστηρίζει ο Spark ανήκουν στην κατηγορία της συσταδοποίησης (clustering). Η κατηγορία αυτή περιέχει τέσσερις αλγορίθμους οι οποίοι είναι:

#### 4.2.2.1 KMeans

Πρόκειται για τον πιο γνωστό αλγόριθμο της Μη-Εποπτευόμενης μάθησης και συγκεκριμένα της συσταδοποίησης, οποίος χωρίζει τα δεδομένα σε k ομάδες, αναζητώντας συνεχώς την εύρεση των σωστών κέντρων ομάδων που ελαχιστοποιούν την ευκλείδεια τετραγωνική απόσταση μεταξύ των παρατηρήσεων.

#### 4.2.2.2 BisectingKMeans

Πρόκειται ουσιαστικά για ένα συνδυασμό του αλγορίθμου ομαδοποίησης του K-Means και της ιεραρχικής συσσώρευσης. Ο αλγόριθμος BisectingKMeans ξεκινάει με όλες τις παρατηρήσεις σε μία μόνο ομάδα και τις διαιρεί διαδοχικά σε k ομάδες.

#### 4.2.2.3 GaussianMixture

Ο συγκεκριμένος αλγόριθμος χρησιμοποιεί k φορές, Gaussian κατανομές με άγνωστες παραμέτρους για την ανάλυση του συνόλου δεδομένων.

#### 4.2.2.4 LDA

Ο αλγόριθμος LDA χρησιμοποιείται για τη μοντελοποίηση εφαρμογών που κάνουν χρήση της φυσικής γλώσσας.

# 5.ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Στο κεφάλαιο αυτό θα αναλύσουμε τα δεδομένα της εργασίας, τα μοντέλα και τα πειράματα που διεξήγαμε.

## 5.1 Αισθητήρες

Τα δεδομένα προς ανάλυση για την παρούσα εργασία προέρχονται από αισθητήρες κυκλοφορίας οχημάτων, που βρίσκονται τοποθετημένοι στη Δανία και πιο συγκεκριμένα στην πόλη Ώρχους. Η λήψη των δεδομένων γίνεται ανάμεσα σε δύο αισθητήρες που ορίζουν μία διαδρομή προς παρατήρηση στην πόλη του Ώρχους. Τα δεδομένα που καταγράφουν αφορούν την εισροή και εκροή των οχημάτων ανάμεσα σε αυτούς τους δύο αισθητήρες. Οι ξεχωριστές αυτές διαδρομές της δειγματοληψίας εκτείνονται σε ολόκληρη την πόλη με σκοπό την ορθότερη και ακριβέστερη λήψη δεδομένων.


**Εικόνα 9 Θέσεις Αισθητήρων**

Κάθε φορά που ένα όχημα, στο οποίο ο αισθητήρας της συσκευής bluetooth του κινητού τηλεφώνου είναι ενεργοποιημένος, εκτελεί τη διαδρομή που ορίζουν τα δύο σημεία των αισθητήρων, τότε συλλέγονται και οι απαραίτητες πληροφορίες. Η διαδικασία της δειγματοληψίας γίνεται σταθερά κάθε 5 λεπτά.

Πέρα των αισθητήρων κυκλοφορίας, έχει γίνει μία αυθαίρετη προσομοίωση αισθητήρων που συλλέγουν πληροφορίες σχετικά με τους ατμοσφαιρικούς ρύπους. Η προσομοίωση έχει γίνει στην ακριβή θέση όπου βρίσκονται και οι αντίστοιχοι αισθητήρες κίνησης.

## 5.2 Δεδομένα

Τα δεδομένα προέρχονται από ένα πλήθος αισθητήρων κίνησης που ορίζουν τις διαδρομές και των αντίστοιχων εικονικών αισθητήρων για τους ατμοσφαιρικούς ρύπους [10], όπως αναφέρθηκαν παραπάνω. Τα δεδομένα και των δύο κατηγοριών αφορούν το διάστημα από την ‘1 Αυγούστου 2014’ έως και την ‘31 Σεπτεμβρίου 2014’.

Ο τεράστιος όγκος των δεδομένων, η ταχύτητα παραγωγής των δεδομένων, η δομημένη μορφή των δεδομένων καθώς και η αξία που περιέχουν τα δεδομένα είναι οι λόγοι που εντάσσονται στην κατηγορία Big Data. Δεν μπορούν να επεξεργαστούν από ένα υπολογιστικό σύστημα με παραδοσιακές μεθόδους καθώς χρειάζονται πολλούς πόρους και υπολογιστική ισχύ για να αναλύονται και να επεξεργάζονται πολύ γρήγορα. Αυτός είναι ο λόγος της επιλογής του Apache Spark για την επεξεργασία των δεδομένων.

### 5.2.1 Δεδομένα Κυκλοφορίας

Στην πρώτη κατηγορία ανήκουν τα δεδομένα που αφορούν την κυκλοφορία των οχημάτων του Ώρχους. Τα δεδομένα αυτά περιέχουν πληροφορίες σχετικά με το πόσα οχήματα διέσχισαν την διαδρομή, τη μέση ταχύτητα των οχημάτων και τον μέσο χρόνο που έκαναν να διασχίσουν την διαδρομή σε δευτερόλεπτα. Πιο αναλυτικά το σχήμα των δεδομένων κίνησης είναι το εξής:

```
root
 |-- status: string (nullable = true)
 |-- avgMT: integer (nullable = true)
 |-- avgSpd: integer (nullable = true)
 |-- time: timestamp (nullable = true)
 |-- vCount: integer (nullable = true)
 |-- _id: integer (nullable = true)
 |-- report_id: integer (nullable = true)
```

**Εικόνα 10 Δομή Δεδομένων Κίνησης**

Η δομή των δεδομένων και ο τύπος κάθε στήλης, όπως αυτά διαβάστηκαν από την πλατφόρμα του Spark απεικονίζονται στην εικόνα 10.

Το στοιχείο “status” περιγράφει την κατάσταση του αισθητήρα, με την τιμή “OK” να σημειώνεται αν είναι ενεργός κι αν έγινε η δειγματοληψία εκείνη τη στιγμή. Σε όλες τις εγγραφές έχει την τιμή “OK”. Το “avgMT”, όπως νωρίτερα αναφέρθηκε, περιγράφει τον μέσο χρόνο, σε δευτερόλεπτα, που χρειάστηκαν τα διερχόμενα οχήματα για να εκτελέσουν τη διαδρομή. Το “avgSpd” καθορίζει τη μέση ταχύτητα των οχημάτων. Το “vCount” αναφέρεται στο πλήθος των οχημάτων που εκτέλεσαν τη διαδρομή. Η χρονική στιγμή που έγινε η εγγραφή περιγράφεται από την οντότητα “time”, η οποία περιέχει την ημερομηνία, την ώρα, τα λεπτά και τα δευτερόλεπτα. Το “_id” είναι το μοναδικό χαρακτηριστικό κάθε εγγραφής και τέλος με το “report_id” περιγράφεται ο κωδικός ονομασίας κάθε διαδρομής.

| status | avgMT | avgSpd | extID | medianMT | time | vCount | _id | report_id |
|---|---|---|---|---|---|---|---|---|
| OK | 64 | 18 | 989 | 64 | 2014-08-01 08:00:00 | 1 | 20747044 | 190073 |
| OK | 41 | 29 | 989 | 41 | 2014-08-01 08:05:00 | 2 | 20747477 | 190073 |
| OK | 60 | 19 | 989 | 60 | 2014-08-01 08:10:00 | 3 | 20747866 | 190073 |
| OK | 91 | 13 | 989 | 91 | 2014-08-01 08:15:00 | 4 | 20748315 | 190073 |
| OK | 117 | 10 | 989 | 117 | 2014-08-01 08:20:00 | 2 | 20748764 | 190073 |
| OK | 71 | 16 | 989 | 71 | 2014-08-01 08:25:00 | 1 | 20749213 | 190073 |
| OK | 71 | 16 | 989 | 71 | 2014-08-01 08:30:00 | 1 | 20749662 | 190073 |
| OK | 71 | 16 | 989 | 71 | 2014-08-01 08:35:00 | 0 | 20750111 | 190073 |
| OK | 76 | 15 | 989 | 76 | 2014-08-01 08:40:00 | 1 | 20750560 | 190073 |
| OK | 76 | 15 | 989 | 76 | 2014-08-01 08:45:00 | 1 | 20751009 | 190073 |
| OK | 128 | 9 | 989 | 128 | 2014-08-01 08:50:00 | 1 | 20751458 | 190073 |
| OK | 128 | 9 | 989 | 128 | 2014-08-01 08:55:00 | 1 | 20751907 | 190073 |
| OK | 14 | 85 | 989 | 14 | 2014-08-01 09:00:00 | 1 | 20752356 | 190073 |
| OK | 14 | 85 | 989 | 14 | 2014-08-01 09:05:00 | 0 | 20752805 | 190073 |
| OK | 14 | 85 | 989 | 14 | 2014-08-01 09:10:00 | 0 | 20753254 | 190073 |
| OK | 14 | 85 | 989 | 14 | 2014-08-01 09:20:00 | 0 | 20754152 | 190073 |
| OK | 14 | 85 | 989 | 14 | 2014-08-01 09:25:00 | 0 | 20754601 | 190073 |
| OK | 14 | 85 | 989 | 14 | 2014-08-01 09:30:00 | 0 | 20755050 | 190073 |
| OK | 74 | 16 | 989 | 74 | 2014-08-01 09:35:00 | 1 | 20755499 | 190073 |
| OK | 74 | 16 | 989 | 74 | 2014-08-01 09:40:00 | 1 | 20755948 | 190073 |

**Εικόνα 11 Δείγμα δεδομένων κίνησης**

### 5.2.2 Δεδομένα Ατμοσφαιρικής ρύπανσης

Στη δεύτερη κατηγορία ανήκουν τα δεδομένα της ατμοσφαιρικής ρύπανσης που δημιουργήθηκαν για να συμπληρώσουν τα δεδομένα κυκλοφορίας των οχημάτων που αναλύθηκαν νωρίτερα. Το σύνολο των ατμοσφαιρικών δεδομένων ακολουθεί το μοντέλο Ποιότητας του Αέρα (AQI) [11]. Ο τρόπος με τον οποίο δημιουργήθηκαν τα δεδομένα είναι αρκετά απλός και δεν έχουν κάποια μαθηματική συσχέτιση με τα δεδομένα κίνησης.

«Ο τρόπος με τον οποίο γεννήθηκαν τα δεδομένα ρύπανσης για κάθε αισθητήρα γίνεται ως εξής: για παράδειγμα, το όζον, αρχικά έχει μία τυχαία ακέραια τιμή μεταξύ του 25 και του 100 και κάθε 5 λεπτά ενημερώνεται σύμφωνα με τους εξής κανόνες:

Αν η προηγούμενη τιμή ήταν μικρότερη από 20, τότε η τελευταία τιμή θα είναι η παλιά τιμή “+” ένα τυχαίο ακέραιο αριθμό μεταξύ του 1 και του 10.

Αν η προηγούμενη τιμή ήταν μεγαλύτερη από 210, τότε η τελευταία τιμή θα είναι η παλιά τιμή “-” ένα τυχαίο ακέραιο αριθμό μεταξύ του 1 και του 10.

Αλλιώς η τελευταία τιμή θα είναι η παλιά τιμή “+” ένα τυχαίο ακέραιο αριθμό μεταξύ του -5 και του 5.» [11]

```
root
 |-- ozone: integer (nullable = true)
 |-- part_ma: integer (nullable = true)
 |-- carbon: integer (nullable = true)
 |-- sulfure: integer (nullable = true)
 |-- nitro: integer (nullable = true)
 |-- lon: double (nullable = true)
 |-- lat: double (nullable = true)
 |-- time: string (nullable = true)
```

**Εικόνα 12 Η δομή των ατμοσφαιρικών δεδομένων**

Η δομή των ατμοσφαιρικών δεδομένων και ο τύπος κάθε στήλης που χρησιμοποιήθηκαν για να διαβαστούν τα δεδομένα στον Spark απεικονίζεται στην εικόνα 11.

Οι οντότητες “ozone”, “part_ma”, “carbon”, “sulfure” και “nitro” αναφέρονται στις τιμές του όζοντος, των αιωρούμενων μικροσωματιδίων, του μονοξειδίου του άνθρακα, του διοξειδίου του θείου και του διοξειδίου του αζώτου, αντίστοιχα.

| ozone | part_ma | carbon | sulfure | nitro | lon | lat | time |
|---|---|---|---|---|---|---|---|
| 101 | 94 | 49 | 44 | 87 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 00:05:00 |
| 106 | 97 | 48 | 47 | 86 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 00:10:00 |
| 107 | 95 | 49 | 42 | 85 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 00:15:00 |
| 103 | 90 | 51 | 44 | 87 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 00:20:00 |
| 105 | 94 | 49 | 39 | 82 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 00:25:00 |
| 106 | 92 | 48 | 42 | 77 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 00:30:00 |
| 110 | 87 | 50 | 40 | 81 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 00:35:00 |
| 106 | 91 | 52 | 36 | 82 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 00:40:00 |
| 106 | 88 | 50 | 40 | 85 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 00:45:00 |
| 110 | 90 | 48 | 42 | 82 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 00:50:00 |
| 115 | 91 | 48 | 38 | 82 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 00:55:00 |
| 114 | 88 | 45 | 42 | 81 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 01:00:00 |
| 118 | 91 | 44 | 43 | 81 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 01:05:00 |
| 113 | 89 | 49 | 48 | 81 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 01:10:00 |
| 114 | 90 | 50 | 49 | 79 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 01:15:00 |
| 115 | 86 | 49 | 52 | 76 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 01:20:00 |
| 115 | 87 | 50 | 47 | 73 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 01:25:00 |
| 120 | 84 | 51 | 44 | 72 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 01:30:00 |
| 120 | 83 | 46 | 42 | 76 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 01:35:00 |
| 115 | 88 | 42 | 47 | 73 | 10.104986076057457 | 56.23172069428216 | 2014-08-01 01:40:00 |

**Εικόνα 13 Δείγμα δεδομένων ρύπανσης**

## 5.3 Bokeh

Ακόμα και η πιο πλούσια ανάλυση δεδομένων δεν θα είχε κανένα νόημα αν δεν υπήρχε  οπτική απεικόνιση των αποτελεσμάτων. Ίσως να είναι και η νούμερο ένα ανάγκη, η οπτική επαφή όταν γίνεται λόγος για ανάλυση δεδομένων. Αυτός είναι και ο λόγος χρήσης της βιβλιοθήκης του bokeh.

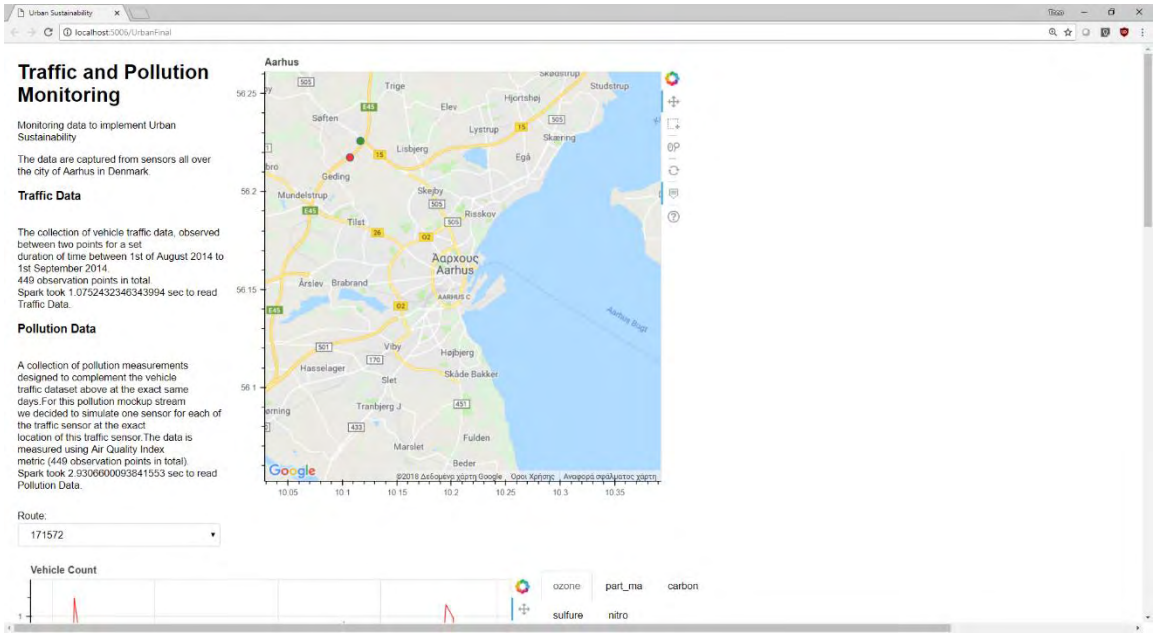
Η βιβλιοθήκη bokeh της python ειδικεύεται στην διαδραστική απεικόνιση δεδομένων σε σύγχρονους περιηγητές. Στόχος της βιβλιοθήκης είναι η κομψότητα, η ευελιξία και η απλότητα που μπορεί να προσφέρει στον χρήστη για να δημιουργήσει διαδραστικά γραφήματα, πίνακες δεδομένων καθώς του δίνει την δυνατότητα να διαλέξει μέσα από μία μεγάλη γκάμα γραφικών εργαλείων. Επεκτείνει βέβαια αυτή την κομψότητα και την ευελιξία στην υψηλή απόδοση που προσφέρει με την αλληλεπίδραση της βιβλιοθήκης με τα μεγάλα σύνολα δεδομένων.

Η βασική ιδέα της βιβλιοθήκης, είναι να δέχεται πολύ απλό κώδικα σε python και αυτή να τον μεταφράζει σε τμήμα κώδικα html και javascript ώστε να μπορεί να προβάλλεται από τους σύγχρονους περιηγητές. Ταυτόχρονα, προσφέρει τη δυνατότητα να δημιουργεί τοπικά έναν server και να σηκώνει σαν ιστοσελίδα τον κώδικα της python, στον τοπικό δίκτυο. Αυτή τη μέθοδο εφαρμόσαμε και εμείς για την δημιουργία της σελίδας παρακολούθησης των δεδομένων.

### 5.3.1 Διαδικτυακή απεικόνιση

Με τη χρήση της βιβλιοθήκης bokeh δημιουργήθηκε μία διεπαφή η οποία συνδέεται άμεσα με τον Apache Spark και μπορεί να διαβάζει και να αναλύει τα δεδομένα της κίνησης καθώς και της ατμοσφαιρικής ρύπανσης. Πιο συγκεκριμένα, δημιουργήθηκε ένας τοπικός server με τη βιβλιοθήκη, ο οποίος εκτελεί την εντολή εκκίνησης του Apache Spark και διαβάζει τα δεδομένα ως Dataframes.

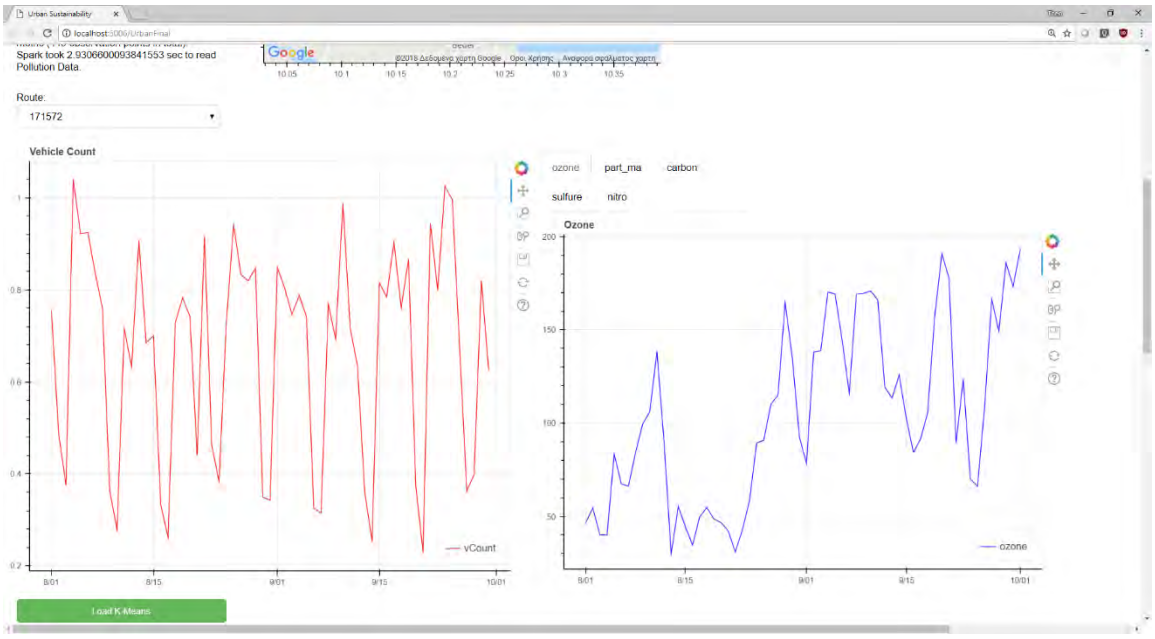
Στη διαδικτυακή εφαρμογή προβάλλονται οι ακριβείς θέσεις των αισθητήρων ανά διαδρομή, την οποία την επιλέγει ο επισκέπτης, καθώς και οι σχετικές πληροφορίες, όπως είναι η γεωγραφική τοποθεσία, το όνομα της οδού και του αριθμού, ο ταχυδρομικός κώδικας και το όνομα της περιοχής όπου είναι τοποθετημένοι.




**Εικόνα 14 Ο χάρτης κάθε διαδρομής**

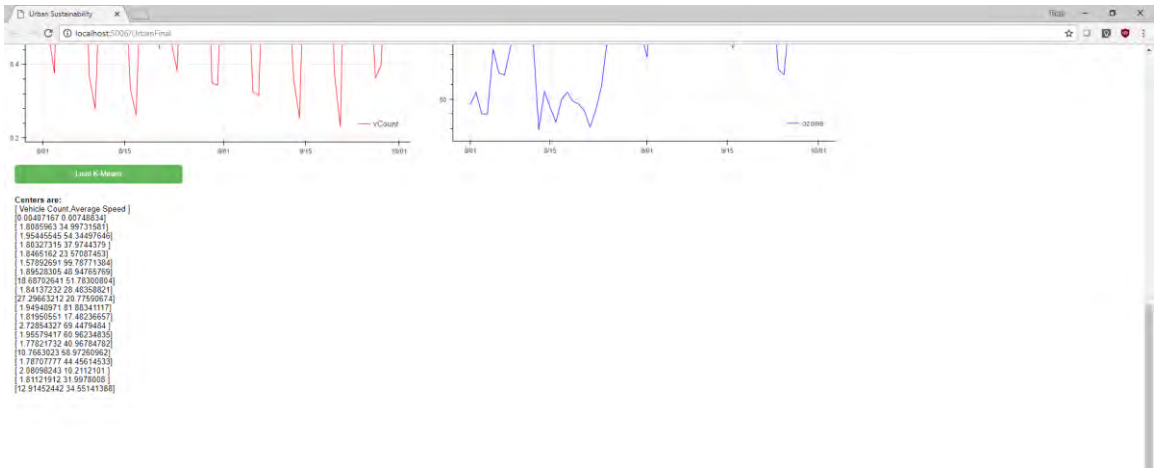
Οι τοποθεσίες μαζί με το γεωγραφικό στίγμα, την οδό, τον αριθμό, την πόλη, τον ταχυδρομικό κώδικα και το αναγνωριστικό κάθε αισθητήρα που αποτελεί μία διαδρομή τυπώνονται πάνω στον χάρτη της πόλης. Με πράσινο χρώμα είναι ο αισθητήρας που βρίσκεται στην αφετηρία της διαδρομής ενώ με κόκκινο ο αισθητήρας που ορίζει και το τέλος της διαδρομής όπως απεικονίζεται στην εικόνα 12.

Παράλληλα με την προβολή των τοποθεσιών των αισθητήρων κάθε διαδρομής γίνεται γραφική απεικόνιση του ιστορικού της κίνησης και των εκπομπές ρύπων.



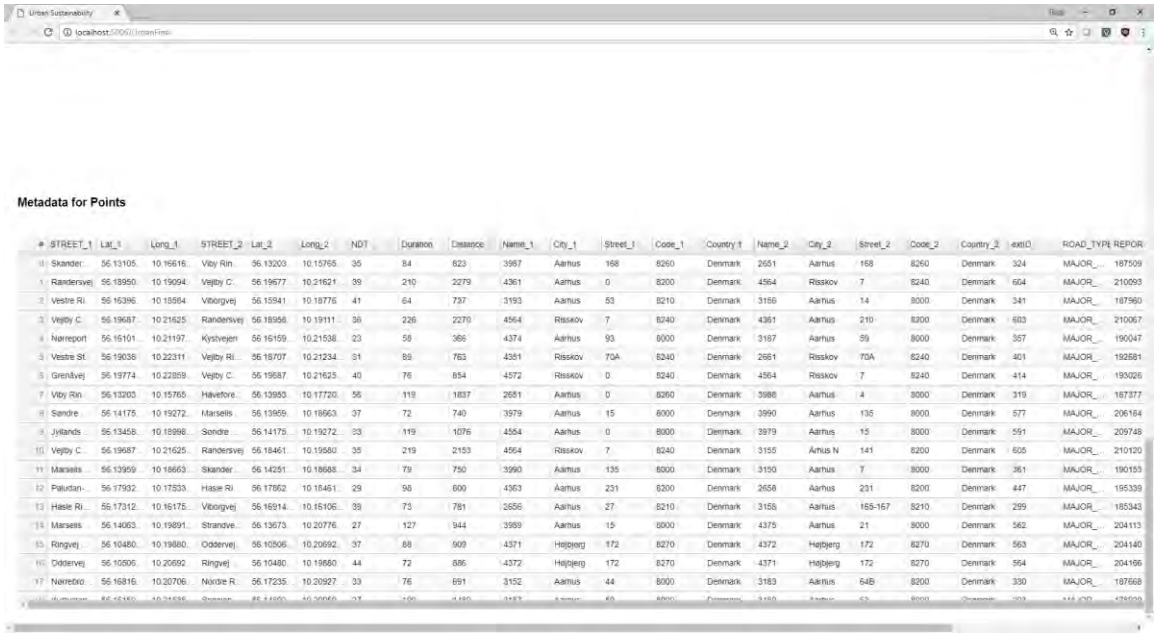

**Εικόνα 15 Γραφικές απεικονίσεις**

Οι γραφικές απεικονίσεις παριστάνουν την εξέλιξη που έχουν στον χρόνο οι τιμές των ρύπων και ο αριθμός των οχημάτων  ανά αισθητήρα, όπως είναι εμφανές από την εικόνα 13.




**Εικόνα 16 Φόρτωση του μοντέλου K-Means**

Στην διαδικτυακή εφαρμογή διαβάζεται το μοντέλο του K-Means και εμφανίζονται οι 20 ομάδες που περιγράφουν την κυκλοφορία, όπως απεικονίζεται στην εικόνα 16.




**Εικόνα 17 Μεταδεδομένα διαδρομών**

Στο τέλος της διαδικτυακής εφαρμογής υπάρχει και ο πίνακας με την αναλυτική περιγραφή των διαδρομών που προέρχεται από τα μεταδεδομένα των συνόλων. Αναγράφονται αναλυτικά όλες οι πληροφορίες για κάθε διαδρομή. Το σημείο εκκίνησης και το σημείο εξόδου, τα ονόματα τα οδών, ο αριθμός, η περιοχή, ο ταχυδρομικός κώδικας κάθε αισθητήρα ανάλογα με το που είναι τοποθετημένος, η μεταξύ τους απόσταση καθώς και το όριο ταχύτητας της διαδρομής σε km/h, όπως απεικονίζεται στην εικόνα 16.

## 5.4 Χρήση της Μηχανικής μάθησης

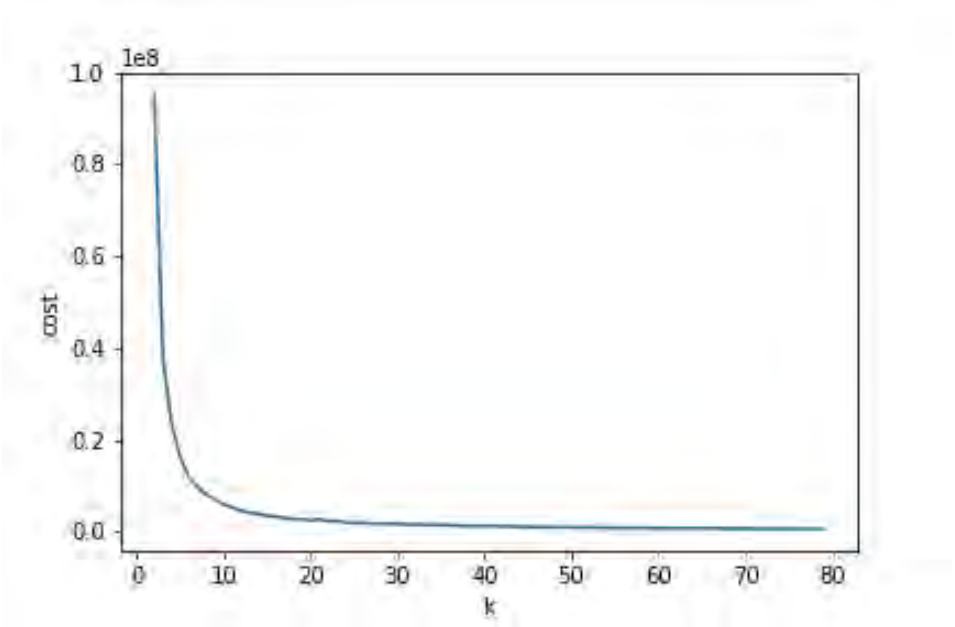
Δεν γινόταν να μην αξιοποιηθεί η βιβλιοθήκη μηχανικής μάθησης που ενσωματώνει η πλατφόρμα του Spark, ιδιαίτερα όταν γίνεται λόγος για Big Data. Αξιοποιήθηκαν και οι δύο κατηγορίες μηχανικής μάθησης, εποπτευόμενη και μη εποπτευόμενη μάθηση, που έχει η βιβλιοθήκη του Spark. Σαν είσοδο και στις δύο κατηγορίες δόθηκαν μόνο τα δεδομένα κυκλοφοριακής κίνησης, επειδή είναι αληθινά δεδομένα από πραγματικούς αισθητήρες.

### 5.4.1 Μη εποπτευόμενη μάθηση

Στην μη εποπτευόμενη μάθηση και πιο συγκεκριμένα στην συσταδοποίηση έγινε μοντελοποίηση του αλγορίθμου K-Means. Σαν είσοδο στον αλγόριθμο δόθηκαν η μέση ταχύτητα των οχημάτων και το πλήθος των διερχόμενων οχημάτων.

Το αρχικό στάδιο πριν ξεκινήσει η διαδικασία ομαδοποίησης είναι ο καθαρισμός των δεδομένων από τυχόν θορύβους. Στα συγκεκριμένα δεδομένα, υπάρχουν εγγραφές όπου η μέση ταχύτητα δεν είναι μηδέν όταν το πλήθος των διερχόμενων οχημάτων είναι ίσο με το μηδέν. Αυτό γίνεται επειδή κρατάει την αμέσως προηγούμενη μέση ταχύτητα όπου ο αριθμός των οχημάτων ήταν μη-μηδενικός.

Μετά τη διαδικασία του καθαρισμού, ακολουθεί το δεύτερο σημαντικό στάδιο το οποίο είναι η εύρεση του καλύτερου k. Το k είναι ο αριθμός των ομάδων που θα δημιουργήσει ο αλγόριθμος K-Means. Ξεκινώντας τις δοκιμές από k=2 έως το k=80, υπολογίστηκαν όλα τα κόστη για κάθε μία δοκιμή.




**Εικόνα 18 Κόστη του K-Means**

Είναι ξεκάθαρο πως η καμπύλη κόστους στο σημείο k=20 παύει να φθίνει και αρχίζει να σταθεροποιείται το κόστος των ομάδων, όπως διακρίνεται στην εικόνα 14.

Στο τελικό στάδιο της συσταδοποίησης, γίνεται η μοντελοποίηση και εκπαίδευση του αλγορίθμου K-Means για k=20.

| Vehicle Count | Average Speed |
|---|---|
| 0.00407167 | 0.00748834 |
| 1.8085963 | 34.99731581 |
| 1.95445545 | 54.34497646 |
| 1.80327315 | 37.9744379 |
| 1.8465162 | 23.57087453 |
| 1.57892691 | 99.78771384 |
| 18.68702641 | 51.78300804 |
| 1.84137232 | 28.48358821 |
| 27.29663212 | 20.77590674 |
| 1.94948971 | 81.88341117 |
| 1.81950551 | 17.48236657 |
| 2.72854327 | 69.4479484 |
| 1.95579417 | 60.96234835 |
| 1.77821732 | 40.96784782 |
| 10.7663023 | 58.97260962 |
| 1.78707777 | 44.45614533 |
| 2.08098243 | 10.2112101 |
| 1.81121912 | 31.9978008 |
| 12.91452442 | 34.55141388 |

**Πίνακας 1 Κέντρα Ομάδων**

Επομένως, ο αλγόριθμος δημιουργεί 20 ομάδες, με τα αντίστοιχα κέντρα όπως αυτά φαίνονται στον πίνακα 1.

Οι 20 ομάδες του αλγορίθμου, αποτελούν 20 προφίλ που μπορούν να περιγράψουν άψογα την κίνηση σε κάθε μία διαδρομή. Για παράδειγμα, όταν ο αριθμός των οχημάτων και η μέση ταχύτητα είναι μηδέν τότε η διαδρομή είναι τελείως άδεια ενώ στην άλλη άκρη, όταν ο αριθμός των οχημάτων είναι υψηλός και η μέση ταχύτητα είναι μηδέν τότε υπάρχει έντονη κυκλοφοριακή συμφόρηση.

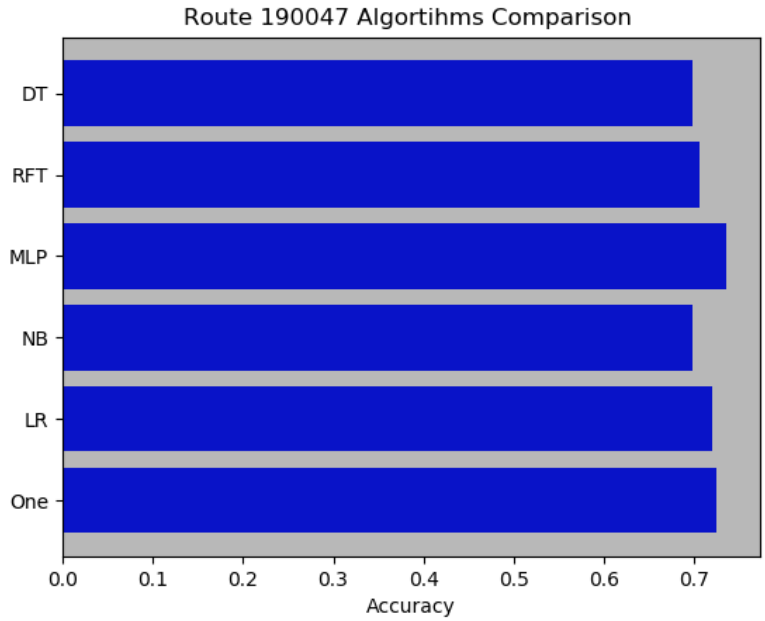
### 5.4.2 Εποπτευόμενη Μάθηση

Στην κατηγορία της εποπτευόμενης μάθησης έπρεπε να τεθεί ποια θα είναι η είσοδος και ποια η επιθυμητή έξοδος των δεδομένων για να δημιουργηθούν τα μοντέλα πρόβλεψης. Επομένως, σαν είσοδος τέθηκε η ημέρα της εβδομάδας, η ώρα και τα λεπτά από τη στήλη time των δεδομένων και ως επιθυμητή έξοδος ο αριθμός των οχημάτων, για κάθε αισθητήρα ξεχωριστά. Σκοπός είναι να μπορεί ο κάθε αλγόριθμος που μοντελοποιήθηκε στην κατηγορία αυτή να προβλέψει τον

αριθμό των οχημάτων με βάση την ημέρα. Όλες οι τιμές των δεδομένων που δίνονται σαν ορίσματα στους αλγορίθμους είναι διακριτές τιμές, γεγονός που τα εντάσσει στην υποενότητα της ταξινόμησης και πιο συγκεκριμένα στην πολλαπλή ταξινόμηση καθώς τα δεδομένα εξόδου δεν είναι μόνο δύο.

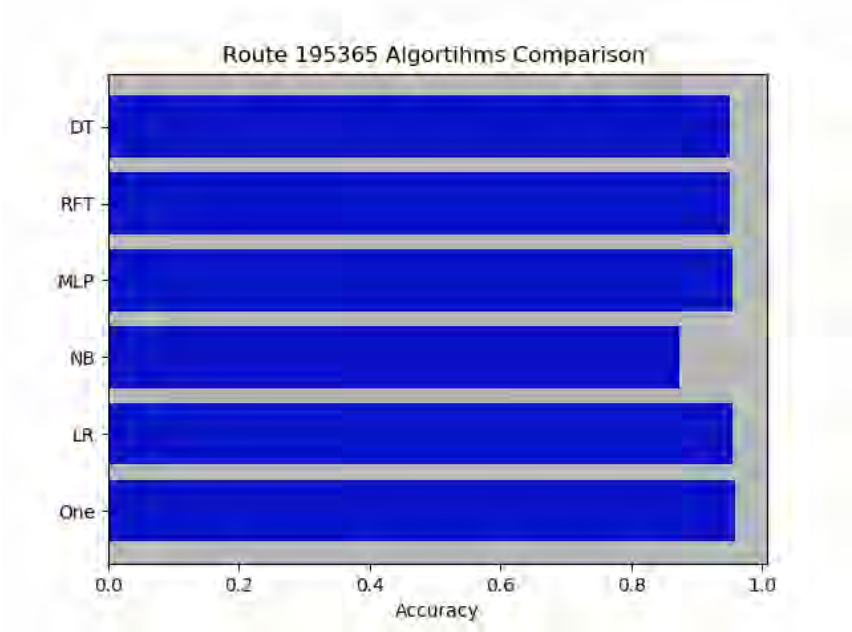
Υλοποιήθηκαν όλοι οι αλγόριθμοι που εξυπηρετούν την πολλαπλή ταξινόμηση. Οι αλγόριθμοι αυτοί είναι α) ο Decision Tree, β) ο Random Forest Tree, γ) ο Multilayer Perception Classifier με το οποίο κατασκευάστηκε ένα νευρωνικό δίκτυο, δ) ο Naive Bayes, ε) ο Logistic Regression Classifier και τέλος στ) ο OneVsRest Classifier.

Στο δέντρο αποφάσεων η μόνη παράμετρος που τροποποιήθηκε είναι η maxDepth, η οποία τέθηκε ίση με 30. Το 30 αποτελεί και τη μέγιστη τιμή βάθους που δέχεται ο αλγόριθμος κατά την υλοποίησή του στον Spark. Στον Random Forest Tree τροποποιήθηκαν 2 από τις παραμέτρους που δέχεται, η πρώτη αφορά τα πλήθη των δέντρων που θα κατασκευάσει η οποία τέθηκε ίση με 100και η δεύτερη είναι το μέγιστο βάθος των δέντρων που όμοια με πριν τέθηκε στο μέγιστο. Στο νευρωνικό δίκτυο δόθηκαν τα επίπεδα με τον αριθμό των νευρώνων και ο μέγιστος αριθμός πράξεων. Πρόκειται για 3 επίπεδα, εκ των οποίων το πρώτο είναι το επίπεδο εισόδου που τέθηκε ίσο με 3, όσα είναι δηλαδή και τα χαρακτηριστικά εισόδου, στο δεύτερο επίπεδο δόθηκαν 2 κρυφοί νευρώνες από 9 κόμβους έκαστος και το τελευταίο επίπεδο, το επίπεδο εξόδου τέθηκε ίσο με τη μέγιστη τιμή οχημάτων, που αποτελεί και τη μέγιστη τάξη πρόβλεψης. Στον OneVsRest σαν παράμετρος classifier δόθηκε ο αλγόριθμος logistic Regression με την παράμετρο regParam ίση με 0. Στον αλγόριθμο Logistic Regression η παράμετρος family που αφορά το είδος της ταξινόμησης, τέθηκε ίση με ‘multinomial’ επειδή τα δεδομένα που δέχεται έχουν περισσότερες από μία τάξεις. Ο αλγόριθμος Naive Bayes μοντελοποιήθηκε με τα προκαθορισμένα ορίσματα που του παρέχει η πλατφόρμα.




**Εικόνα 19 Ακρίβεια διαδρομής 190047**

Στην διαδρομή “190047” ο αλγόριθμος που παρουσίασε την μεγαλύτερη ακρίβεια είναι ο MultiLayer Perception με ακρίβεια 73%, τον ακολουθεί ο αλγόριθμος OneVsRest με 72%, ο Logistic Regression με 71%, ο Random Forest Tree με ακρίβεια 70%  και τελευταίοι είναι οι αλγόριθμοι Decision Tree και Naive Bayes με ακρίβεια 69%, όπως περιγράφεται στην εικόνα 15.




**Εικόνα 20 Ακρίβεια διαδρομής 195365**

Στην διαδρομή “195365” οι αλγόριθμοι Decision Tree, Random Forest Tree, MultiLayer Perception, Logistic Regression και OneVsRest ισοβαθμίζουν στην πρώτη θέση με ακρίβεια 95%. Τελευταίος είναι αλγόριθμος Naive Bayes με ακρίβεια 87%, όπως περιγράφεται στην εικόνα 16.

Μπορούμε να συμπεράνουμε από τις συγκρίσεις των δύο διαδρομών πως ο αλγόριθμος Naive Bayes δίνει την μικρότερη ακρίβεια για τα δεδομένα κίνησης. Την καλύτερη ακρίβεια την δίνει ο Multilayer Perception. Αν και αυτό εξαρτάται κάθε φορά από την διαδρομή και το κατά πόσο επαναλαμβάνεται το μοτίβο των οχημάτων.

Η ακρίβεια των μοντέλων ταξινόμησης στην εποπτευόμενη μάθηση υπολογίστηκε από το πακέτο μετρήσεων που δίνεται από τον Spark κάτω από το πακέτο pyspark.ml.evaluation στην Python.

$$ACC = \frac{1}{N}\sum_{L=0}^{N-1} \hat{\delta}(\hat{y}_i - y_i)$$

**Εξίσωση 1 Ακρίβεια**

Πιο συγκεκριμένα, ο αλγόριθμος υπολογισμού της ακρίβειας των δεδομένων περιγράφεται από την εξίσωση 1. Όπου το N είναι το πλήθος των δεδομένων και τα $y_i,\ \hat{y}_i$ η επιθυμητή έξοδος και η έξοδος που πρόβλεψε ο αλγόριθμος, αντίστοιχα. Η συνάρτηση $\hat{\delta}$ είναι μία βηματική συνάρτηση η οποία επιστρέφει 1 αν η διαφορά των δύο τιμών είναι 0 και 0 σε άλλη περίπτωση.

# 6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Σκοπός της εργασίας ήταν να δείξει πως τα μεγάλα δεδομένα και η ανάλυση τους μπορούν να προσφέρουν ευεργετικές ιδιότητες στις έξυπνες πόλεις και να αυξήσουν την αστική αειφορία προκειμένου να καταπολεμήσουν τα προβλήματα που την υποβαθμίζουν, όπως είναι η κίνηση και η ατμοσφαιρική ρύπανση.

Οι ομάδες που προκύπτουν από τη μη εποπτευόμενη μάθηση είναι 20 γιατί περιγράφουν αναλυτικά τα δεδομένα κίνησης όλων των διαδρομών μαζί. Σκοπός βέβαια είναι πώς η έξυπνη πόλη θα τα ερμηνεύσει προκειμένου να περιγράψει την κίνηση σε κάθε περιοχή αλλά και συνολικά. Μπορεί για παράδειγμα, μία πόλη να συμπτύξει τις 20 ομάδες σε 5 επίπεδα κίνησης, υπολογίζοντας τον μέσο όρο των κέντρων των ομάδων. Αυτό είναι στην κρίση των αναλυτών κάθε πόλης καθώς πρέπει να ληφθούν και άλλοι παράγοντες υπόψιν, όπως για παράδειγμα οι ώρες αιχμής κάθε διαδρομής ανάλογα την εποχή και την ώρα, το μέγεθος κάθε διαδρομής τόσο σε μήκος όσο και σε πλάτος γιατί παίζει καθοριστικό ρόλο στο πλήθος των οχημάτων που μπορεί να εξυπηρετήσει ταυτόχρονα, την απόσταση των δύο σημείων καθώς και το επιτρεπτό όριο ταχύτητας.

Οι αλγόριθμοι ταξινόμησης δεν μπορούν να αναλύσουν και να προβλέψουν των αριθμό των οχημάτων ολόκληρης της πόλης γιατί κάθε διαδρομή έχει ένα μοναδικό επαναλαμβανόμενο μοτίβο οχημάτων που θα την διαπεράσουν. Επομένως, έγινε ανάλυση και μοντελοποίηση για κάθε διαδρομή ξεχωριστά. Είναι προφανές, πως τα αποτελέσματα διαφέρουν μεταξύ τους και κάποια ίσως να διαφέρουν κατά πολύ. Καθοριστικό παράγοντα παίζει η περίοδος που κάλυψαν τα δεδομένα κίνησης. Δύο μήνες δεν είναι αρκετοί για να περιγράψουν την κίνηση, αν τα δεδομένα αφορούσαν έναν ολόκληρο έτος για παράδειγμα, οι αλγόριθμοι θα δέχονταν μία πιο πλούσια περιγραφή της κίνησης με αποτέλεσμα να εξάγουν και μία πιο σωστή πρόβλεψη.

Επίσης θα πρέπει να ληφθούν υπόψιν και άλλοι παράγοντες που παίζουν καθοριστικό ρόλο, όπως για παράδειγμα οι αργίες και οι περίοδοι διακοπών. Επίσης, μία άλλη λύση θα ήταν στα δεδομένα να υπάρχει και μια στήλη tags η οποία θα περιγράφει τον λόγο που κάθε όχημα εκτελεί τη συγκεκριμένη διαδρομή. Για παράδειγμα, τα οχήματα που εισέρχονται και εξέρχονται από τα δύο αυτά σημεία με σκοπό να πάνε στη δουλειά τους αποτελούν μέρος του επαναλαμβανόμενου μοτίβου επομένως θα έπρεπε να έχουν μεγαλύτερο βάρος στους αλγορίθμους μηχανικής μάθησης, ενώ οχήματα που εκτελούν σπάνια αυτή τη διαδρομή για να πάνε σε μία επίσκεψη ή για να πάνε σε σημεία ψυχαγωγίας θα έπρεπε να έχουν μικρότερο βάρος. Έτσι, το σύστημα θα μπορεί να γνωρίζει καλύτερα την πιθανότητα των οχημάτων κάθε διαδρομής.

Με τα επίπεδα κίνησης οι αρμόδιες αρχές θα μπορούν να ενημερώνουν τους πολίτες για την κίνηση σε κάθε διαδρομή. Αυτό θα βοηθήσει σημαντικά την ευεξία της πόλης. Οι οδηγοί των οχημάτων θα μπορούν να χαράξουν διαφορετική

πορεία ανάλογα με το ποιοι δρόμοι είναι γεμάτοι και ποιοι όχι. Αυτό έχει σημαντικό αντίκτυπο στις ανθρώπινες μετακινήσεις, στα μέσα μαζικής μεταφοράς, σε μεταφορές, εταιρείες ανεφοδιασμού αλλά κυρίως σε ασθενοφόρα, περιπολικά και οχήματα της πυροσβεστικής που θα μπορούν να μειώσουν δραματικά τον χρόνο άφιξής τους. Παράλληλα με αυτό, οι πεζοί ή οι ποδηλάτες θα είναι σε θέση να αλλάξουν πορεία προκειμένου να διαλέξουν μία διαδρομή με λιγότερη κίνηση και επομένως μικρότερα επίπεδα ρύπων. Αναβαθμίζεται έτσι και η ανθρώπινη υγεία αποφεύγοντας την έκθεση σε υψηλά ποσοστά βλαβερών ουσιών και καυσαερίου.

Με την πρόβλεψη των οχημάτων και την εξαγωγή της κίνησης, θα μπορούν οι πολίτες να ενημερώνονται για μελλοντικές τιμές ώστε να οργανώσουν διαφορετικά το πρόγραμμα τους. Επίσης, με την πρόβλεψη της κίνησης η πόλη θα μπορεί να ρυθμίζει διαφορετικά τους σηματοδότες εξισορροπώντας τα επίπεδα σε ολόκληρη την πόλη.

Τέλος, η διαδικτυακή εφαρμογή διαβάζει το ιστορικό των δεδομένων και κάνει γραφική απεικόνιση της εξέλιξης των οχημάτων και των ρύπων στον χρόνο. Σκοπός είναι να αποτελέσει ένα πρότυπο σελίδας το οποίο θα παρακολουθεί τους αισθητήρες σε πραγματικό χρόνο και θα τρέχει τους αλγορίθμους μηχανικής μάθησης συνέχεια στα νέα δεδομένα. Στο σύστημα παρακολούθησης, θα εμφανίζονται και οι εβδομαδιαίες προβλέψεις σύμφωνα με την κίνηση που έχουν καταγράψει οι αισθητήρες όλο το χρόνο και θα μπορεί να ενημερώνει τους πολίτες για τα επίπεδα κίνησης καθώς και για τους ρύπους στην κάθε διαδρομή, έτσι ώστε να είναι σε θέσει να αναδιαμορφώσουν την διαδρομή τους.

# ΒΙΒΛΙΟΓΡΑΦΙΑ


1. Yunhe Pan, Yun Tian, Xiaolong Liu, Dedao Gu, Gang Hua. (2016) “Urban Big Data and the Development of City Intelligence”.
2. Yuzhe Wu , Weiwen Zhang, Jiahui Shen a, Zhibin Mo, Yi Peng. (2017) “Smart city with Chinese characteristics against the background of big data: Idea, action and risk”.
3. Li-MinnAng, Kah PhooiSeng (2016) “Big Sensor Data Applications in Urban Environments”.
4. Xiao Luo, Liang Dong, Yi Dou, Ning Zhang, Jingzheng Ren, Ye Li, Lu Sun, Shengyong Yao.(2016) “Analysis on spatial-temporal features of taxis emissions from big data informed travel patterns: a case of Shanghai, China”.
5. Ibrahim Abaker Targio Hashem, Victor Chang, Nor Badrul Anuara, Kayode Adewole, Ibrar Yaqoob, Abdullah Gani, Ejaz Ahmed, Haruna Chiromaca. (2016) “The role of big data in smart city”.
6. M. Mazhar Rathore , Awais Ahmad, Anand Paul, Seungmin Rho. (2016) “Urban planning and building smart cities based on the Internet of Things using Big Data analytics”.
7. Y.Lakshmi Prasad. “Big Data Analytics Made Easy” (p.10-11)
8. Tomasz Jach, Ewa Magiera, Wojciech Froelich.(2015) “Application of HADOOP to Store and Process Big Data Gathered from an Urban Water Distribution System”.
9. Tomasz Drabas, Denny Lee (2017) “Learning Pyspark”.
10. http://iot.ee.surrey.ac.uk:8080/datasets.html
11. https://en.wikipedia.org/wiki/Air_quality_index
12. http://iot.ee.surrey.ac.uk:8080/datasets/pollution/readme.txt
13. http://iot.ee.surrey.ac.uk:8080/datasets.html#pollution
14. Francisco R. Klauser, Anders Albrechtslund. (2014) From self-tracking to smart urban infrastructures: Towards an interdisciplinary research agenda on Big Data.
15. Rencai Dong, Siyuan Li, Yonglin Zhang, Nana Zhang, Tao Wang, Xinrui Tan, Xiao Fu. (2016) Analysis of urban environmental problems based on big data from the urban municipal supervision and management information system.
16. Qi Shi, Mohamed Abdel-Aty. (2015) Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways.
17. Rashid Mehmooda, Gary Graham. (2015) Big data logistics: a health-care transport capacity sharing model.
18. T.Kato. (2016) Chapter 4.3 - Consensus Building for a Resilient Society: Utilization of Big Data.

19. Jinwei Hao, Jin Zhu, Rui Zhong. (2015) The rise of big data on urban studies and planning practices in China: Review and open research issues.
20. Rob Kitchin, Tracey P. Lauriault & Gavin McArdle. (2015) Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards.

## ΠΑΡΑΡΤΗΜΑ

| | Dt | RFT | MLP | NB | LR | One |
|---|---|---|---|---|---|---|
| 171572 | 0.588064 | 0.610102 | 0.605068 | 0.563915 | 0.605615 | 0.615834 |
| 172329 | 0.589619 | 0.595892 | 0.59061 | 0.509774 | 0.58677 | 0.594915 |
| 173011 | 0.934616 | 0.936989 | 0.940556 | 0.846136 | 0.936664 | 0.940451 |
| 173118 | 0.631315 | 0.629709 | 0.647983 | 0.574649 | 0.658589 | 0.655075 |
| 178929 | 0.722523 | 0.72698 | 0.73939 | 0.714567 | 0.74027 | 0.737724 |
| 178955 | 0.70781 | 0.731676 | 0.73193 | 0.714588 | 0.738031 | 0.7324 |
| 178983 | 0.675751 | 0.674946 | 0.679957 | 0.629308 | 0.679698 | 0.679983 |
| 179009 | 0.669857 | 0.669493 | 0.680911 | 0.608026 | 0.690869 | 0.687034 |
| 180898 | 0.660885 | 0.66858 | 0.675319 | 0.676371 | 0.678541 | 0.68823 |
| 180952 | 0.636346 | 0.64777 | 0.647959 | 0.611067 | 0.659562 | 0.650904 |
| 184919 | 0.757284 | 0.761622 | 0.789474 | 0.768712 | 0.785428 | 0.789439 |
| 184945 | 0.813509 | 0.813848 | 0.817431 | 0.790967 | 0.827311 | 0.81649 |
| 184972 | 0.61244 | 0.61567 | 0.617705 | 0.53076 | 0.619595 | 0.636763 |
| 185025 | 0.604851 | 0.609814 | 0.616075 | 0.5979 | 0.62755 | 0.636461 |
| 185343 | 0.596724 | 0.597391 | 0.608609 | 0.610266 | 0.615231 | 0.623153 |
| 185369 | 0.631568 | 0.656876 | 0.665683 | 0.667658 | 0.661617 | 0.667125 |
| 186953 | 0.621515 | 0.618518 | 0.644128 | 0.604677 | 0.652369 | 0.648127 |
| 186979 | 0.769099 | 0.761246 | 0.787705 | 0.76696 | 0.783303 | 0.778972 |
| 187059 | 0.604821 | 0.596845 | 0.618838 | 0.561987 | 0.615925 | 0.608146 |
| 187085 | 0.626078 | 0.629923 | 0.624041 | 0.620807 | 0.636037 | 0.630422 |
| 187165 | 0.603729 | 0.612271 | 0.605899 | 0.585403 | 0.631478 | 0.626107 |
| 187191 | 0.57333 | 0.590153 | 0.587384 | 0.577641 | 0.580617 | 0.573205 |
| 187377 | 0.85932 | 0.869872 | 0.888617 | 0.805986 | 0.887906 | 0.882686 |
| 187403 | 0.88104 | 0.883531 | 0.891545 | 0.76533 | 0.886179 | 0.894062 |
| 187483 | 0.823058 | 0.821555 | 0.835552 | 0.835374 | 0.837854 | 0.836531 |
| 187509 | 0.851158 | 0.850818 | 0.851419 | 0.76306 | 0.85837 | 0.857385 |
| 187589 | 0.744332 | 0.743898 | 0.770322 | 0.707542 | 0.777625 | 0.776019 |
| 187615 | 0.668847 | 0.67368 | 0.706359 | 0.677479 | 0.696571 | 0.698394 |
| 187668 | 0.627646 | 0.630267 | 0.634022 | 0.6036 | 0.646662 | 0.642903 |
| 187748 | 0.652533 | 0.650571 | 0.648128 | 0.600042 | 0.670092 | 0.683817 |
| 187774 | 0.570216 | 0.573058 | 0.558215 | 0.507971 | 0.581772 | 0.578533 |
| 187801 | 0.824498 | 0.826004 | 0.836726 | 0.827197 | 0.853388 | 0.841771 |
| 187827 | 0.995583 | 0.995051 | 0.994011 | 0.995488 | 0.996319 | 0.994925 |
| 187854 | 0.593862 | 0.60309 | 0.604541 | 0.549704 | 0.606025 | 0.612059 |
| 187880 | 0.612722 | 0.615079 | 0.629708 | 0.603111 | 0.613531 | 0.610583 |
| 187907 | 0.5593 | 0.570041 | 0.572051 | 0.495616 | 0.564409 | 0.5662 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 187933 | 0.557622 | 0.559224 | 0.574703 | 0.550924 | 0.56555 | 0.570255 |
| 187960 | 0.669878 | 0.663497 | 0.680634 | 0.637546 | 0.694676 | 0.68777 |
| 187986 | 0.690913 | 0.706823 | 0.714079 | 0.718425 | 0.708134 | 0.72066 |
| 188066 | 0.926097 | 0.920997 | 0.927868 | 0.913966 | 0.928505 | 0.926044 |
| 188092 | 0.938816 | 0.943242 | 0.946766 | 0.94619 | 0.942493 | 0.938982 |
| 188145 | 0.59152 | 0.582501 | 0.589082 | 0.563809 | 0.604769 | 0.600428 |
| 189941 | 0.582217 | 0.581868 | 0.574752 | 0.5573 | 0.597034 | 0.592727 |
| 189967 | 0.597285 | 0.615174 | 0.635132 | 0.608604 | 0.621202 | 0.633728 |
| 190047 | 0.697814 | 0.705959 | 0.736019 | 0.697877 | 0.719845 | 0.72408 |
| 190073 | 0.665721 | 0.668253 | 0.690279 | 0.674732 | 0.698211 | 0.698476 |
| 190100 | 0.739586 | 0.732349 | 0.756624 | 0.676411 | 0.754679 | 0.749311 |
| 190126 | 0.77108 | 0.783789 | 0.785629 | 0.775697 | 0.786439 | 0.782339 |
| 190153 | 0.805699 | 0.810508 | 0.831373 | 0.718551 | 0.831241 | 0.835145 |
| 190179 | 0.898745 | 0.905528 | 0.908718 | 0.915145 | 0.902898 | 0.91275 |
| 192520 | 0.656766 | 0.656131 | 0.658134 | 0.660102 | 0.649353 | 0.658437 |
| 192546 | 0.672678 | 0.682312 | 0.692724 | 0.619224 | 0.697437 | 0.692966 |
| 192681 | 0.647708 | 0.649201 | 0.66127 | 0.573617 | 0.656687 | 0.656713 |
| 192707 | 0.613185 | 0.614226 | 0.639513 | 0.620158 | 0.62647 | 0.625456 |
| 192734 | 0.637726 | 0.628578 | 0.641456 | 0.649209 | 0.653734 | 0.648457 |
| 192760 | 0.644361 | 0.655741 | 0.641262 | 0.634731 | 0.673073 | 0.666274 |
| 192787 | 0.578457 | 0.570839 | 0.572403 | 0.480968 | 0.572798 | 0.567699 |
| 192813 | 0.585181 | 0.579627 | 0.586074 | 0.518118 | 0.597006 | 0.577965 |
| 192840 | 0.545978 | 0.551505 | 0.53349 | 0.454839 | 0.519428 | 0.525306 |
| 192866 | 0.569227 | 0.572156 | 0.580417 | 0.488975 | 0.531149 | 0.547681 |
| 192893 | 0.84088 | 0.837832 | 0.856931 | 0.84377 | 0.859131 | 0.854577 |
| 192919 | 0.855628 | 0.861232 | 0.866781 | 0.868352 | 0.868433 | 0.870311 |
| 193000 | 0.647739 | 0.635593 | 0.632536 | 0.648676 | 0.646019 | 0.643524 |
| 193026 | 0.599253 | 0.589333 | 0.596126 | 0.487381 | 0.602159 | 0.608642 |
| 193053 | 0.723617 | 0.730495 | 0.755996 | 0.763434 | 0.756072 | 0.75874 |
| 193079 | 0.748969 | 0.749627 | 0.774326 | 0.76458 | 0.7721 | 0.774255 |
| 193106 | 0.968697 | 0.96663 | 0.96744 | 0.970098 | 0.971946 | 0.970645 |
| 193132 | 0.98531 | 0.986705 | 0.986902 | 0.90506 | 0.986631 | 0.987731 |
| 193159 | 0.743398 | 0.739879 | 0.768608 | 0.727526 | 0.769959 | 0.772934 |
| 193185 | 0.801814 | 0.803772 | 0.818658 | 0.819815 | 0.823425 | 0.823274 |
| 193213 | 0.950101 | 0.954379 | 0.951844 | 0.897068 | 0.955954 | 0.953298 |
| 193239 | 0.895202 | 0.89774 | 0.904486 | 0.906672 | 0.908975 | 0.905872 |
| 193268 | 0.563636 | 0.570673 | 0.598982 | 0.502843 | 0.595781 | 0.612583 |
| 193294 | 0.567186 | 0.579172 | 0.612371 | 0.531956 | 0.609926 | 0.615154 |
| 194960 | 0.566244 | 0.564123 | 0.553105 | 0.435918 | 0.553023 | 0.552928 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 194986 | 0.56615 | 0.563216 | 0.578834 | 0.524662 | 0.568209 | 0.564692 |
| 195233 | 0.569066 | 0.57382 | 0.560256 | 0.484388 | 0.550613 | 0.570191 |
| 195259 | 0.596812 | 0.606621 | 0.612532 | 0.558083 | 0.604933 | 0.604717 |
| 195339 | 0.943765 | 0.94106 | 0.949182 | 0.874243 | 0.945925 | 0.950049 |
| 195365 | 0.950497 | 0.950385 | 0.95533 | 0.872462 | 0.954482 | 0.959359 |
| 195392 | 0.874106 | 0.8815 | 0.885326 | 0.857713 | 0.881725 | 0.889725 |
| 195418 | 0.929087 | 0.924206 | 0.934391 | 0.828468 | 0.933434 | 0.932214 |
| 195446 | 0.571826 | 0.581969 | 0.586942 | 0.545979 | 0.598129 | 0.595134 |
| 195499 | 0.667093 | 0.669625 | 0.688047 | 0.61859 | 0.691224 | 0.688375 |
| 195525 | 0.593437 | 0.59859 | 0.61225 | 0.518047 | 0.616572 | 0.616883 |
| 195764 | 0.611385 | 0.626922 | 0.659566 | 0.611244 | 0.654871 | 0.644413 |
| 195790 | 0.624439 | 0.623574 | 0.644363 | 0.51372 | 0.634659 | 0.641398 |
| 195817 | 0.774902 | 0.780362 | 0.791777 | 0.770902 | 0.798325 | 0.790374 |
| 195843 | 0.814114 | 0.818064 | 0.830638 | 0.778473 | 0.838538 | 0.83661 |
| 195870 | 0.749125 | 0.757311 | 0.780483 | 0.745748 | 0.779213 | 0.775776 |
| 195896 | 0.806944 | 0.808449 | 0.831621 | 0.773418 | 0.830769 | 0.827726 |
| 197302 | 0.660623 | 0.665172 | 0.688064 | 0.641843 | 0.697179 | 0.69595 |
| 197328 | 0.78961 | 0.790009 | 0.808931 | 0.782572 | 0.810517 | 0.809119 |
| 197355 | 0.945161 | 0.942821 | 0.943659 | 0.945999 | 0.950546 | 0.94435 |
| 197381 | 0.909996 | 0.911309 | 0.917508 | 0.921564 | 0.916326 | 0.917796 |
| 197734 | 0.552654 | 0.564263 | 0.532451 | 0.431172 | 0.549935 | 0.534241 |
| 197760 | 0.558785 | 0.562472 | 0.550254 | 0.509401 | 0.554257 | 0.551416 |
| 197842 | 0.554134 | 0.565697 | 0.604445 | 0.535456 | 0.597751 | 0.604071 |
| 197868 | 0.577858 | 0.565273 | 0.600425 | 0.463404 | 0.607893 | 0.601997 |
| 197896 | 0.570784 | 0.593001 | 0.60492 | 0.551937 | 0.614586 | 0.617591 |
| 197922 | 0.5872 | 0.597596 | 0.63252 | 0.549762 | 0.621107 | 0.63116 |
| 198330 | 0.625175 | 0.637889 | 0.647494 | 0.571485 | 0.645212 | 0.645842 |
| 201183 | 0.609427 | 0.615134 | 0.619498 | 0.562462 | 0.634218 | 0.629554 |
| 201749 | 0.739711 | 0.746827 | 0.754889 | 0.634749 | 0.762716 | 0.756041 |
| 201775 | 0.728246 | 0.736366 | 0.755551 | 0.753743 | 0.761451 | 0.749391 |
| 201961 | 0.59912 | 0.610069 | 0.599203 | 0.538957 | 0.615899 | 0.61081 |
| 203530 | 0.622497 | 0.64817 | 0.633982 | 0.566064 | 0.637125 | 0.643819 |
| 203557 | 0.903291 | 0.905022 | 0.907039 | 0.906972 | 0.912888 | 0.909726 |
| 203583 | 0.919842 | 0.920803 | 0.921163 | 0.922504 | 0.916265 | 0.917921 |
| 203610 | 0.521154 | 0.526102 | 0.517711 | 0.446097 | 0.514212 | 0.521046 |
| 203663 | 0.780673 | 0.797139 | 0.809379 | 0.749882 | 0.808746 | 0.817781 |
| 203689 | 0.785544 | 0.788324 | 0.806752 | 0.79549 | 0.802803 | 0.806663 |
| 203716 | 0.989962 | 0.992689 | 0.991694 | 0.992408 | 0.991688 | 0.991811 |
| 203742 | 0.94786 | 0.951872 | 0.947368 | 0.953063 | 0.950914 | 0.950538 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 203769 | 0.530736 | 0.538955 | 0.537667 | 0.478316 | 0.539873 | 0.553672 |
| 203795 | 0.605105 | 0.59589 | 0.61919 | 0.525108 | 0.6254 | 0.62255 |
| 203822 | 0.947817 | 0.949089 | 0.948873 | 0.949699 | 0.947889 | 0.943452 |
| 203848 | 0.96364 | 0.969697 | 0.969908 | 0.945406 | 0.96914 | 0.968863 |
| 203875 | 0.990898 | 0.988781 | 0.990909 | 0.992722 | 0.990607 | 0.990257 |
| 203901 | 0.983148 | 0.986608 | 0.986995 | 0.941558 | 0.986553 | 0.987945 |
| 203928 | 0.881387 | 0.896769 | 0.896996 | 0.895833 | 0.891616 | 0.89719 |
| 203954 | 0.92311 | 0.924191 | 0.925503 | 0.829646 | 0.923795 | 0.923438 |
| 204060 | 0.590936 | 0.576087 | 0.600321 | 0.568689 | 0.606389 | 0.601616 |
| 204087 | 0.626981 | 0.632572 | 0.650617 | 0.598606 | 0.657028 | 0.658395 |
| 204113 | 0.632906 | 0.633604 | 0.639422 | 0.58838 | 0.649393 | 0.651866 |
| 204140 | 0.542691 | 0.556466 | 0.558307 | 0.441851 | 0.522722 | 0.537623 |
| 204166 | 0.526912 | 0.541399 | 0.543553 | 0.451592 | 0.52701 | 0.522752 |
| 204300 | 0.850031 | 0.861602 | 0.866492 | 0.863776 | 0.868533 | 0.868188 |
| 205997 | 0.933703 | 0.93436 | 0.941897 | 0.911312 | 0.936439 | 0.935428 |
| 206078 | 0.959247 | 0.963249 | 0.962125 | 0.960911 | 0.968653 | 0.964317 |
| 206104 | 0.981125 | 0.983259 | 0.982213 | 0.938516 | 0.984121 | 0.981574 |
| 206184 | 0.938727 | 0.940577 | 0.942099 | 0.884229 | 0.945513 | 0.941474 |
| 206210 | 0.91478 | 0.918942 | 0.930321 | 0.901087 | 0.929334 | 0.919164 |
| 206237 | 0.801286 | 0.807495 | 0.823948 | 0.747543 | 0.825488 | 0.824166 |
| 206263 | 0.857621 | 0.869514 | 0.876642 | 0.802488 | 0.874213 | 0.870961 |
| 206343 | 0.775705 | 0.768546 | 0.802832 | 0.718992 | 0.801575 | 0.796304 |
| 206369 | 0.741409 | 0.745813 | 0.760563 | 0.765511 | 0.75091 | 0.766341 |
| 206502 | 0.716039 | 0.709087 | 0.735857 | 0.668176 | 0.731037 | 0.731377 |
| 209748 | 0.955063 | 0.96069 | 0.958989 | 0.894715 | 0.955417 | 0.953503 |
| 209774 | 0.964721 | 0.964033 | 0.965907 | 0.928833 | 0.967105 | 0.966765 |
| 209933 | 0.585249 | 0.582217 | 0.590104 | 0.487002 | 0.59182 | 0.587071 |
| 210013 | 0.730599 | 0.740503 | 0.769917 | 0.7153 | 0.763787 | 0.760365 |
| 210040 | 0.83636 | 0.845514 | 0.861777 | 0.862136 | 0.851408 | 0.855313 |
| 210067 | 0.945256 | 0.942165 | 0.942666 | 0.939176 | 0.938206 | 0.939718 |
| 210093 | 0.9079 | 0.906077 | 0.91845 | 0.918022 | 0.91656 | 0.910687 |
| 210120 | 0.812698 | 0.811896 | 0.828578 | 0.753228 | 0.828066 | 0.82591 |
| 210146 | 0.832557 | 0.836687 | 0.853233 | 0.813849 | 0.84444 | 0.852613 |
| 210173 | 0.904453 | 0.905669 | 0.921324 | 0.860553 | 0.912027 | 0.917361 |
| 210199 | 0.90966 | 0.908153 | 0.915689 | 0.885109 | 0.917209 | 0.91665 |

**Πίνακας 2 Ακρίβειες Διαδρομών**

Στον πίνακα 2 απεικονίζονται τα ποσοστά ακρίβειας των αλγορίθμων για κάθε διαδρομή ξεχωριστά.