ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ - WEB INTELLIGENCE

Ανάλυση και εξόρυξη δεδομένων χαρακτηριστικών καινούργιων αυτοκινήτων της Ελληνικής αγοράς με την υποστήριξη web service για την ανάκτηση δεδομένων

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΝΙΚΟΛΑΟΥ ΠΑΠΑΣΤΕΡΓΙΟΥ

Επιβλέπων : Στέφανος Ουγιάρογλου
Επ. Καθηγητής, ΔΙ.ΠΑ.Ε

Θεσσαλονίκη, Φεβρουάριος 2023

Η σελίδα αυτή είναι σκόπιμα λευκή.

ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ – WEB INTELLIGENCE

Τίτλος Μεταπτυχιακής Διπλωματικής Εργασίας

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΝΙΚΟΛΑΟΥ ΠΑΠΑΣΤΕΡΓΙΟΥ

Επιβλέπων : Στέφανος Ουγιάρογλου
Επ. Καθηγητής. ΔΙ.ΠΑ.Ε.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις Choose a date.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Όνομα Επώνυμο	Όνομα Επώνυμο	Όνομα Επώνυμο
Choose an item. ΔΙ.ΠΑ.Ε.	Choose an item. ΔΙ.ΠΑ.Ε.	Choose an item. ΔΙ.ΠΑ.Ε.

Θεσσαλονίκη, Φεβρουάριος 2023

(Υπογραφή)

.....

Click here to enter text.

Click here to enter text.

Περίληψη

Τα τελευταία χρόνια πολλαπλοί παράγοντες έχουν συμβάλλει ώστε να αλλάξει γενικότερα ο τρόπος ζωής και η καθημερινότητα πολλών ανθρώπων. Οι επιπτώσεις της παγκόσμιας οικονομικής κρίσης, η πανδημία του κορονοϊού, η ενεργειακή κρίση, ο πόλεμος και άλλα έχουν δρομολογήσει αξιοσημείωτες αλλαγές τόσο στην Ελλάδα όσο και στην Ε.Ε. Ένας από τους κλάδους που επηρεάστηκε καθοριστικά από τα παραπάνω γεγονότα είναι και η αυτοκινητοβιομηχανία. Μελετώντας τις παραμέτρους που επηρεάζουν τις πωλήσεις αυτοκινήτων, αρκετές κυβερνήσεις έχοντας ανθρώπους ικανούς και αξιόλογους σε συγκεκριμένες θέσεις, διαθέτουν τη δυνατότητα να προτείνουν καινοτόμους τρόπους στήριξης των πωλήσεων καινούργιων οχημάτων που δεν είναι επιζήμια για το περιβάλλον, όχι μόνο στην Ελλάδα, αλλά στο σύνολο των χωρών της Ε.Ε. Παράλληλα, παρατηρείται μια έντονη επιθυμία από υποψήφιους αγοραστές για καινούργιας γενιάς αυτοκίνητα. Κάποιοι εστιάζουν στα ηλεκτρικά, αρκετοί στα υβριδικά, όπως επίσης και στα βενζινοκίνητα, λιγότεροι πλέον στα πετρελαιοκίνητα. Όμως, λόγω της έντονης ζήτησης καινούργιων αυτοκινήτων, τα εργοστάσια έχουν έλλειψη σημαντικών προϊόντων για τα νέα αυτοκίνητα, με αποτέλεσμα να δημιουργούνται σημαντικές καθυστερήσεις στην παράδοση του αυτοκινήτου. Σκοπός της παρούσας διπλωματικής είναι να δημιουργηθεί ένα σύνολο δεδομένων από χαρακτηριστικά καινούργιων αυτοκινήτων, με τη φιλοδοξία να συγκροτηθεί ένας χρήσιμος και ευέλικτος οδηγός γνώσεων και πληροφοριών για τους επίδοξους αγοραστές. Με τη βοήθεια της εξόρυξης γνώσης αλλά και της μηχανικής μάθησης, εφαρμόστηκαν αλγόριθμοι συσταδοποίησης για εύρεση ομάδων αυτοκινήτων με παρόμοια χαρακτηριστικά. Έπειτα για τον αυτόματο προσδιορισμό τύπου καυσίμου εφαρμόστηκαν αλγόριθμοι κατηγοριοποίησης με τα χαρακτηριστικά των αυτοκινήτων αρχικά και στη συνέχεια με εκείνα που επηρεάζουν την απόδοση του αυτοκινήτου. Για αυτό το λόγο πραγματοποιήθηκαν πειράματα που δύνανται να εφαρμοστούν από ειδικούς για την εξαγωγή πολύτιμων συμπερασμάτων. Αξιοποιήθηκαν αρκετές τεχνολογίες και βιβλιοθήκες μηχανικής μάθησης βάσει των οποίων προβήκαμε σε προβλέψεις και καταλήξαμε σε επωφελή συμπεράσματα.

Λέξεις Κλειδιά: αυτοκίνητο, ελληνική αγορά, εξόρυξη γνώσης, ανάκτηση δεδομένων, αλγόριθμοι εξόρυξης δεδομένων

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

In recent years, multiple factors have contributed to change the way of life and everyday life of many people in general. The effects of the global financial crisis, the coronavirus, the energy crisis and the war have set in motion remarkable changes both in Greece and in general in the EU. One of the sectors that was decisively affected by the above events is the automotive industry. By studying the parameters that affect car sales, several governments put competent and worthy people in several positions who have the ability to propose innovative ways to support sales of new vehicles that are harmless for the environment, not only in Greece but in all EU countries. At the same time, there is a strong desire from prospective buyers for new generation cars. Some focus on electrics, several on hybrids as well as gasoline engines, fewer now on diesel engines. Due to the strong demand for new cars, the factories are short of important products for the new cars, resulting in significant delays in the delivery of the car. The purpose of this thesis is to create a data set with features of new cars, with the ambition to build a useful and flexible knowledge for prospective buyers. With the help of knowledge mining as well as machine learning, clustering algorithms were applied to find cars with similar characteristics. Then, for the automatic determination of fuel type, classification algorithms are applied with the characteristics of the cars and those that affect the performance of the car. Experiments were carried out that can be applied by experts to draw valuable conclusions. Several machine learning technologies and libraries were leveraged so that we were able to make predictions and conclude to beneficial results.

Keywords: cars, greek market, knowledge mining, data retrieval, data mining algorithms

Η σελίδα αυτή είναι σκόπιμα λευκή

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Εικόνων

Κατάλογος Διαγραμμάτων

1 *Εισαγωγή*

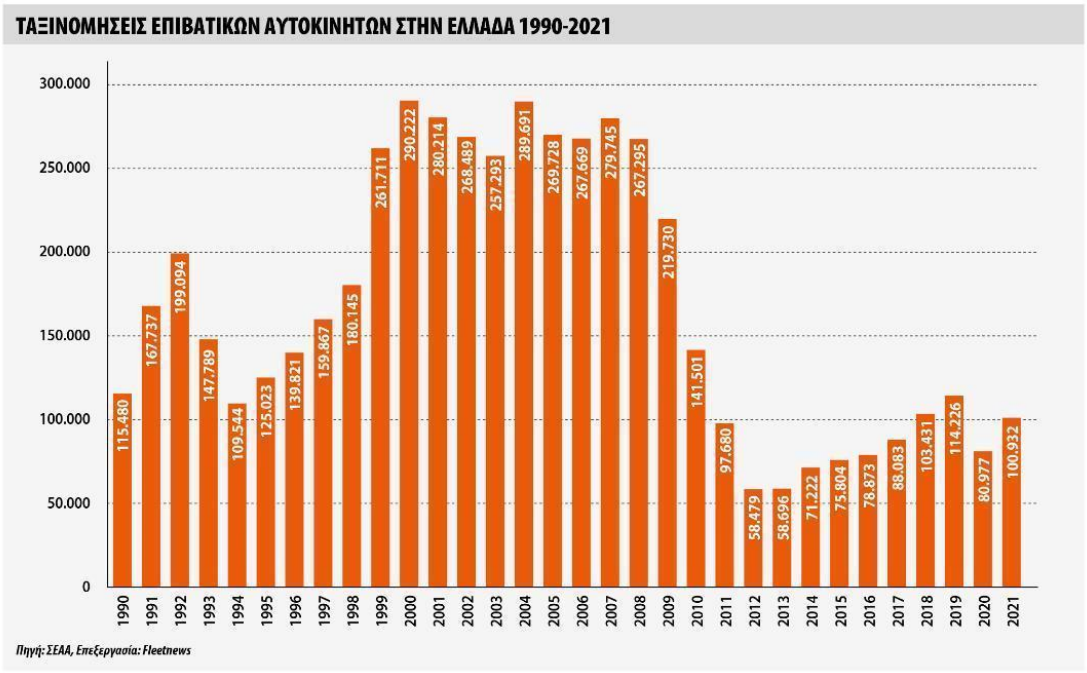
1.1 Η αγορά αυτοκινήτων στην Ελλάδα

Η αυτοκινητοβιομηχανία αποτελεί πρωταρχικό παράγοντα για την οικονομία της ΕΕ καθώς προσφέρει 2% του ΑΕΠ της ΕΕ και διατηρεί το 1,5% των εργαζόμενων της ΕΕ[1]. Παράλληλα, η προεπεξεργασία ενός αυτοκινήτου χρειάζεται έναν ικανοποιητικό αριθμό επαγγελματιών που μπορούν να ασχοληθούν με πληθώρα επιλογών παροχής υπηρεσιών, όπως η χρηματοδότηση αυτοκινήτου, η ασφάλιση, η εργασία σε αντιπροσωπεία ή και να δουλεύουν ως εξειδικευμένοι μηχανικοί αυτοκινήτων. Η μείωση της ζήτησης στην αγορά της ΕΕ έχει καταστροφικές επιπτώσεις στις ευρωπαϊκές πωλήσεις οχημάτων. Τονίζεται πως οι εταιρείες που έδωσαν βάση στις αγορές που ζημιώθηκαν περισσότερο από τη κρίση (Ελλάδα, Ιταλία και Ισπανία) ήρθαν αντιμέτωπες με τα απότοκα της οικονομικής κρίσης. Οι μειωμένες πωλήσεις διόγκωσαν το ζήτημα της παραγωγικής ικανότητας με αποτέλεσμα κάποιες κατασκευαστικές εταιρείες να απολύσουν εργαζομένους αλλά ακόμα και να βάλουν λουκέτο ολόκληρες εγκαταστάσεις παραγωγής. Όσοι όμως, απευθύνθηκαν σε εξωτερικές αγορές πέρα από τα ευρωπαϊκά πλαίσια είδαν τις επιχειρήσεις να κατορθώνουν να ισορροπούν τα έσοδα με τα έξοδα και τελικά να βγαίνουν κερδισμένες. Παράγοντες όπως η παγκόσμια οικονομική κρίση, η πανδημία του κορονοϊού, ο πόλεμος και η ενεργειακή κρίση σε συνδυασμό με την δυσχερή καθημερινότητα έφεραν σαρωτικές αλλαγές στην κοινωνία με σοβαρό αντίκτυπο και στις πωλήσεις αυτοκινήτων. Ο τομέας των πωλήσεων αυτοκινήτων είναι ένα πεδίο σημαντικό για την ελληνική οικονομία, αφού αναπαριστά ένα μεγάλο μερίδιο των κρατικών εσόδων, κυρίως από τα τέλη ταξινόμησης που εφαρμόζονται απευθείας με την πώληση του αυτοκινήτου, αλλά και από τεκμήρια που καλείται να πληρώσει ο αγοραστής κάθε έτος. Η υπάρχουσα κατάσταση λειτουργεί ως κίνητρο για τους Έλληνες πολιτικούς που επιδιώκουν να βρουν έναν ενδεδειγμένο τρόπο ώστε να διαθέτουν την ευχέρεια να αντιμετωπίσουν τις συνέπειες της τρέχουσας κατάστασης στον τομέα των μεταφορών.

Τα εκκρεμή ζητήματα στην αυτοκινητοβιομηχανία είναι τεράστια και πολυσύνθετα. Αρχικά, η ζήτηση είναι έκδηλα ανεβασμένη κυρίως μετά τον κορονοϊό, αλλά και ο πόλεμος υπήρξε ανασταλτική παράμετρος παραγωγής αγαθών και υπηρεσιών. Πολλοί κατασκευαστές δε μπορούν να αντιμετωπίσουν με αποτελεσματικότητα το παραπάνω ζήτημα και συνεχώς δημιουργούνται χρονοβόρες καθυστερήσεις στους χρόνους παράδοσης των αυτοκινήτων. Ο εκτιμώμενος χρόνος ήταν 2 ή 3 μήνες το πολύ και αυτό συνέβαινε συνήθως σε κατασκευαστικές που είχαν πολλές παραγγελίες να διευθετήσουν. Σήμερα αυτό έχει επιδεινωθεί δραματικά και ο εκτιμώμενος χρόνος είναι το λιγότερο 6 μήνες. Αυτά που ίσχυαν σε παλαιότερα έτη με τις 15 ημέρες αναμονής έχουν εξαφανιστεί και πλέον πρέπει να περάσει ένα μεγάλο χρονικό διάστημα για να επαναλειτουργήσουν με ταχύτερους ρυθμούς οι εταιρίες. Επιπρόσθετα, δεν είναι λίγες οι αντιπροσωπείες που πλέον στις παραγγελίες έχουν επιβάλει περιορισμούς, είτε στη επιλογή περισσότερων χαρακτηριστικών στο αυτοκίνητο είτε ακόμη και στην επιλογή βασικών χαρακτηριστικών, π.χ. διαθεσιμότητα συγκεκριμένων χρωμάτων. Ενδεικτικά, στοιχεία όπως ηλιοροφή, ηλεκτρικά ρυθμιζόμενο τιμόνι και κοτσαδόρος δε συμπληρώνονται στις παραγγελίες, γιατί γνωρίζουν ότι θα ανεβεί περισσότερο ο χρόνος αναμονής.

Ένα πρόβλημα που καλούνται οι υποψήφιοι πελάτες να διαχειριστούν είναι η τιμή του αυτοκινήτου τους. Η ραγδαία αύξηση, σε συνδυασμό με την αργοπορία εκτέλεσης της παραγγελίας δημιουργεί μια τεράστια διαφορά στην τελική τιμή του αυτοκινήτου. Δηλαδή, ο υποψήφιος πελάτης δεν μπορεί να ξέρει αν ανέβει η τιμή στο ενδιάμεσο χρονικό διάστημα παραλαβής του αυτοκινήτου. Στην Ασία, πολλές χώρες όπου κατασκευάζονται οι ημιαγωγοί προκαλείται μέγιστο πρόβλημα λόγω της υφιστάμενης έλλειψης. Εργοστάσια που βάζουν λουκέτο, πολλαπλασιάζουν το πρόβλημα μεταφοράς των ημιαγωγών, γιατί είναι ζόρικο να ανταπεξέλθουν στις συνθήκες που επικρατούν. Οι ημιαγωγοί πολύ συχνά καθυστερούν και αυτό έχει ως αποτέλεσμα να μη μπορούν να κατασκευάζονται καινούργια αυτοκίνητα. Αυτό επηρεάζει πολλούς καταναλωτές που τελικώς ακυρώνουν τις παραγγελίες τους ή προβληματίζονται και αναβάλλουν εν τέλει την αγορά ενός καινούργιου αυτοκινήτου. Η έλλειψη των ημιαγωγών φαίνεται και στις χαμηλές πωλήσεις σε όλη την Ευρώπη και στην Ελλάδα. Οι αναλυτές της συμβουλευτικής εταιρείας AlixPartners[2] πιστεύουν ότι θα προκληθεί μείωση της παγκόσμιας παραγωγής 7,7 εκατομμύρια φέτος, με κόστος 210 δισ. δολάρια. Η παρούσα κατάσταση δυσκολεύει και τις εταιρείες μίσθωσης που αδυνατούν να αναπτύξουν το στόλο τους, με συνέπεια να εξαναγκάζονται να ανανεώσουν τις συμβάσεις με τους πελάτες τους για μεγάλο διάστημα.

Με όλα αυτά που διαδραματίζονται σήμερα η αγορά του αυτοκινήτου στην Ευρωπαϊκή Ένωση δείχνει από την αρχή του έτους 2022 σημαντική μείωση. Σύμφωνα με τα επίσημα στοιχεία του ΣΕΑΑ[3] (Σύνδεσμος Εισαγωγέων Αντιπροσώπων Αυτοκινήτων), σημειώθηκε αύξηση της τάξης του 4,3% σε ταξινομήσεις αυτοκινήτων στην Ελλάδα, σύμφωνα με τα απολογιστικά στοιχεία του Συνδέσμου Εισαγωγέων Αντιπροσώπων Αυτοκινήτων (ΣΕΑΑ). Το 2022 ταξινομήθηκαν 105.283 αυτοκίνητα (ένα μεγάλο ποσοστό βέβαια από εταιρείες μίσθωσης) έναντι 100.911 το 2021, 80.977 το 2020 και 114.109 το 2019. Η περσινή επίδοση, με βάση τα στελέχη της εγχώριας αγοράς, αποτυπώνεται "σε κάποιο βαθμό ικανοποιητική", αν σκεφτεί κανείς ότι, στο 11μηνο του 2022, η Ελλάδα βρισκόταν στις 8 χώρες της ΕΕ, στις οποίες η αγορά αυτοκινήτου σημείωνε πρόοδο, οταν ο ευρωπαϊκός μέσος όρος ανερχόταν στο -6,1%.



Εξίσωση 1 Διάγραμμα ταξινόμησης αυτοκινήτων 1990-2021[4]

Σύμφωνα με στοιχεία του ΣΕΑΑ[5], για το 2022 πρώτη στη λίστα προτιμήσεων ήταν άλλη μια φορά η Toyota με μερίδιο αγοράς 14,1% (13% το 2021), δεύτερη η Hyundai με 8,7% (το ίδιο μερίδιο το 2021) και τρίτη η Peugeot με 8,6% (10,5% το 2021). Η Tesla, με τα εξ ολοκλήρου ηλεκτρικά αυτοκίνητα, πούλησε πέρσι 589 οχήματα έναντι 598 το 2021, κρατώντας το ίδιο ποσοστό (0,6%). Το 2022, το μερίδιο αγοράς των επαναφορτιζόμενων αυτοκινήτων (BEV και PHEV) στην Ελλάδα ανήλθε στο 7,9% έναντι 6,9% το 2021. Σε

αριθμούς, το 2022 ταξινομήθηκαν 8.337 επαναφορτιζόμενα αυτοκίνητα έναντι 6.961 το 2021. Τονίζεται πως οι πωλήσεις αυτοκινήτων στην Ελλάδα είναι γενικότερα σημαντικά μειωμένες, γεγονός που φαίνεται την τελευταία χρόνια πως επηρεάστηκε από την οικονομική κρίση, την πανδημία και τον πόλεμο. Δηλαδή μία αγορά που θα μπορούσε να βρίσκεται με 250.000-280.000 καινούργια αυτοκίνητα τώρα βρίσκεται στις 100.000, αριθμός που είναι πολύ μακριά από το επιδιωκόμενο.

Τα μερίδια της ελληνικής αγοράς αυτοκινήτου (2022)

Κατάταξη	Μάρκα	Ταξινομήσεις	Μερίδιο αγορά (%)
1	TOYOTA	14.828	14,1
2	HYUNDAI	9.116	8,7
3	PEUGEOT	9.093	8,6
4	VW	7.664	7,3
5	OPEL	5.808	5,5
6	CITROEN	5.349	5,1
7	MERCEDES	4.791	4,6
8	FIAT	4.738	4,5
9	KIA	4.551	4,3
10	BMW	4.447	4,2

Πηγή: ΣΕΑΑ

Πίνακας 1 Πωλήσεις καινούργιων αυτοκινήτων στη Ελλάδα το 2022[5]

Παρακάτω, θα παρουσιαστεί ένας πίνακας με στοιχεία της ΣΕΑΑ[6] (Σύνδεσμος Εισαγωγέων Αντιπροσώπων Αυτοκινήτων) σε σχέση με τις ταξινομήσεις αυτοκινήτων στη Ελλάδα μόνο για το 2021, καθώς ακόμη δεν έχει αναρτηθεί σχετικός πίνακας για το 2022.

Πίνακας 4. Ταξινομήσεις Αυτοκινήτων ανά καύσιμο στην Ελληνική Αγορά για το 2021

Τύπος Καυσίμου/Ενέργειας	2021	ΜΕΡΙΔΙΟ	2020	Δ% 2021 vs 2020
Βενζίνη	49625	49,18%	43060	15,25%
Πετρέλαιο	17549	17,39%	22251	-21,13%
Υβριδικά (HEV)	23382	23,17%	11751	98,98%
Εναλλακτικής Ισχύος (APV)	3394	3,36%	1780	90,67%
Ηλεκτρικά Φορτιζόμενα (ECV)	6961	6,90%	2135	226,04%

Πηγή: www.seaa.gr

Πίνακας 2 Ταξινομήσεις καινούργιων αυτοκινήτων ανά καύσιμο στη Ελλάδα το 2021[6]

1.2 Κίνητρο και συνεισφορά

Λόγω της υπερβολικής πληροφορίας αλλά και του ανεπεξέργαστου όγκου δεδομένων που υπάρχει για τα αυτοκίνητα, εκ των πραγμάτων το πρώτο πράγμα που θα μπορούσε εύκολα να σκεφτεί κάποιος είναι να αναζητήσει και να κατηγοριοποιήσει συγκεκριμένα χαρακτηριστικά αυτοκινήτων, καθώς και να τα συσταδοποίησει. Όπως έχει προαναφερθεί και παραπάνω, πολλά εργοστάσια αδυνατούν να ανταπεξέλθουν στον αριθμό παραγγελιών από αγοραστές, με αποτέλεσμα να σημειώνονται πολύ μεγάλες καθυστερήσεις τόσο στην υλοποίηση, όσο και στην παράδοση ενός αυτοκινήτου. Με τα δεδομένα αυτά, ο καταναλωτής συνήθως βλέπει την τιμή εκκίνησης του αυτοκινήτου που επιθυμούσε να αυξάνεται ραγδαία και εύλογα προβληματίζεται. Με την αυξημένη ζήτηση που υπάρχει, αρκετοί επιλέγουν να απευθυνθούν σε μεταχειρισμένα αυτοκίνητα, που ωστόσο και εκεί οι τιμές έχουν αυξηθεί.

Στόχος της διπλωματικής είναι να συλλέξει τα πιο σημαντικά χαρακτηριστικά από όλα τα καινούργια αυτοκίνητα της ελληνικής αγοράς και βάσει αυτών μετέπειτα μπορούν να πραγματοποιηθούν αξιόλογες αλλαγές. Μέσω της εν λόγω πλατφόρμας(autotriti) συλλέχθηκαν τα δεδομένα-χαρακτηριστικά καινούργιων αυτοκινήτων και χρειάστηκε σε πολλές περιπτώσεις, λόγω της έλλειψης πληροφορίας που υπάρχει, να γίνει δυναμικά εισχώρηση τιμών με βάση παρόμοια χαρακτηριστικά διαφορετικών αυτοκινήτων. Συμπληρωματικά, γι' αυτά τα δεδομένα επειδή είναι διαφορετικής κλίμακας και δεν έπρεπε να επηρεαστούν αργότερα οι αλγόριθμοι μηχανικής μάθησης, επιβάλλεται να εφαρμοστεί κανονικοποίηση των δεδομένων.

Αρχικά, με την συσταδοποίηση πολλά αυτοκίνητα με αρκετά κοινά χαρακτηριστικά, όπως π.χ. κυβικά, ισχύς, ροπή, κατανάλωση, ανεξαρτήτως κατασκευαστή, χρειάστηκε να ενταχθούν σε μια κοινή συστάδα. Με τον ίδιο τρόπο, συστάθηκαν και άλλες συστάδες με πανομοιότυπα χαρακτηριστικά, πράγμα που βοηθά τον υποψήφιο αγοραστή, αφού δε θέλει να πληρώσει ακριβά ένα αυτοκίνητο (π.χ. mercedes), αλλά προτιμάει να πάρει ένα toyota με σχεδόν ίδια χαρακτηριστικά. Στη σύγχρονη εποχή, για πολλούς ανθρώπους η τιμή (κόστος) είναι ίσως και το πρωταρχικό, ανασταλτικό στοιχείο για μια απόφαση αγοράς. Στη συνέχεια, με την κατηγοριοποίηση μπορούμε να προβλέψουμε με συγκεκριμένα χαρακτηριστικά που διαθέτουμε, π.χ. τον τύπο καυσίμου σε ένα αυτοκίνητο, κι έτσι να καταλάβουμε αρχικά με τι καύσιμο λειτουργεί για να είναι συμβατό με τις ανάγκες-επιθυμίες του αγοραστή, αλλά και παράλληλα να διαπιστώσει ποια αυτοκίνητα είναι eco-friendly για το περιβάλλον.

Λαμβάνοντας υπόψη τα προβλήματα που προαναφέρθηκαν, δηλαδή τις σαρωτικές και ραγδαίες αλλαγές που έχουν συντελεστεί παγκοσμίως, είναι απολύτως φυσιολογικό αλλά και αναμενόμενο να διαμορφώνεται μια καινούργια πραγματικότητα σε πολλούς κλάδους, όπως

και στον κλάδο της αυτοκινητοβιομηχανίας με τραγικά αρνητικές συνέπειες. Γι' αυτόν τον λόγο προτιμήθηκε να χρησιμοποιηθεί ένα καινούργιο σύνολο δεδομένων που αφορά μόνο αυτοκίνητα που κυκλοφορούν στην ελληνική αγορά. Για να υλοποιηθεί όμως κάτι τέτοιο, χρειάζονται αλλαγές που προϋποθέτουν να τις έχει σκεφτεί ο δημιουργός, όπως ότι φαντάζει προφανές ότι πρέπει να προηγηθεί μια προεπεξεργασία, διότι δεν υπάρχει ολοκληρωμένο σύνολο δεδομένων.

Η συνεισφορά της διπλωματικής εργασίας συνοψίζεται στα εξής:

1. Μελετήθηκαν λεπτομερώς όλοι οι πιθανοί τρόποι εύρεσης συνόλου δεδομένων και συλλέχθηκαν με επιμέλεια κάποια χρήσιμα στοιχεία.. Καταλήξαμε στο συμπέρασμα ότι θα ήταν ιδιαίτερα ενδιαφέρουσα και όχι τόσο χρονοβόρα η δημιουργία ενός καινούργιου συνόλου δεδομένων, μέσω των οποίων θα μπορούσαν να υλοποιηθούν νέα πειράματα πάνω στα ίδια δεδομένα.
2. Μέσω της ιστοσελίδας autotriti και του πακέτου της python beautifulsoup καταφέραμε με επιτυχία να ανακτήσουμε δεδομένα και να δημιουργήσουμε το δικό μας πακέτο δεδομένων.
3. Όπως αναφέρθηκε και παραπάνω, χρειάστηκε να γίνει σε τιμές που δεν υπήρχαν imputation και κανονικοποίηση των δεδομένων.
4. Χρησιμοποιήθηκαν 3 βασικοί αλγόριθμοι για την συσταδοποίηση των δεδομένων.
5. Χρησιμοποιήθηκαν 2 βασικοί αλγόριθμοι για την κατηγοριοποίηση των δεδομένων και την αυτόματη πρόβλεψη καυσίμου. Επίσης, σε άλλο πείραμα δοκιμάστηκαν μόνο οι στήλες που έχουν σχέση με την απόδοση του αυτοκινήτου.
6. Με όλες τις παραπάνω ενέργειες καταλήξαμε σε χρήσιμα συμπεράσματα.

1.4 Οργάνωση κειμένου

Εδω θα διαβάσετε μια μικρή περίληψη σχετικά με τι θα επακολουθήσει στα επόμενα κεφάλαια της εργασίας για να μπορέσετε να καταλάβετε ακριβώς με ποια σειρά αναπτύχθηκε η εργασία αλλα και χωρίς αυτό να διαβάσετε κάτι συγκεκριμένο που επιθυμείτε.

Αρχικά για το πρώτο κεφάλαιο που διαβασετε ήδη περιέχει μια μικρή εισαγωγή στο θέμα αλλα και στις δυσκολίες που υπάρχουν στον κλάδο της αυτοκινητοβιομηχανίας.

Στο 2ο κεφάλαιο θα προσεγγίσουμε θεωρητικά την μεθοδολογία αλλά και χρήσιμα εργαλεία για την καλύτερη κατανόηση του θέματος αλλά και για ποιο λόγο χρησιμοποιήθηκαν οι τεχνολογίες και οι αλγόριθμοι που επιλέχθηκαν να χρησιμοποιηθούν.

Στο 3ο κεφάλαιο με την ίδια λογική πάλι θα προσεγγιστούν θεωρητικά οι τεχνολογίες που χρησιμοποιήθηκαν για την ανάπτυξη του project οι οποίες γενικότερα χρησιμοποιούνται κατά κόρον στον κόσμο της τεχνητής νοημοσύνης και γενικότερα της επιστήμης.

Στο 4ο κεφάλαιο θα αναφερθεί ο τρόπος επιλογής της πλατφόρμας από την οποία ανακτήθηκαν τα δεδομένα και με ποιον τρόπο, δηλαδη τι λογισμικό. Στη συνέχεια θα φανεί η επιλογή κάποιων συγκεκριμένων χαρακτηριστικών από την σελίδα αυτών που θεωρήθηκαν τα πιο σημαντικά. Επίσης θα γίνει και μια τεκμηρίωση στο πως ακριβώς υλοποιήθηκε αυτο το λογισμικό. Τέλος αφου εξηγηθούν τα παραπάνω θα παρουσιαστει το σύνολο δεδομένων που δημιουργήθηκε.

Στο 5ο κεφάλαιο αφου έχει δημιουργηθεί το σύνολο δεδομένων με την βοήθεια των αλγορίθμων κατηγοριοποίησης αλλά και συσταδοποίησης θα παρουσιαστούν αποτελέσματα από τα πειράματα. Για κάθε ένα αλγόριθμο αυτό θα γίνει ξεχωριστά.

Στο 6ο κεφάλαιο θα καταλήξουμε, σύμφωνα με τα αποτελέσματα που πήραμε απο τους αλγόριθμους, σε χρήσιμα συμπεράσματα.

2 *Αλγόριθμοι συσταδοποίησης και κατηγοριοποίησης*

2.1 Συσταδοποίηση

Τον τελευταίο καιρό η αυτοματοποίηση της αναζήτησης και συλλογής πληροφοριών κατέληξε να καταλαμβάνει έναν τεράστιο όγκο πληροφοριών για αρκετά διαφορετικά είδη συστημάτων. Επομένως, αρκετοί τρόποι υλοποίησης έχουν δημιουργηθεί με σκοπό την οργάνωση και τη μοντελοποίηση. Τέτοιοι τρόποι υλοποίησης έχουν ως στόχο τη μετουσίωσή τους στη διάγνωση, την εκπαίδευση, την πρόβλεψη, και πολλούς άλλους κλάδους. Ο ορισμός, η αξιολόγηση και η εφαρμογή αυτών των τρόπων υλοποίησης αποτελούν αναπόσπαστο κομμάτι του τομέα της μηχανικής μάθησης, η οποία τελικά εξελίχθηκε σε ένα σοβαρό υποτομέα της επιστήμης των υπολογιστών και της στατιστικής, καθ' όσον έπαιξε πρωτεύοντα ρόλο στην εξέλιξη της ανθρωπότητας. Η μηχανική μάθηση περιέχει διάφορα πεδία όπως ανάλυση παλινδρόμησης, συσταδοποίηση, κατηγοριοποίηση και μεθόδους επιλογής χαρακτηριστικών. Η κατηγοριοποίηση χωρίς επίβλεψη, δηλαδή η συσταδοποίηση, χρησιμοποιείται για να μπορούν να μπουν τα δεδομένα σε κλάσεις χωρίς να γνωρίζουν την ετικέτα τους.

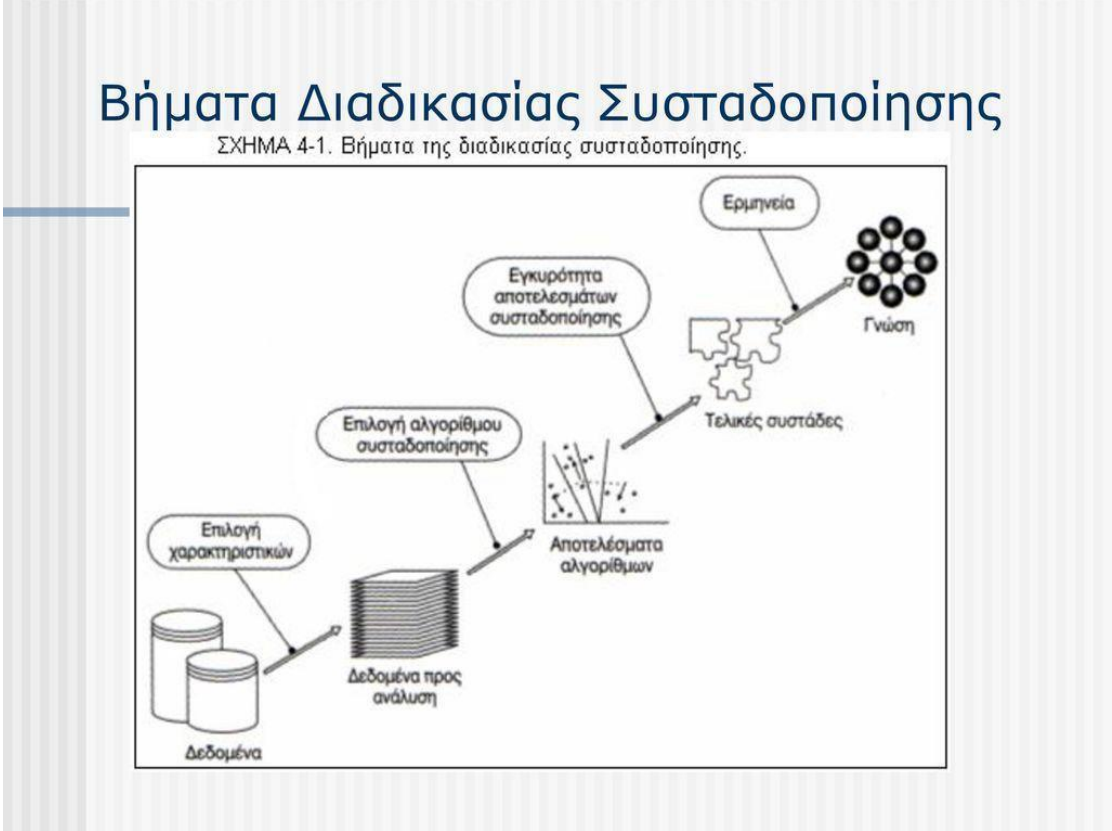
Η συσταδοποίηση αντικειμένων είναι απαραίτητη για ποικίλους λόγους σε διαφορετικούς κλάδους της μηχανικής, της επιστήμης και της τεχνολογίας, των ανθρωπιστικών επιστημών,

της ιατρικής επιστήμης, ακόμη και της καθημερινότητάς μας[8]. Ας υποθέσουμε ότι κάποιοι άνθρωποι που ταλαιπωρούνται από κάποια αρρώστια, παρουσιάζουν κάποια παρεμφερή συμπτώματα και μπαίνουν σε μια συστάδα με ετικέτα τις περισσότερες φορές το όνομα της ασθένειας. Τα άτομα που δεν παρουσιάζουν παρόμοια συμπτώματα θα μπουν εύλογα σε άλλη συστάδα. Ουσιαστικά, συσταδοποίηση είναι η διαδικασία όπου ένα σύνολο οντοτήτων (ή εγγραφών) χωρίζεται σε πολλές συστάδες και τα άτομα κάθε συστάδας είναι «παρόμοια» μεταξύ τους, αλλά ταυτόχρονα διαφέρουν από τα άτομα των άλλων συστάδων. Η διαδικασία της συσταδοποίησης διαμορφώνεται από τα παρακάτω στάδια. Τονίζεται ότι πριν την υλοποίηση της διαδικασίας χρίζεται αναγκαία η προεπεξεργασία των δεδομένων συσταδοποίησης[9].

1)Διαλέγουμε τα πιο σημαντικά χαρακτηριστικά γνωρίσματα που μπορεί να χρησιμοποιήσουμε στην συσταδοποίηση, με κίνητρο την επίτευξη της καλύτερης ομοιογένειας σε κάθε συστάδα.

2) Επιλέγουμε το βέλτιστο αλγόριθμο συσταδοποίησης για το σύνολο δεδομένων με σκοπό την επίτευξη ενός καλού σχήματος συσταδοποίησης με βάση τα δεδομένα μας.

3) Μετρώνται τα παραγόμενα αποτελέσματα των αλγορίθμων συσταδοποίησης με βέλτιστες προδιαγραφές επικύρωσης. Τέλος, ερμηνεύονται και παρουσιάζονται τα εξαγόμενα αποτελέσματα τα οποία προέκυψαν από την παραπάνω διαδικασία συσταδοποίησης.



Εικόνα 1 Βήματα Διαδικασίας Συσταδοποίησης [9]

Το Webster εκφράζει την πεποίθηση ότι η συσταδοποίηση[8] είναι «μια στατιστική μέθοδος κατηγοριοποίησης για να συμπεράνουμε αν κάποιο από τα μοτίβα εισχωρούν σε διάφορες συστάδες δημιουργώντας ποσοτικές συγκρίσεις διάφορων χαρακτηριστικών». Η ομοιότητα είναι ο βασικός μοχλός μιας συστάδας, παρομοίως και στη διαδικασία συσταδοποίησης. Συνήθως, ο αριθμός των συστάδων που θα φτιαχτούν, ορίζεται από τον χρήστη, αφού υπάρχουν διαθέσιμα αποκλειστικά δεδομένα αριθμητικού τύπου για την απεικόνιση χαρακτηριστικών από τα μοτίβα σε μια συστάδα. Ένας τρόπος εξαγωγής πληροφορίας που βασίζεται στη σχέση ανάμεσα στα πρότυπα, θα μπορούσε να είναι η δημιουργία της αριθμητικής τιμής. Απεικονίζονται τα στοιχεία των αντικειμένων με αριθμητικές τιμές. Μια προσέγγιση για τον ορισμό της ομοιότητας, συμπεριλαμβάνει ως μέτρο της απόστασης μεταξύ των μοτίβων, όσο μικρότερη είναι η απόσταση (π.χ. Ευκλείδεια απόσταση) ανάμεσα σε δύο αντικείμενα, τόσο μεγαλύτερη θα είναι η ομοιότητα και το αντίστροφο.

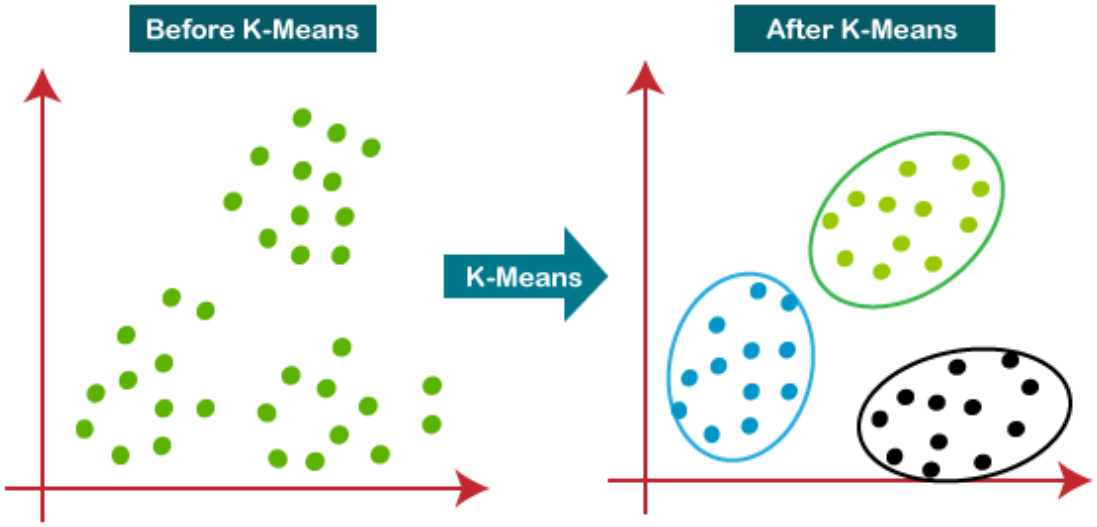
Οι μέθοδοι συσταδοποίησης είναι κατά κύριο λόγο πιο δύσκολες από τις εποπτευόμενες προσεγγίσεις, αλλά προσφέρουν αρκετές πληροφορίες με βάση πολύπλοκα δεδομένα[10].

Ουσιαστικά, κάθε μέθοδος συσταδοποίησης χρησιμοποιείται με άλλο τρόπο και συνήθως φέρνει αποτελέσματα που δεν προκύπτουν με άλλη μέθοδο. Η επιλογή μιας μεθόδου συσταδοποίησης γίνεται με βάση διάφορους παράγοντες. Η κατανάλωση πόρων διαφέρει μεταξύ των διαφόρων μεθόδων, και ορισμένες μέθοδοι δεν ταιριάζουν ούτε χρησιμεύουν για όλα τα σύνολα δεδομένων. Κάποιες μέθοδοι είναι αμιγώς πιο βέλτιστες στον καθορισμό τύπων συστάδων. Η ποιότητα αξιολόγησης των μεθόδων συσταδοποίησης έχει σχέση με το μέτρο ομοιότητας που λαμβάνεται υπόψη, αλλά και από τον τρόπο με τον οποίο εισέρχεται διαφορετικός τύπος χαρακτηριστικών που υπάρχουν στα δεδομένα όπως: κατηγορικά, αριθμητικά, δυαδικά. Τα αποτελέσματα που λαμβάνονται από ορισμένες μεθόδους θα πρέπει να μπουν με την σωστή σειρά με την οποία διαβάζονται ή επεξεργάζονται τα αρχεία δεδομένων. Αυτό συμβαίνει περισσότερο σε ορισμένες μη ιεραρχικές μεθόδους, όπου κάποιες μεγάλες συστάδες προσπαθούν να δημιουργηθούν στην αρχή της διαδικασία και τελικά φαίνονται πολύ μικρότερες κατά την επεξεργασία των τελευταίων εγγραφών.

2.1.1 K-Means

Ο αλγόριθμος k-means είναι ευρέως γνωστός και παράλληλα απλοϊκός αλγόριθμος, καθώς υπήρξε από τους πρώτους αλγορίθμους συσταδοποίησης όπου κυρίως χρησιμοποιείται για να επιλύσει ζητήματα συσταδοποίησης. Στην τμηματική συσταδοποίηση τα δεδομένα εισέρχονται σε k-cluster δίχως ιεραρχική δομή, βελτιστοποιώντας την συνάρτηση κριτηρίου. Το κριτήριο που χρησιμοποιείται κατά κόρον είναι η Ευκλείδεια απόσταση[8], που ανιχνεύει την μικρότερη απόσταση ανάμεσα σε σημεία με όλες τις υπάρχοντες συστάδες και τοποθετεί το σημείο στη συστάδα.

Ο αλγόριθμος K-means τοποθετεί κάθε σημείο στη συστάδα του οποίου το κέντρο είναι το πιο κοντινό. Το κέντρο είναι ο μέσος όρος όλων των σημείων ανά συστάδα και οι συντεταγμένες είναι ο αριθμητικός μέσος όρος για κάθε διάσταση χωριστά σε όλα τα σημεία της συστάδας[11]. Η λογική του K-means είναι η πληροφόρηση του κέντρου της συστάδας η οποία εκπροσωπείται από το σημεία με επαναληπτικό υπολογισμό και η επαναληπτική διαδικασία θα συνεχιστεί μέχρι κάποια πληρούνται τα κριτήρια σύγκλισης.



Εικόνα 2 K-means αλγόριθμος [12]

Άθροισμα τετραγώνου σφάλματος (SSE)

Το SSE[8] είναι το απλούστερο και πιο ευρέως χρησιμοποιούμενο μέτρο κριτηρίου για συσταδοποίηση. Το SSE είναι το μέτρο της απόστασης όλων των σημείων από το κέντρο των συστάδων τους Υπολογίζεται ως:

$$SSE = \sum_{k=1}^{K} \sum_{\forall x_i \in C_k} \|x_i - \mu_k\|^2$$

όπου Ck είναι το σύνολο των περιπτώσεων των συστάδων k. Το μk είναι ο διανυσματικός μέσος όρος της συστάδας k.

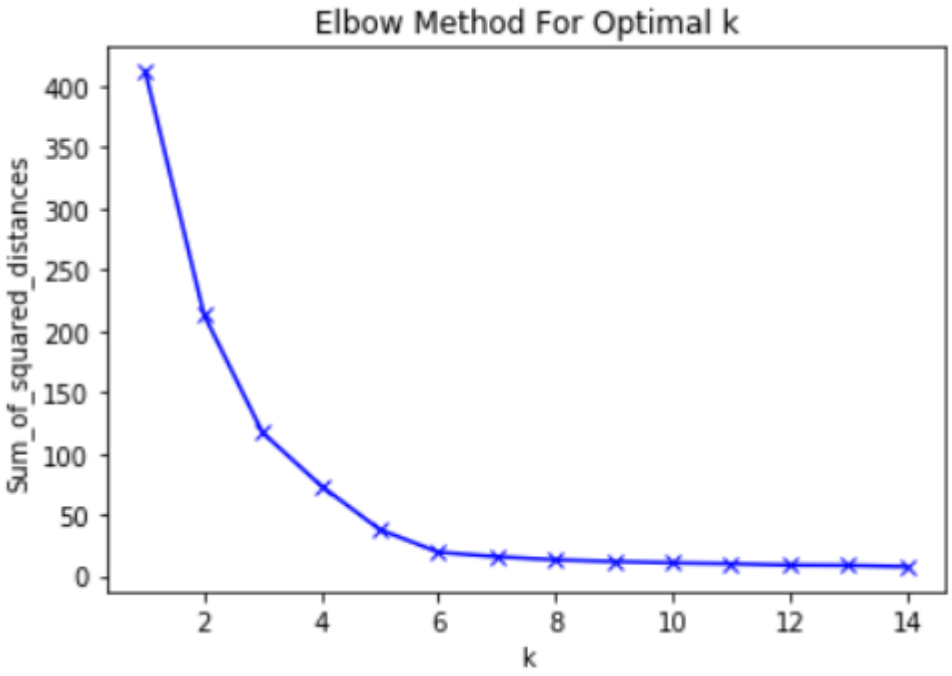
Η συνάρτηση κριτηρίου SSE είναι κατάλληλη για περιπτώσεις στις οποίες οι συστάδες σχηματίζουν συμπαγή σύννεφα που είναι καλά διαχωρισμένα το ένα από το άλλο

Μέθοδος αγκώνα(Elbow curve)

Η μέθοδος του αγκώνα είναι μια γνωστή μέθοδος κατανοητή και εύχρηστη στην εφαρμογή που χρησιμοποιείται για να οριστεί ο κατάλληλος αριθμός συστάδων των δεδομένων σε ένα σύνολο δεδομένων. Εφαρμόζεται για τον προσδιορισμό της παραμέτρου k. Υλοποιώντας τον αλγόριθμο συσταδοποίησης k-means, η μέθοδος του αγκώνα κάνει τις κατάλληλες τροποποιήσεις και επιλέγει την καμπύλη του αγκώνα για να πάρει τον βέλτιστο αριθμό συστάδων. Η λογική του είναι πως όσο μεγαλώνει ο αριθμός των συστάδων, το άθροισμα τετραγωνικού σφάλματος (SSE) μειώνεται. Η μέθοδος elbow χρησιμοποιεί το SSE και τον

αριθμο των συστάδων για να μπορέσει να καταλήξει στη καλύτερη επιλογή του k. Το σημείο που πραγματοποιείται αυτό αποκαλείται σημείο του αγκώνα και εφαρμόζεται για τον ορισμό του βέλτιστου αριθμού συστάδων.

Είναι απαραίτητο να ειπωθεί πως η μέθοδος αγκώνα αρκετές φορές δεν είναι η πιο ακριβής μέθοδος για τον ορισμό του βέλτιστου αριθμού συστάδων και το σημείο του αγκώνα είναι πιθανό να μην μπορεί κάποιος να το διακρίνει, ειδικά όταν τα δεδομένα είναι θορυβώδη ή όταν οι συστάδες έχουν διαφορετικές πυκνότητες. Συνήθως χρησιμοποιείται σε περιπτώσεις με μικρό αριθμό συστάδων. Όπως φαίνεται και στο παρακάτω γράφημα(Εξίσωση 2) δύο καλές τιμές για το k είναι το 3 και το 5. Όμως η βέλτιστη τιμή για τον κατάλληλο προσδιορισμό των συστάδων είναι το 3 διότι εκεί πραγματοποιείται η μεγαλύτερη καμπύλη.



Εξίσωση 2 Γράφημα Elbow curve(Μέθοδος αγκώνα)[14]

2.1.2 DBSCAN

Οι αλγόριθμοι συσταδοποίησης που βασίζονται στην πυκνότητα αποτελούνται από δύο σημαντικά χαρακτηριστικά[15], την ανθεκτικότητα τους σε ζητήματα εμφάνισης θορύβου και ακραίων τιμών, αλλά και τη δυνατότητα εύρεσης διαφόρων συχνοτήτων στο σύνολο των σημείων και δημιουργούνται για την ανάπτυξη συστάδων αυθαίρετου σχήματος. Γι' αυτό το λόγο, μια συστάδα αναπαρίσταται ως μια περιοχή, όπου η πυκνότητα των αντικειμένων

στοιχείων περνάει ένα όριο. Από τους πιο γνωστούς και ευρέως χρησιμοποιούμενους αλγορίθμους συσταδοποίησης είναι η χωρική συσταδοποίηση εφαρμογών με θόρυβο σύμφωνα με την πυκνότητα (DBSCAN)

Ο DBSCAN είναι ο πρώτος αλγόριθμος συσταδοποίησης σύμφωνα με την πυκνότητα. Ο αλγόριθμος DBSCAN αναπτύχθηκε πρώτη φορά από τον Ester[16] το 1996. Οι συστάδες προβάλλονται παρακολουθώντας την πυκνότητα των σημείων. Περιοχές με μια μεγάλη πυκνότητα σημείων αναπαριστώνται όταν υπάρχουν συστάδες, ενώ οι περιοχές με μικρή πυκνότητα σημείων απεικονίζουν συστάδες θορύβου ή ακραίες τιμές. Ο DBSCAN δημιουργήθηκε για να μπορέσει να διαχειριστεί μεγάλο όγκο δεδομένων, με θόρυβο, και μπορεί να καταλάβει πότε οι συστάδες έχουν διαφορετικά μεγέθη και σχήματα.

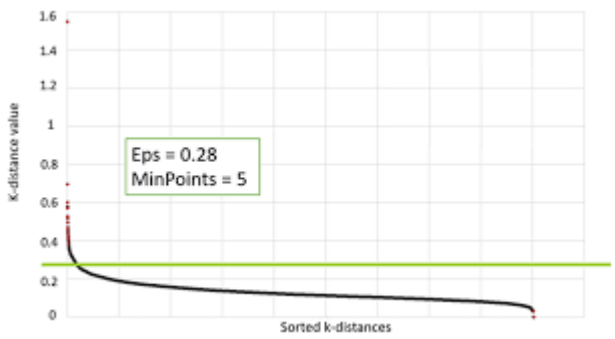
Η λογική του DBSCAN[11] είναι ότι, για όλα τα σημεία μιας συστάδας, η γειτονιά μιας ακτίνας πρέπει να περιλαμβάνει το ελάχιστο ένα αριθμό πόντων, επομένως η πυκνότητα στη γειτονιά χρειάζεται να περάσει αυτό το όριο. Παρ' όλα αυτά, ο αλγόριθμος δε λειτουργεί αποδοτικά με συστάδες διαφορετικών πυκνοτήτων. Αυτός ο αλγόριθμος έχει δύο παραμέτρους εισόδου:

- Eps, οριοθετεί την μέγιστη απόσταση δύο σημείων που ανήκουν στη ίδια γειτονιά

- MinPts, ο ελάχιστος αριθμός σημείων για να δημιουργηθεί μια συστάδα.

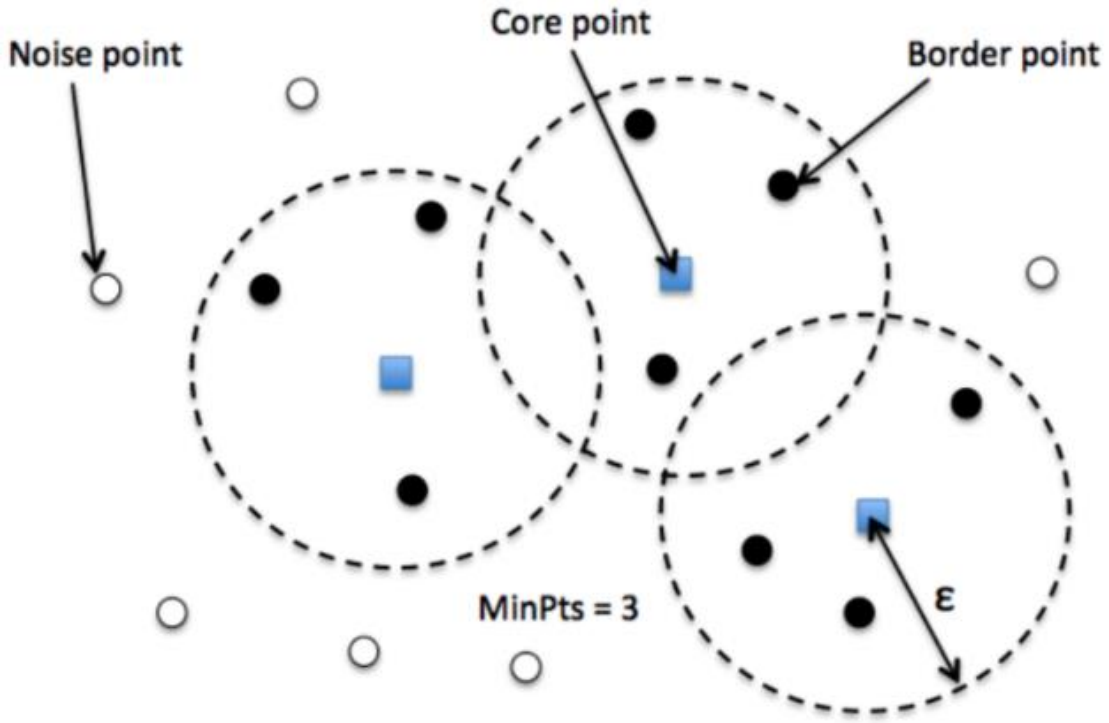
Για να οριστεί η τιμή MinPts για το DBSCAN[17] δεν είναι αναγκαίο να γνωρίζει ο χρήστης το θέμα αλλά είναι και να κατάλληλα εξοικειωμένος με το σύνολο δεδομένων. Για δισδιάστατα δεδομένα, δίνεται η προεπιλεγμένη τιμή MinPts = 4 του DBSCAN (Ester et al., 1996).

Καθώς διαλέξετε την τιμή MinPts, έχετε την δυνατότητα να υπολογίσετε το ε. Μια τεχνική για τον αυτόματο υπολογισμό της βέλτιστης τιμής ε βρίσκει τη μέση απόσταση ανάμεσα στα σημεία και των k πλησιέστερων γειτόνων τους, όπου k = η τιμή MinPts. Έπειτα, οι μέσες αποστάσεις k αναπαριστώνται με αύξουσα σειρά σε ένα γράφημα k-απόστασης. Όπως διακρίνεται και στο παρακάτω γράφημα τα ελάχιστα σημεία για να σχηματιστεί μια συστάδα είναι 5 και η καμπύλη στο γράφημα όπως φαίνεται γίνεται στη τιμή 0.28



Εξίσωση 3 Παράμετροι Dbscan[18]

Η διαδικασία συσταδοποίησης στηρίζεται στην κατηγοριοποίηση των στοιχείων στο σύνολο δεδομένων ως στοιχεία πυρήνα, στοιχεία συνόρων και στοιχεία θορύβου. Στοιχεία πυρήνα είναι τα στοιχεία τα οποία έχουν σε ακτίνα eps τουλάχιστον minpts γείτονες. Τα στοιχεία συνόρων είναι αυτά που δεν είναι πυρήνα αλλά ανήκουν στην γειτονία ενός στοιχείου πυρήνα. Στοιχεία θόρυβου είναι αυτα που δεν είναι ούτε πυρήνα ούτε συνόρου.

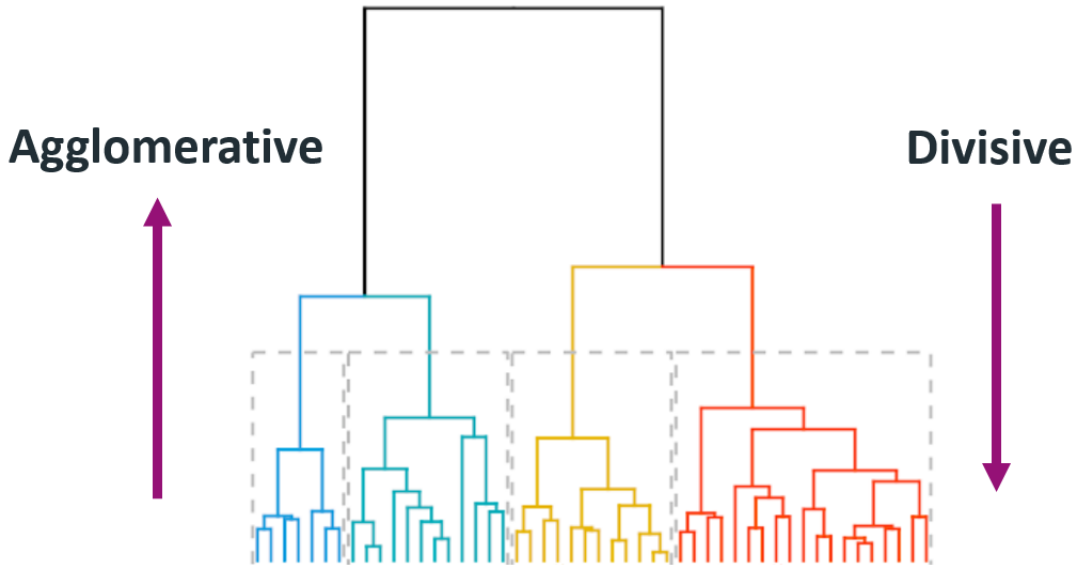


Εικόνα 3 Αναπαράσταση σημείων DBSCAN[19]

2.1.3 Ιεραρχική συσταδοποίηση

Οι αλγόριθμοι ιεραρχικής συσταδοποίησης ορίζουν τα δεδομένα σε μια ιεραρχική δομή με βάση τον πίνακα εγγύτητας[20]. Τα αποτελέσματα της ιεραρχικής συσταδοποίησης κατά κύριο λόγο αναπαριστώνται με ένα δυαδικό δέντρο ή δενδρογράφημα. Ο ριζικός κόμβος του δενδρογράμματος περιλαμβάνει όλα τα δεδομένα και κάθε κόμβος φύλλου είναι αντικείμενο δεδομένων. Οι ενδιάμεσοι κόμβοι δείχνουν τον χώρο όπου τα αντικείμενα υπάρχουν κοντά μεταξύ τους και το ύψος του δενδρογράμματος αναπαριστά την απόσταση ανάμεσα σε όλα τα

ζεύγη αντικειμένων ή συστάδων, ή ένα αντικείμενο και μια συστάδα. Τα ακριβή αποτελέσματα συσταδοποίησης διακρίνονται με την κοπή του δενδρογράμματος σε διάφορα επίπεδα. Αυτή η απεικόνιση προσφέρει αρκετά συγκεκριμένη ανάλυση και οπτικοποίηση για τις πιθανές δομές συσταδοποίησης δεδομένων. Οι αλγόριθμοι ιεραρχικής συσταδοποίησης κατηγοριοποιούνται ως συσσωρευτικές και διαιρετικές μέθοδοι. Η συσσωρευτική συσταδοποίησης αρχίζει με συστάδες και όλα έχουν μόνο ένα αντικείμενο. Μετά αρχίζουν οι πράξεις συγχώνευσης και βάζουν όλα τα αντικείμενα στην ίδια συστάδα. Η διαιρετική συσταδοποίησης αρχίζει αντίθετα. Όταν αρχίζει, όλα τα δεδομένα βρίσκονται σε μια συστάδα και μια διαδικασία το διαιρεί διαδοχικά, μέχρι όλες οι συστάδες να είναι μονότονες. Μια συστάδα με Ν αντικείμενα, συνήθως είναι πολύ ακριβό για να υλοποιηθεί. Οπότε, η διαιρετική συσταδοποίησης σε πολλές περιπτώσεις δεν προτιμάται.



Εικόνα 4 Ιεραρχική Συσταδοποίηση [21]

Σύμφωνα με διαφορετικούς ορισμούς για την απόσταση μεταξύ δύο συστάδων, χρησιμοποιούνται αρκετοί αλγόριθμοι συσσωρευτικής συσταδοποίησης. Οι πιο απλοϊκές, αλλά και παράλληλα πιο γνωστές μέθοδοι περιέχουν μονή και πλήρη τεχνική σύνδεσης. Για τη μονή μέθοδο σύνδεσης, η απόσταση ανάμεσα σε δύο συστάδες ανακαλύπτεται από τα δύο πιο κοντινά αντικείμενα των διαφορετικών συστάδων, οπότε αποκαλείται μέθοδος κοντινότερου γείτονα. Από την άλλη μεριά, η μέθοδος πλήρους σύνδεσης εφαρμόζει την πιο μακρινή απόσταση ενός ζεύγους αντικειμένων για τον ορισμό της απόστασης μεταξύ συστάδων. Ο αλγόριθμος πλήρους σύνδεσης[8] εξάγει στενά συνδεδεμένες ή συμπαγής συστάδες. Ο αλγόριθμος μονής σύνδεσης εξάγει συστάδες που είναι στραβές ή επιμήκεις.

Φαίνεται πως ο πλήρης αλγόριθμος σύνδεσης εξάγει πιο σημαντικές ιεραρχίες σε πολλές εφαρμογές από το αλγόριθμο μονής σύνδεσης. Αρκετά πιο πολύπλοκοι αλγόριθμοι συσσωρευτικής συσταδοποίησης, είναι η μέση σύνδεση, η διάμεση σύνδεση, η κεντροειδής σύνδεση και η μέθοδος του Ward και έχουν την ευχέρεια να αναπτυχθούν διαλέγοντας τους καλύτερους συντελεστές. Η μέθοδος Ward[9] ξεχωρίζει από τις υπόλοιπες μεθόδους και είναι φτιαγμένη με σκοπό την ελαχιστοποίηση της διακύμανσης μέσα στις συστάδες. Η μέθοδος αυτή αναπτύσσει συστάδες με περίπου ίδιο αριθμό παρατηρήσεων και αναγνωρίζεται από αρκετά χρήσιμες ιδιότητες. Σήμερα χρησιμοποιείται σε πολλές πρακτικές εφαρμογές, όπως και στην παρούσα διπλωματική. Η μονή σύνδεση, η πλήρης σύνδεση και η μέση σύνδεση παίρνουν κάθε σημείο ενός ζεύγους συστάδων, όταν μετριέται η απόσταση ανάμεσα στις συστάδες τους, και ονομάζονται μέθοδοι γραφήματος. Οι άλλες αποκαλούνται γεωμετρικές μέθοδοι, αφού παίρνουν γεωμετρικά κέντρα για να απεικονίσουν συστάδες και να ορίσουν τις αποστάσεις τους.

Οι κλασικοί αλγόριθμοι ιεραρχικής συσταδοποίησης έχουν το πρόβλημα της έλλειψης στιβαρότητας και επηρεάζονται αρκετά από τον θόρυβο και τα ακραία σημεία. Η υπολογιστική πολυπλοκότητα για τους περισσότερους αλγόριθμους ιεραρχικής συσταδοποίησης είναι ακριβή και γι' αυτόν τον λόγο η εφαρμογή «ζορίζεται» σε σύνολα δεδομένων μεγάλης κλίμακας. Άλλα μειονεκτήματα ιεραρχικής συσταδοποίησης είναι η επιθυμία σχηματισμού σφαιρικών σχημάτων και το φαινόμενο αναστροφής, στο οποίο η κανονική αρχική δομή παραμορφώνεται. Οι ιεραρχικοί αλγόριθμοι είναι πιο χρήσιμοι από τους επιμέρους αλγόριθμους. Για παράδειγμα, ο αλγόριθμος συσταδοποίησης μονής σύνδεσης [20]είναι αποδοτικός σε σύνολα δεδομένων που περιλαμβάνουν αρκετά διαχωρισμένες, αλυσιδωτές και ομόκεντρες συστάδες. Από την άλλη η πολυπλοκότητα του χρόνου και ο χώρος των επιμέρους αλγορίθμων είναι μικρότερα συγκριτικά με ιεραρχικούς αλγόριθμους. Μπορούν όμως τα σημαντικά γνωρίσματα αυτών των δύο κατηγοριών να δημιουργήσουν ένα υβριδικό αλγόριθμο.

2.2 Κατηγοριοποίηση

Η λειτουργικότητα και η ακρίβεια των αλγορίθμων εξόρυξης δεδομένων είναι ένα αρκετά σοβαρό ζήτημα που χρήζει περισσότερης έρευνας και γι' αυτό το λόγο έχει τραβήξει την προσοχή του ακαδημαϊκού χώρου και της βιομηχανίας[22]. Για την εξόρυξη δεδομένων

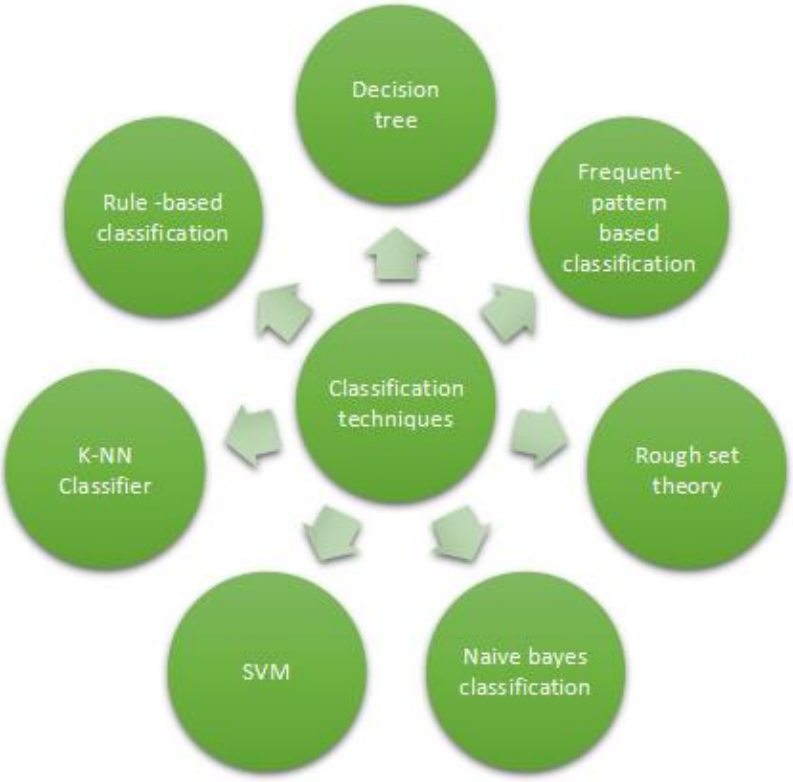
βασική προϋπόθεση είναι η κατηγοριοποίηση. Η κατηγοριοποίηση παίρνει κάθε στοιχείο από ένα σύνολο δεδομένων σε μία από προκαθορισμένες κλάσεις ή κατηγορίες. Η κατηγοριοποίηση των εργασιών ανάλυσης δεδομένων ενός μοντέλου ή ενός κατηγοριοποιητή δημιουργείται για να μαντέψει κατηγορικές ετικέτες (τα χαρακτηριστικά ετικέτας κλάσης). Η κατηγοριοποίηση[23] χρησιμοποιείται σαν συνάρτηση εξόρυξης δεδομένων, που εισάγει στοιχεία σε μια συλλογή για πρόβλεψη κατηγοριών ή κλάσεων. Ο σκοπός της κατηγοριοποίησης είναι να στοχεύσει με ακρίβεια τη κλάση στόχου κάθε φορά. Οι εργασίες κατηγοριοποίησης έχουν την δυνατότητα να εφαρμόσουν οποιαδήποτε στρατηγική. Εάν τα στοιχεία εκπαίδευσης κατηγοριοποιούνται χωρίς να γνωρίζουμε τα δεδομένα εκπαίδευσης αυτό ονομάζεται κατά προτεραιότητα. Από την άλλη πλευρά αν τα δεδομένα κατηγοριοποιήθηκαν με βάση των δεδομένων εκπαίδευσης αυτό ονομάζεται εκ των υστέρων κατηγοριοποίηση.

Στην κατηγοριοποίηση[24], τα δεδομένα τα γνωρίζουμε εξ αρχής και είναι στηριζόμενα σε υποθέσεις. Η κατηγοριοποίηση περιέχει την στόχευση ενός ορισμένου αποτελέσματος σύμφωνα με μια συγκεκριμένη εισροή. Για να μαντέψουμε το αποτελέσματα, πρέπει να αποκτήσουμε τα δεδομένα που μπορούμε να συλλέξουμε. Αυτά τα δεδομένα κατηγοριοποιούν τις εγγραφές. Οι πηγές δεδομένων χωρίζονται σε σετ εκπαίδευσης και σετ δοκιμών. Το σετ εκπαίδευσης περιέχει τα δεδομένα που κατηγοριοποιούνται πριν και χρησιμοποιείται για σκοπούς κατηγοριοποίησης. Τα χαρακτηριστικά μπορούν να στοχεύουν σε αποτελέσματα. Έπειτα, τα δεδομένα δοκιμής εισέρχονται στον αλγόριθμο και ελέγχονται με το χαρακτηριστικό που είναι αποθηκευμένο από πρίν και σύμφωνα με αυτές τις εκδοχές κατηγοριοποιούνται. Ο αλγόριθμος αναλύει τα δεδομένα δίνεται και προβλέπει τα αποτελέσματα.

Δύο κύριες κατηγορίες αλγορίθμων έχουν οι κατηγοριοποιητές, τους πρόθυμους και τους οκνηρούς κατηγοριοποιητές. Και οι δύο έχουν τον ίδιο σκοπό, να μπορέσουν να μαντέψουν σωστά, παρά ταύτα λειτουργούν διαφορετικά. Το διαθέσιμο σετ εκπαίδευσης επηρεάζει την ακρίβεια του αλγορίθμου. Ένας πρόθυμος κατηγοριοποιητής[23] προ-επεξεργάζεται τα διαθέσιμα δεδομένα εκπαίδευσης και φτιάχνει ένα μοντέλο κατηγοριοποίησης, όπου εφαρμόζεται για την κατηγοριοποίηση νέων, μη κατηγοριοποιημένων ειδών. Από την άλλη πλευρά, οι οκνηροί κατηγοριοποιητές δε δημιουργούν κανένα μοντέλο κατηγοριοποίησης. Παίρνουν ως μοντέλο κατηγοριοποίησης το σύνολο δεδομένων εκπαίδευσης. Ένας οκνηρός αλγόριθμος κατηγοριοποιεί ένα καινούργιο αντικείμενο εξετάζοντας το σετ εκπαίδευσης. Αφού οι πρόθυμοι κατηγοριοποιητές αναπτύσσουν ένα μοντέλο κατηγοριοποίησης πριν την εμφάνιση οποιουδήποτε καινούργιου στοιχείου, η διαδικασία κατηγοριοποίησης είναι πολύ

γρήγορη. Οι οκνηροί κατηγοριοποιητές δε σπαταλούν χρόνο για την δημιουργία μοντέλων κατηγοριοποίησης, απλώς η διαδικασία κατηγοριοποίησης τους είναι αρκετά πιο αργή σε σχέση με τους πρόθυμους κατηγοριοποιητές. Ένα μειονέκτημα των πρόθυμων κατηγοριοποιητών είναι ότι χρειάζεται να κατασκευάσουν μια γενική υπόθεση ότι περιλαμβάνεται όλο το εκπαιδευτικό σετ. Όταν γίνεται κάτι τέτοιο είναι πιθανό να μεταβληθεί η ακρίβεια κατηγοριοποίησης και να κριθεί η κατασκευή του μοντέλου αρκετά χρονοβόρα και πολύπλοκη εργασία προεπεξεργασίας. Από την άλλη πλευρά, οι οκνηροί κατηγοριοποιητές έχουν ολόκληρο το σετ προπόνησης γι' αυτό έχουν την ευχέρεια να προσθέσουν πιο δύσκολες υποθέσεις στα δεδομένα. Οπότε, μπορούν να καλυτερέψουν την ακρίβειας κατηγοριοποίησης. Ένα μειονέκτημα των οκνηρών κατηγοριοποιητών είναι ότι αναγκάζουν όλα τα δεδομένα εκπαίδευσης να είναι πάντα διαθέσιμα, πράγμα που σημαίνει μεγάλες απαιτήσεις αποθήκευσης. Η πρόθυμη κατηγοριοποίηση, μετά την κατασκευή του μοντέλου κατηγοριοποίησης, μπορεί να σβήσει τα δεδομένα εκπαίδευσης για να μην έχει πρόβλημα χώρου. Τον τελευταίο καιρό, το πρόβλημα της κατηγοριοποίησης δείχνει να απασχολεί αρκετούς επιστήμονες από διάφορα ερευνητικά πεδία της επιστήμης των υπολογιστών.

Τα δέντρα απόφασης είναι ευρέως αναγνωρισμένη υποκατηγορία πρόθυμων κατηγοριοποιητών. Τα διαθέσιμα δεδομένα εκπαίδευσης αναπτύσσουν μια δομή δέντρου που χρησιμοποιείται για την κατηγοριοποίηση νέων στοιχείων. Άλλοι πρόθυμοι κατηγοριοποιητές στηρίζονται σε τεχνητά νευρωνικά δίκτυα. Ένα νευρωνικό δίκτυο εκπαιδεύεται από τα εκπαιδευτικά αντικείμενα και στη συνέχεια υλοποιεί κατηγοριοποιήσεις. Οι πιθανολογικοί αλγόριθμοι κατηγοριοποίησης ανήκουν επίσης στην κατηγορία των πρόθυμων κατηγοριοποιητών. Χτίζουν ένα μοντέλο που βασίζεται σε πιθανότητες. Ένα παράδειγμα ενός πιθανοτικού κατηγοριοποιητή είναι ο Bayes. Στην κατηγορία των οκνηρών κατηγοριοποιητών βρίσκεται ο k-Nearest Neighbor(KNN)



Εικόνα 5 Αλγόριθμοι κατηγοριοποίησης [25]

2.2.1 KNN

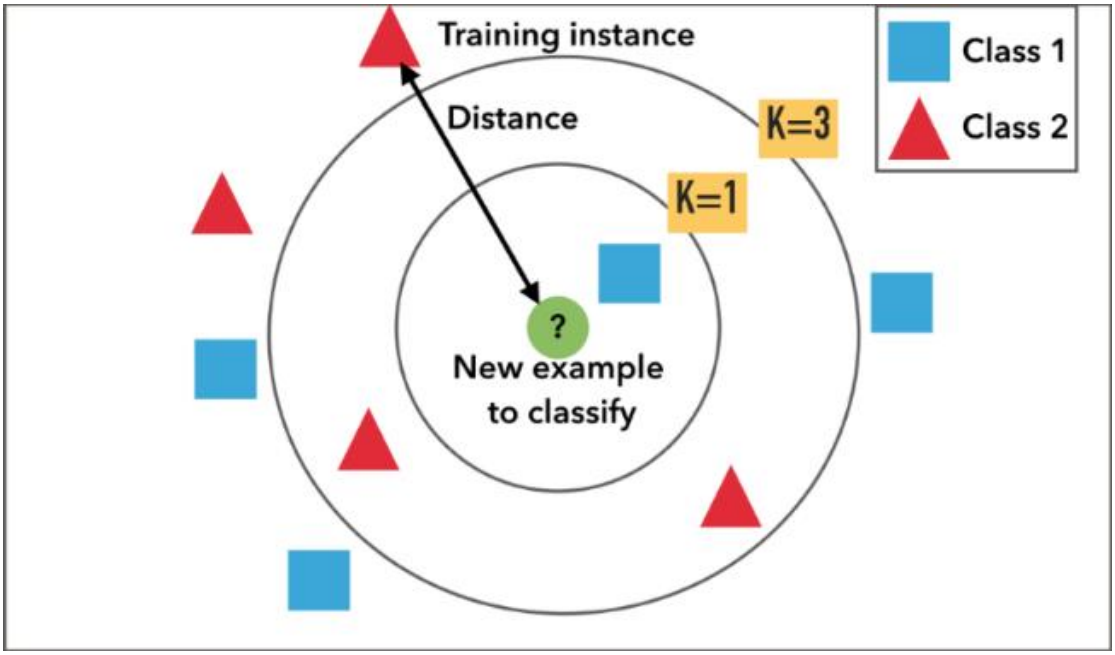
Ο αλγόριθμος κατηγοριοποίησης kNN με την πάροδο του χρόνου έχει βελτιώσει αισθητά την απόδοση του σε δεδομένα μεγάλου μεγέθους, όπως η προσέγγιση του άπειρου[26]. Ωστόσο, η λειτουργικότητα του kNN είναι εύκολο να επηρεαστεί από κάποια θέματα. Για παράδειγμα, η επιλογή της τιμής k, η επιλογή μέτρων απόστασης και άλλα. Αυτοί οι κατηγοριοποιητές στηρίζονται στη μάθηση από δείγματα εκπαίδευσης. Κάθε δείγμα παρουσιάζεται σαν ένα σημείο σε έναν ν-διάστατο χώρο. Όλα τα δείγματα εκπαίδευσης κρατούνται σε μοτίβο n-διαστάσεων και η "εγγύτητα" βρίσκεται από την ευκλείδεια απόσταση[27]. Οι κατηγοριοποιητές του πλησιέστερου γείτονα διαμερίζουν ίσο βάρος σε όλα τα χαρακτηριστικά. Για την τεχνική του πλησιέστερου γείτονα υπάρχουν δύο κατηγορίες, η δομημένη τεχνική KNN και η λιγότερο δομημένη τεχνική KNN[28]. Η δομημένη τεχνική παρακολουθεί την κύρια δομή των δεδομένων, όμως πολλές φορές δημιουργούνται προβλήματα λόγω της έλλειψης κατάλληλου μηχανισμού που επεξεργάζεται τα δεδομένα εκπαίδευσης. Στη λιγότερο δομημένη τεχνική, όλα τα δεδομένα κατηγοριοποιούνται σε δείγματα σημείων δεδομένων και δεδομένων εκπαίδευσης. Η απόσταση μετριέται ανάμεσα στα σημεία δειγμάτων και όλων των σημείων εκπαίδευσης και το σημείο με τη μικρότερη απόσταση είναι γνωστός ως ο πλησιέστερος γείτονας. Οι γείτονες απεικονίζονται από

αντικείμενα τα οποία γνωρίζουν που θα κατηγοριοποιηθούν. Αυτό είναι το σύνολο εκπαίδευσης για τον αλγόριθμο, αν και δεν υπάρχει κάποιος προκαθορισμένος τρόπος στη διαδικασία της εκπαίδευσης. Για να βρεθούν οι γείτονες, τα αντικείμενα απεικονίζονται σαν διανύσματα θέσης σε ένα πολυδιάστατο χαρακτηριστικό χώρο. Κατά κύριο λόγο, εφαρμόζεται η Ευκλείδεια απόσταση[29] αλλά και άλλα μέτρα απόστασης, για παράδειγμα η απόσταση του Μανχάταν, που χρησιμοποιούνται εξίσου αν και λιγότερες φορές. Ο αλγόριθμος k-πλησιέστερου γείτονα επηρεάζει την τοπική δομή των δεδομένων. Αφού είναι τύπος μεθόδων οκνηρής μάθησης που βασίζεται σε περιπτώσεις, τα μη κατηγοριοποιημένα σημεία δεδομένων εντοπίζονται και εισχωρούν σε συγκεκριμένη ετικέτα με βάση τον προηγούμενο k πλησιέστερο γείτονα[30] (KNN) και εφαρμόζεται μηχανισμός ψηφοφορίας για τον ορισμό αντικειμένου στόχου. Ένα από τα βασικά πλεονεκτήματα της τεχνικής KNN[28] είναι ότι είναι εύχρηστη και ακριβής για μεγάλα δεδομένα εκπαίδευσης και ισχυρή σε θορυβώδη δεδομένα εκπαίδευσης

Εφαρμόζεται συχνά στην κατηγοριοποίηση κειμένου, αναγνώριση προτύπων, μοντέλων κατάταξης, εφαρμογές αναγνώρισης αντικειμένων και αναγνώρισης συμβάντων. Ο M. Cover και ο P. E. Hart[29] πρότειναν ο k πλησιέστερος γείτονας (kNN) στον οποίο ο πλησιέστερος γείτονας μετριέται σύμφωνα με την τιμή του k να ορίζει πλησιέστερους γείτονες που χρειάζεται να υπολογιστούν για να δημιουργηθεί κλάση ενός δείγματος σημείου δεδομένων. Για να προσδιοριστεί μια κλάση στην οποία ανήκει ένα σημείο, είναι αναγκαίο να υπάρχουν πλησιέστεροι γείτονες και αυτό ονομάζεται K-NN. Αυτά τα δείγματα δεδομένων είναι απαραίτητα και πρέπει να καταχωρούνται στη μνήμη όταν εκτελείται ο αλγόριθμος και χρησιμοποιούνται ως τεχνική που στηρίζεται στη μνήμη. Οι T. Bailey και A. K. Jain προτείνουν λύσεις για το kNN που στηρίζεται σε βάρη. Τα σημεία εκπαίδευσης παίρνουν τιμές βάρους σύμφωνα με τις αποστάσεις τους από τα σημεία δεδομένων του δείγματος. Η υπολογιστική πολυπλοκότητα και οι απαιτήσεις μνήμης όμως είναι τις περισσότερες φορές το πρωταρχικό ζήτημα. Για να λύσουμε το πρόβλημα της μνήμης πρέπει τα δεδομένα να μειώνονται. Πολλές φορές όταν επαναλαμβάνουμε την διαδικασία δεν προσφέρουμε πρόσθετες πληροφορίες, και αυτές εξαφανίζονται από τα δείγματα της προπόνησης. Αν θέλουμε καλύτερη επίδοση, τα σημεία δεδομένων που στη ουσία δεν προσφέρουν κάποιο καλύτερο αποτέλεσμα πρέπει να εξαλείφονται από τα δεδομένα του σετ εκπαίδευσης. Το σύνολο δεδομένων εκπαίδευσης NN έχει την δυνατότητα να δομηθεί εφαρμόζοντας αρκετές τεχνικές για τη βελτίωση του περιορισμού της μνήμης kNN.

Ο αλγόριθμος k-πλησιέστερων γειτόνων είναι ένας απλός αλγόριθμος μηχανικής μάθησης. Ένας κατηγοριοποιητής πρέπει να έχει ένα ακέραιο θετικό k, ένα σύνολο δεδομένων εκπαίδευσης και μια μέτρηση για την εγγύτητα. Όταν εισέρχεται ένα νέο στοιχείο το οποίο θέλουμε να κατηγοριοποιήσουμε θα βρούμε τα πλησιέστερα σημεία και θα το

καταχωρήσουμε σε μία κλάση ανάλογα με την απόσταση του με αυτά. Για παράδειγμα όπως φαίνεται και στο παρακάτω σχήμα(Εικόνα 6) αν το K=1 θα εισαχθεί στη κλάση 1, αν όμως το K=3 θα εισαχθεί στη κλάση 2. Σε προβλήματα κατηγοριοποίησης δύο κλάσεων, είναι επιθυμητό να διαλέξετε ένα περιττό αριθμό για το k, γιατί με αυτό τον τρόπο θα μπορέσουν να εξαλειφθούν οι ισοψηφίες. Εφαρμόζεται για την μέτρηση του συνολικού αριθμού γειτόνων που καθορίζουν την κατηγοριοποίηση[28]. Ένα αντικείμενο εισέρχεται σε μια συστάδα, όταν η πλειοψηφία των γειτόνων του αντικείμενου που εισέρχεται στην κλάση είναι όμοια με τους k πλησιέστερους γείτονες[29]. Είναι σίγουρα σημαντική η εύρεση k-πλησιέστερων γειτόνων, γιατί όλοι αυτοί ρυθμίζονται σωστά για υπολογιστεί ένας ικανοποιητικός μέσος όρος.



Εικόνα 6 KNN Αλγόριθμος [31]

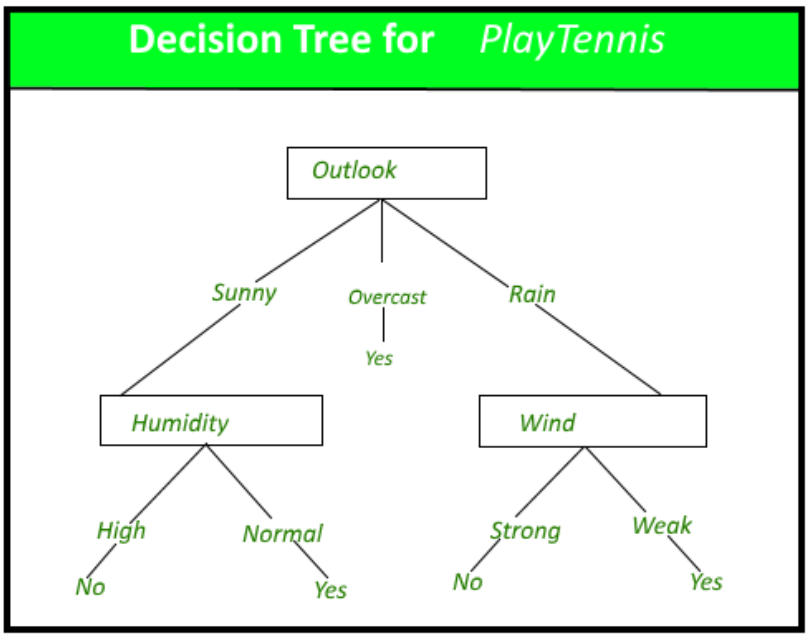
2.2.2 Decision Trees

Τα δέντρα απόφασης[32] είναι πολύ γνωστή μέθοδος και εφαρμόζεται σε διάφορους κλάδους της επιστήμης, όπως μηχανική μάθηση, επεξεργασία εικόνας και αναγνώριση προτύπων. Εχει την δυνατότητα να χρησιμοποιηθεί για να εφαρμοστούν υποθέσεις με βάση κατηγορηματικά ονόματα κλάσεων, για να κατηγοριοποιήσει τη γνώση σύμφωνα με το σετ εκπαίδευσης και ετικέτες της κλάσης αλλά επίσης να κατηγοριοποιήσει και τα καινούργια διαθέσιμα δεδομένα. Ένα δέντρο απόφασης είναι μια δομή δέντρου που μοιάζει με διάγραμμα ροής[33], όπου κάθε εσωτερικός κόμβος υποδηλώνει μια πρόβλεψη σε ένα

χαρακτηριστικό, κάθε κλαδί αντιπροσωπεύει ένα αποτέλεσμα της πρόβλεψης και κάθε κόμβος φύλλου (τερματικός κόμβος) έχει μια ετικέτα κλάσης.

Για να μπορέσει να κάποιος να καταλάβει ένα δέντρο αποφάσεων θα χρειαστεί να χωρίσει τα δεδομένα σε υποσύνολα με βάση μια πρόβλεψη τιμής χαρακτηριστικών. Αυτή η διαδικασία επαναλαμβάνεται[34] για κάθε υποσύνολο και τερματίζει όταν το υποσύνολο σε έναν κόμβο έχει την ίδια τιμή με τη μεταβλητή που στοχεύει ή όταν ο διαχωρισμός δεν έχει ουσία στις προβλέψεις. Η κατασκευή ενός δέντρου αποφάσεων δεν προϋποθέτει γνώσεις πάνω στο κλάδο αυτό ή επεξεργασία παραμέτρων. Για αυτό το λόγο είναι βέλτιστη για ανακάλυψη γνώσης. Τα δέντρα αποφάσεων μπορούν να χειριστούν δεδομένα υψηλών διαστάσεων.

Ένα στιγμιότυπο κατηγοριοποιείται ξεκινώντας από τον ριζικό κόμβο του δέντρου, στοχεύοντας το χαρακτηριστικό που καθορίζεται από αυτόν τον κόμβο και μετά μετακινώντας προς τα κάτω το κλαδί του δέντρου το συσχετίζει με την τιμή του χαρακτηριστικού. Αυτή η διαδικασία στη συνέχεια επαναλαμβάνεται για το υπόδεντρο που έχει ρίζες στον νέο κόμβο. Για παράδειγμα όπως φαίνεται στο παρακάτω γράφημα(Εξίσωση 3) κατηγοριοποιείται ένα πρωινό σύμφωνα με αν θα μπορεί κάποιος να παίξει τένις και επιστρέφει την κατηγοριοποίηση που σχετίζεται με το συγκεκριμένο φύλλο. (σε αυτήν την περίπτωση Ναι ή Όχι).



Εξίσωση 4 Διάγραμμα An example of a Decision tree[34]

Δύο είναι οι βασικές μετρήσεις στα δέντρα απόφασης, η εντροπία και το κέρδος πληροφορίας Η εντροπία[32] βρίσκει την μέτρηση των προβλημάτων ενός συνόλου δεδομένων ή την ύπαρξη τυχαίων στοιχείων. Η τιμή της εντροπίας πάντα κυμαίνεται ανάμεσα στο 0 και στο 1. Η τιμή του είναι πιο βέλτιστη, όταν ισούται με 0 ενώ αυτό είναι χειρότερη όταν ισούται με 1, επομένως όσο περισσότερη αβεβαιότητα στα σημεία υπάρχουν στα δεδομένα τόσο μεγαλύτερη και η εντροπία. Το κέρδος πληροφοριών είναι μια μέτρηση που εφαρμόζεται για την τμηματοποίηση και συχνά αποκαλείται αμοιβαία πληροφόρηση. Αυτό δείχνει πόση γνώση μπορεί να έχει η τιμή μιας τυχαίας μεταβλητής. Εδώ όσο μεγαλύτερη είναι η τιμή της τόσο το καλύτερο. Χρησιμοποιείται για να πάρει αποφάσεις και να αποφασίσει το βέλτιστο διαχωρισμό.

Τα δέντρα απόφασης χωρίζονται σε Κατηγορικά Μεταβλητά Δέντρα Αποφάσεων (CVDT) και Συνεχή Μεταβλητά Δέντρα Αποφάσεων[35] (CVDT). Πολύ σημαντικοί όροι ενός δέντρου απόφασης είναι ο ριζικός κόμβος, ο διαχωρισμός, κόμβος απόφασης, το φύλλο ή τερματικός κόμβος, το κλάδεμα, το υπόδεντρο και ο κόμβος γονέα και παιδιού. Είναι αρκετά απλή η κατανόηση της εξόδου του δέντρου αποφάσεων, λόγω ότι η γραφικής απεικόνισης και ο χρήστης είναι απλό να βρει την λύση. Ένα δέντρο αποφάσεων είναι γρήγορος τρόπος για να εντοπιστεί η σχέση μεταξύ των μεταβλητών και να αναπτυχθεί καινούργια μεταβλητή, αρκετά ακριβής για τον προσδιορισμό της μεταβλητής στόχου. Το δέντρο αποφάσεων χρειάζεται λιγότερα δεδομένα σε σχέση με άλλες τεχνικές μοντελοποίησης. Το πιο σημαντικό σε αυτή την τεχνική είναι ότι έχει την δυνατότητα να λειτουργήσει και με αριθμητικές αλλά και κατηγορικές μεταβλητές. Για να αναπτύξετε το μοντέλο δέντρου αποφάσεων δύο μέθοδοι είναι αρκετά γνωστές όπως ο ID3 κατηγοριοποιητής (Quinlan, 2020) και ο C4.5 κατηγοριοποιητής (Salzberg, 1994) μηχανικής μάθησης.

Ο αλγόριθμος ID3[28] εμφανίστηκε το 1986 και είναι ένας αρκετά γνωστός αλγόριθμος στην επιστήμη της εξόρυξης δεδομένων και μηχανικής μάθησης, επειδή είναι ακριβής αλλά και εύχρηστος. Ο αλγόριθμος ID3 στηρίζεται στο κέρδος πληροφοριών. Τα πλεονεκτήματά του είναι ότι είναι εύκολα κατανοητός από τον χρήστη και συμπεριλαμβάνει όλο το σετ εκπαίδευσης, ενώ στα μειονεκτήματα του εντάσσεται η αδυναμία παρακολούθησης προηγούμενων ανιχνεύσεων, δε μπορεί να διαχειριστεί τις πληροφορίες που δεν υπάρχουν και δεν παρέχει συνολική βελτιστοποίηση. Ο πιο δημοφιλής αλγόριθμος στη για ανάπτυξη δέντρων απόφασης είναι ο C4.5[36](Quinlan 1993). Είναι επέκταση του αλγορίθμου ID3 και εξαλείφει τις αδυναμίες του ID3. Μια έρευνα που συγκρίνει τα δέντρα αποφάσεων και άλλα

είδη μάθησης αλγορίθμων έχει πραγματοποιηθεί από τον (Tjen-Sien et al. 2000)[36] και παρουσιάζει τον C4.5 να παρέχει αρκετά υψηλή ταχύτητα και πολύ χαμηλό ποσοστό σφάλματος. Το C4.5 θεωρεί ότι τα δεδομένα εκπαίδευσης μπορούν να τοποθετηθούν στη μνήμη. Λίγο αργότερα, ο Gehrke et al. (2000) εισήγαγε το Rainforest, ένα πλαίσιο για την δημιουργία ταχύτερων και επεκτάσιμων αλγόριθμοι για την ανάπτυξη δέντρων αποφάσεων που υλοποιούνται λόγω της μνήμης.

Table 1 Comparisons between different Decision Tree Algorithms

Algorithms	ID3	C4.5	C5.0	CART
Type of data	Categorical	Continuous and Categorical	Continuous and Categorical, dates, times, timestamps	continuous and nominal attributes data
Speed	Low	Faster than ID3	Highest	Average
Pruning	No	Pre-pruning	Pre-pruning	Post pruning
Boosting	Not supported	Not supported	Supported	Supported
Missing Values	Can't deal with	Can't deal with	Can deal with	Can deal with
Formula	Use information entropy and information Gain	Use split info and gain ratio	Same as C4.5	Use Gini diversity index

Πίνακας 3 Comparison of DT Algorithms[37]

3 *ΤΕΧΝΟΛΟΓΙΕΣ*

3.1 PYTHON

Η Python είναι μια ερμηνευμένη, διαδραστική, αντικειμενοστραφής γλώσσα προγραμματισμού[38]. Περιλαμβάνει δομές δεδομένων υψηλού επιπέδου, δυναμική πληκτρολόγηση και δέσμευση, επεκτάσεις, κλάσεις, εξαιρέσεις, αυτόματη διαχείριση μνήμης κ.λπ. Αναπτύχθηκε το 1990 από τον Guido van Rossum. Εφαρμόζεται στην απλή επεξεργασία κειμένου, σε προγράμματα περιήγησης w.w.w. και σε ανάπτυξη παιχνιδιών[39]. Η γλώσσα ερμηνεύεται σαν μια μορφή αγγλικής διαλέκτου, καθιστώντας πιο εύκολη τη δημιουργία πολύπλοκων προγραμμάτων. Ένα πρόγραμμα python μεταγλωττίζεται από τον διερμηνέα σε ανεξάρτητο κώδικα byte, όπου αυτός με την σειρά του μεταφράζεται. Η Python λειτουργεί ως μια γλώσσα, είναι επεκτάσιμη που έχει συμπυκνωμένο και μπορεί να επεκταθεί προσθέτοντας επεκτάσεις. Η γλώσσα προγραμματισμού Python[40] είναι η πιο βέλτιστη γλώσσα για να μπορέσει κάποιος άνθρωπος να μάθει να προγραμματίζει, διότι παρέχει ισχυρά εργαλεία που δείχνουν τον τρόπο σκέψης των ανθρώπων και πως αυτά εισέρχονται στον κώδικα. Σήμερα αρκετοί καθηγητές από πολλά αναγνωρισμένα πανεπιστήμια, όπως το MIT, το UC Berkeley, το UC Davis, Sonoma State University, University of Washington, University of Waterloo, Luther College, και το Swarthmore College, το εφαρμόζουν για την εκπαίδευση πρωτοετών φοιτητών στα τμήματα πληροφορικής.

Μια από βασικότερες αιτίες για την γρήγορη ανάπτυξη της Python[41] είναι η απλότητα της σύνταξής της και στην συνέχεια θα παρουσιαστούν οι λόγοι που συμβάλλουν στη δημοφιλία της συγκεκριμένης γλώσσας.

Υποστηρικτική κοινότητα: Η Python υπάρχει εδώ και τρεις δεκαετίες, περιέχει ανεπτυγμένη και υποστηρικτική κοινότητα η οποία κάθε μέρα αυξάνεται όλο και περισσότερο.

Ανάπτυξη Ιστού: Ο προγραμματισμός Ιστού με την python μπορεί να χρησιμοποιηθεί με ποικίλους τρόπους, καθώς η python περιλαμβάνει αρκετά frameworks για την ανάπτυξη ιστοσελίδων, όπως το Django, το flask, και άλλα. Είναι μια γλώσσα που εφαρμόζεται συχνά στην ανάπτυξη Ιστού.

Εφαρμογή σε μεγάλα δεδομένα και μηχανική μάθηση: Τα μεγάλα δεδομένα και η μηχανική μάθηση είναι αρκετά διαδομένα και ευρέως χρησιμοποιούμενα στην επιστήμη των υπολογιστών, εξελίσσοντας με αυτό τον τρόπο τις εταιρείες του χώρου να τροποποιήσουν τις ενέργειες και την γενικότερη λειτουργία τους. Η Python χρησιμοποιείται κατά κόρον στον τομέα αυτό και είναι αυτή την στιγμή από τις διαδεδομένες γλώσσες για αυτήν τη δουλειά.

Δωρεάν και ανοιχτού κώδικα: Η Python και οι βιβλιοθήκες της είναι διαθέσιμες και ανοιχτού κώδικα. Η Python, επίσης, προστατεύεται από πνευματικά δικαιώματα και ο πηγαίος κώδικας της Python είναι διαθέσιμος υπό τη Γενική Δημόσια Άδεια GNU (GPL).

Υψηλού επιπέδου: Η Python είναι μια γλώσσα προγραμματισμού με ισχυρή αφαίρεση από τα στοιχεία της υποκείμενης πλατφόρμας ή του μηχανήματος. Σε αντίθεση με γλώσσες προγραμματισμού χαμηλού επιπέδου, λειτουργεί με στοιχεία φυσικής γλώσσας, είναι εύχρηστη και αυτοματοποιεί τους κύριους τομείς των υπολογιστών, όπως η κατανομή πόρων.

Επεκτάσιμο: Η python έχει την ευχέρεια να συνδυαστεί με διαφορετική γλώσσα προγραμματισμού, όπως η C ή C++ για ένα συγκεκριμένο σκοπό και αυτό προσφέρει δυνατότητα πειραματισμού, αλλά και επίτευξη ενός ισχυρού προγράμματος.

Εκτεταμένες βιβλιοθήκες: Οι βιβλιοθήκες της Python είναι αρκετές και περιλαμβάνουν ένα ευρύ φάσμα εγκαταστάσεων. Περιλαμβάνει έτοιμο κώδικα γραμμένο σε C που δίνει την δυνατότητα εισόδου σε λειτουργίες συστήματος, όπως λειτουργίες I/O καθώς και επεκτάσεις γραμμένες σε Python που λύνουν συχνά προβλήματα που επηρεάζουν τους προγραμματιστές.

Ως **δυναμικά πληκτρολογημένη γλώσσα**, η Python είναι εύκαμπτη. Επιπλέον, η Python είναι αρκετά ανεκτή σε σφάλματα[39], εκτός αν υπάρχει σφάλμα προγραμματιστικό του χρήστη. Όμως, λόγω αυτού πολλές φορές λόγω της διαφορετικότητας ενός προγράμματος μπορεί να διαφέρει η σημασία ενός αντικειμένου και αυτό δυσκολεύει αρκετά την απόδοση της Python

Φορητότητα: Αυτό σημαίνει ότι μπορεί να λειτουργήσει σε διαφορετικά λειτουργικά συστήματα όπως: Windows, Linux, UNIX, Amigo, Mac OS κ.λπ. Δεν υπάρχει καμία διαφοροποίηση σε όλα τα λειτουργικά συστήματα.

Αντικειμενοστραφής: Ο Αντικειμενοστραφής προγραμματισμός (OOP) σας παρέχει λύσεις σε σύνθετα προβλήματα. Υπάρχει δυνατότητα να χωριστούν τα δύσκολα προβλήματα σε μικρότερα δημιουργώντας αντικείμενα.

3.2 BEAUTIFUL SOUP

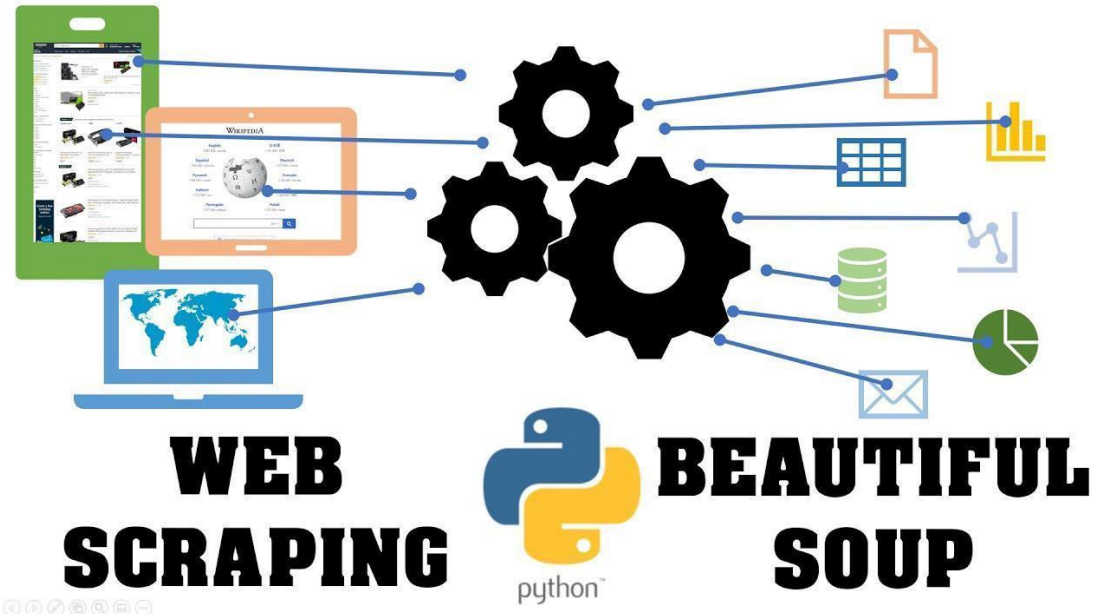
Τα τελευταία χρόνια αρκετοί επιστήμονες δουλεύουν για την εξαγωγή πληροφοριών, με γνώμονα είδη γεγονότων, οντοτήτων ή σχέσεων από δεδομένα κειμένου. Η εξαγωγή πληροφοριών[42] εφαρμόζεται σε μηχανές αναζήτησης, βιβλιοθήκες ειδήσεων, εγχειρίδια κειμένου για κάποιο κλάδο ή λεξικά. Ένα είδος εξαγωγής πληροφοριών είναι η εξόρυξη κειμένου, μια λειτουργία ανάκτησης πληροφοριών που προσπαθεί να εντοπίσει καινούργιες πληροφορίες με αυτόματη εξαγωγή τους από διάφορες πηγές κειμένου. Η εξόρυξη κειμένου λειτουργεί για να εξαλείψει πληροφορίες από αρχεία κειμένου με την βοήθεια γλωσσικών και στατιστικών αλγορίθμων. Η αναζήτηση στον Ιστό και η εξαγωγή πληροφοριών πραγματοποιούνται κατά κύριο λόγο από προγράμματα ανίχνευσης Ιστού.

Ο ανιχνευτής Ιστού[43] είναι πρόγραμμα υπολογιστή που επιτρέπει την εξερεύνηση στον παγκόσμιο ιστό (WWW) αυτόματα και αποτελεσματικά. Με την φοβερή ανάπτυξη που έχουν τα τελευταία χρόνια οι σελίδες στο Διαδίκτυο, είναι αρκετά πολύπλοκο και αργό για τους ανθρώπους να ερευνήσουν για πληροφορίες. Οι ανιχνευτές βοηθούν τους ανθρώπους να βρουν αυτό που ψάχνουν. Ένας ανιχνευτής Ιστού αντικαθιστά την διαδικασία ενός ανθρώπου να δημιουργεί αιτήματα στον ιστότοπο. Ερευνητές πιστεύουν ότι ένας ανιχνευτής ιστού είναι ικανός να τραβήξει πληροφορίες από αρκετές ιστοσελίδες σε κάποιο ορισμένο χρόνο μέσω των συνδέσμων URL και των υπερσυνδέσμων της ιστοσελίδας. Δηλαδή, μετά τη λήψη πολλών URL, ο ανιχνευτής επιλέγει τις ιστοσελίδες που είναι σχετικές με τη διεύθυνση URL, εξάγει τους απαραίτητους υπερσυνδέσμους και κάνει το ίδιο με τις ιστοσελίδες που περιλαμβάνουν υπερσυνδέσμους Το scrapping είναι πολύ σημαντικό σε περιπτώσεις που δεν υπάρχουν δεδομένα σε μορφή κατανοητή από μηχανή, όπως JSON ή XML[44]. Ο ανιχνευτής χρησιμοποιείται από αρκετές διαδικτυακές εφαρμογές, όπως η μηχανή αναζήτησης, η εξόρυξη δεδομένων και τα μεγάλα δεδομένα. Οι ανιχνευτές ιστού είναι κατανεμημένοι, έχουν προσαρμοστικότητα, απόδοση και αποτελεσματικότητα, ποιότητα, φρεσκάδα και επεκτασιμότητα.

Όλο και περισσότερος κόσμος σήμερα ασχολείται με την ανάπτυξη προγράμματος ανίχνευσης ιστού και γι' αυτόν το λόγο δημιουργούνται δυνατές βιβλιοθήκες ανιχνευτών, που δίνουν την δυνατότητα στους προγραμματιστές να έχουν αρκετές επιλογές και βοήθεια στην ανάπτυξη ανιχνευτή. Ένας πλήρης ανιχνευτής είναι αναγκαίος στη λήψη δεδομένων και την αποθήκευση σε βάση δεδομένων. Οι κύριες λειτουργικές μονάδες λήψης εγγράφων HTML σύμφωνα με τη διεύθυνση URL υπάρχουν στη βιβλιοθήκη αιτήματος (request). Για παράδειγμα, μπορεί να πραγματοποιηθεί λήψη αρχείων από το διακομιστή, διατήρηση της

σύνδεσης, μεταφόρτωση αρχείων, κ.λπ.. Η ανίχνευση Ιστού ήταν ο μόνος τρόπος να συλλεχθούν δεδομένα από το Διαδίκτυο μέχρι την ανακάλυψη των API. Τα API είναι πολύ χρήσιμα εργαλεία που βοηθούν στη εξαγωγή πληροφορίας με οργανωμένο τρόπο.

Το Beautifulsoup είναι ένα πακέτο Python που μπορεί να ανακτήσει δομημένα δεδομένα από μια ιστοσελίδα και αναπτύχθηκε από τον Leonard Richardson. Το Beautifulsoup εφαρμόζεται για ανάλυση XML και HTML εγγράφων. Είναι πιο εύχρηστο σε σχέση με την κανονική έκφραση, διότι παρέχει λιγότερα βήματα για πλοήγηση, εξέταση και ενημέρωση ενός δέντρου ανάλυσης. Το BeautifulSoup μπορεί να μετατρέψει αυτόματα ένα εξερχόμενο έγγραφο σε UTF-8 και ένα εισερχόμενο σε Unicode, οπότε δεν είναι απαραίτητο να βλέπει κάποιος τις κωδικοποιήσεις, παρά μόνο αν το έγγραφο δεν προσδιορίζει ένα. Ένα αντικείμενο στο BeautifulSoup έχει δύο ορίσματα την πηγή της ιστοσελίδας και τον αναλυτή. Οι διάφοροι αναλυτές μεταξύ των html.parser, lxml και html5lib μπορούν να διαχειριστούν ένα τέτοιο αντικείμενο. Το html.parser[45] εγκαθίσταται με την python και περιλαμβάνει μια κλάση HTMLParser που εφαρμόζεται ως βασικός HTML και XHTML αναλυτής. Η lxml περιέχει αρκετά χαρακτηριστικά και εύχρηστη βιβλιοθήκη για την επεξεργασία XML και HTML. Το html5lib χρησιμοποιείται με τη WHATWF HTML, προδιαγραφή που λειτουργεί αναλύοντας περιεχόμενο HTML, όπως ένα πρόγραμμα περιήγησης ιστού. Είναι ανοιχτού κώδικα και παρέχει άδεια σύμφωνα με το Άδεια BSD.



Εικόνα 7 Web-scraping Beautifulsoup [46]

3.3 SCIKIT-LEARN

Το scikit-learn είναι η πιο γνωστή βιβλιοθήκη Python για μηχανική μάθηση και εφαρμόζεται για χαρακτηριστικά μηχανικής και κλασικής μοντελοποίησης σε δεδομένα μικρής ή και μεσαίου κλίμακας διότι διαθέτει ένα συνεπές API[47]. Επιπλέον βοηθάει τη αποδοτικότητα των NumPy και SciPy[48] με πολυαριθμητικούς αλγορίθμους εξόρυξης δεδομένων. Παρέχει το matplotlib πακέτο για σχεδίαση γραφημάτων. Ένα από τα σημαντικά στοιχεία του είναι μια εμπεριστατωμένη ηλεκτρονική τεκμηρίωση για κάθε αλγόριθμο που χρησιμοποιεί. Η εμπεριστατωμένη τεκμηρίωση είναι αναγκαία για τους υποψήφιους ενδιαφερόμενους και χρειάζεται πολύ περισσότερο από όχι αρκετά τεκμηριωμένες υλοποιήσεις αλγορίθμων. Το πακέτο παρέχει ένα πολύ μεγάλο κομμάτι του πυρήνα αλγορίθμων εξόρυξης δεδομένων. Ωστόσο, ομάδες σημαντικών αλγορίθμων εξόρυξης δεδομένων δεν έχουν συμπεριληφθεί, όπως κανόνεςκατηγοριοποίησης (classification rules) και κανόνες σύνδεσης (association rules).

Πέρα από τους εκτιμητές που χρειάζονται για επεξεργασία και μοντελοποίηση δεδομένων, το Scikit-learn παρέχει ακόμα ένα API πρώτης κατηγορίας[49] για την σύνδεση της δημιουργίας και της εκτέλεσης της μηχανικής μάθησης το Pipeline API. Δίνει τη δυνατότητα στους εκτιμητές να συμπεριλάβουν την επεξεργασία δεδομένων, χαρακτηριστικών μηχανικής και εκτιμητών μοντελοποίησης, που θα ενωθούν για την λειτουργία από άκρο σε άκρο. Το Scikit-learn έχει επίσης ένα API για την αξιολόγηση εκπαιδευομένων μοντέλων εφαρμόζοντας γνωστές τεχνικές για επικύρωση. Το Scikit-learn εκμεταλλεύεται το τεράστιο περιβάλλον για να διαθέτει εφαρμογές πολύ σύγχρονης τεχνολογίας δημοφιλών αλγορίθμων μηχανικής μάθησης, παρέχοντας ταυτόχρονα μια αρκετά απλή διεπαφή στη χρήση από την Python. Με το παραπάνω παρέχονται λύσεις για στατιστική ανάλυση δεδομένων από αρχάριους στις βιομηχανίες λογισμικού και ιστού[50], ή και σε κλάδους πέρα από το κλάδο επιστήμης των υπολογιστών, όπως η βιολογία ή η φυσική. Πάραυτα το πακέτο είναι δυνατό σε μεθόδους που στηρίζονται σε λειτουργίες όπως γραμμικά μοντέλα (linear models) και υλοποιήσεις μηχανών διανυσμάτων υποστήριξης (support vector machines). Διαθέτει ταχύτητα παρόλο που είναι γραμμένο σε διερμηνευμένη γλώσσα. Αυτό συμβαίνει συνήθως διότι οι συνεισφέροντες χρειάζεται να βελτιστοποιήσουν τον κώδικα. Όμως για να μπορέσει κάποιος να χρησιμοποιήσει το scikit-learn χρειάζεται να είναι καλός προγραμματιστής της Python λόγω της διεπαφής της γραμμής εντολών. Έτσι, οποιοσδήποτε θέλει να πειραματιστεί με τέτοια είδους προβλήματα να μάθει την γλώσσα και έπειτα την βιβλιοθήκη scikit-learn.

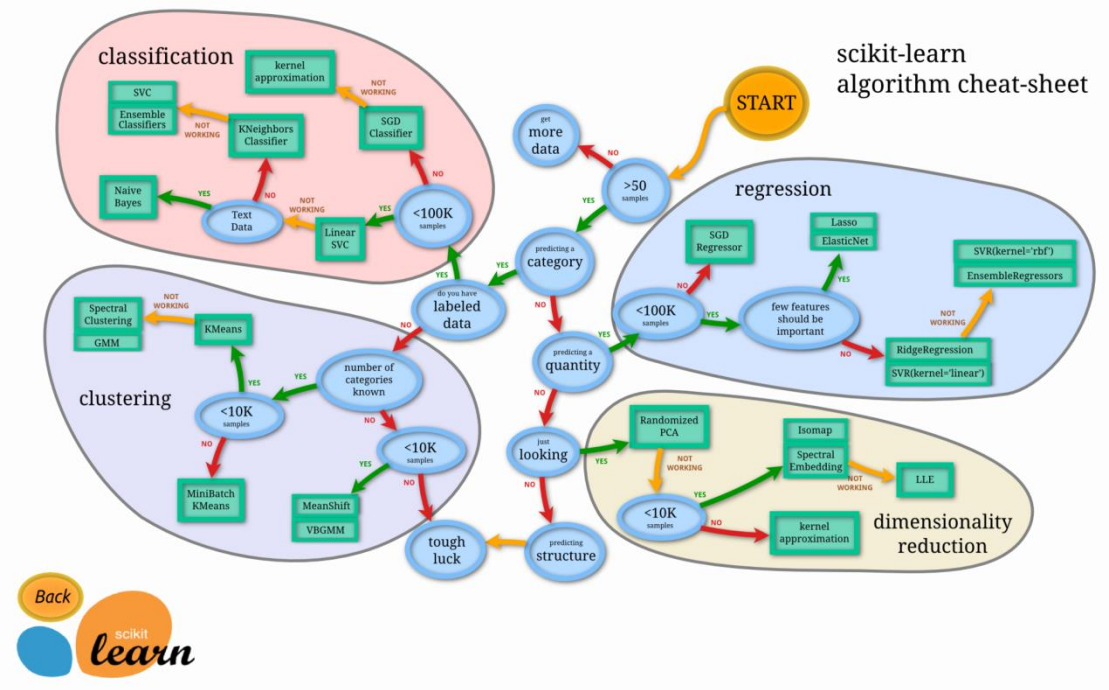
Το Scikit-learn συμπεριλαμβάνει τέσσερις βασικές θεματολογίες[51] που συνδέονται με τη μηχανική μάθηση. Αυτά είναι ο μετασχηματισμός δεδομένων, η εποπτευόμενη μάθηση, η μάθηση χωρίς επίβλεψη και αξιολόγηση και επιλογή μοντέλου.

Μετασχηματισμός δεδομένων. Ο μετασχηματισμός δεδομένων είναι αρκετά σημαντικός στα δεδομένα ανάλυσης. Κάποιοι μετασχηματισμοί εφαρμόζονται συνήθως για διάφορες μεταβλητές εισόδου. Το Scikit-learn διαθέτει αρκετές εύχρηστες λειτουργίες για την υλοποίηση των μεθόδων μετασχηματισμού και την προεπεξεργασία δεδομένων. Η κύρια δομή δεδομένων που εφαρμόζονται οι περισσότερες συναρτήσεις του Scikit-learn είναι ο πίνακας NumPy.

Επίβλεψη μάθησης. Η εποπτευόμενη μάθηση προσδιορίζεται σε ένα κομμάτι της μηχανικής εκμάθησης αλγορίθμων που αναπτύσσουν μια σύνδεση ανάμεσα στα μεταβλητά χαρακτηριστικά και στου μεταβλητούς στόχους τους. Η ανάγκη για τη εφαρμογή εποπτευόμενων μεθόδων μάθησης είναι να γνωρίζει κάποιος εκ των προτέρων τα χαρακτηριστικά και τις ετικέτες τους. Η εποπτευόμενη μάθηση μπορεί να διαμεριστεί σε δύο κατηγορίες σύμφωνα με τη ετικέτα. Στην παλινδρόμηση υπάρχουν οι συνεχείς ετικέτες και στη κατηγοριοποίηση υπάρχουν διακριτές ετικέτες.

Η μάθηση χωρίς επίβλεψη είναι ικανή να εξάγει πληροφορίες όπως λειτουργικά δίκτυα ή κομμάτια του ανθρώπινου εγκεφάλου από δεδομένα σε κατάσταση ηρεμίας. Ο κώδικας Python για την μηχανική μάθηση είναι σχετικά εύκολος. Οι δυσκολίες εμφανίζονται στην χρησιμοποίηση της βέλτιστης προεπεξεργασίας στα δεδομένα, στην επιλογή του σωστού μοντέλου για το ζήτημα και τα συμπεράσματα των αποτελεσμάτων.

Αξιολόγηση και επιλογή μοντέλου. Οι μέθοδοι μηχανικής μάθησης συχνά επηρεάζονται από υπερβολική προσαρμογή (overfitting). Γι' αυτό το λόγο, η μέτρηση της λειτουργίας μιας μεθόδου μηχανικής μάθησης για ένα ζήτημα κατηγοριοποίησης είναι η εμπεριστατωμένη επικύρωση. Πρέπει να διαμεριστούν τα δεδομένα σε αρκετά τμήματα τυχαία και να εφαρμοστεί σετ εκπαίδευσης (train set) σε κάποιο από αυτά και άλλα ως σετ δοκιμής (test set). Βέβαια, κάποιες φορές για να αξιολογηθεί πιο βέλτιστα η μοντελοποίηση χρειάζεται να εφαρμοστεί πριν από το σετ δοκιμής και ένα σετ επικύρωσης (validation set). Το Scikit-learn διαθέτει απλές μετρήσεις για να αξιολογήσει την ομοιότητα ανάμεσα των προβλεπόμενων ετικετών και των πραγματικών ετικετών.



Εικόνα 8 Διάφοροι Αλγόριθμοι εκτιμητών που χρησιμοποιούνται με το scikit-learn[52]

3.4 Mysql

Πολλές εφαρμογές στο διαδίκτυο λαμβάνουν δεδομένα από βάσεις δεδομένων και αυτό επηρεάζει άμεσα την αποδοτικότητα του ιστού και την εμπειρία του χρήστη. Ο χρόνος που χρειάζεται για να εξαχθούν τα κατάλληλα δεδομένα από τη βάση δεδομένων είναι σημαντικός διότι ανάλογα με τον χρόνο απόκρισης ο χρήστης δείχνει αν έμεινε ευχαριστημένος με την υπηρεσία. Σε αρκετές περιπτώσεις μπορεί να μην απεικονιστεί το επιθυμητο για έναν μεμονωμένο χρήστη επειδή τα αποτελέσματα είναι μοναδικά για κάθε αίτημα χρήστη. Αυτό είναι πιθανό να μεταβληθεί από την άδεια χρήστη ή συγκεκριμένες απαιτήσεις χρήστη για την επιλογή δεδομένων.

Η MySQL είναι ένα πολύ γνωστό και ευρέως χρησιμοποιούμενο σύστημα διαχείρισης βάσεων δεδομένων ανοιχτού κώδικα SQL που υλοποιείται, διανέμεται και υποστηρίζεται από την Oracle Corporation[53]. Μια βάση δεδομένων είναι μια δομημένη συλλογή δεδομένων. Για να αποκτήσετε πρόσβαση να καταχωρίσετε και να επεξεργαστείτε δεδομένα που είναι αποθηκευμένα σε μια βάση δεδομένων υπολογιστή, είναι απαραίτητο ένα σύστημα διαχείρισης βάσης δεδομένων όπως ο MySQL Server. Οι υπολογιστές είναι κατάλληλοι να διαχειρίζονται τεράστιο όγκο δεδομένων, τα συστήματα διαχείρισης βάσεων δεδομένων είναι

βασικά για τους υπολογιστές, ως ξεχωριστά βοηθητικά προγράμματα ή ως κομμάτια άλλων εφαρμογών.

Οι βάσεις δεδομένων MySQL είναι σχεσιακές. Μια σχεσιακή βάση δεδομένων αποθηκεύει δεδομένα σε πίνακες χωρίς να τα καταχωρεί σε μια μεγαλύτερη αποθήκη. Οι βάσεις δεδομένων διαμορφώνονται σε φυσικά αρχεία κατάλληλα για γρήγορη απόκριση. Το λογικό μοντέλο, με στοιχεία όπως βάσεις δεδομένων, πίνακες, προβολές, γραμμές και στήλες, αναπτύσσει ένα προσαρμόσιμο περιβάλλον προγραμματισμού. Μέσω κανόνων που δείχνουν τις σχέσεις μεταξύ διαφορετικών πεδίων δεδομένων οργανώνεται ορθά μια βάση δεδομένων, και η εφαρμογή που υλοποιεί ο χρήστης δεν θα έχει ασυνεπή, διπλότυπα, μη ενημερωμένα ή ελλιπή δεδομένα. . Η SQL είναι η πιο γνωστή τυποποιημένη γλώσσα που εφαρμόζεται για την είσοδο σε βάσεις δεδομένων. Με βάση το περιβάλλον προγραμματισμού ύπαρχει δυνατότητα να γράψει απευθείας SQL, ή να εισάγει SQL σε κώδικα γραμμένο σε διαφορετική γλώσσα ή να εφαρμόσει ένα API συγκεκριμένης γλώσσας που έχει ενσωματωμένη τη σύνταξη SQL. Η SQL καθορίζεται από το πρότυπο ANSI/ISO SQL.

Το λογισμικό MySQL είναι ανοιχτού κώδικα. Αυτό δίνει την δυνατότητα σε κάθε χρήστη να υλοποιήσει και να επεξεργαστεί το λογισμικό. Το λογισμικό MySQL μπορεί να γίνει λήψη δωρεάν. Υπάρχει τρόπος απεικόνισης του πηγαίου κώδικα και τροποποιήσει ανάλογα με τις ανάγκες. Το λογισμικό MySQL διαθέτει την GPL (GNU General Public License), ορίζοντας τις δυνατότητες που έχει ο κάθε ενδιαφερόμενος.

Ο διακομιστής βάσης δεδομένων MySQL διαθέτει ταχύτητα αξιοπιστία επεκτασιμότητα και ευχρηστία. Ο MySQL Server σχεδιάστηκε με την ιδέα ότι θα διαχειρίζεται τεράστιες βάσεις δεδομένων με μεγαλύτερη ευκολία από τις πρότερες λύσεις και έχει χρησιμοποιηθεί σε πολύ απαιτητικά περιβάλλοντα παραγωγής. Ο MySQL Server διαθέτει πολλές και χρήσιμες λειτουργίες. Η συνδεσιμότητα, η ταχύτητα και η ασφάλειά δείχνουν οτι ο MySQL Server είναι βέλτιστος για είσοδο σε βάσεις δεδομένων. Το λογισμικό βάσης δεδομένων MySQL είναι ένα σύστημα πελάτη/διακομιστή που απαρτίζεται από έναν διακομιστή SQL πολλών νημάτων που παρέχει διαφορετικά back end, διαφορετικά προγράμματα πελατών και βιβλιοθήκες, εργαλεία διαχείρισης και ένα πληθώρα διεπαφών προγραμματισμού εφαρμογών (API).

Εικόνα 9 Λογότυπο Mysql [54]

4 *Δημιουργία του Συνόλου δεδομένου*

4.1 Διαδικτυακοί τόποι παρουσίασης αυτοκινήτων

Με την πάροδο του χρόνου και την ταχεία ανάπτυξη της τεχνολογίας πολλά πράγματα που χρειαζόταν στο παρελθόν να κάνουν οι άνθρωποι, έχουν απλουστευθεί με τη χρήση του διαδικτύου. Παλαιότερα ένας άνθρωπος για να αγοράσει ένα αυτοκίνητο ήταν αναγκαίο να μεταβεί ο ίδιος σε μια αντιπροσωπεία αυτοκινήτων για να επιλέξει το αυτοκίνητο της αρεσκείας του, αλλά μόνο από την συγκεκριμένη εταιρεία.

Βέβαια, η αγορά ενός καινούργιου αυτοκινήτου με την πληθώρα των νέων μοντέλων που κυκλοφορούν σήμερα, είναι μια δύσκολη απόφαση. Σίγουρα χρειάζεται καθαρό μυαλό, γνώσεις, διαθέσιμος χρόνος και το ανάλογο τίμημα. Αυτό που πρέπει να γνωρίζει ένας υποψήφιος αγοραστής είναι ποιες ανάγκες έχει. Η αξιολόγηση των αναγκών ποικίλει από άνθρωπο σε άνθρωπο. Γενικότερα, όμως, τα σημαντικά κριτήρια είναι τα εξής.

- Να ορίσει τις βασικές ανάγκες του

Ο αγοραστής θα πρέπει να έχει σκεφτεί τη σκοπιμότητα χρήσης για την οποία θα αγοράσει ένα καινούργιο αυτοκίνητο και με αυτό τον τρόπο επιβάλλεται να επιλέξει την πλέον συμφέρουσα αγορά . Με αυτό τον τρόπο, θα διαλέξει τον κατάλληλο τύπο αυτοκινήτου.

- Κόστος

Ένας πολύ σημαντικός παράγοντας για την αγορά ενός καινούργιου αυτοκινήτου είναι η τιμή του. Αρκετοί άνθρωποι, λόγω αυτού, καταφεύγουν σε λύσεις μικρότερου κόστους άσχετα με την αρχική προτίμησή τους. Επίσης, κάποιοι χρειάζεται να λάβουν δανειοδότηση και αυτό μπορεί στο μέλλον να τους δυσκολέψει την καθημερινότητα.

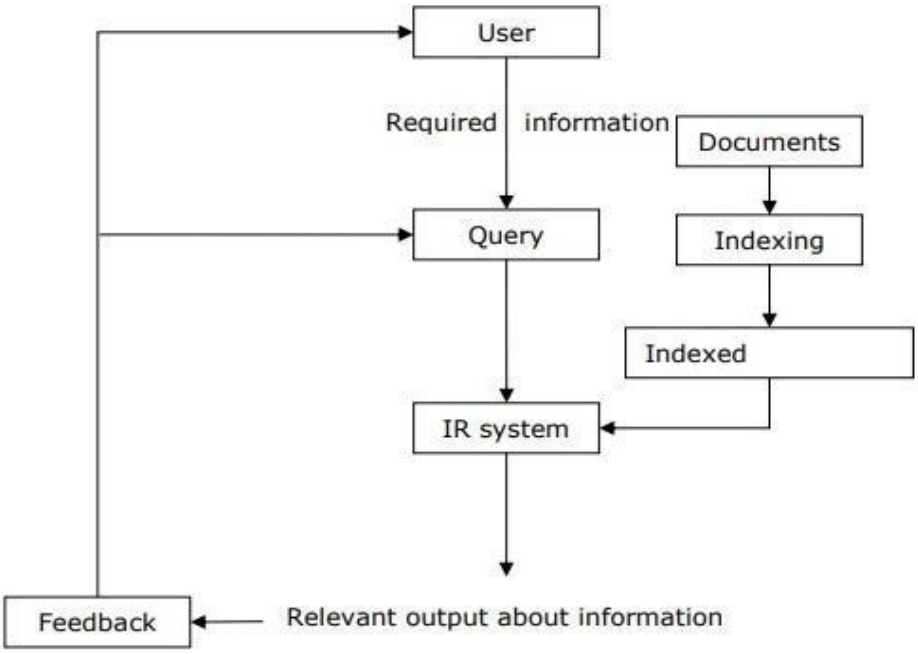
- Εγγύηση

Κάθε καινούργιο αυτοκίνητο διαθέτει εγγύηση αρκετών χρόνων. Έτσι, πολλοί αγοραστές δε χρειάζεται να ανησυχούν για οποιοδήποτε πρόβλημα παρουσιαστεί στο αυτοκίνητό τους.

Όπως καταλαβαίνει κανείς, για όλους τους προαναφερόμενους λόγους η αγορά ενός αυτοκινήτου είναι άκρως σημαντική. Γι' αυτόν το λόγο έχουν αναπτυχθεί αρκετοί ιστότοποι όπου με το πάτημα ενός κουμπιού, ένας ενδιαφερόμενος μπορεί να μάθει ότι χρειάζεται. Πολλοί από αυτούς παρουσιάζουν όλους τους κατασκευαστές που κατασκευάζουν τα αυτοκίνητα αλλά παραθέτουν και σχεδόν όλα τα χαρακτηριστικά που μπορεί να διαθέτει ένα καινούργιο αυτοκίνητο. Δεκάδες ιστοσελίδες όπως το car.gr, το newsauto.gr, το autotriti.gr, το autoblog.gr, το gocar.gr, το 4troxoi.gr, το traction.gr κ.λ.π. Οι διαδικτυακοί τόποι παρουσίασης δεδομένων παρέχουν αρκετές πληροφορίες για έναν ενδιαφερόμενο. Αρχικά, κάποιος μπορεί να αναζητήσει οποιοδήποτε αυτοκίνητο επιθυμεί επιλέγοντας τη μάρκα και το μοντέλο της. Επίσης, υπάρχει ένα διαδραστικό μενού όπου αναφέρονται σημαντικά πράγματα σχετικά με τα καινούργια αυτοκίνητα, όπως το πότε έρχονται τα νέα μοντέλα, μια σύγκριση μεταξύ τους, δοκιμή, καλύτερη αγορά, πωλήσεις κ.λ.π. Επιπροσθέτως όπως φαίνεται και στην παρακάτω εικόνα, υπάρχει μία ροή ειδήσεων για τις τελευταίες ειδήσεις πάνω στο τομέα της αυτοκινητοβιομηχανίας. Επίσης, στην αρχική σελίδα φαίνεται μια ροή ειδήσεων σχετικά με τα δημοφιλή άρθρα και όσο κατεβαίνει η σελίδα χωρίζονται ανά κατηγορία. Όλα αυτά, μαζί με άρθρα που θα διαβάσει κάποιος μέσα από τον ιστότοπο για να πάρει τη συμφερότερη απόφαση κάνουν τους ιστότοπους αρκετά απαραίτητους για όλους τους ενδιαφερόμενους. Για τις ανάγκες της παρούσας διπλωματικής, χρησιμοποιήθηκε το σάιτ autotriti.gr που παρέχει μια πληθώρα χαρακτηριστικών καινούργιων αυτοκινήτων που χρησιμοποιήθηκαν για τη δημιουργία του συνόλου δεδομένων.

4.2 Ανάπτυξη λογισμικού ανάκτησης δεδομένων

Με την τεράστια άνοδο της πληροφορίας στο διαδίκτυο, έχει γίνει αναγκαία η εφαρμογή αυτοματοποιημένων εργαλείων για αναζήτηση κατάλληλων πληροφοριών και για επίβλεψη ή προσδιορισμό των προτύπων χρήσης τους[55]. Η ανάκτηση πληροφοριών (IR) είναι η επιστήμη της εύρεσης πληροφοριών σε σχεσιακές βάσεις δεδομένων, εγγράφων, κείμενο, αρχεία πολυμέσων και παγκόσμιου ιστού. Η ανακάλυψη της αναζήτησης πληροφοριών πραγματοποιήθηκε από τον Vannevar Bush το 1945[56]. Αρκετά χρόνια μετά, τα πρώτα λειτουργικά συστήματα ανέπτυξαν λογισμικό ανάκτησης πληροφοριών. Ως το 1990, διάφορες τεχνικές προέκυψαν όπου θα επεξεργαζόντουσαν μερικές χιλιάδες έγγραφα. Το Διαδίκτυο και οι μηχανές αναζήτησης ιστού έχουν παροτρύνει τους επιστήμονες και τις μεγάλες επιχειρήσεις να υλοποιήσουν συστήματα ανάκτησης τεράστιας κλίμακας, για να μπορέσουν να διαχειριστούν τον όγκο των διαδικτυακών δεδομένων. Αρκετοί είναι οι ενδιαφερόμενοι που συμβάλλουν στο τομέα της ανάκτησης δεδομένων, όπως επαγγελματίες ερευνητές, πολιτικοί αναλυτές, κυβερνητικοί ερευνητές και μετεωρολόγοι. Οι εφαρμογές του IR διαφέρουν, γιατί περιέχουν εξαγωγή πληροφοριών από έγγραφα, αναζήτηση σε ψηφιακές βιβλιοθήκες, φιλτράρισμα πληροφοριών, φιλτράρισμα ανεπιθύμητων μηνυμάτων, εξαγωγή αντικειμένων από εικόνες, αυτόματη σύνοψη, κατηγοριοποίηση και συσταδοποίηση εγγράφων και αναζήτηση στο διαδίκτυο. Η αποδοτικότητα ενός συστήματος ανάκτησης αξιολογείται κατά πόσο βοηθάει πρακτικά τους χρήστες του συστήματος[57].



Εξίσωση 5 Διάγραμμα αρχιτεκτονικής ενός συστήματος ανάκτησης δεδομένων[58]

Η διαδικασία ανάκτησης πληροφοριών μετατρέπει έγγραφα σε κατάλληλες απεικονίσεις[59] για να έχουν τη δυνατότητα τα σημαντικά έγγραφα να ανακτηθούν με ακρίβεια. Υπάρχουν διάφορες στρατηγικές που χρησιμοποιούν ειδικά μοντέλα για τη διαδικασία απεικόνισης εγγράφων. Χρησιμοποιούνται τέσσερα βασικά μοντέλα ανάκτησης δεδομένων: το Boolean μοντέλο, το μοντέλο διανυσματικού χώρου, το γλωσσικό μοντέλο και το πιθανολογικό μοντέλο. Αρκετά από αυτά που εφαρμόζεται συχνά σε συστήματα ανάκτησης πληροφοριών και στον ιστό είναι το Boolean μοντέλο, το μοντέλο διανυσματικού χώρου, το γλωσσικό μοντέλο. Αυτά τα μοντέλα αναπαριστούν διαφορετικά έγγραφα και ερωτήματα, αλλά εφαρμόζουν το ίδιο πλαίσιο. Όλοι αυτοί συμπεριλαμβάνουν κάθε έγγραφο ή ερώτημα ως ένα μια μεγάλη εισροή λέξεων ή όρων. Δεν ενδιαφερόμαστε για την σειρά και τη θέση των όρων σε μια πρόταση ή σε ένα έγγραφο. Επομένως, ένα έγγραφο αναλύεται από ένα σύνολο διακριτών όρων. Ο όρος είναι μια λέξη που συμβάλλει στην υπενθύμιση των βασικών θεμάτων του εγγράφου.

Θέματα ενός συστήματος ανάκτησης πληροφοριών[60] είναι η αποδοτικότητα, ή αποτελεσματικότητα και ο χρόνος εκτέλεσης. Τα στάδια βελτιστοποίησης της ανάκτησης πληροφοριών είναι πολύ σημαντικό χαρακτηριστικό των μηχανών αναζήτησης και έχει τεράστιο αντίκτυπο στη καλύτερη ακρίβεια αναζήτησης. Λόγω της δύναμης του διαδικτύου, διάφορες εταιρείες εφαρμόζουν συγκεκριμένο ανιχνευτή για να εξάγουν τις κατάλληλες πληροφορίες. Τη λύση στη ζήτημα εύρεσης κατάλληλης αντιστοίχισης για μια συγκεκριμένη λέξη-κλειδί στον ιστό έδωσαν οι συμβατικές μηχανές αναζήτησης που υλοποίησαν ευρετικές μεθόδους. Παρατηρώντας την διαδικασία ανάκτησης πληροφοριών, φαίνεται πως η ευρετηρίαση αποτελεί ένα βασικό μέρος του συστήματος ανάκτησης. Οι πληροφορίες που χρειάζεται να εισάγουν οι χρήστες είναι πολλές φορές λανθασμένες είτε ανακριβείς και γι' αυτόν τον λόγο πολλές φορές το ερώτημα είναι πιθανό να μετατραπεί κατά τη διαδικασία της ανάκτησης, για να λειτουργούν ασαφή συστήματα χειρισμού αβεβαιότητας. Το σύστημα ανάκτησης πληροφοριών είναι βασικό να αναπτυχθεί με σκοπό τη λύση προβλημάτων, να παίρνει αποφάσεις και να υλοποιεί σωστά αυτό που ζητάει ο χρήστης, ασχέτως το πού βρήκε τις πληροφορίες που ζητήθηκαν. Ο βασικός σκοπός των συστημάτων ανάκτησης είναι να μπορέσει να λειτουργήσει ως βοήθημα στο χρήστη για να αποθηκεύσει και να οργανώσει τις πληροφορίες, αλλά και να ανακτήσει παρεμφερή έγγραφα που καλύπτουν ανάγκες των χρηστών. Τα δύο κύρια μέτρα που εφαρμόζονται κατά κόρον για την αξιολόγηση της ακριβείας ενός συστήματος ανάκτησης πληροφοριών είναι η Ακρίβεια και η Ανάκληση.

4.2.1 Γνωρίσματα

Η διαδικασία υλοποίησης των περισσότερων συστημάτων ανάκτησης πληροφοριών διαφοροποιήθηκε σε μια δομή λειτουργίας τεσσάρων σταδίων. Όπως έχει αναφερθεί και παραπάνω, τα περισσότερα από αυτά είναι παρεμφερή για ένα μεγάλο πλήθος συστημάτων ανάκτησης. Ένα σύστημα ανάκτησης δεδομένων[61] ευρετηριάζει τα δεδομένα που ταιριάζουν στο ερώτημα και το προσθέτει σε μια βάση δεδομένων. Έπειτα, χωρίζει κάθε λέξη της σελίδας σε στοιχεία, ανάλογα με τη μορφολογία. Εάν ένας χρήστης παραπέμπει στον επεξεργαστή αναζήτησης, μετά το ερώτημα εφαρμόζεται το τρίτο και το τέταρτο στάδιο της διαδικασίας του συστήματος ανάκτησης. Το σύστημα διαλέγει κάθε έγγραφο που υπάρχει στη βάση δεδομένων που ταιριάζει στο ερώτημα που καταχωρήθηκε, και έπειτα τα παρουσιάζει με βάση το σημαντικό βάρος αναζήτησης, την ακρίβεια της φράσης, τον χρόνος ενημέρωσης κ.λπ.

Αυτό αποκαλείται κατάταξη των αποτελεσμάτων αναζήτησης. Οι πρώτοι μηχανισμοί για κατάταξη εγγράφων εστιάζουν στην αποτελεσματικότητα εισαγωγής του κειμένου στο ερώτημα αναζήτησης. Όμως αυτός δεν ήταν ο ενδεδειγμένος τρόπος, διότι η καταγραφή μιας φράσης που ζητήθηκε, η θέση της στο έγγραφο ή η λίστα των λέξεων-κλειδιών δε διέθετε ακριβή απεικόνιση του εγγράφου. Η πιο γνωστή κατηγορία τύπων για την μέτρηση του βάρους ενός εγγράφου, σύμφωνα με ένα ερώτημα αποκαλείται TF*IDF[62]. Το TF*IDF είναι ένα αριθμητικό μέτρο του αντιστοιχίας για λέξεις και έγγραφα. Το TF είναι ο δείκτης εμφάνισης μιας λέξης μέσα σε ένα έγγραφο και το IDF είναι η σπανιότητα της λέξης στη συλλογή εγγράφων και όσο αυτά διαφέρουν τόσο υψηλότερη η αξία του TF*IDF.

Οι διαφορές ανάμεσα σε συστήματα ανάκτησης εμφανίζονται στους μηχανισμούς κατάταξής τους. Όλα τα συστήματα ανάκτησης βοηθούν το μηχανισμό της κατάταξης εγγράφων να μπορέσει να καθοδηγεί την αγορά αναζήτησης. Αρκετά συστήματα ανάκτησης παρουσιάζουν αποτελέσματα της αναζήτησης που μεταβάλλονται από πολλούς παράγοντες, οι οποίοι είναι τοποθετημένα από ανταγωνιστές. Με αυτό το τρόπο η Google στα τέλη της δεκαετίας του 1990 κατάφερε να εκθρονίσει το σύστημα της Alta Vista που ήταν το κυρίαρχο στην παγκόσμια αγορά αναζήτησης εκείνη την περίοδο.

Η χρήση των μηχανών αναζήτησης συνήθως πραγματοποιείται από την ανάγκη απόκτησης πληροφοριών που απαντούν στις ερωτήσεις του χρήστη και όχι τόσο από την επιθυμία εύρεσης ενός εγγράφου. Αυτή είναι η κύρια διαφορά ανάμεσα σε πληροφορίες και δεδομένα ως μέσο ενημέρωσης. Η υλοποίηση κατάλληλων παραλλαγών σύγκλισης του περιεχομένου

και των δεδομένων είναι μία αυξανόμενη τάση της έρευνας στο πεδίο της ανάκτησης πληροφοριών.

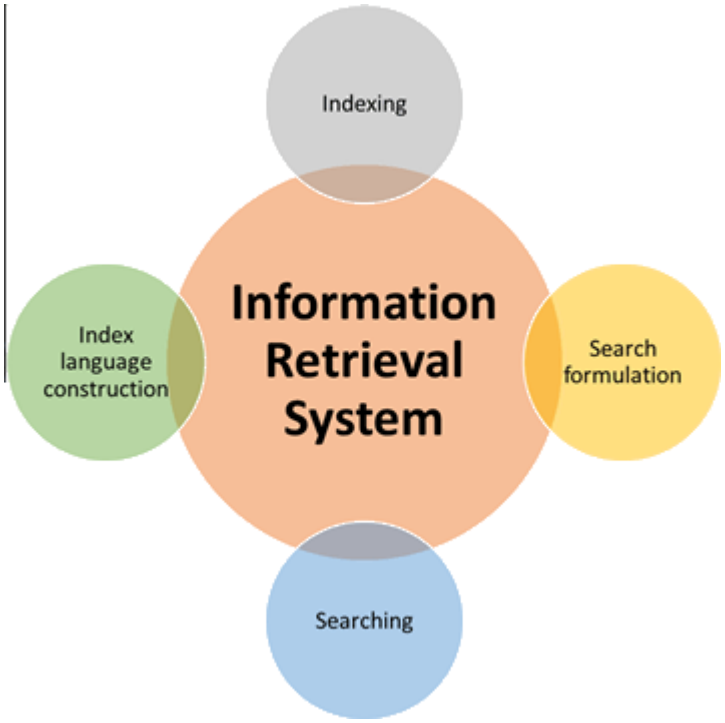
Με βάση τους ειδικούς μια ακριβής πληροφορία ενός συστήματος ανάκτησης πληροφορίας είναι απαραίτητο να πληροί τις παρακάτω απαιτήσεις:

• ευχρηστία(usability).

• το περιεχόμενο πρέπει να είναι οργανωμένο με ακρίβεια και επικαιροποιημένο.

• ταχεία αναζήτηση σε βάση δεδομένων και γρήγορη απάντηση.

• αξιοπιστία και ακρίβεια στα αποτελέσματα αναζήτησης.



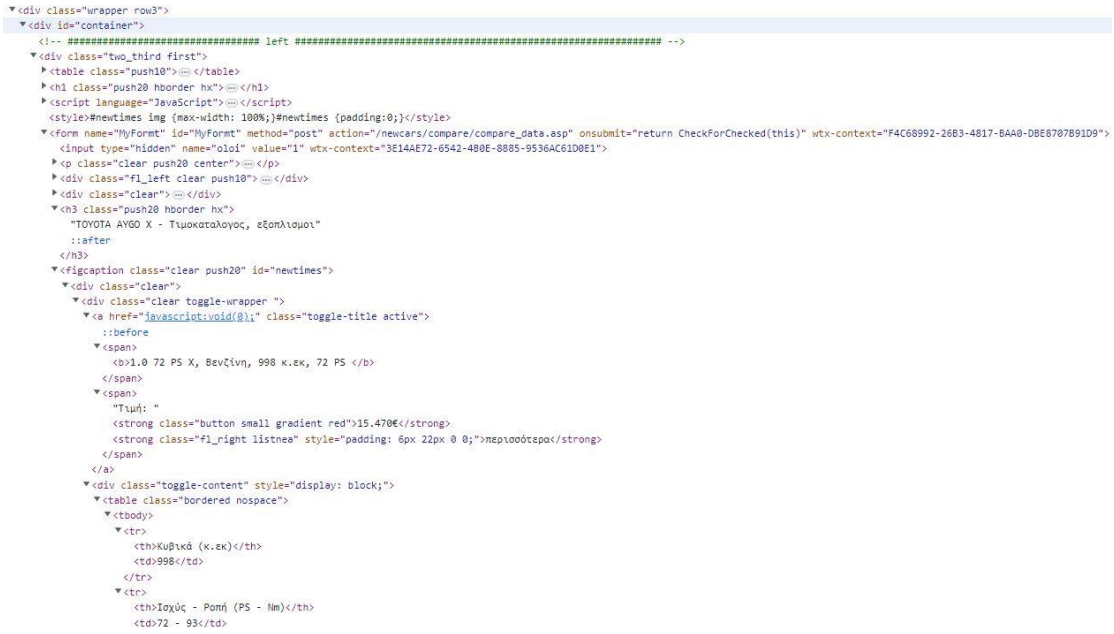
Εικόνα 10 Απεικόνιση Συστήματος Ανάκτησης Δεδομένων[63]

4.2.2 Τεκμηρίωση κώδικα

Στα πλαίσια της παρούσας διπλωματικής υλοποιήθηκε ένα λογισμικό ανάκτησης δεδομένων με το οποίο θα ασχοληθούμε ευθύς αμέσως, δίνοντας έμφαση στον κώδικα. Για τις ανάγκες της εργασίας χρησιμοποιήθηκε το Visual Code studio. Σταδιακά, θα αναλυθούν όλα τα σημαντικά κομμάτια που αναπτύχθηκαν και θα παρουσιαστούν με την σειρά.

Αρχικά πριν ξεκινήσουμε να δείχνουμε τον κώδικα που αναπτύχθηκε είναι αναγκαίο να κάνουμε μια αναφορά σχετικά με το DOM(Document Object Model) της ιστοσελίδας

autotriti όπου εκεί μέσω του εργαλείου Beautifulsoup και των HTTP requests ανακτήθηκαν τα δεδομένα για να δημιουργηθεί το σύνολο δεδομένων. Ουσιαστικά δημιουργείται μια επανάληψη για να ανακτήθουν όλες οι μάρκες αυτοκινήτων με την χρήση του συνδέσμου("https://www.autotriti.gr/data/newcars/times/montela/times_"+marka+".asp"). Για κάθε μάρκα μέσω της κεφαλίδας h3 (‘class="push20 hborder hx"’) καταχωρείται το μοντέλο του αυτοκινήτου. Έπειτα για όλα τα μοντέλα της κάθε μάρκας μέσω της κλάσης “clear toggle-wrapper” ανακτήθηκαν όλα τα δεδομένα-χαρακτηριστικά των καινούργιων αυτοκινήτων που μας ενδιαφέρουν.



Εικόνα 11 DOM ιστοσελίδας Autotriti

Θα χρειαστεί να προστεθούν οι βιβλιοθήκες που θα χρησιμοποιήσει το λογισμικό. Αυτές είναι το beautifulSoup στην οποία έχουμε αναφερθεί και παραπάνω αναλυτικά για αυτό και τις δυνατότητες του. Επίσης, είναι αναγκαία η χρήση της βιβλιοθήκης αιτημάτων (Http requests) για να μπορέσουμε να κάνουμε αιτήματα στην ιστοσελίδα που θέλουμε να αντλήσουμε δεδομένα. Στη συνέχεια, θα χρησιμοποιήσουμε ένα csv (comma separated values) αρχείο όπου εκεί θα καταχωρηθούν τα δεδομένα χαρακτηριστικά των καινούργιων αυτοκινήτων που έχουν επιλεχθεί για την υλοποίηση των αλγορίθμων μηχανικής μάθησης. Τέλος, η βιβλιοθήκη numpy της python για την δημιουργία πολυδιάστατων πινάκων.

```
from bs4 import BeautifulSoup
import requests
import csv
import numpy as np
```

Εικόνα 12 Εισαγωγή Βιβλιοθηκών

Στη συνέχεια αφού έχουμε τοποθετήσουμε τις μάρκες των αυτοκινήτων σε μία λίστα θα ανακτήσουμε τα δεδομένα που υπάρχουν σε κάθε μάρκα ξεχωριστά με την βοήθεια του beautifulsoup.

```
for marka in markes:
    print(marka)
    URL="https://www.autotriti.gr/data/newcars/times/montela/times_"+marka+".asp"
    page = requests.get(URL)
    results = BeautifulSoup(page.content, "lxml")

    job_elements = results.find_all("div", class_="clear toggle-wrapper")
```

Εικόνα 13 Ανάκτηση δεδομένων

Κάθε φορά που ο ανιχνευτής εισέρχεται στα στοιχεία των μαρκών είναι αναγκαίο να εφαρμοστούν κάποιες τροποποιήσεις για να συλλεχθούν τα δεδομένα στην μορφή που επιθυμούμε. Η κάθε επανάληψη του παρακάτω κώδικα πραγματοποιείται για τον λόγο ότι κάθε φορά επιλέγεται διαφορετικό μοντέλο αυτοκινήτου. Όμως λόγω της διαμόρφωσης της σελίδας όπως φαίνεται στη παρακάτω εικόνα(τυχαίο παράδειγμα) χρειάζεται να τραβήξουμε τα δεδομένα αυτούσια.

Εικόνα 14 Παρουσίαση χαρακτηριστικών από την ιστοσελίδα autotriti

Πηγή:https://www.autotriti.gr/data/newcars/times/montela/times_TOYOTA.asp

Αφού πραγματοποιηθούν οι τροποποιήσεις που χρειάζονται τότε προσθέτουμε τα δεδομένα σε συγκεκριμένη θέση για κάθε στήλη σε μια λίστα car_details. Στη λίστα που δημιουργείται όμως παρατηρήθηκε ότι κάποιες τιμές στη στήλη κυβικά δεν ανακτώνται σωστά οπότε πολλαπλασιαζεται η τιμή τους επι 1000.

```
for job_element in job_elements:
    brands=job_element.parent.parent.find_previous_sibling('h3')
    value=brands.text.strip().replace("- Τιμοκαταλογος, εξοπλισμοι"," ")


    car_details=[]
    car_details.append(marka)
    car_details.append(value)

    title_element = job_element.find("a", class_="toggle-title")
    title=title_element.find('b',)
    titl=str(title)
    titl=titl.replace(", ","-#-")


    cleantitle = BeautifulSoup(titl, "lxml").text

    cleantitle=cleantitle.replace("κ.εκ","")
    cleantitle=cleantitle.replace("PS","")
    cleantitle=cleantitle.replace("Βενζίνη","benz")
    cleantitle=cleantitle.replace("Ηλεκτρικό","electric")

    x = cleantitle.split("-#-")

    val1=0
    try:

        print(x[2])
        val1=float(x[2])

        if  val1<10:
            val1=val1*1000
    except:
        print("ch")
```

Εικόνα 15 Καταχώρηση 4 χαρακτηριστικών στη λίστα car_details

Με τον ίδιο τρόπο εισάγουμε την τιμή στη λίστα car_details εφαρμόζοντας τις παρακάτω τροποποιήσεις.

```
price= job_element.find("strong", class_="button small gradient red")
pric=str(price)
cleanprice = BeautifulSoup(pric, "lxml").text

cleanprice=cleanprice.replace("€","")
cleanprice=cleanprice.replace(".","")
cleanprice=cleanprice.replace("ΚΠ","1")

car_details.append(cleanprice)
```

Εικόνα 16 Καταχώρηση τιμής στη λίστα car_details

Αφού έχουν καταχωρηθεί επιτυχώς οι τιμές που θέλουμε, ψάχνουμε για όλα τα μοντέλα στις γραμμές ενός πίνακα που διαθέτει η ιστοσελίδα autotriti. Επειδή ο πίνακας είναι πολυδιάστατος μέσω της ευρετηρίασης της κάθε γραμμής και βρίσκουμε το περιεχόμενο της κάθε στήλης και ξανά με τις κατάλληλες τροποποιήσεις, προσθέτουμε τις τιμές στην λίστα που έχουμε δημιουργήσει. Επίσης, λόγω της μορφοποίησης του πίνακα να παρέχονται στη ίδια στήλη ισχύς και ροπή, τις χωρίζουμε για να πάρουμε τις τιμές του ξεχωριστά. Τέλος, όπως θα δούμε και παρακάτω, τα χαρακτηριστικά που χρησιμοποιούμε απ' τον πίνακα είναι οι πρώτες 7 στήλες, οπότε εκεί τερματίζει και η επανάληψη. Αυτό γίνεται για όλα τα αυτοκίνητα που υπάρχουν στην ιστοσελίδα.

```
body_element=job_element.find_all("tr",)
index=0
for s in body_element:
    sa=str(s)
    sa=sa.replace("<td>","#")
    cleantext = BeautifulSoup(sa, "lxml").text
    try:
        if index!=0:

            car_det=cleantext.split("#")

            if index==1:
                ropi = car_det[1].split("-")
                car_det[1]=ropi[1]

            car_det[1]=car_det[1].replace("Δεν πληρώνει","0")
            car_det[1]=car_det[1].replace(",",".")
            car_det[1]=car_det[1].replace("zone","-1")
            car_det[1]=car_det[1].replace("Ναι","yes")
            car_det[1]=car_det[1].replace("Οχι","no")
            car_det[1]=car_det[1].replace(" ","")


            car_details.append(car_det[1])

        index=index+1
        if index==7:
            break
    except:
        print("")

ola.append(car_details)

print()
```

Εικόνα 17 Καταχώριση των χαρακτηριστικών που χρειάστηκαν απο τον πίνακα του autotriti

Αφού ολοκληρωθεί η διαδικασία ανάκτησης προσθέτουμε όλα τα χαρακτηριστικά και τις ονομασίες από τις στήλες σε ένα csv για να χρησιμοποιηθεί αργότερα με τους αλγορίθμους μηχανικής μάθησης.

```
with open(file, 'w', encoding='UTF8', newline='') as f:
    writer = csv.writer(f)
    writer.writerow(header)
    writer.writerows(ola)
```

Εικόνα 18 Εισαγωγή χαρακτηριστικών σε csv

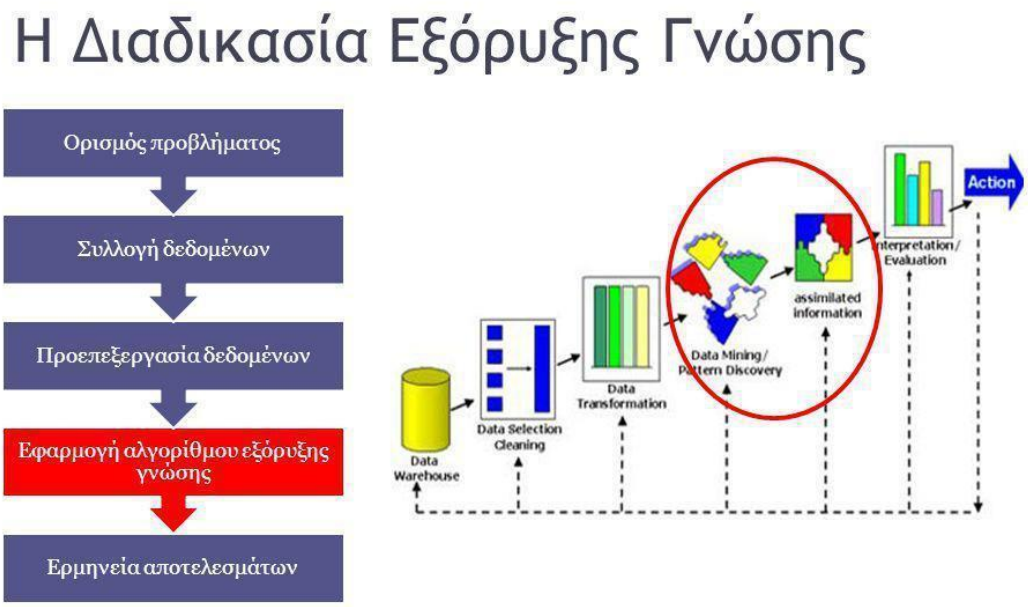
Τέλος τα δεδομένα που έχουν ανακτηθεί καταχωρούνται σε μία βάση Mysql με την χρήση της βιβλιοθήκης mysql. Κάθε φορά που εκτελείται το λογισμικό καταχωρούνται τα νέα μοντέλα με τα χαρακτηριστικά τους που παρέχονται από την ιστοσελίδα.

```
import mysql.connector as mysql

conn = mysql.connect(host='localhost', database='cars', user='root', password='aris1234')
if conn.is_connected():
    cursor = conn.cursor()
    cursor.execute("select database();")
    record = cursor.fetchone()
    cursor.execute('truncate cardata;')


    for i,row in empdata.iterrows():
        cartable = "INSERT INTO cardata(marka,modelo,ekdosi,kaysimo,cc,isxis,timi,ropi,kmh,katanalosi,kausaeria,autonomia,teli) VALUES (%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s)"
        #print(tuple(row))
        cursor.execute(cartable, tuple(row))
        conn.commit()
```

Εικόνα 19 Εισαγωγή Δεδομένων στη βάση δεδομένων

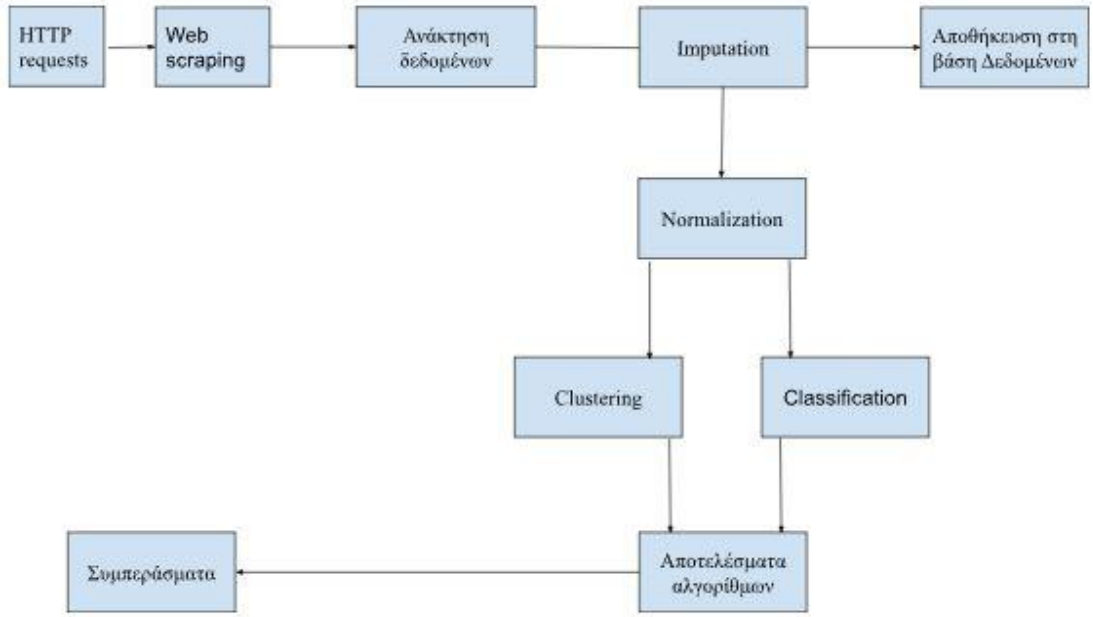
4.3 *Το σύνολο δεδομένων*

Η συλλογή δεδομένων είναι μια χρονοβόρα διαδικασία ασχέτως επιστημονικού τομέα. Αρκετές φορές το διαδίκτυο λόγω τεράστιου όγκου δεδομένων, έχει έλλειψη δεδομένων σε διάφορα σύνολα δεδομένων που παρέχονται. Αρχικά, για να ξεκινήσει κάποιος να συλλέξει δεδομένα θα πρέπει να έχει παρατηρήσει κάποιο πρόβλημα που χρήζει επίλυσης. Πάρα πολλά τα παραδείγματα όπως η ανίχνευση συγκεκριμένης ασθένειας, το πρόβλημα των μεταφορών στις πόλεις ή την επικοινωνία των ανθρώπων που ζουν μακριά. Αφού λοιπόν έχει οριστεί το πρόβλημα, σχεδόν πάντα είναι απαραίτητη η προεπεξεργασία των δεδομένων για να καταλήξει ο χρήστης σε αυτό που πραγματικά επιθυμεί. Έπειτα, αφού καταλήξει στα δεδομένα που χρειάζεται, εφαρμόζει τεχνικές όπως υλοποίηση αλγορίθμων εξόρυξης γνώσης, όπως ήδη έχουμε αναφέρει και παραπάνω. Αφού χρησιμοποιηθούν οι αλγόριθμοι εξάγονται χρήσιμα συμπεράσματα που βοηθούν στην κατανόηση του προβλήματος και πιθανότητα και στη λύση του.



Εικόνα 20 Διαδικασία Εξόρυξης Δεδομένων[64]

Στο παρακάτω διάγραμμα φαίνεται ολόκληρη η διαδικασία στα πλαίσια της παρούσας διπλωματικής βήμα βήμα έτσι ώστε να καταλήξουμε σε χρήσιμα συμπεράσματα.



Εξίσωση 6 Σταδια Υλοποιήσης Εργασίας

Η ιστοσελίδα autotriti παρέχει μια πληθώρα χαρακτηριστικών καινούργιων αυτοκινήτων. Ενδεικτικά παρακάτω φαίνονται και στη εικόνα κάποια από αυτά.

1.0 72 PS X, Βενζίνη, 998 κ.εκ, 72 PS Τιμή: 15.470€ περισσότερα	
Κυβικά (κ.εκ)	998
Ισχύς - Ροπή (PS - Nm)	72 - 93
0-100 (χλμ./ώρα)	15,6
Κατανάλωση (λτ/100 χλμ.)	4,7
Εκπομπές CO2 (γρ./χλμ.)	107
Αυτονομία με ένα ρεζερβουάρ (χλμ.)	85
Τέλη κυκλοφορίας (€/έτος)	Δεν πληρώνει
ΕΞΟΠΛΙΣΜΟΣ ΑΝΕΣΗΣ	
Air Condition	Ναι
Κλιματισμός	ΟΧΙ
Ηλεκτρικά παράθυρα εμπρός	Ναι
Ηλεκτρικά παράθυρα πίσω	ΟΧΙ
Ηλεκτρικοί καθρέφτες	ΟΧΙ
Θερμαινόμενοι καθρέπτες	ΟΧΙ
Κεντρικό κλείδωμα	Ναι
Keyless entry	ΟΧΙ

Πίνακας 4 Μερικά χαρακτηριστικά αυτοκινήτου Toyota

Πηγή:https://www.autotriti.gr/data/newcars/times/montela/times_TOYOTA.asp

Για το σκοπό της παρούσας διπλωματικής χρησιμοποιήθηκαν τα εξής χαρακτηριστικά λόγω του ότι ήταν αριθμητικά και μπορούμε να εφαρμόσουμε τους αλγόριθμους εξόρυξης γνώσης. Μάρκα, μοντέλο, έκδοση, κυβικά, ισχύς, τιμή, ροπή, χλμ/ώρα, κατανάλωση, εκπομπές CO2, αυτονομία. τέλη κυκλοφορίας.

Ενδεικτίκα τα 10 πρώτα αυτοκίνητα που έχουν καταχωρηθεί στη βάση δεδομένων.

id	marka	modelo	ekdosi	kaysimo	cc	isxis	timi	ropi	kmh	katanalosi	kausaeria	autonomia	teli
1	ABARTH	ABARTH 595	1.4 T-Jet 165 BASIC	benz	1368.0	165	23900	230	10.819999999999999	4.779999999999999	139.0	209.0	88.96
2	ABARTH	ABARTH 595	1.4 T-Jet 180 BASIC	benz	1368.0	180	27900	250	6.7	6.0	139.0	196.6	88.96
3	ALFA ROMEO	ALFA ROMEO GIULIA	2.0 T 200 AT SUPER	benz	1995.0	200	47400	330	6.6	6.4	144.0	75.0	100.8
4	ALFA ROMEO	ALFA ROMEO GIULIA	2.0 T 200 AT SPRINT	benz	1995.0	200	49400	330	6.6	6.4	144.0	75.0	100.8
5	ALFA ROMEO	ALFA ROMEO GIULIA	2.0 T 280 AT 4X4 TI	benz	1995.0	280	58200	400	5.7	6.6	151.0	80.0	105.7
6	ALFA ROMEO	ALFA ROMEO GIULIA	2.0 T 280 AT 4X4 VELOCE	benz	1995.0	280	61700	400	5.7	6.6	151.0	80.0	105.7
7	ALFA ROMEO	ALFA ROMEO GIULIA	2.2 190 AT SUPER	Diesel	2143.0	190	47600	380	7.1	4.7	126.0	294.2	80.64
8	ALFA ROMEO	ALFA ROMEO GIULIA	2.2 190 AT SPRINT	Diesel	2143.0	190	49600	380	7.1	4.7	126.0	186.6	80.64
9	ALFA ROMEO	ALFA ROMEO GIULIA	2.2 210 AT 4X4 TI	Diesel	2143.0	210	55200	470	6.8	5.1	136.0	81.4	87.04
10	ALFA ROMEO	ALFA ROMEO GIULIA	2.2 210 AT 4X4 VELOCE	Diesel	2143.0	210	58700	470	6.8	5.1	136.0	85.0	87.04

Πίνακας 5 Mysql Βάση δεδομένων αυτοκινήτων

5 *Προ επεξεργασία Δεδομένων*

5.1 Data imputation

Η προ επεξεργασία[65] είναι ένα βασικό κομμάτι στη εξόρυξη δεδομένων. Υπάρχουν διαφορετικές εργασίες εξόρυξης και μια απ' αυτές είναι η αντικατάσταση ελλιπών τιμών. Η αντικατάσταση ελλιπών τιμών είναι βασικό μέρος στη διαδικασία μηχανικής μάθησης και στη εξόρυξη δεδομένων. Η έλλειψη τιμών[66] είναι πιθανό να μην παρέχει σημαντικές γνώσεις για ένα σύνολο δεδομένων. Η συλλογή γνώσεων από μια αποθήκη δεδομένων με ελλιπείς τιμές μπορεί να εμφανίσει σημαντικά ζητήματα, πόσο μάλλον όταν έχουν χρησιμοποιηθεί αλγόριθμοι εξόρυξης δεδομένων. Τα σύνολα δεδομένων στον πραγματικό κόσμο επηρεάζονται εύκολα σε τέτοια κατάσταση και είναι πιθανό να οδηγηθούν σε διαφορετικά αποτελέσματα από τα πραγματικά. Επηρεάζουν την εποπτευόμενη μαθησιακή διαδικασία, μειώνουν την ακρίβεια ανάλυσης δεδομένων και τους αλγόριθμους κατηγοριοποίησης και αλλάζουν την τιμή των δεδομένων. Ζητήματα που εμφανίζονται σε ελλιπείς τιμές στις εργασίες εξόρυξης δεδομένων είναι: αποτελεσματικότητα, τα προβλήματα στην επεξεργασία και ανάλυση δεδομένων και η προκατάληψη. Οι ερευνητές πρέπει να διαχειριστούν σωστά τα παραπάνω αν θέλουν να βγάλουν μια αποτελεσματική και αποδεκτή ανάλυση δεδομένων. Η ακρίβεια των μοντέλων πρόβλεψης μπορεί να μειωθεί αρκετά όταν υπάρχουν τιμές που λείπουν.

Υπάρχουν διάφορες μέθοδοι που μπορούν να ληφθούν, να αγνοηθούν, να διαγραφούν και να καταλογιστούν[67]. Η επιλογή μιας μεθόδου αντικατάστασης καθορίζεται από το σύνολο δεδομένων, το μηχανισμό δεδομένων που λείπουν, μοτίβα, και μεθόδους χειρισμού τιμών που

λείπουν. Το θέμα με αυτές οι τεχνικές είναι ότι ορισμένες τεχνικές είναι πιθανό να λειτουργούν πολύ καλά σε ορισμένους τύπους δεδομένων, ενώ σε άλλους όχι τόσο. Παρουσιάζονται κάποιες από τις τεχνικές αντικατάστασης τιμών που εφαρμόζονται για να δημιουργηθούν τεχνητά δεδομένα[68].

Μέσος όρος: Ο ευκολότερος τρόπος για να αντικατασταθούν οι ελλιπείς τιμές είναι να αντικαταστήσετε κάθε τιμή που λείπει με τον μέσο όρο των τιμών για τη συγκεκριμένη μεταβλητή. Ο μέσος όρος του χαρακτηριστικού μετριέται παίρνοντας τις τιμές που δεν λείπουν και υλοποιείται για να αντικαταστήσει τις τιμές που λείπουν με αυτές των χαρακτηριστικών.

Κ-Πλησιέστερος γείτονας: Το K-Nearest Neighbor είναι μια μέθοδος προ-αντικατάστασης που αντικαθιστά τις τιμές, πριν από τη διαδικασία εξόρυξης δεδομένων. κατηγοριοποιεί τα δεδομένα σε συστάδες και στη συνέχεια αντικαθιστά τις τιμές που λείπουν με την αντίστοιχη τιμή από τον πλησιέστερο γείτονα. Ο πλησιέστερος γείτονας είναι η πλησιέστερη τιμή σύμφωνα με την Ευκλείδεια απόσταση. Οι τιμές που λείπουν αντικαθιστούν από δεδομένα που είναι παρόμοια.

Μεγιστοποίηση προσδοκιών Η μεγιστοποίηση προσδοκιών διαθέτει εκτιμήσεις των μέσων και πίνακες συνδιακύμανσης που έχουν την ευχέρεια να εφαρμοστούν για να υπάρξει συνέπεια στις εκτιμήσεις των τιμών που λείπουν. Στηρίζεται σε ένα βήμα προσδοκίας και βήμα μεγιστοποίησης, τα οποία επαναλαμβάνονται πολλές φορές μέχρι να παρθεί η εκτίμηση μέγιστης πιθανότητας. Αυτή η μέθοδος απαιτεί μεγάλο όγκο δεδομένων.

Hot-Deck: Όταν μια τιμή λείπει αντικαθίσταται με μια παρατηρούμενη τιμή που βρίσκεται πιο κοντά στους όρους απόστασης. Δηλαδή το Hot-Deck διαλέγει τυχαία μια παρατηρούμενη τιμή από μια συστάδα που είναι παρόμοια σύμφωνα τις επιλεγμένες μεταβλητές. Το Hot-Deck χωρίζεται σε δύο φάσεις. Στη πρώτη φάση, τα δεδομένα είναι χωρισμένα σε συστάδες. Στη δεύτερη φάση, κάθε τιμή που λείπει ταιριάζει μόνο με μια συστάδα. Όλες οι περιπτώσεις σε μια συστάδα υλοποιούνται για τη καταχώρηση των τιμών που λείπουν. Αυτό γίνεται από έναν πίνακα συσχέτισης που χρησιμοποιείται για τον προσδιορισμό των υψηλά συσχετισμένων μεταβλητών.

Για την ανάγκες της παρούσας διπλωματικής, χρησιμοποιήθηκε ο KNN imputer με τις κατάλληλες τροποποιήσεις που εφαρμόστηκαν για αντικατάσταση των τιμών που λείπουν από το σύνολο δεδομένων. Επίσης, παρατηρούμε ότι καταχωρούνται αριθμητικές τιμές σε κατηγορικά δεδομένα, έτσι ώστε να έχουμε περισσότερα δεδομένα για τους αλγόριθμους και

καλύτερες συσχετίσεις σε τιμές που λείπουν. Αυτό γίνεται μόνο για να εφαρμοστούν οι αλγόριθμοι εξόρυξης δεδομένων διότι στη βάση δεδομένων τα στοιχεία της στήλης καυσίμου εισέρχονται ως κατηγορικά.

```
from sklearn.impute import KNNImputer
empdata=empdata.replace('yes',1)
empdata = empdata.replace(r'^\s*$', np.nan, regex=True)
NaN= np.nan
empdata = empdata.dropna( how='any',
                    subset=['kaysimo'])
empdata=empdata.replace(1,np.nan)
cols=['cc','isxis','timi','ropi','kmh','katanalosi','kausaeria','autonomia']
empdata['cc']=empdata['cc'].replace("0",np.nan)
empdata['isxis']=empdata['isxis'].replace(0,np.nan)
empdata['ropi']=empdata['ropi'].replace(0,np.nan)
empdata['katanalosi']=empdata['katanalosi'].replace("0",np.nan)
empdata['kmh']=empdata['kmh'].replace("0",np.nan)
empdata['kausaeria']=empdata['kausaeria'].replace("0",np.nan)
empdata['autonomia']=empdata['autonomia'].replace("0",np.nan)
empdata=empdata.replace("benz",1)
empdata=empdata.replace("Diesel",2)
empdata=empdata.replace("Mild hybrid",3)
empdata=empdata.replace("Plug-in hybrid",4)
empdata=empdata.replace("electric",5)
empdata=empdata.replace("Diesel Mild Hybrid",6)
empdata=empdata.replace("Lpg",7)
empdata=empdata.replace("Mild Hybrid",3)
empdata=empdata.replace("Hybrid",8)
empdata=empdata.replace("Plug-In Hybrid",4)
empdata=empdata.replace("Cng",9)

imputer= KNNImputer(n_neighbors=5, weights="uniform")
print(imputer.fit_transform(empdata.iloc[:,3:13]))
empdata.iloc[:,3:13] =imputer.fit_transform(empdata.iloc[:,3:13] )
empdata.to_csv(file, sep=',', encoding='utf-8',index=False)
```

Εικόνα 21 KNN Imputation

5.2 Normalization

Ο μετασχηματισμός δεδομένων[69] όπως η κανονικοποίηση έχει την δυνατότητα να αυξήσει την ακρίβεια και την αποτελεσματικότητα αλγορίθμων εξόρυξης που περιέχουν νευρωνικά δίκτυα, πλησιέστερο γείτονα και αλγόριθμους συσταδοποίησης. Αυτές οι μέθοδοι διαθέτουν βέλτιστα αποτελέσματα αν τα δεδομένα έχουν κανονικοποιηθεί. Οι αναλυτές δεδομένων[70]

χρειάζεται να κανονικοποιήσουν τις αριθμητικές μεταβλητές, καθώς καταχωρούν κάθε μεταβλητή στη ίδια κλίμακα. Η κανονικοποίηση όλων των μεταβλητών στην ίδια κλίμακα είναι ζωτικής σημασίας κατά την υλοποίηση πράξεων που είναι ευάλωτα σε διακύμανση ή εξάπλωση δεδομένων για να μπορέσουν οι μεταβλητές που έχουν μικρότερες διακυμάνσεις μεταβλητές να υπερισχύουν από τις υψηλότερες. Ένα χαρακτηριστικό κανονικοποιείται κλιμακώνοντας τις τιμές του σε ένα μικρό εύρος, συνήθως από 0 ως 1. Εάν υλοποιείται ο αλγόριθμος επαναδιάδοσης του νευρικού δικτύου για κατηγοριοποίηση, κανονικοποιώντας τις τιμές εισόδου για κάθε χαρακτηριστικό που υπολογίζεται στα δείγματα εκπαίδευσης συμβάλλει στην εγρήγορση της μάθησης. Για μεθόδους που στηρίζονται στη απόσταση, η κανονικοποίηση συμβάλλει στην πρόληψη χαρακτηριστικών με τεράστια εύρη από τα αντισταθμιστικά χαρακτηριστικά με μικρότερα εύρη.

Διάφοροι τύποι τεχνικών κανονικοποίησης είναι διαθέσιμες όπως[71].

Κανονικοποίηση ελάχιστης μέγιστης (Min-Max): Η κανονικοποίηση ελάχιστης μέγιστης υλοποιεί μια γραμμική επεξεργασία στα αρχικά δεδομένα. Οι τιμές κανονικοποιούνται εντός ενός καθορισμένου εύρους. Το πλεονέκτημα της κανονικοποίησης Min-Max είναι ότι όλες οι τιμές είναι μέσα σε αυτό το εύρος.

Κανονικοποίηση βαθμολογίας Z: Η κανονικοποίηση βαθμολογίας Z, αποκαλείται επιπροσθέτως μηδενική μέση κανονικοποίηση. Τα δεδομένα κανονικοποιούνται με βάση τη μέση και τη τυπική απόκλιση.

Κανονικοποίηση δεκαδικής κλίμακας: Η κανονικοποίηση δεκαδικής κλίμακας στηρίζεται στην κίνηση του δεκαδικού σημείου της τιμής του χαρακτηριστικού. Οι αριθμοί υποδιαστολής μετακινούνται βασιζόμενοι από τις μέγιστες απόλυτες τιμές του χαρακτηριστικού.

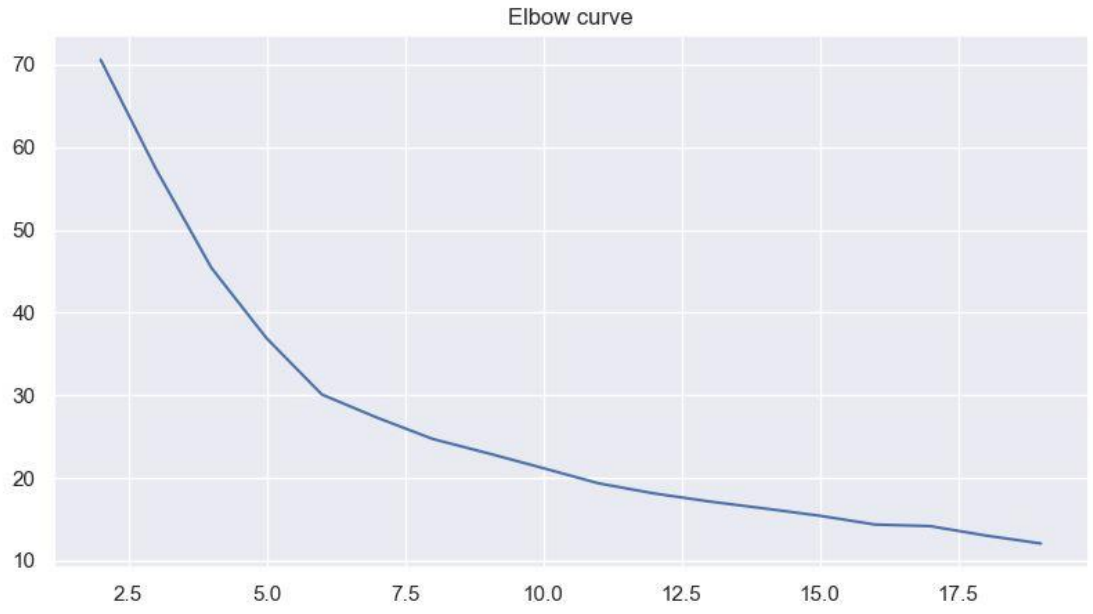
Επιλέχθηκε από’ τις παραπάνω τεχνικές κανονικοποίησης η μέθοδος μέγιστης ελάχιστης, διότι θα έχει πολύ καλύτερα αποτελέσματα από τις υπόλοιπες.

5.3 Συσταδοποίηση

Αφού έχει υλοποιηθεί σωστά η προ επεξεργασία των δεδομένων και έχει δημιουργηθεί το σύνολο δεδομένων με τα χαρακτηριστικά που επιθυμούμε, θα προχωρήσουμε στη ανάπτυξη των αλγορίθμων εξόρυξης δεδομένων και συγκεκριμένα της συσταδοποίησης αρχικά για να βγάλουμε χρήσιμα συμπεράσματα.

5.3.1 Αποτελέσματα K-means

Αρχικά υλοποιήθηκε ο K-means αλγόριθμος με την βοήθεια των κατάλληλων βιβλιοθηκών και των κατάλληλων τροποποιήσεων. Δημιουργούμε με τον k-means το σχεδιάγραμμα της μεθόδου του αγκώνα όπως φαίνεται στη Εικόνα. Φαίνεται λοιπόν ότι το κατάλληλο k είναι το 4 καθώς βλέπουμε πως η γραμμή κάνει μεγάλη καμπύλη σε εκείνο το σημείο.



Εξίσωση 7 Γράφημα Elbow Curve

Στη συνέχεια καθορίζουμε τον αριθμό των clusters και βγάζουμε για το καθένα το τετραγωνικό σφάλμα (SSE).

Συστάδα 0	30.71587614771997
Συστάδα 1	6.91322736919531
Συστάδα 2	5.163109429051768
Συστάδα 3	2.6283315648048546

Αφού έχουμε εφαρμόσει data imputation και normalization θα παρουσιαστεί το παρακάτω διάγραμμα με εύρος 0 έως 1. Η ακόλουθη ανάλυση των συστάδων απ' το γράφημα γίνεται κατά προσέγγιση, διότι μέσα σε αυτά υπάρχουν και στοιχεία που διαφέρουν.

Η συστάδα 0 έχει αυτοκίνητα μεσαίου και μεγάλου κυβισμού, μεσαία και μεγάλη ισχύ, σχετικά χαμηλή τιμή, σχετικά μικρή ροπή, χαμηλή απόκριση χλμ/ώρα, σχετικά χαμηλή

κατανάλωση, υψηλή εκπομπή καυσαερίων, μεσαία αυτονομία (εδώ εννοούμε το ντεπόζιτο στο ρεζερβουάρ) και υψηλά τέλη κυκλοφορίας.

Η συστάδα 1 έχει αυτοκίνητα μικρού κυβισμού, μεσαία ισχύ, χαμηλή τιμή, σχετικά μικρή ροπή, χαμηλή απόκριση στα χλμ/ώρα, μέτρια και υψηλή κατανάλωση, σχετικά χαμηλή εκπομπή καυσαερίων, μικρή αυτονομία και χαμηλά τέλη κυκλοφορίας.

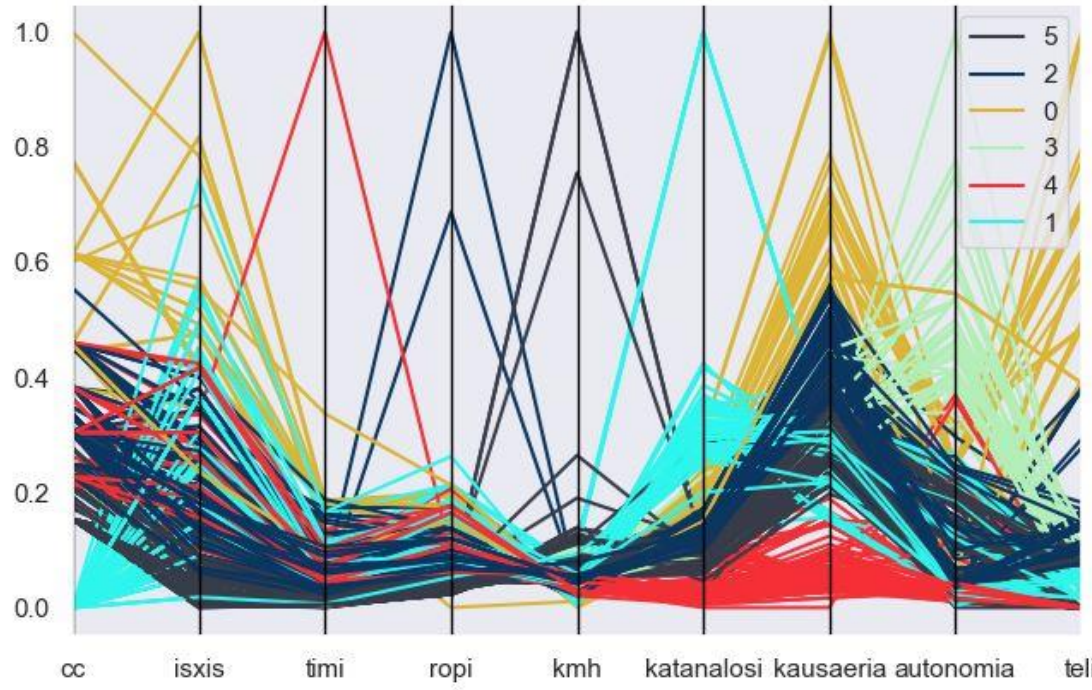
Η συστάδα 2 έχει αυτοκίνητα μεσαίου κυβισμού, σχετικά μικρή ισχύ, σχετικά χαμηλή τιμή, σχετικά μικρή ροπή, χαμηλή απόκριση χλμ/ώρα, σχετικά χαμηλή κατανάλωση, μέτρια εκπομπή καυσαερίων, σχετικά μικρή αυτονομία και χαμηλά και μέτρια τέλη κυκλοφορίας.

Η συστάδα 3 έχει αυτοκίνητα σχετικά μικρού και μεσαίου κυβισμού, μικρή ισχύ, χαμηλή τιμή, μικρή ροπή, χαμηλή απόκριση χλμ/ώρα, σχετικά χαμηλή κατανάλωση, μέτρια εκπομπή καυσαερίων, μεσαία και μεγάλη αυτονομία και χαμηλά τέλη κυκλοφορίας.

Η συστάδα 4 έχει αυτοκίνητα μεσαίου κυβισμού, μεσαία ισχύ, χαμηλή και υψηλή τιμή, σχετικά μικρή ροπή, χαμηλή απόκριση χλμ/ώρα, χαμηλή κατανάλωση, χαμηλή εκπομπή καυσαερίων, μικρή και μεσαία αυτονομία και χαμηλά τέλη κυκλοφορίας.

Η συστάδα 5 έχει αυτοκίνητα μικρού και μεσαίου κυβισμού, μικρή και μεσαία ισχύ, χαμηλή τιμή, μικρή ροπή, σχετικά χαμηλή και υψηλή απόκριση χλμ/ώρα, σχετικά χαμηλή κατανάλωση, μετρία εκπομπή καυσαερίων, σχετικά μικρή αυτονομία, και σχετικά χαμηλά τέλη κυκλοφορίας.

Στον ακόλουθο πίνακα φαίνεται ο αριθμός των στοιχείων που καταχωρήθηκαν ανά συστάδα.

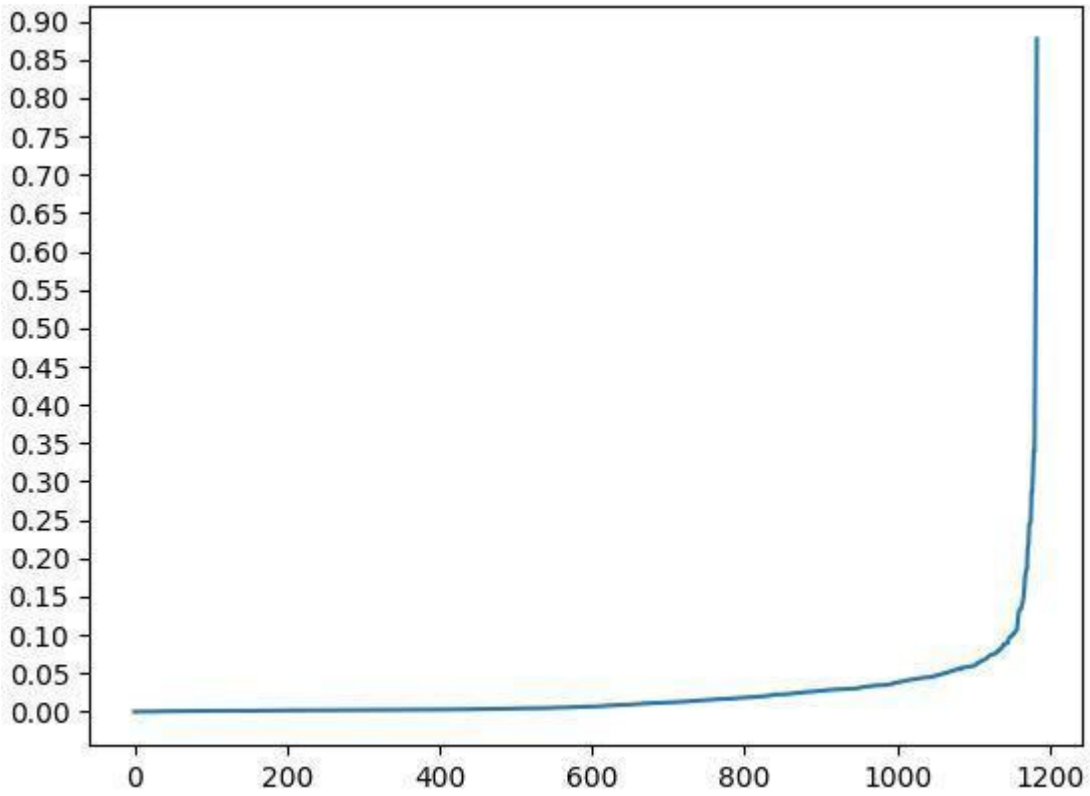


Εξίσωση 8 Γράφημα αναπαράστασης στοιχείων των συστάδων

Συστάδα 0	24
Συστάδα 1	127
Συστάδα 2	236
Συστάδα 3	68
Συστάδα 4	94
Συστάδα 5	646

5.3.2 Αποτελέσματα DBSCAN

Αφού ολοκληρώθηκε ο αλγόριθμος K-means στη συνέχεια θα εξεταστεί ο αλγόριθμος DBSCAN. Με παρόμοιο τρόπο όπως και στον kmeans, τοποθετούμε τα δεδομένα που επιθυμούμε να χρησιμοποιήσουμε για συσταδοποίηση και με τις κατάλληλες επεξεργασίες βρίσκουμε το κατώφλι eps που μας ενδιαφέρει. Στο σχήμα αναπαριστώνται οι πλησιέστεροι γείτονες από τους οποίους μπορεί να υπολογιστεί το eps, δηλαδή η μέγιστη απόσταση δύο σημείων για να θεωρηθούν γείτονες. Μετά από αρκετές δοκιμές καταλήξαμε στο συμπέρασμα πως για να ύπαρχει μια ορθή ανακατανομή των αυτοκινητών σε συστάδες πως το eps=0.10 και minpts=10.



Εξίσωση 9 Γράφημα Πλησιέστερων Γειτόνων

Αφού εφαρμόσαμε τον αλγόριθμο DBSCAN με την βοήθεια της βιβλιοθήκης parallel coordinates παρατηρούμε την συσταδοποίηση των συστάδων σε γράφημα.

Αρχικά, όπως φαίνεται και στον παρακάτω πίνακα, η συστάδα 0 λόγω της συλλογής των περισσότερων στοιχείων μονοπωλεί το ενδιαφέρον του γραφήματος, καθώς υπάρχουν αρκετές παράλληλες με το χρώμα της συστάδας.

Η συστάδα 0 έχει αυτοκίνητα μικρού και μεσαίου κυβισμού, μικρή και μεσαία ισχύ, χαμηλή τιμή, σχετικά μικρή ροπή, χαμηλή απόκριση χλμ/ώρα, χαμηλή και μέτρια κατανάλωση, μέτρια εκπομπή καυσαερίων, μικρή και μεσαία αυτονομία και χαμηλά τέλη κυκλοφορίας.

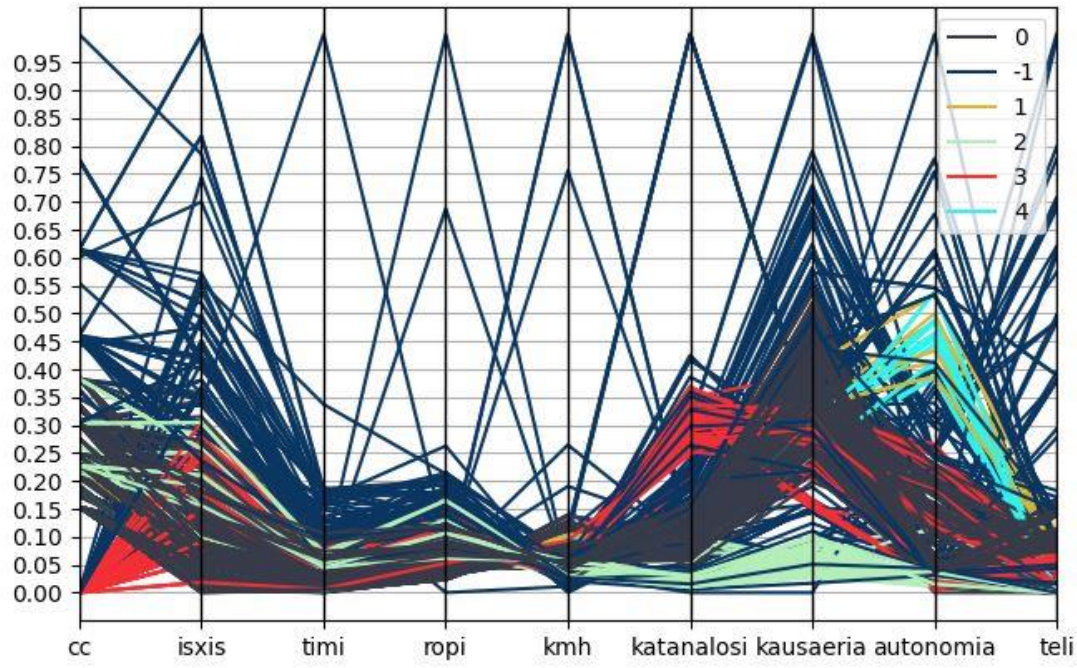
Η συστάδα 1 έχει αυτοκίνητα σχετικά μικρού κυβισμού, σχετικά μικρή ισχύ, χαμηλή τιμή, μικρή ροπή, σχετικά χαμηλή απόκριση χλμ/ώρα, σχετικά χαμηλή κατανάλωση, μέτρια εκπομπή καυσαερίων, μεσαία αυτονομία και χαμηλά τέλη κυκλοφορίας.

Η συστάδα 2 έχει αυτοκίνητα μεσαίου κυβισμού, μεσαία ισχύ, χαμηλή τιμή, σχετικά μικρή ροπή, χαμηλή απόκριση χλμ/ώρα, χαμηλή κατανάλωση, χαμηλή εκπομπή καυσαερίων, μικρή αυτονομία και χαμηλά τέλη κυκλοφορίας.

Η συστάδα 3 έχει αυτοκίνητα μικρού κυβισμού, μεσαία ισχύ, χαμηλή τιμή, σχετικά μικρή ροπή, χαμηλή απόκριση χλμ/ώρα, μέτρια κατανάλωση, μέτρια εκπομπή καυσαερίων, μικρή και μεσαία αυτονομία και χαμηλά τέλη κυκλοφορίας.

Η συστάδα 4 έχει αυτοκίνητα σχετικά μικρού κυβισμού, μικρη ισχύ, σχετικά χαμηλή τιμή, σχετικά μικρή ροπή, χαμηλή απόκριση χλμ/ώρα, μέτρια κατανάλωση, μέτρια εκπομπή καυσαερίων, μεσαία αυτονομία και χαμηλά τέλη κυκλοφορίας.

Τα ακραία σημεία(outliers) έχει αυτοκίνητα μεσαίου και μεγάλου κυβισμού, μεγάλη ισχύ, χαμηλή και υψηλή τιμή, μικρή και μεγάλη ροπή, υψηλή απόκριση χλμ/ώρα, μέτρια κατανάλωση, υψηλή εκπομπή καυσαερίων, μεσαία και μεγάλη αυτονομία και σχετικά υψηλά τέλη κυκλοφορίας.



Εξίσωση 10 Γράφημα αναπαράστασης στοιχείων των συστάδων

Στον ακόλουθο πίνακα φαίνεται ο αριθμός των στοιχείων που καταχωρήθηκαν ανά συστάδα. Τέλος, έχουμε και τα σημεία που δεν έχουν κανένα σημείο σαν γείτονα και θεωρούνται ακραία σημεία (outliers).

Συστάδα 0	854
Συστάδα 1	13
Συστάδα 2	66
Συστάδα 3	96
Συστάδα 4	32
Ακραία Σημεία	132

Ενδεικτικά φαίνονται τα αυτοκίνητα που θεωρήθηκαν outliers.

OUTLIERS

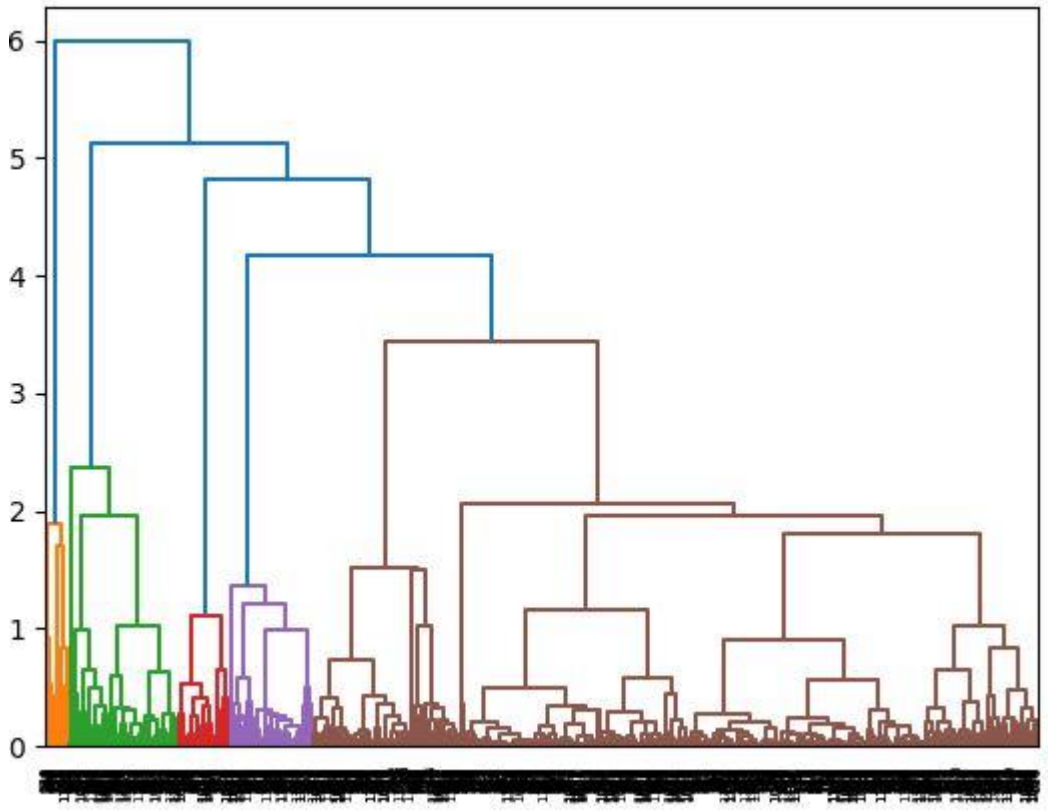
	marka	modelo	ekdosi	kaysimo	...	kausaeria	autonomia	teli	anomaly_score
11	ALFA ROMEO	ALFA ROMEO GIULIA	2.9 510 QV AT QUADRIFOGLIO	1.0	...	227.0	529.4	499.400	-1
22	ALFA ROMEO	ALFA ROMEO STELVIO	2.9 510 QV 4X4 QUADRIFOGLIO	1.0	...	267.0	520.8	720.900	-1
44	AUDI	AUDI e-tron GT	GT quattro	5.0	...	106.4	94.4	34.176	-1
45	AUDI	AUDI e-tron GT	RS GT	5.0	...	106.4	98.2	34.176	-1
55	AUDI	AUDI Q3	SPORTBACK 35 TDI S tronic BASIC	2.0	...	133.0	1160.0	85.120	-1
...	...	...	...	...	...	...	...	...	...
1173	VOLVO	VOLVO S90	T8 2.0 455 AWD PLUS BRIGHT	4.0	...	18.0	74.0	0.000	-1
1174	VOLVO	VOLVO XC40	RECHARGE PURE ELECTRIC P8 AWD 408 ULTIMATE	5.0	...	113.2	72.6	44.772	-1
1175	VOLVO	VOLVO XC40	RECHARGE PURE ELECTRIC P8 AWD 408 PLUS	5.0	...	89.2	77.4	17.280	-1
1190	VOLVO	VOLVO XC90	B5 AWD 235 PLUS BRIGHT 7-S	2.0	...	179.0	425.0	152.150	-1
1191	VOLVO	VOLVO XC90	T8 AWD 2.0 455 PLUS BRIGHT 7-S	4.0	...	30.0	94.4	0.000	-1

Εικόνα 22 Outliers Αυτοκίνητα

5.3.3 Αποτελέσματα Ιεραρχικής Συσταδοποίησης

Τελευταίος αλγόριθμος συσταδοποίησης της παρούσας διπλωματικής είναι ο agglomerative (ιεραρχική συσταδοποίηση). Σύμφωνα με το κριτήριο σύνδεσης (linkage), καθορίζονται οι αποστάσεις μεταξύ των σημείων. Συγχωνεύονται τα ζευγάρια των συστάδων που ελαχιστοποιούν αυτό το κριτήριο. Υπάρχουν 4 τρόποι σύνδεσης: ward, complete, single, average. Χρησιμοποιήθηκε μόνο η ward μέθοδος για εξαγωγή συμπερασμάτων και θα φανεί από τα παρακάτω γραφήματα ο λόγος.

Όπως φαίνεται και στο παρακάτω γράφημα, αν κόψουμε το δέντρο περίπου στο 4.5 δημιουργούνται 4 συστάδες, από τις οποίες θα βγάλουμε τα συμπεράσματα. Αν κοιτάξει κάποιος τις υπόλοιπες 3 μεθόδους θα παρατηρήσει ότι δεν υπάρχει κάποιο καλό σημείο κοψίματος, ώστε να δημιουργηθούν ομοιόμορφα οι συστάδες, γι' αυτόν το λόγο και απορρίπτονται.



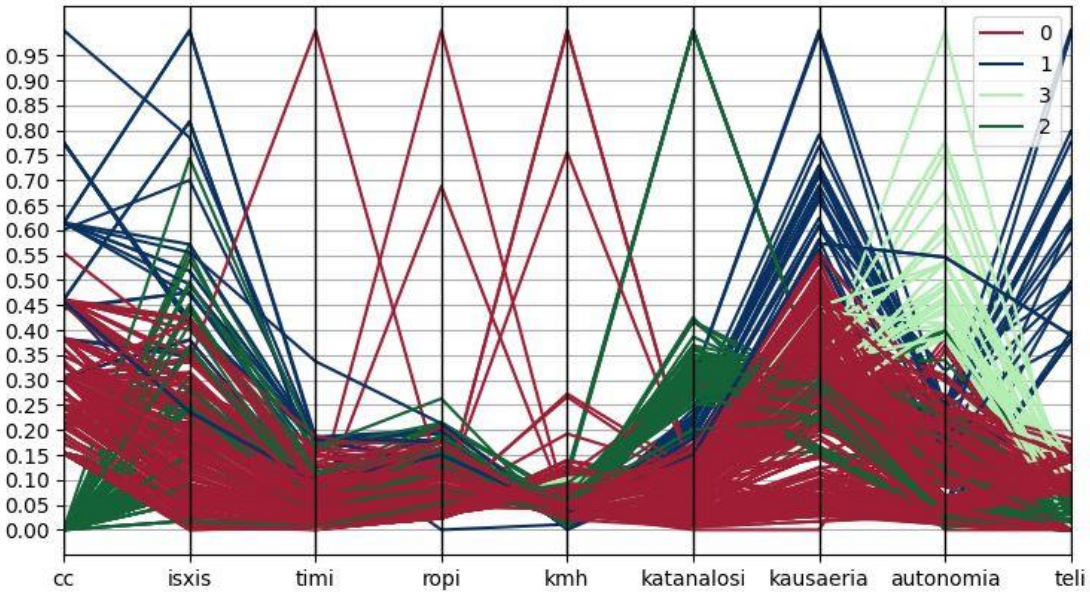
Εξίσωση 11 Γράφημα Linkage Ward

Η συστάδα 0 έχει αυτοκίνητα μεσαίου κυβισμού, μικρή και μεσαία ισχύ, σχετικά χαμηλή τιμή, σχετικά μικρή αλλά και μεγάλη ροπή, χαμηλή και υψηλή απόκριση χλμ/ώρα, χαμηλή κατανάλωση, χαμηλή και μέτρια εκπομπή καυσαερίων, μικρή και μεσαία αυτονομία και χαμηλά τέλη κυκλοφορίας.

Η συστάδα 1 έχει αυτοκίνητα μεσαίου και μεγάλου κυβισμού, μεσαία και μεγάλη ισχύ, σχετικά χαμηλή τιμή, σχετικά μικρή ροπή, χαμηλή απόκριση χλμ/ωρα, χαμηλή κατανάλωση, υψηλή εκπομπή καυσαερίων, μικρή αυτονομία και υψηλά τέλη κυκλοφορίας.

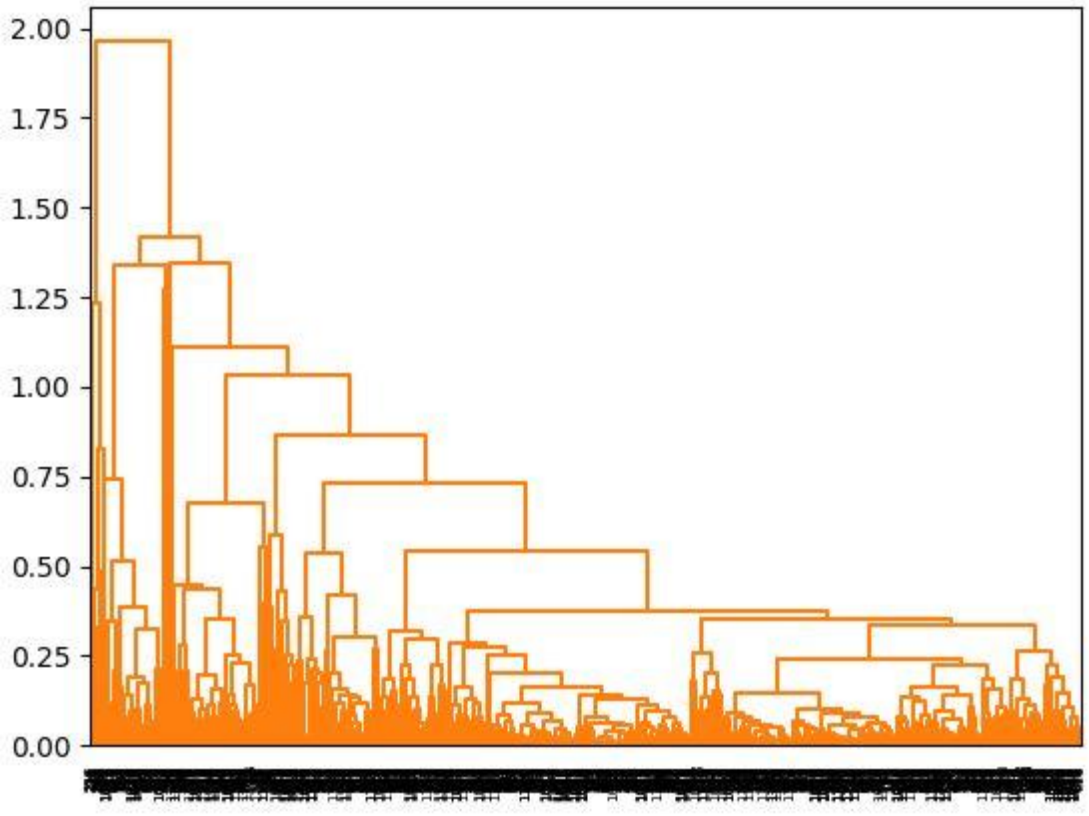
Η συστάδα 2 έχει αυτοκίνητα μικρού κυβισμού, μεσαία ισχύ, σχετικά χαμηλή τιμή, σχετικά μικρή ροπή, χαμηλή απόκριση χλμ/ώρα, μεσαία και υψηλή κατανάλωση, μέτρια εκπομπή καυσαερίων, μικρή αυτονομία και σχετικά χαμηλά τέλη κυκλοφορίας.

Η συστάδα 3 έχει αυτοκίνητα μικρού κυβισμού, σχετικά μικρή ισχύ, χαμηλή τιμή, μικρή ροπή, σχετικά χαμηλή απόκριση χλμ/ώρα, χαμηλή κατανάλωση, μέτρια εκπομπή καυσαερίων, μεσαία και μεγάλη αυτονομία, και σχετικά χαμηλά τέλη κυκλοφορίας.

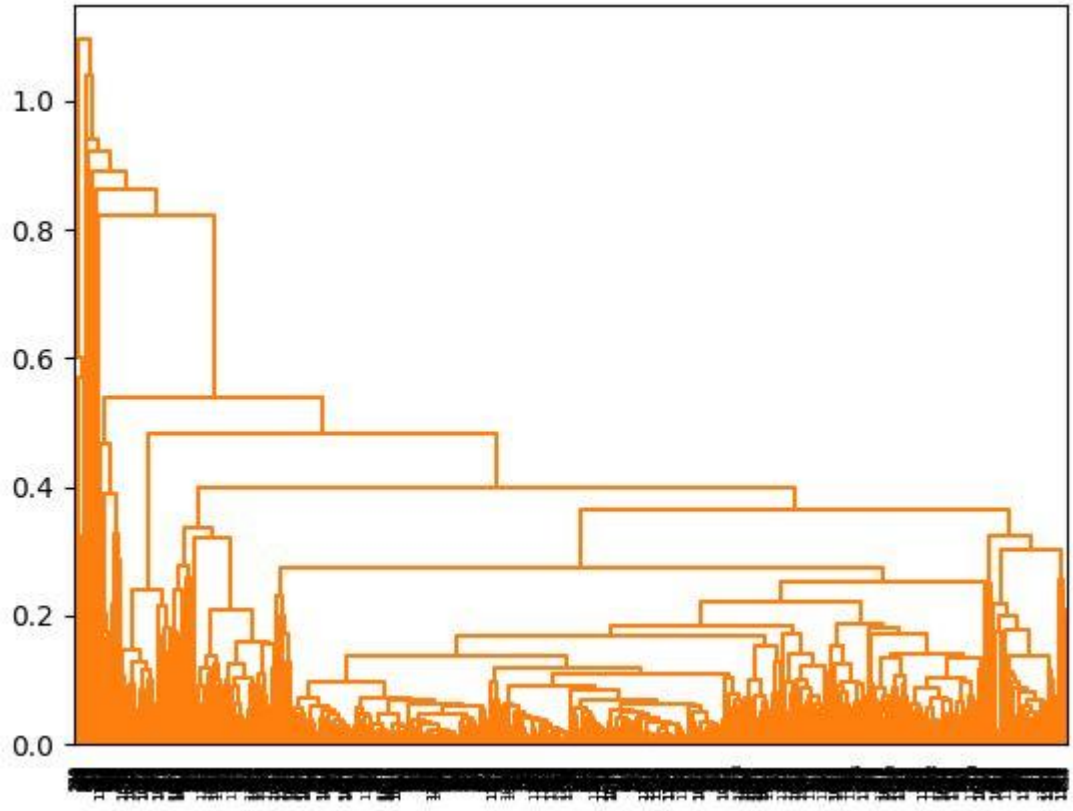


Εξίσωση 12 Γράφημα αναπαράστασης στοιχείων των συστάδων

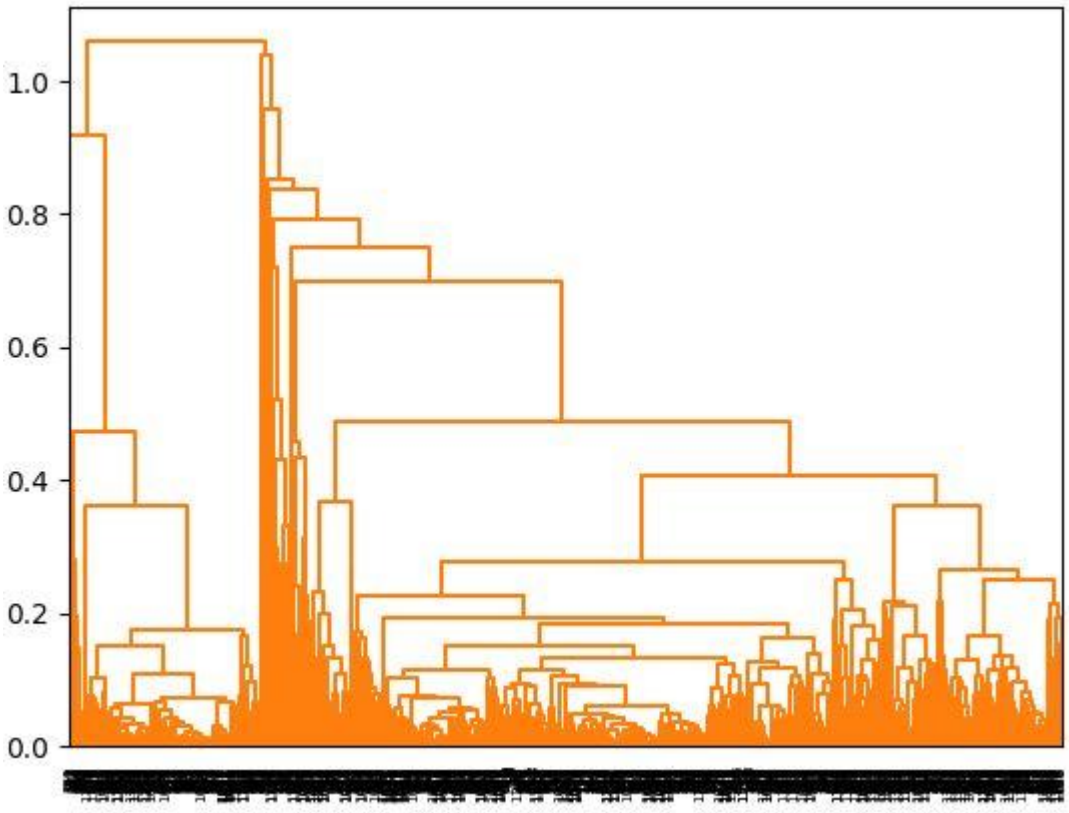
Συστάδα 0	966
Συστάδα 1	130
Συστάδα 2	64
Συστάδα 3	29



Εξίσωση 13 Γράφημα Complete Linkage



Εξίσωση 14 Γράφημα Single Linkage



Εξίσωση 15 Γράφημα Average Linkage

5.4 Κατηγοριοποίηση

Αφού έχει υλοποιηθεί σωστά η προεπεξεργασία των δεδομένων και οι αλγόριθμοι συσταδοποίησης, στη συνέχεια εξετάζονται οι αλγόριθμοι κατηγοριοποίησης. Με την χρήση κατάλληλων βιβλιοθηκών και κατάλληλων τροποποιήσεων, τα δεδομένα χωρίζονται με τη μέθοδο split σε σετ εκπαίδευσης 70%(train set) και σετ πρόβλεψης 30%(test set). Δηλαδή, από τις 1184 καταχωρήσεις, 832 είναι για εκπαίδευση και 352 για πρόβλεψη. Με τον ίδιο τρόπο, υλοποιήθηκαν οι δύο αλγόριθμοι Decision tree και KNN. Υπάρχουν και για τους 2 αλγόριθμους δύο πειράματα, όπου στο πρώτο παίρνουμε όλα τα αριθμητικά χαρακτηριστικά (κυβικά, ισχύς, τιμή, ροπή, χλμ/ώρα, κατανάλωση, καυσαέρια, αυτονομία, τέλη κυκλοφορίας) και στο δεύτερο μόνο τα χαρακτηριστικά που έχουν σχέση με την απόδοση του αυτοκινήτου.

5.4.1 Αυτόματη αναγνώριση τύπου καυσίμου

Η επιλογή τύπου καυσίμου σε ένα καινούργιο αυτοκίνητο είναι αρκετά σημαντική για τους ενδιαφερόμενους. Υπάρχει πλέον πληθώρα επιλογών στη αγορά ανάλογα για τις ανάγκες του καθενός. Οι κατασκευαστές παρέχουν διαφορετικές κατηγορίες καυσίμου ανάλογα με τις

ανάγκες. Η προστασία του περιβάλλοντος και η αυστηροποίηση των νόμων σχετικά με τον τομέα αυτόν, προσπαθούν να στρέψουν την αγορά σε αναζήτηση εναλλακτικών καυσίμων για τα οχήματα, τόσο για τις αυτοκινητοβιομηχανίες όσο και για τους καταναλωτές. Πέρα όμως απ' την προστασία του περιβάλλοντος, συντελούν και πολλοί άλλοι παράγοντες που επηρεάζουν την απόφαση αγοράς ενός αυτοκινήτου. Ο ενδιαφερόμενος μπορεί να επιλέξει αυτοκίνητο με πολύ υψηλές αποδόσεις ταχύτητας και έτσι προβαίνει στην αγορά ενός βενζινοκίνητου αυτοκινήτου. Αυτός που θέλει διάρκεια ζωής του κινητήρα θα επιλέξει πετρελαιοκίνητο. Μια ,επίσης, σημαίνουσα παράμετρος είναι το κόστος, οπότε αρκετοί καταφεύγουν στην λύση του φυσικού αερίου ή υγραερίου. Γι' αυτούς τους λόγους ακριβώς, αποφασίστηκε να εφαρμοστούν αλγόριθμοι κατηγοριοποίησης με τα χαρακτηριστικά των αυτοκινήτων, ώστε να εξαχθούν χρήσιμα συμπεράσματα στην επιλογή τύπου καυσίμου.

5.4.2 Αποτελέσματα Decision Tree

Αρχικά, παρουσιάζονται το μοντέλο του DT με όλα τα αριθμητικά δεδομένα. Όπως φαίνεται και στο Confusion Matrix, εξάγονται αξιόπιστα αποτελέσματα με την χρήση του αλγορίθμου για τον τύπο καυσίμου. Σύμφωνα με την θεωρία, η διαγώνιος αντιπροσωπεύει τις σωστά προβλεπόμενες τιμές, ενώ τα υπόλοιπα δείχνουν σε ποια κατηγορία τα καταχωρεί και θεωρούνται λάθος. Ο άξονας του χ είναι οι σωστές προβλέψεις ενώ ο άξονας του y αντιπροσωπεύει όλες τις προβλέψεις.

M.h←Mild Hybrid

P.I.h←Plug In-Hybrid

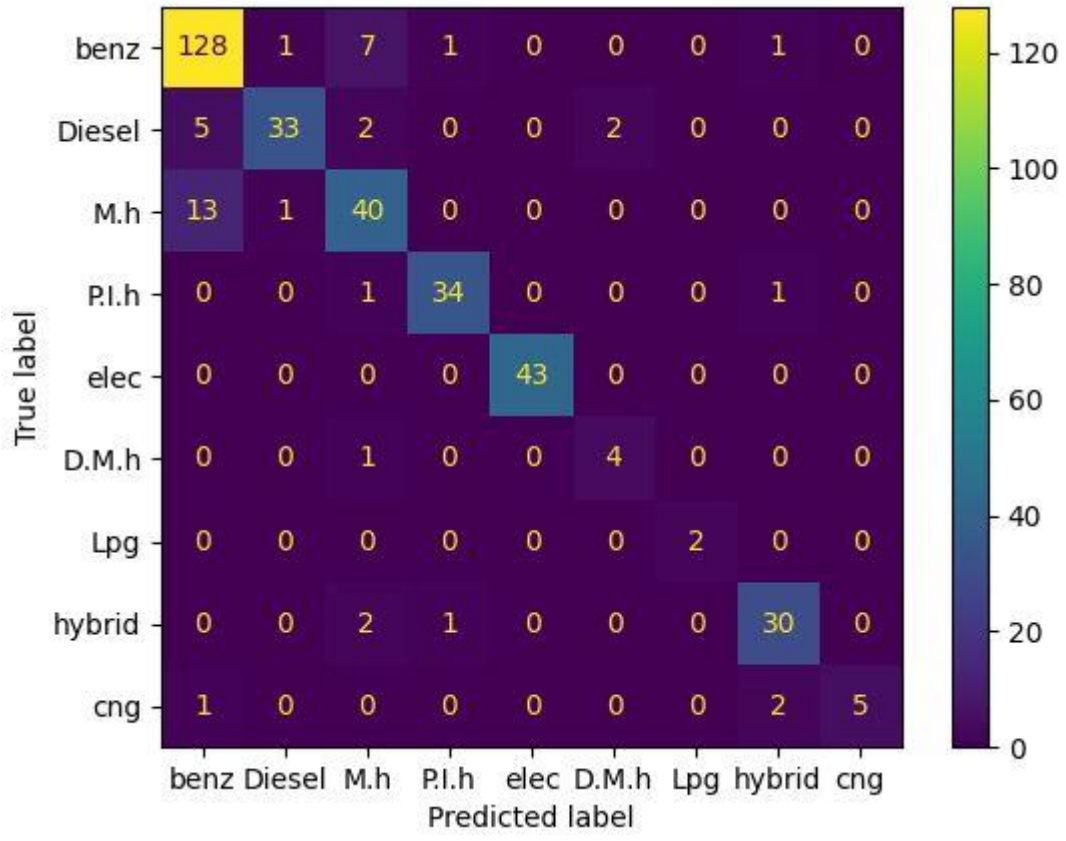
D.M.h←Diesel Mild Hybrid

Ακρίβεια(Accuracy): Η ακρίβεια[72] είναι ένα σύστημα μέτρησης σε δυαδική κατηγοριοποίηση που ορίζει η πληροφορία που ανακτήθηκε είναι σωστή ή λάθος. Εφαρμόζεται ως διαφορετική μέθοδος της βαθμολογίας F και υπολογίζεται ως: (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives).

Ακρίβεια(Precision): Η Ακρίβεια είναι η ποσοστιαία τιμή που δείχνει πόσα αποτελέσματα είναι σωστά που στηρίζονται στις ελπίδες μιας συγκεκριμένης εφαρμογής. Χρησιμοποιηθεί σε κάθε κατηγορία ενός συστήματος πρόβλεψης τεχνητής νοημοσύνης.

Ανάκληση: Η ανάκληση είναι η ποσοστιαία τιμή που δείχνει πόσα σωστά αποτελέσματα επιστράφηκαν που στηρίζονται στις ελπίδες μιας συγκεκριμένης εφαρμογής. Χρησιμοποιείται σε κάθε κατηγορία ενός συστήματος πρόβλεψης τεχνητής νοημοσύνης.

F-score(F-measure): Το F-score είναι ο μέσος όρος των τιμών ακρίβειας(precision) και Ανάκλησης(recall) ενός συστήματος και βρίσκεται από τον τύπο: 2 x [(Ακρίβεια x Ανάκληση) / (Ακρίβεια + Ανάκληση)].



Πίνακας 6 Confusion Matrix DT με όλα τα αριθμητικά χαρακτηριστικά

Παρακάτω φαίνονται σε πίνακες οι μετρικές που βρέθηκαν για τον κάθε τύπο καυσίμου.

Accuracy	0.88

Precision

benz	diesel	M.h	P.I.h	elec	D.M.h	Lpg	hybrid	cng
0.87	0.94	0.75	0.94	1	0.66	1	0.88	1

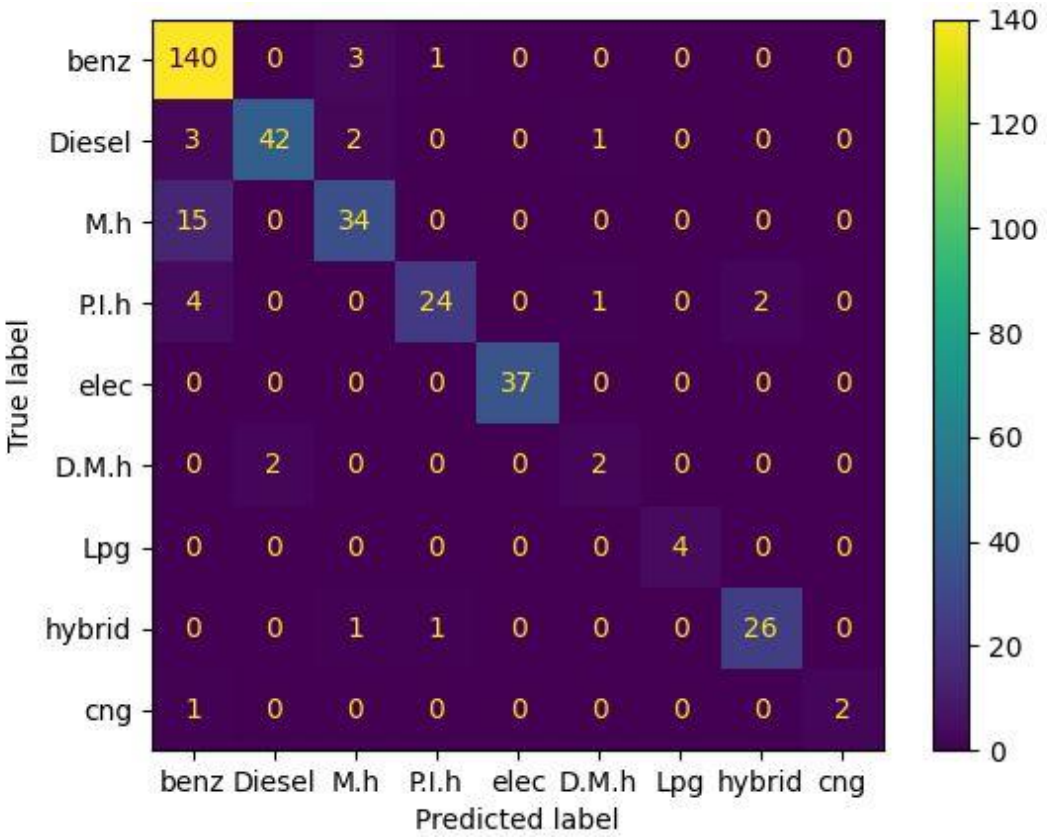
Recall

benz	diesel	M.h	P.I.h	elec	D.M.h	Lpg	hybrid	cng
0.92	0.78	0.74	0.94	1	0.8	1	0.90	0.62

F-score

benz	diesel	M.h	P.I.h	elec	D.M.h	Lpg	hybrid	cng
0.89	0.85	0.74	0.94	1	0.72	1	0.89	0.76

Στη συνέχεια παρουσιάζονται τα αποτελέσματα του DT με βάση τα χαρακτηριστικά που έχουν σχέση με την απόδοση των αυτοκινήτων(κυβικά,ισχύς,ροπή,χλμ/ωρα).



Πίνακας 7 Confusion Matrix DT με 4 χαρακτηριστικά

Παρακάτω φαίνονται σε πίνακες οι μετρικές που βρέθηκαν για τον κάθε τύπο καυσίμου.

Accuracy	0.89

Precision

benz	diesel	M.h	P.I.h	elec	D.M.h	Lpg	hybrid	cng
0.85	0.95	0.85	0.92	1	0.5	1	0.92	1

Recall

benz	diesel	M.h	P.I.h	elec	D.M.h	Lpg	hybrid	cng

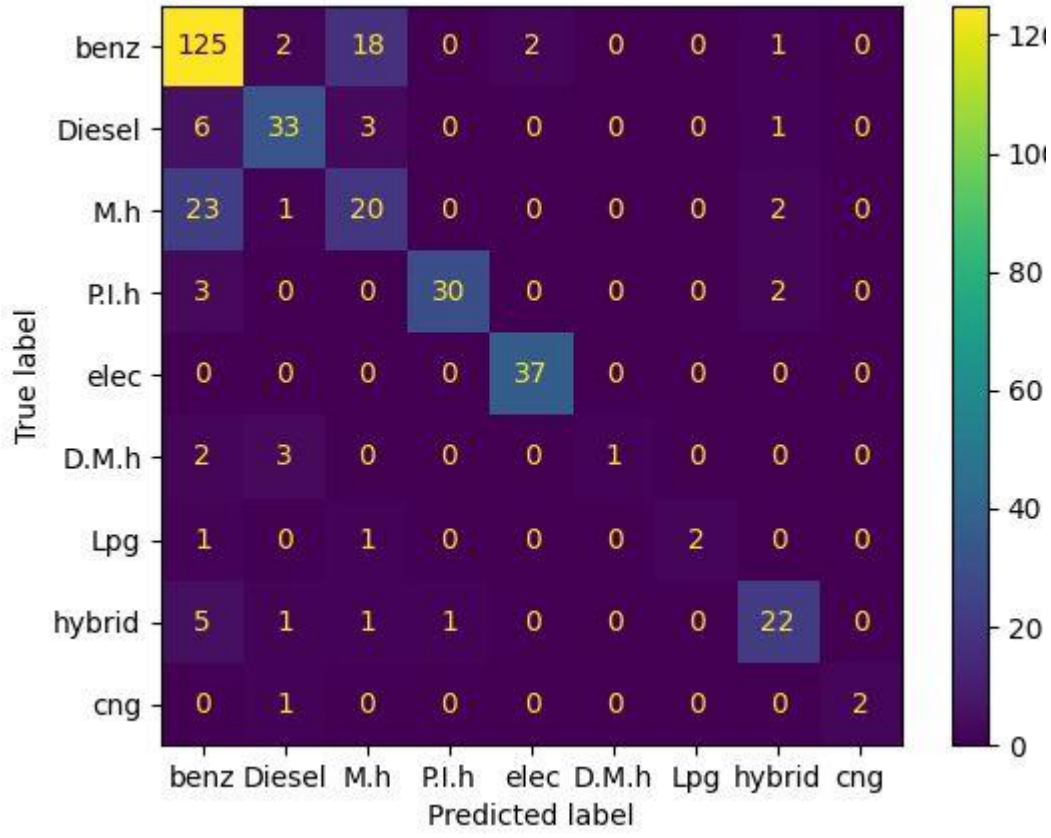
0.97	0.875	0.69	0.77	1	0.5	1	0.92	0.66

F-score

benz	diesel	M.h	P.I.h	elec	D.M.h	Lpg	hybrid	cng
0.91	0.91	0.76	0.84	1	0.5	1	0.92	0.8

5.4.3 Αποτελέσματα KNN

Με τον ίδιο τρόπο φτιάχνουμε το confusion matrix με τον KNN όπως και με το Decision Tree.



Πίνακας 8 Confusion Matrix KNN

Παρακάτω φαίνονται σε πίνακες οι μετρικές που βρέθηκαν για τον κάθε τύπο καυσίμου.

Accuracy	0.77

Precision

benz	diesel	M.h	P.I.h	elec	D.M.h	Lpg	hybrid	cng
0.75	0.80	0.46	0.96	0.94	1	1	0.78	1

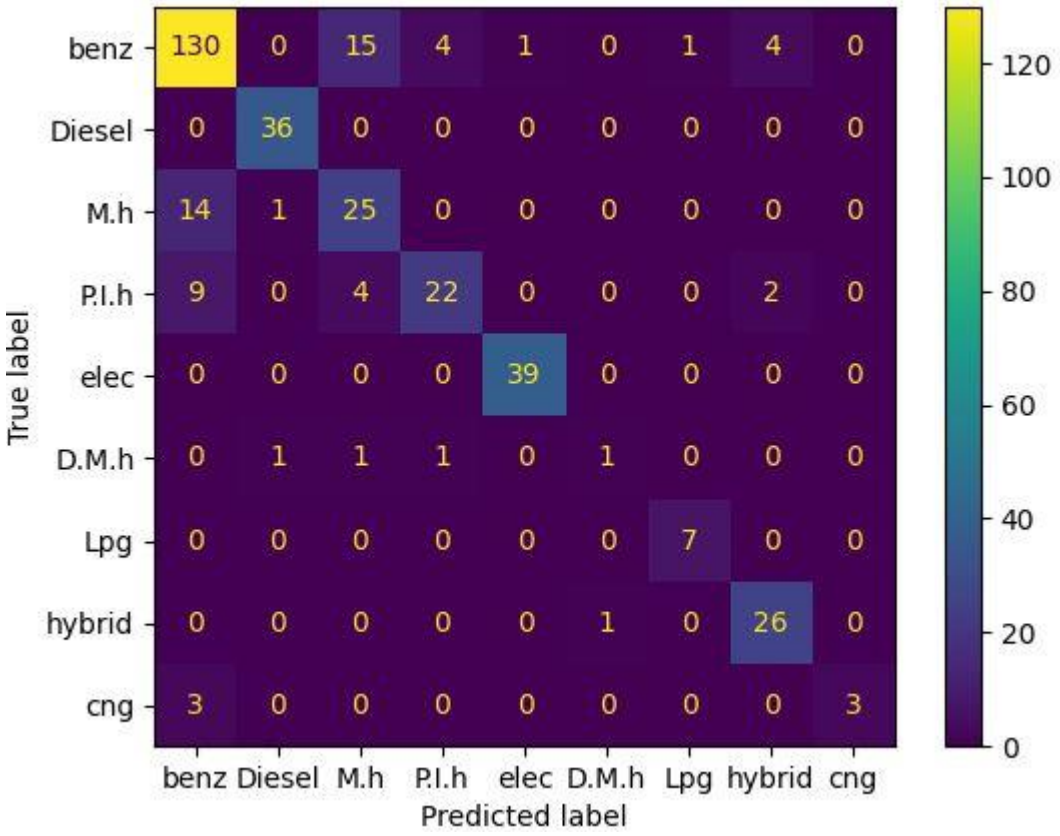
Recall

benz	diesel	M.h	P.I.h	elec	D.M.h	Lpg	hybrid	cng
0.84	0.76	0.43	0.85	1	0.16	0.5	0.73	0.66

F-score

benz	diesel	M.h	P.I.h	elec	D.M.h	Lpg	hybrid	cng
0.79	0.78	0.44	0.90	0.97	0.28	0.66	0.75	0.8

Τέλος εφαρμόζουμε με τον KNN για τα τα 4 χαρακτηριστικά απόδοσης των αυτοκινήτων.



Πίνακας 9 Confusion Matrix KNN με 4 χαρακτηριστικά

Παρακάτω φαίνονται σε πίνακες οι μετρικές που βρέθηκαν για τον κάθε τύπο καυσίμου.

Accuracy	0.82

Precision

benz	diesel	M.h	P.I.h	elec	D.M.h	Lpg	hybrid	cng
0.83	0.94	0.55	0.81	0.975	0.5	0.87	0.81	1

Recall

benz	diesel	M.h	P.I.h	elec	D.M.h	Lpg	hybrid	cng
0.83	1	0.62	0.59	1	0.25	1	0.96	0.5

F-score

benz	diesel	M.h	P.I.h	elec	D.M.h	Lpg	hybrid	cng
0.75	0.97	0.58	0.68	0.98	0.33	0.93	0.88	0.66

6 *Επίλογος*

6.1 Σύνοψη και συμπεράσματα

Στην παρούσα διπλωματική εργασία διερευνήθηκαν εκ πρώτης οι συνθήκες αγοράς αυτοκινήτου στην ελληνική αγορά. Η έρευνα ξεκίνησε με τη συλλογή δεδομένων που σχετίζονταν με την αγορά καινούργιων αυτοκινήτων. Όμως, λόγω της αυξανόμενης ζήτησης και των προβλημάτων που υφίστανται, όπως η έλλειψη των microchips από τους κατασκευαστές, διόγκωσαν το ζήτημα με πολύ μεγάλες καθυστερήσεις παράδοσης των αυτοκινήτων. Γι' αυτόν το λόγο αποφασίστηκε να ανακτηθούν δεδομένα-χαρακτηριστικά καινούργιων αυτοκινήτων απ' την ιστοσελίδα autotriti και να δημιουργηθεί ένα σύνολο δεδομένων με τα πιο βασικά χαρακτηριστικά. Αυτό βέβαια ήταν αρκετά χρονοβόρο διότι το σύνολο δεδομένων που επιλέξαμε από την ιστοσελίδα χρειαζόταν αρκετή προ-επεξεργασία. Η ενέργεια αυτή θα βοηθήσει αρκετά τους υποψήφιους αγοραστές έχοντας πλέον την γνώση να πάρουν την καλύτερη απόφαση, διότι ξέρουν ούτως η άλλως ότι υπάρχει καθυστέρηση στη παράδοση. Η κατάσταση επιδεινώνεται περαιτέρω αν δεν έχουν κάνει την ενδεδειγμένη επιλογή, με βάση τις ανάγκες τους. Για την ανάλυση και τη στατιστική επεξεργασία των δεδομένων που συλλέχθηκαν, αναπτύχθηκαν αρχικά 3 μοντέλα συσταδοποίησης. Όλα τα μοντέλα δημιούργησαν συστάδες με παρόμοια χαρακτηριστικά των αυτοκινήτων. Για παράδειγμα, αν κάποιος θέλει να αγοράσει ένα mercedes πλέον μπορεί, σύμφωνα με την συσταδοποίηση, να επιλέξει ένα αυτοκίνητο με παρόμοια χαρακτηριστικά αλλά λιγότερο κοστοβόρο στη αγορά. Φαίνεται εκ του αποτελέσματος πως ο kmeans για το σύνολο δεδομένων που συλλέξαμε είχε τα καλύτερα αποτελέσματα. Στη συνέχεια, με τη βοήθεια των αλγορίθμων κατηγοριοποίησης, παρέχεται δυνατότητα αυτοματοποιημένης εύρεσης τύπου καυσίμου στο αυτοκίνητο, κάτι που σίγουρα συμβάλλει αποφασιστικά ως κριτήριο στην τελική απόφαση του ενδιαφερομένου, ανάλογα με τις ανάγκες που έχει. Απ' ό,τι φαίνεται ο

decision tree έχει ελαφρώς καλύτερα αποτελέσματα από' τον KNN, όπως επίσης και συγκριτικά με το πείραμα που αναπτύχθηκε με τα χαρακτηριστικά απόδοσης του αυτοκινήτου, υπήρχαν διαφορές αλλά και πάλι ο DT ήταν καλύτερος από τον KNN. Το οξύμωρο της υπόθεσης ήταν ότι στα πειράματα με τα χαρακτηριστικα που αφορούσαν την απόδοση παρατηρήθηκε καλύτερο αποτέλεσμα συγκριτικά με τα πειράματα με όλα τα αριθμητικα χαρακτηριστικά του συνόλου δεδομένων.

6.2 Μελλοντικές επεκτάσεις

Το σύνολο δεδομένων που έχει αναπτυχθεί είναι καινοτόμο και μπορεί να λύσει αρκετά προβλήματα, είτε να δημιουργήσει μια διαθεματική ιστοσελίδα που θα συλλέγει τα δεδομένα της βάσης ανάλογα με τις προτιμήσεις του χρήστη και θα του δείχνει προτεινόμενα αυτοκίνητα.

Επίσης, λόγω έλλειψης χρόνου δεν προλάβαμε να αναπτύξουμε ένα πείραμα ακόμα κατηγοριοποίησης, βρίσκοντας την μάρκα των αυτοκινήτων και πιθανώς σε ποια χώρα κατασκευάζεται.

Συμπληρωματικά, θα μπορούσε να χρησιμοποιηθεί εξίσου η μέθοδος της παλινδρόμησης για να βρίσκονται τυχαία τα τέλη κυκλοφορίας του αυτοκινήτου και να φαίνεται αν είναι χαμηλά η υψηλά. Επιπροσθέτως, θα μπορούσε να αναζητείται η κατανάλωση των αυτοκινήτων.

Πέρα από τα παραπάνω, θα ήταν εύκολο να εφαρμοστούν και οι κανόνες σύνδεσης (association rules) που λειτουργούν στη εξόρυξη δεδομένων σαν "if-then", δηλαδή αν η μια κατάσταση Α ισχύει τότε συνεπάγεται και η κατάσταση Β. Για παράδειγμα, λειτουργεί ως μια δήλωση ενός ενδιαφερομένου που ψάχνει αν υπάρχουν κυβικά, τιμή και ροπή σε ένα αυτοκίνητο, θα αναζητήσει τότε την ισχύ και την κατανάλωση.

7 Βιβλιογραφία

1. Nanaki, E. (2018). Measuring the impact of economic crisis to the Greek vehicle market. *Sustainability*, *10*(2), 510. https://doi.org/10.3390/su10020510
2. *Καθυστερήσεις πολλών μηνών στις παραδόσεις καινούργιων αυτοκινήτων*. (2021, December 15). LiFO. https://www.lifo.gr/now/world/kathysteriseis-pollon-minon-stis-paradoseis-kainoyrion-aytokiniton
3. Σαρημπαλίδης, Β. (2022, October 17). *Η ελληνική αγορά αυτοκινήτου καλά κρατεί…*. NewsAuto.Gr. https://www.newsauto.gr/news/i-elliniki-agora-aftokinitou-kala-krati/
4. *Οι ταξινομήσεις το 2021 και η πορεία από το 1990 στην Ελλάδα*. (2022, January 17). FleetNews. https://fleetnews.gr/oi-taxinomiseis-to-2021-kai-i-poreia-apo-to-1990-stin-ellada/
5. *Οι Έλληνες προτιμούν την Toyota (14,1% μερίδιο αγοράς), αλλά αγόρασαν και 589 αυτοκίνητα Tesla*. (n.d.). Capital.Gr. Retrieved February 7, 2023, from https://www.capital.gr/oikonomia/3693233/oi-ellines-protimoun-tin-toyota-14-1-meridio-agoras-alla-agorasan-kai-589-autokinita-tesla
6. *Η ελληνική αγορά αυτοκινήτου στα χρόνια της πανδημίας*. (n.d.). Capital.Gr. Retrieved February 7, 2023, from https://www.capital.gr/me-apopsi/3626312/i-elliniki-agora-autokinitou-sta-xronia-tis-pandimias
7. Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. da F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PloS One*, *14*(1), e0210236. https://doi.org/10.1371/journal.pone.0210236

8. Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., & Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664–681. https://doi.org/10.1016/j.neucom.2017.06.053
9. Καρανίκας Ι. (2016). Αξιολόγηση αλγορίθμων συσταδοποίησης ακολουθιακών χωροχρονικών δεδομένων με χρήση διαφορετικών συναρτήσεων απόστασης. ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ.https://dione.lib.unipi.gr/xmlui/handle/unipi/10134
10. Zaït, M., & Messatfa, H. (1997). A comparative study of clustering methods. *Future Generations Computer Systems: FGCS*, *13*(2–3), 149–159. https://doi.org/10.1016/s0167-739x(97)00018-6
11. Madhulatha, T. S. (2012). An overview on clustering methods. *IOSR Journal of Engineering*, *02*(04), 719–725. https://doi.org/10.9790/3021-0204719725
12. Sharma, P. (2021, April 26). *K means clustering simplified in python*. Analytics Vidhya.https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python/
13. Γράτσος Κ. (2023). Εφαρμογή ιστού για την εκτέλεση συσταδοποίησης κ-μέσων και τον βέλτιστο προσδιορισμό της παραμέτρου κ με τη μέθοδο του αγκώνα. ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΑΣ
14. Alade, T. (2018, May 27). *Tutorial: How to determine the optimal number of clusters for k-means clustering*. Cambridge Spark. https://blog.cambridgespark.com/how-to-determine-the-optimal-number-of-clusters-for-k-means-clustering-14f27070048f
15. Chakraborty, S., & Nagwani, N. K. (2014). Analysis and study of incremental DBSCAN clustering algorithm. In *arXiv [cs.DB]*. http://arxiv.org/abs/1406.4754
16. Rehman, S. U., Asghar, S., Fong, S., & Sarasvady, S. (2014). DBSCAN: Past, present and future. *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*.
17. Mullin, T. (2020, July 10). *DBSCAN parameter estimation using python*. Medium. https://medium.com/@tarammullin/dbscan-parameter-estimation-ff8330e3a3bd
18. Cerquitelli, T., Di Corso, E., Proto, S., Bethaz, P., Mazzarelli, D., Capozzoli, A., Baralis, E., Mellia, M., Casagrande, S., & Tamburini, M. (2020). A data-driven energy platform: From energy performance certificates to human-readable knowledge through dynamic high-resolution geospatial maps. *Electronics*, *9*(12), 2132. https://doi.org/10.3390/electronics9122132
19. Chauhan, N. S. (n.d.). *DBSCAN clustering algorithm in machine learning*. KDnuggets. Retrieved February 8, 2023, from https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html

20. Xu, R., & Wunsch, D., 2nd. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645–678. https://doi.org/10.1109/TNN.2005.845141
21. Sharma, H. (2021, April 23). *Hierarchical Clustering - Himanshu Sharma*. Medium. https://harshsharma1091996.medium.com/hierarchical-clustering-996745fe656b
22. Kesavaraj, G., & Sukumaran, S. (2013). A study on classification techniques in data mining. *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*.
23. Ουγιάρογλου, Σ. (2021). *Algorithms and techniques for efficient and effective nearest neighbours classification*. Πανεπιστήμιο Μακεδονίας. Σχολή Επιστημών Πληροφορίας. Τμήμα Εφαρμοσμένης Πληροφορικής.
24. Umadevi, S., & Marseline, K. S. J. (2017). A survey on data mining classification algorithms. *2017 International Conference on Signal Processing and Communication (ICSPC)*.
25. *Classification-based approaches in data mining*. (2021, August 1). GeeksforGeeks. https://www.geeksforgeeks.org/classification-based-approaches-in-data-mining/
26. Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning *k* for kNN classification. *ACM Transactions on Intelligent Systems and Technology*, *8*(3), 1–19. https://doi.org/10.1145/2990508
27. A study of some data mining classification techniques. (2017). *International Journal of Modern Trends in Engineering & Research*, *4*(1), 210–215. https://doi.org/10.21884/ijmter.2017.4031.zt9tv
28. Soofi, A. A., & Awan, A. (2017). Classification Techniques in Machine Learning: Applications and Issues. *Journal of Basic & Applied Sciences*, *13*, 459–465. https://doi.org/10.6000/1927-5129.2017.13.76
29. Archana, S., & Elangovan, K. (n.d.). *Survey of classification techniques in data mining*. Cloudfront.net. Retrieved March 1, 2023, from https://d1wqtxts1xzle7.cloudfront.net/73504805/V2I214-libre.pdf?1635052212=&response-content-disposition=inline%3B+filename%3DSurvey_of_Classification_Techniques_in_D.pdf&Expires=1677667872&Signature=dUaejiuGc6qWnsHFUTwTwhGLC15-zc05lStWP9RfDCTRKBjkOPgZjk86ydnc6YUhtZ4Eg9ZpKJlnaKJFGX0tfa-ZcW2xIN3bR-dAOr5llz9oJmIX2FXMItX6ZcS4jlUGwSHNMUUAyTTEV7rr0pzQWFh2qMb-HSs4DcNsUQUEL3aMgcm5qsIO~I9ilYOU1OlylLlZQBA~vU0qWqJjNLI10DQitLlqtV4GtLPOaHJhNEc9J81pGR6bQxSIQimsaT70zYkMk5nacm2M2mq9-x57NeASuinYd-2DqatU89c4ubnJK-

zQdOyV9pngFh4BI9iWfRl6pgo2Bbvmj022dM1SWw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

30. Lan, K., Wang, D.-T., Fong, S., Liu, L.-S., Wong, K. K. L., & Dey, N. (2018). A survey of data mining and deep learning in bioinformatics. *Journal of Medical Systems*, *42*(8), 139. https://doi.org/10.1007/s10916-018-1003-9
31. Soni, A. (2020, July 3). *Advantages And Disadvantages of KNN*. Medium. https://medium.com/@anuuz.soni/advantages-and-disadvantages-of-knn-ee06599b9336
32. Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165
33. F.y, O., Babcock University, J.e.t, A., O, A., J. O, H., O, O., & J, A. (2017). Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology*, *48*(3), 128–138. https://doi.org/10.14445/22312803/ijctt-v48p126
34. *Decision tree*. (2017, October 16). GeeksforGeeks. https://www.geeksforgeeks.org/decision-tree/
35. Shafiq, M., Tian, Z., Bashir, A. K., Jolfaei, A., & Yu, X. (2020). Data mining and machine learning methods for sustainable smart cities traffic classification: A survey. *Sustainable Cities and Society*, *60*(102177), 102177. https://doi.org/10.1016/j.scs.2020.102177
36. Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3
37. Venkata, S., Kumar, K., & Kiruthika, P. (n.d.). *An Overview of Classification Algorithm in Data mining*. https://doi.org/10.17148/IJARCCE.2015.41259
38. Sanner, M. F. (1999). Python: a programming language for software integration and development. *Journal of Molecular Graphics & Modelling*, *17*(1), 57–61.
39. Python current trend applications- an overview. (2019). *Ijaerd*. https://www.academia.edu/41143260/PYTHON_CURRENT_TREND_APPLICATIONS_AN_OVERVIEW
40. Bogdanchikov, A., Zhaparov, M., & Suliyev, R. (2013). Python to learn programming. *Journal of Physics. Conference Series*, *423*, 012027. https://doi.org/10.1088/1742-6596/423/1/012027
41. Srinath, K. R. (2017). *Python – the fastest growing programming language*. https://www.semanticscholar.org/paper/72dbcf413ded66553a27eefe6f6a1acef494bdbd

42. Vargiu, E., & Urru, M. (2012). Exploiting web scraping in a collaborative filtering-based approach to web advertising. *Artificial Intelligence Research*, *2*(1). https://doi.org/10.5430/air.v2n1p44
43. Chang, Z. (2022). A survey of modern crawler methods. *The 6th International Conference on Control Engineering and Artificial Intelligence*.
44. Khder, M. (2021). Web scraping or Web Crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing and Its Applications*, *13*(3), 145–168. https://doi.org/10.15849/ijasca.211128.11
45. Uzun, E., Yerlikaya, T., & Kirat, O. (n.d.). *Comparison of python libraries used for web data extraction*. Erdincuzun.com. Retrieved February 7, 2023, from https://erdincuzun.com/wp-content/uploads/download/plovdiv_2018_01.pdf
46. Monica. (2020, July 12). *Build your own dataset with beautiful soup*. The Startup. https://medium.com/swlh/build-your-own-dataset-with-beautiful-soup-583717e3dad7
47. Jovic, A., Brkic, K., & Bogunovic, N. (2014). An overview of free software tools for general data mining. *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*.
48. Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *Information (Basel)*, *11*(4), 193. https://doi.org/10.3390/info11040193
49. Hao, J., & Ho, T. K. (2019). Machine learning made easy: A review of Scikit-learn package in Python programming language. *Journal of Educational and Behavioral Statistics: A Quarterly Publication Sponsored by the American Educational Research Association and the American Statistical Association*, *44*(3), 348–361. https://doi.org/10.3102/1076998619832248
50. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. In *arXiv [cs.LG]*. http://arxiv.org/abs/1201.0490
51. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*, 14. https://doi.org/10.3389/fninf.2014.00014
52. *Choosing the right estimator*. (n.d.). Scikit-Learn. Retrieved February 8, 2023, from https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html
53. (N.d.). Mysql.com. Retrieved February 7, 2023, from https://dev.mysql.com/doc/refman/8.0/en/what-is-mysql.html

54. Ashiq, K. S. (2019, January 12). *Working with MySQL - Ashiq KS*. Medium. https://medium.com/@ashiqgiga07/working-with-mysql-dae8f149aa57
55. Saini, C., & Arora, V. (2016). Information retrieval in web crawling: A survey. *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*.
56. Bassil, Y. (2012). A survey on Information Retrieval, text categorization, and web crawling. In *arXiv [cs.IR]*. http://arxiv.org/abs/1212.2065
57. Raghavan, P. (n.d.). *Information retrieval algorithms: A survey*. Psu.edu. Retrieved February 7, 2023, from https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=410b5ae67caaa0c9049aa5747083cbc38adfe045
58. *Natural language processing - quick guide*. (2021, July 28). TutorialsPoint. https://www.tutorialspoint.com/natural_language_processing/natural_language_processing_quick_guide.htm
59. Fridah Nyamisa, M. (2017). A survey of information retrieval techniques. *Advances in Networks*, *5*(2), 40. https://doi.org/10.11648/j.net.20170502.12
60. Irfan, S., & Babu, B. V. (2016). Information retrieval in big data using evolutionary computation: A survey. *2016 International Conference on Computing, Communication and Automation (ICCCA)*.
61. Melnikov, V. O., Melikyan, G. S., & Maksimov, O. A. (2009). Characteristics of information retrieval systems on the internet: Theoretical and practical aspects. *Automatic Documentation and Mathematical Linguistics*, *43*(1), 42–50. https://doi.org/10.3103/s0005105509010063
62. *Information retrieval system: Introduction and information retrieval*. (n.d.). Flexiprep.com. Retrieved February 9, 2023, from https://www.flexiprep.com/NIOS-Notes/Senior-Secondary/Library-Science/NIOS-Library-Science-Unit-16-Information-Retrieval-System-Part-1.html
63. *Information retrieval system: Introduction and information retrieval*. (n.d.). Flexiprep.com. Retrieved February 15, 2023, from https://www.flexiprep.com/NIOS-Notes/Senior-Secondary/Library-Science/NIOS-Library-Science-Unit-16-Information-Retrieval-System-Part-1.html
64. *Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης*. (n.d.). Slideplayer.Gr. Retrieved February 8, 2023, from https://slideplayer.gr/slide/1881801/
65. Suthar, B., Patel, H., Goswami, A., & Scholar, M. (2012). *A survey: Classification of imputation methods in data mining*. https://www.semanticscholar.org/paper/1eba6555d2a3c6ec232ac5ce216e9b47d130bdd4

66. Caparino, E. T., Sison, A. M., & Medina, R. P. (2019). Application of the modified imputation method to missing data to increase classification performance. *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*.
67. Susanti, S. P., & Azizah, F. N. (2017). Imputation of missing value using dynamic Bayesian network for multivariate time series data. *2017 International Conference on Data and Software Engineering (ICoDSE)*.
68. Aljuaid, T., & Sasi, S. (2016). Proper imputation techniques for missing values in data sets. *2016 International Conference on Data Science and Engineering (ICDSE)*.
69. Al Shalabi, L., & Shaaban, Z. (2006). Normalization as a preprocessing engine for data mining and the approach of preference matrix. *2006 International Conference on Dependability of Computer Systems*.
70. *Solutions to chapter 1 AN INTRODUCTION TO DATA MINING Prepared by James Cunningham, Graduate Assistant*. (n.d.). Docplayer.net. Retrieved February 7, 2023, from https://docplayer.net/187271731-Solutions-to-chapter-1-an-introduction-to-data-mining-prepared-by-james-cunningham-graduate-assistant.html
71. Saranya#, C., & Manikandan, G. (n.d.). *A study on normalization techniques for privacy preserving data mining*. Enggjournals.com. Retrieved February 7, 2023, from https://www.enggjournals.com/ijet/docs/IJET13-05-03-273.pdf
72. *Precision and Recall, F-score and Accuracy – measuring NLP performance*. (n.d.). Expert.Ai. Retrieved February 15, 2023, from https://community.expert.ai/articles-showcase-56/precision-and-recall-f-score-and-accuracy-measuring-nlp-performance-191

Παράρτημα

Παράρτημα I Clustering Algorithms

Αρχικά θα αναφερθούμε στον Αλγόριθμο K-means. Τοποθετούνται οι κατάλληλες βιβλιοθήκες της python που θα χρησιμοποιηθούν για την υλοποίηση του αλγορίθμου.

```
from sklearn.cluster import KMeans
from matplotlib import pyplot as plt
import numpy as np
import pandas as pd
from sklearn.impute import KNNImputer
import seaborn as sns
from pandas.plotting import parallel_coordinates
from sklearn.preprocessing import MinMaxScaler
from scipy.constants.constants import point
```

Εικόνα 23 Εισαγωγή βιβλιοθηκών για συσταδοποίηση

Στη συνέχεια αφού εισαχθούν οι βιβλιοθήκες πραγματοποιούμε τις κατάλληλες τροποποιήσεις. Όπως φαίνεται και στη εικόνα χρησιμοποιείται ο Knn imputer και normalization με την μέθοδο Min-Max. Για να πραγματοποιηθεί το imputation χρειάζονται μόνο αριθμητικά δεδομένα και για αυτό παίρνουμε δεδομένα από την 4 στήλη και μετά.

```
sns.set()
data= pd.read_csv("cars.csv")

X= data.iloc[:,4:13]
#=X.replace('yes',1)
X = X.replace(r'^\s*$', np.nan, regex=True)
NaN= np.nan

print(X.head())
imputer= KNNImputer(n_neighbors=5, weights="uniform")
print(imputer.fit_transform(X))
X=imputer.fit_transform(X)

scaler = MinMaxScaler()
X=scaler.fit_transform(X)
```

Εικόνα 24 Imputation-Normalization

Για να βρούμε το βέλτιστο K χρησιμοποιούμε την μέθοδο Αγκώνα(elbow curve). Από το σχήμα που έχει παρουσιαστεί πιο πάνω φαίνεται ότι το κατάλληλο K είναι το 4 δηλαδή ο αριθμός των συστάδων. Μετά εφαρμόζουμε τον Kmeans στα δεδομένα.

```
distorsions = []
for k in range(2, 20):
    kmeans = KMeans(n_clusters=k,random_state = 11)
    labels=kmeans.fit(X)
    distorsions.append(kmeans.inertia_)

fig = plt.figure(figsize=(15, 5))
plt.plot(range(2, 20), distorsions)
plt.grid(True)
plt.title('Elbow curve')
plt.show()
noclusters=4


kmeans = KMeans(n_clusters=noclusters,random_state = 11)
labels=kmeans.fit(X)
kmeans2 = KMeans(n_clusters=noclusters,random_state = 11).fit(X)
```

Εικόνα 25 Αλγόριθμος Kmeans

Για να βρούμε το τετραγωνικό σφάλμα των συστάδων χρειάζεται να βρούμε το κέντρο των συστάδων και μετά από το τύπο της θεωρίας υπολογίζεται το τετραγωνικό σφάλμα.

```
cluster_centers = [X[kmeans2.labels_ == i].mean(axis=0) for i in range(noclusters)]
print(cluster_centers)
clusterwise_sse = [0, 0,0]
clusterwise_sse = [0] *noclusters


for point, label in zip(X, kmeans2.labels_):
    clusterwise_sse[label] += np.square(point - cluster_centers[label]).sum()

print(clusterwise_sse)
```

Εικόνα 26 Τετραγωνικό Σφάλμα

Τέλος δημιουργούμε με την βιβλιοθήκη parallel coordinates τα γραφήματα που παρουσιάζονται παραπάνω και το κάθε στοιχείο σε ποια συστάδα ανήκει.

```
df2 = pd.DataFrame(X ,columns = ["cc","isxis","timi","ropi","kmh","katanalosi","kausaeria","autonomia","teli"])

df2['Clusters']=labels.labels_
print(df2)
parallel_coordinates(df2, 'Clusters',color=('#0a3161','#dcac36','#b4eeb4','#f62e36'))
plt.show()
print(labels.labels_)
print(df2['Clusters'].value_counts())

for i in range(0,data.__len__()):
    print(str(data.loc[i][1])+" "+str(data.loc[i][2])+" is in cluster "+str(labels.labels_[i]))
```

Εικόνα 27 Αναπαράσταση Συστάδων με Kmeans

DBSCAN

Ο δεύτερος αλγόριθμος που υλοποιήθηκε είναι ο DBSCAN. Αρχικά εισάγουμε τις κατάλληλες βιβλιοθήκες όπως και στον kmeans με την προσθήκη του DBSCAN αντί για τον K-means. Επίσης την προσθήκη της βιβλιοθήκης Counter λόγω των Outliers.

Στη συνέχεια κάνουμε τις ίδιες τροποποιήσεις με τον K-means μαζί με το imputation και το normalization και τοποθετούμε ένα dataframe με τις στήλες που μας ενδιαφέρουν. Εισάγουμε την βιβλιοθήκη κοντινότερων γειτόνων και βρίσκουμε το Eps.

```
data2 = data.iloc[:, 4:]
data2 = pd.DataFrame(X ,columns = ["cc","isxis","timi","ropi","kmh","katanalosi","kausaeria","autonomia","teli"])

from sklearn.neighbors import NearestNeighbors

neighb = NearestNeighbors(n_neighbors=5)
nbrs=neighb.fit(data2)
distances,indices=nbrs.kneighbors(data2)
plt.yticks(np.arange(0, 1, 0.05))
distances = np.sort(distances, axis = 0)
distances = distances[:, 1]

plt.plot(distances)
plt.show()
```

Εικόνα 28 Εγγύτεροι γείτονες

Αφού έχει βρεθεί το Eps ορίζουμε σαν γείτονες 5 σημεία τουλάχιστον και το τοποθετούμε στα δεδομένα που δημιουργήσαμε εκ νέου και βάζουμε τις ετικέτες σε λίστα. Με την βοήθεια ξανά του parallel coordinates βρίσκουμε τα στοιχεία των συστάδων και τα ακραία σημεία. Τέλος τα εμφανίζουμε τα ακραία σημεία για να καταλήξουμε σε επιπλέον συμπεράσματα.

```
labels = DBSCAN(eps = 0.25, min_samples = 5).fit(data2).labels_
clusters = len(Counter(labels))
print(f"Number of clusters: {clusters}")
print(f"Number of outliers: {Counter(labels)[-1]}")


dbscan = DBSCAN(eps = 0.25, min_samples = 5)
dbscan.fit(data2)
data2['Clusters']= dbscan.labels_.tolist()
print(data2)
parallel_coordinates(data2, 'Clusters',color=('#9f1d35','#0a3161','#dcac36','#b4eeb4','#156238'))

plt.yticks(np.arange(0, 1, 0.05))
plt.show()
print("OUTLIERS")
data["anomaly_score"] = labels
valid=data[data.anomaly_score!=-1]
anomalies = data[data.anomaly_score == -1]
print(anomalies)
print(valid)

print(data['anomaly_score'].value_counts())
```

Εικόνα 29 Αναπαράσταση Συστάδων DBSCAN

Τελευταίος αλγόριθμος συσταδοποίησης που ασχολήθηκε η παρούσα διπλωματική είναι η ιεραρχική συσταδοποίηση.

Εδώ χρησιμοποιούμε μεθόδους απόστασης Linkage και από τα πειράματα που φαίνονται και παραπάνω αυτή που τελικά θα χρησιμοποιήθει είναι η μέθοδος Ward.

```
data2 = data.iloc[:, 4:-1]
data2 = pd.DataFrame(X ,columns = ["cc","isxis","timi","ropi","kmh","katanalosi","kausaeria","autonomia","teli"])

linkarray=['ward','complete','single','average']
linkarray=['ward']
for link in linkarray:
    linkage_data = linkage(data2, method=link, metric='euclidean')
    dendrogram(linkage_data,color_threshold=3.5)
    plt.show()
```

Εικόνα 30 Μέθοδοι Linkage

Εφαρμόζουμε τον αλγόριθμο Agglomerative και ορίζουμε τον αριθμό των συστάδων με το κατάλληλο κλάδεμα. Όπως και στον DBSCAN δημιουργούμε το γράφημα με τις συστάδες και τέλος βγάζουμε τα ακριβή αποτελέσματα για το που ανήκει το κάθε αυτοκίνητο.

```
for link in linkarray:
    hierarchical_cluster = AgglomerativeClustering(n_clusters=4,linkage=link)
    labels = hierarchical_cluster.fit_predict(data2)
    model = AgglomerativeClustering(distance_threshold=0, n_clusters=None)
    model=model.fit(data2)

    data2['Clusters']= hierarchical_cluster.labels_.tolist()
    print(data2)
    data['Clusters']=hierarchical_cluster.labels_.tolist()

    parallel_coordinates(data2, 'Clusters',color=('#9f1d35','#0a3161','#b4eeb4','#156238'))
    plt.yticks(np.arange(0, 1, 0.05))

    plt.show()
    print(data)
    print(data['Clusters'].value_counts())
    data["anomaly_score"] = labels
    valid=data[data.anomaly_score!=-1]
    print(valid)
    print(data['anomaly_score'].value_counts())
```

Εικόνα 31 Αναπαράσταση Συστάδων Agglomerative

Παράρτημα Β Classification Algorithms

Αφού έχουν ολοκληρωθεί οι αλγόριθμοι συσταδοποίησης θα παρουσιαστούν και οι αλγόριθμοι κατηγοριοποίησης.

Αρχικά θα αναφερθεί ο αλγόριθμος απόφασης δέντρου(DT) και όπως φαίνεται και στη εικόνα εισάγονται οι κατάλληλες βιβλιοθήκες και χωρίζονται τα δεδομένα με την μέθοδο split τυχαία αφού έχει γίνει το normalization.

```
import pandas as pd
from sklearn.tree import DecisionTreeClassifier
from sklearn import metrics
import matplotlib.pyplot as plt
import numpy as np
from sklearn.preprocessing import MinMaxScaler

train_filename = 'cars.csv'

data2 = pd.read_csv(train_filename,index_col=False)
data=data2.copy()
data2['split'] = np.random.randn(data2.shape[0], 1)

data=data.iloc[:,4:]
scaler = MinMaxScaler()
data=scaler.fit_transform(data)
msk = np.random.rand(len(data)) <= 0.7
```

Εικόνα 32 Εισαγωγή βιβλιοθηκών για κατηγοριοποιήση και normalization

Έπειτα τοποθετούμε τα κατηγορικά δεδομένα μαζί με τα υπόλοιπα και ορίζουμε ως σετ εκπαίδευσης τα πρώτα 70% στοιχεία και τα υπόλοιπα ως σετ πρόβλεψης. Σαν κλάση παίρνουμε το καύσιμο γιατί αυτό αναζητούμε και τα για δεδομένα τις αριθμητικές στήλες.

```
data=pd.DataFrame(data,columns=['cc','isxis','timi','ropi','kmh','katanalosi','kausaeria','autonomia','teli'])

data['marka']=data2['marka']
data['modelo']=data2['modelo']
data['ekdosi']=data2['ekdosi']
data['kaysimo']=data2['kaysimo']
data['split']=data2['split']
print(data)

train = data[msk]
test = data[~msk]
class_train = train.iloc[:, -2]
attr_train = train.iloc[:, :9]

class_test = test.iloc[:, -2]
attr_test = test.iloc[:, :9]
class_names=["benz","Diesel","M.h","P.I.h","elec","D.M.h","Lpg","hybrid","cng"]
```

Εικόνα 33 Δημιουργία σετ εκπαίδευσης και πρόβλεψης

Εφαρμόζουμε τον DT στα δεδομένα εκπαίδευσης και προσπαθούμε να προβλέψουμε τα δεδομένα του σετ πρόβλεψης. Αφού τα προβλέψουμε εφαρμόζουμε τις σημαντικότερες μετρικές των αλγορίθμων κατηγοριοποίησης (accuracy,precision,recall.fscore) και δημιουργούμε το confusion matrix.

```
model = DecisionTreeClassifier()
model.fit(attr_train, class_train)

predictions = model.predict(attr_test)

print(metrics.accuracy_score(class_test, predictions))
precision, recall, fscore, supp = metrics.precision_recall_fscore_support(class_test, predictions,average=None)
print("PRECISION")
print(precision)
print("RECALL")
print(recall)
print("FSCORE")
print(fscore)

confusion_matrix = metrics.confusion_matrix(class_test, predictions)
cm_display = metrics.ConfusionMatrixDisplay(confusion_matrix = confusion_matrix,display_labels=class_names)

cm_display.plot()
plt.show()
```

Εικόνα 34 Αλγόριθμος Decision Tree

Με τον ίδιο τρόπο χρησιμοποιούμε τον αλγόριθμο DT για να υπολογίσουμε τα δεδομένα απλά με 4 στήλες που αφορούν την απόδοση των αυτοκινήτων.

```
del data['timi']
data['marka']=data2['marka']
data['modelo']=data2['modelo']
data['ekdosi']=data2['ekdosi']
data['kaysimo']=data2['kaysimo']
data['split']=data2['split']
print("DT")
print(data)


train = data[msk]
test = data[~msk]

class_train = train.iloc[:, -2]
attr_train = train.iloc[:, :4]

class_test = test.iloc[:, -2]
attr_test = test.iloc[:, :4]
class_names=["benz","Diesel","M.h","P.I.h","elec","D.M.h","Lpg","hybrid","cng"]
```

Εικόνα 35 Δημιουργία σετ εκπαίδευσης με τα χαρακτηριστικά της απόδοσης αυτοκινήτων

Έχοντας ολοκληρώσει τον αλγόριθμο DT θα υπολογίσουμε τον αλγόριθμο KNN με ακριβώς τον ίδιο τρόπο όπως και στον DT με την διαφορά ότι θα αλλάξουμε το μοντέλο από DT σε KNN με κοντινότερους γείτονες να είναι 5.

```
model = KNeighborsClassifier(n_neighbors=5)
model.fit(attr_train, class_train)

predictions = model.predict(attr_test)

print(metrics.accuracy_score(class_test, predictions))
precision, recall, fscore, supp = metrics.precision_recall_fscore_support(class_test, predictions,average=None)
print("PRECISION")
print(precision)
print("RECALL")
print(recall)
print("FSCORE")
print(fscore)

confusion_matrix = metrics.confusion_matrix(class_test, predictions)
cm_display = metrics.ConfusionMatrixDisplay(confusion_matrix = confusion_matrix,display_labels=class_names)

cm_display.plot()
plt.show()
```

Εικόνα 36 Αλγόριμθμος KNN