ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΔΙΠΛΩΜΑ ΕΙΔΙΚΕΥΣΗΣ (MSc)

στην ΑΝΑΠΤΥΞΗ & ΑΣΦΑΛΕΙΑ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

“ Τεχνικές Εξόρυξης Δεδομένων των πυρκαγιών για την περιοχή της Εύβοιας ”

ΓΑΤΟΣ ΔΗΜΗΤΡΙΟΣ

ΑΜ p3312103

ΑΘΗΝΑ, ΣΕΠΤΕΜΒΡΙΟΣ 2023

Περίληψη

Ένα πρόβλημα που απασχολεί όλο και περισσότερες χώρες ανά τον κόσμο είναι οι δασικές πυρκαγιές. Η σημαντικότητα του συγκεκριμένου ζητήματος δεν εξαρτάται μόνο από τις ανθρώπινες απώλειες, αλλά και τις επιπτώσεις στο περιβάλλον και στην οικονομία. Σήμερα υπάρχουν μέθοδοι που έχουν την δυνατότητα να προβλέψουν τις δασικές πυρκαγιές ώστε να σταματήσουν λίγο μετά την εκδήλωση τους ή και ακόμη και πριν συμβούν. Ένας τρόπος είναι με την χρήση της νέας τεχνολογίας και πιο συγκεκριμένα της Τηλεπισκόπησης και των Γεωγραφικών Συστημάτων Πληροφοριών που μέσω δορυφορικών εικόνων η άλλων τρόπων μπορούν να ανιχνεύσουν την φωτιά και να προβλέψουν την πορεία της. Η λύση αυτή όμως απαιτεί ειδική επεξεργασία των δεδομένων και σε πολλές περιπτώσεις το κόστος είναι μεγάλο. Ένας άλλος τρόπος είναι η ανάλυση δεδομένων προκειμένου να προβλεφθούν οι δασικές πυρκαγιές. Τα πλεονεκτήματα της συγκεκριμένης λύσης είναι το χαμηλό κόστος και η εύκολη επεξεργασία των δεδομένων. Στην παρούσα εργασία με την χρήση δύο μεθόδων, της γραμμικής παλινδρόμησης και του τυχαίου δάσους θα προβλέψουμε την έκταση της καμένης γης στην Εύβοια και θα συγκρίνουμε ποια μέθοδο έχει μεγαλύτερη ακρίβεια και καλύτερη απόδοση. Για αυτό το σκοπό συλλέξαμε και αναλύσαμε δεδομένα από την πυροσβεστική και μετεωρολογική υπηρεσία για την περίοδο 2020-2022 που αφορούν την Εύβοια και χωρίσαμε σε δύο βάσεις τα δεδομένα μας με κριτήριο την έκταση της καμένης γης και εφαρμόσαμε τις παραπάνω μεθόδους. Επίσης χωρίσαμε την βλάστηση της Ευβοίας σε 5 μεγάλες κατηγορίες και προσθέσαμε και την συγκεκριμένη μεταβλητή στο μοντέλο μας. Τέλος με σκοπό την υποστήριξη των τοπικών εμπλεκόμενων φορέων και βελτιστοποίηση των αποφάσεων τους σχετικά με την πρόληψη μελλοντικών πυρκαγιών αλλά και την καλύτερη κατανομή των πυροσβεστικών δυνάμεων, αρχικά δημιουργήσαμε ένα δέντρο απόφασης με τον αλγόριθμο j48 με αρκετά καλά αποτελέσματα και στην συνέχεια πραγματοποιήσαμε περιγραφική εξόρυξη (descriptive mining) στα δεδομένα με σκοπό τον εντοπισμό συσχετίσεων ή μοτίβων μεταξύ τους. Οι μέθοδοι που χρησιμοποιήσαμε ήταν η ομαδοποίηση σε 4 συστάδες των δεδομένων με παρόμοια χαρακτηριστικά μεταξύ τους με τον αλγόριθμο k-mean. Με τους αλγορίθμους Apriori και GSP προσπαθήσαμε να ανακαλύψουμε σχέσεις και κανόνες, αρχικά μεταξύ των μεταβλητών με κυριότερο στόχο εάν υπάρχει σχέση αναμεσά στην έκταση της

καμένης γης και την κατεύθυνση του ανέμου και στην συνέχεια σε επίπεδο διαφορετικής συστάδας.

Λέξεις κλειδιά: πυρκαγιές Ευβοίας, συσταδοποίηση, δέντρο απόφασης,ανάλυση κανόνων συσχέτισης και ανακάλυψης ακολουθιών

Abstract

A problem that is increasingly affecting more and more countries around the world is forest fires. The importance of this issue depends not only on the human losses but also on the impact on the environment and the economy. Today some methods have the potential to predict forest fires so that they can be stopped shortly after their occurrence or even before they happen. One way is using new technology, in particular remote sensing and Geographical Information Systems, which through satellite images or other means can detect the fire and predict its course. However, this solution requires special data processing and in many cases the costs are high. Another way is to analyze data to predict forest fires. The advantages of this solution are low cost and easy data processing. In this paper by using two methods, linear regression, and random forest, we will predict the area of burnt land in Evia and we will compare which method has higher accuracy and better performance. For this purpose, we collected and analyzed data from the fire and meteorological service for the period 2020-2022 concerning Evia and divided our data into two databases based on the extent of burnt land and applied the above methods. We also divided the vegetation of Evia into 5 major categories and added this variable to our model. Finally, to support local stakeholders and optimize their decisions regarding the prevention of future fires and the better allocation of firefighting forces, we first created a decision tree with the algorithm j48 with quite good results and then we performed descriptive mining on the data to identify correlations or patterns between them. The methods we used were clustering into 4 clusters of data with similar characteristics between them using the k-mean algorithm. With the Apriori and GSP algorithms, we tried to discover relationships and rules, first between the variables with the main objective of finding out if there is a relationship between the area of burnt land and wind direction and then at the level of different cluster

Keywords: fires in Evia, clustering, decision tree, analysis of association rules and sequence discovery

Περιεχόμενα

Λίστα Πινάκων

Λίστα Εικόνων

1 Εισαγωγή

Διαχρονικά το δέντρο έχει εξέχουσα σημασία και αποτελεί αναπόσπαστο κομμάτι της παράδοσης των λαών όπως για παράδειγμα ο πλάτανος που συμπεριλαμβάνεται σε πάρα πολλούς μύθους της αρχαιότητας. Η προσφορά του δέντρου (και κατ' επέκταση του δάσους) δεν περιορίζεται μόνο στην αισθητική ιδιότητα (Τσαγκάρη, Καρέτσος, & Προύτσος, 2011) αλλά έχει ζωτική σημασία για την επιβίωση του ανθρώπου. Χαρακτηριστικό παράδειγμα αποτελεί ότι παραπάνω από την μισή ποσότητα οξυγόνου που παράγεται κατά την φωτοσύνθεση (δέσμευση διοξειδίου του άνθρακα από την ατμόσφαιρα και μετατροπή σε οξυγόνο), μένει ελεύθερο για να καταναλωθεί από τους ζωντανούς οργανισμούς.

Η απειλή που αντιμετωπίζεται στις μεσογειακές περιοχές, όπως η Νότια Ευρώπη και η Ελλάδα, εντείνεται ακόμη περισσότερο, με αποτέλεσμα να παρουσιάζονται σοβαρές συνέπειες σε οικολογικό, οικονομικό και κοινωνικό επίπεδο. Παρατηρείται ότι το 85% της εκτάσεως που έχει υποστεί πύρινη καταστροφή στην Ευρώπη συμβαίνει ακριβώς στις μεσογειακές χώρες. (Jesús, Jose, & Camia, 2013). Για παράδειγμα στην Εύβοια το 2021 σύμφωνα με ανάλυση του Εθνικού Αστεροσκοπείου Αθηνών/meteo.gr το ένα τρίτο των δασών της καταστράφηκε και πιο συγκεκριμένα 275.000 στρέμματα και 34000 στρέμματα καλλιέργειας ελιάς.

Οι αρνητικές επιπτώσεις των δασικών πυρκαγιών έχουν ήδη καταγραφεί όπως η υποβάθμιση της αισθητικής αξίας του τοπίου, η μεταβολή του μικροκλίματος, η ερημοποίηση του εδάφους και η εμφάνιση πλημμυρικών φαινομένων. Επιπλέον, οι άμεσες και έμμεσες οικονομικές συνέπειες μιας δασικής πυρκαγιάς είναι πολύ σοβαρές καθώς όχι μόνο προκαλείται καταστροφή αρκετών δασικών προϊόντων (όπως ξυλεία και ρητίνη) αλλά παρουσιάζονται απώλειες περιουσιών και ανθρώπινων ζωών. Τέλος δεν είναι αμελητέα και η αίσθηση ανασφάλειας που δημιουργείται στον γενικό πληθυσμό. (Τσαγκάρη, Καρέτσος, & Προύτσος, 2011)

Όσον αφορά την Ελλάδα το πρόβλημα είναι πολύ σημαντικό καθώς όπως δείχνει και ο παρακάτω πίνακας από την έκθεση της WWF, από το 2016-2020 η καμένη έκταση είναι 1.230.062 στρέμματα και το κόστος της δασοπυρόσβεσης είναι 651.453.135,82 €

(πίνακας 1). Λόγω του τεράστιου κόστους που έχει η καταστολή της φωτιάς όλο και περισσότερα κράτη, έχουν αρχίσει να δίνουν σημασία στη πρόβλεψη και στην πρόληψη των πυρκαγιών που μέσω των νέων τεχνολογιών τηλεπισκόπησης και GIS (Geographic Information System), δίνεται η δυνατότητα να ανιχνευτεί και να παρακολουθηθεί η εξέλιξη μιας πυρκαγιάς από τα πρώιμα στάδια της. Χαρακτηριστικό παράδειγμα τέτοιας τεχνολογίας είναι η επεξεργασία των δορυφορικών εικόνων εάν εμφανιστεί καπνός ή όχι (Tutmez, Ozdogan, & Boran, 2016)

Επίσης πολλές μελέτες έχουν επικεντρωθεί στην πρόβλεψη της πορείας μιας φωτιάς μέσω μοντέλων προσομοίωσης που είναι προσαρμοσμένα ανάλογα με τις ιδιαιτερότητες της περιοχής με ικανοποιητικά αποτελέσματα (Alexandridis, Russo, Vakalis, & Siettos, 2011). Οι τεχνικές εξόρυξης δεδομένων μπορούν να βοηθήσουν τους διαχειριστές των πυρκαγιών ή τα κέντρα ελέγχου να διαμορφώσουν καλύτερες στρατηγικές κατανομής πυροσβεστικών πόρων και να δράσουν γρηγορότερα με προληπτικά μέτρα εντοπίζοντας της περιοχές κίνδυνου. Χαρακτηριστικά παραδείγματα τέτοιων τεχνικών είναι η ομαδοποίηση και ταξινόμηση των παρατηρήσεων σε συστάδες με παρόμοια χαρακτηριστικά ή η ανακάλυψη κρυμμένων μοτίβων και κανόνων που υπάρχουν στα ανεπεξέργαστα δεδομένα (Tutmez, Ozdogan, & Boran, 2016). Το αντικείμενο της συγκεκριμένης εργασίας είναι η ανακάλυψη τέτοιων μοτίβων και συσχετίσεων στο νησί της Ευβοίας με την χρήση του αλγορίθμου Apriori και GSP και η ομαδοποίηση των δεδομένων σε συστάδες με κοινά χαρακτηριστικά. Επίσης κατασκευάσαμε ένα δένδρο απόφασης που μπορεί να αποτελέσει σημαντικό εργαλείο για την καλύτερη κατανομή των πυροσβεστικών δυνάμεων. Τέλος, έχοντας γνωστά ορισμένα δεδομένα όπως η ταχύτητα και η κατεύθυνσή του ανέμου, η θερμοκρασία και η βλάστηση δημιουργήσουμε ένα μοντέλο εκτίμησης του μεγέθους της καμένης έκτασης χρησιμοποιώντας τους αλγορίθμους linear regression και Random Forest.

Έτος	Καμένες εκτάσεις (σε στρ.)	Κόστος/στρέμμα (σε ευρώ)	Κόστος δασοπυρόσβεσης
2016	420.011	529,61	222.442.025,71
2017	231.323	529,61	122.510.974,03
2018	193.816	529,61	102.646.891,76
2019	162.758	529,61	86.198.264,38
2020	222.154	529,61	117.654.979,94
Σύνολο περιόδου			**651.453.135,82**

Πίνακας 1.1 : Κόστος Δασοπυρόσβεσης Π.Σ. για την περίοδο 2016 - 2020 σε ευρώ.

1.1 Στόχος της εργασίας

Με βάση την παραπάνω εισαγωγή, στόχος της εργασίας είναι εφαρμόζοντας τεχνικές εξόρυξης δεδομένων στα δεδομένα πυρκαγιών που μας προσφέρει ελεύθερα η πυροσβεστική υπηρεσία, εμπλουτισμένα με τα μετεωρολογικά δεδομένα και την βλάστηση στην περιοχή της Ευβοίας, να δημιουργήσουμε χρήσιμους κανόνες συσχέτισης και να ανακαλύψουμε κρυμμένα μοτίβα αναμεσά στην καμένη γη, στην κατεύθυνση του ανέμου και στις υπόλοιπες μεταβλητές. Επιπλέον με τον αλγόριθμο k-mean θα ομαδοποιήσουμε σε 4 συστάδες τις πυρκαγιές ώστε οι τοπικοί αρμόδιοι φορείς να προσαρμόσουν ανάλογα την στρατηγική τους. Παράλληλα με την χρήση δύο μεθόδων, της γραμμικής παλινδρόμησης και του τυχαίου δάσους θα προβλέψουμε την έκταση της καμένης γης στην Εύβοια και θα συγκρίνουμε ποια μέθοδο έχει μεγαλύτερη ακρίβεια και καλύτερη απόδοση. Τέλος θα προχωρήσουμε και στην κατασκευή ενός δέντρου απόφασης, με απώτερο σκοπό όχι μόνο τον καλύτερο καταμερισμό των δυνάμεων της πυροσβεστικής αλλά και την οργάνωση ενός αποτελεσματικότερου συστήματος πρόληψης.

1.2 Περιγραφή των βασικών στοιχείων μιας Πυρκαγιάς.

Πριν προχωρήσουμε σε οποιαδήποτε μελέτη μιας πυρκαγιάς, είναι σημαντικό να αναφέρουμε ορισμένες βασικές εννοείς. Η καύση (combustion) είναι μια πολύπλοκη

διαδικασία κατά την οποία το καύσιμο θερμαίνεται, αναφλέγεται και οξειδώνεται γρήγορα, απελευθερώνοντας θερμότητα. Η πυρκαγιά είναι ένας συγκεκριμένος τύπος καύσης που αυτοσυντηρείται, εκπέμπει θερμότητα και συνοδεύεται από φλόγα και καπνό. Στη φωτιά, η διαθεσιμότητα του καύσιμου καυσίμου ρυθμίζεται από τη θερμότητα που παράγεται κατά την καύση (Scott, 2012).

Τα σωματίδια στερεών καυσίμων μετατρέπονται σε καύσιμα αέρια μέσω πυρόλυσης (pyrolysis), δηλαδή της διάσπασης σύνθετων μορίων σε απλούστερες, εύφλεκτες ουσίες που προκαλείται από τη θερμότητα. Αυτό δημιουργεί έναν κύκλο, όπου η καύση παράγει θερμότητα, η οποία παράγει περισσότερο καύσιμο, οδηγώντας σε περαιτέρω καύση.

Η φωτιά μπορεί να ταξινομηθεί σε δύο τύπους: φλεγόμενη καύση (flaming combustion) και καύση που σιγοκαίει (smoldering combustion). Η φλεγόμενη καύση συμβαίνει όταν το αέριο καύσιμο αναφλέγεται και παράγει θερμότητα και φως με τη μορφή φλόγας. Από την άλλη πλευρά, η καύση που σιγοκαίει περιλαμβάνει την αργή καύση στερεών καυσίμων χωρίς απαραίτητα να παράγονται ορατές φλόγες.

Εδώ να σημειώσουμε ότι όταν λεμέ φλόγες εννοούμαι την ορατή απόδειξη της ταχείας αντίδρασης μεταξύ καυσίμου και οξυγόνου κατά την παραπάνω διαδικασία.

Οι παράγοντες που είναι κρίσιμης σημασίας για την φωτιά είναι τρεις και πιο συγκεκριμένα το καύσιμο, η θερμότητα και το οξυγόνο, η όπως είναι γνωστοί στην βιβλιογραφία, το τρίγωνο της φωτιάς. Εάν απουσιάζει οποιοσδήποτε από αυτούς τους παράγοντες, η φωτιά σβήνει. Το καύσιμο παρέχει το υλικό για την καύση, η θερμότητα προωθεί την αντίδραση και το οξυγόνο τη συντηρεί (Scott, 2012).

Επίσης υπάρχει και η διάκριση της πυρκαγιάς ανάλογα την φύση του καυσίμου αλλά και το μέρος που λαμβάνει χώρα. Πιο συγκεκριμένα μια πυρκαγιά πάνω ή μέσα σε ένα κτίριο ονομάζεται πυρκαγιά δομής ή πυρκαγιά κτιρίου. Οι πυρκαγιές αυτές μπορούν να προκαλέσουν καταιγίδες πυροκάλυψης (pyrocumulonimbus) και να έχουν σημαντικές περιβαλλοντικές επιπτώσεις. Κατά τη διάρκεια της ιστορίας, οι πυρκαγιές της βλάστησης έχουν διαμορφώσει διακριτά καθεστώτα πυρκαγιάς (fire regimes) που ορίζονται από τη συχνότητα, την έκταση, το πρότυπο, τη συμπεριφορά και τις περιβαλλοντικές επιπτώσεις αυτών των πυρκαγιών. Οι οργανισμοί, συμπεριλαμβανομένων των φυτών, έχουν προσαρμοστεί σε αυτά τα καθεστώτα. Οι ανθρώπινες δραστηριότητες, όπως η διαχείριση

των πόρων και η χρήση της γης, επηρεάζουν επίσης τα πρότυπα πυρκαγιών. (Bowman, et al., 2020)

Οι πυρκαγιές βλάστησης καίνε κατά μέσο όρο 400-500 εκατομμύρια εκτάρια ετησίως, επηρεαζόμενες από το κλίμα, τη βλάστηση και τους ανθρώπινους παράγοντες. Η συχνότητα των πυρκαγιών είναι υψηλότερη σε περιοχές με ενδιάμεση παραγωγικότητα όπως οι τροπικές σαβάνες, ενώ τα τροπικά δάση και τα δάση υψηλής βιομάζας βιώνουν σπάνιες φυσικές πυρκαγιές. Οι άγονες περιοχές έχουν περιορισμένη ικανότητα καύσης, αλλά η δραστηριότητα των πυρκαγιών μπορεί να αυξηθεί υπό ορισμένες συνθήκες. (Bowman, et al., 2020)

Επιπρόσθετα και το περιβαλλοντικό κόστος είναι πολύ υψηλό καθώς αποτελούν σημαντικές πηγές αερίων του θερμοκηπίου, ιδίως CO2. Συμβάλλουν κατά 22% στις παγκόσμιες εκπομπές διοξειδίου του άνθρακα, με τις πυρκαγιές στις σαβάνες και τα λιβάδια να είναι οι κύριοι συντελεστές. Η αποψίλωση των δασών και η καύση γεωργικών αποβλήτων απελευθερώνουν επίσης σημαντικές εκπομπές άνθρακα. Αν και ορισμένες εκπομπές μπορούν να εξισορροπηθούν από την ανάκτηση μετά την πυρκαγιά, η μόνιμη αποψίλωση των δασών και η καύση οργανικών καυσίμων έχουν ως αποτέλεσμα καθαρές πηγές άνθρακα. (Bowman, et al., 2020)

Στην παρούσα εργασία θα αναλύσουμε τις πυρκαγιές βλάστησης. Οι πυρκαγιές βλάστησης, γνωστές με διάφορα ονόματα όπως πυρκαγιές άγριας γης, πυρκαγιές τοπίου κ.α, είναι διαταραχές που επηρεάζουν πολλαπλές πτυχές του συστήματος της Γης όπως για παράδειγμα την βιόσφαιρα, την υδρόσφαιρα, την ατμόσφαιρα κ.α. Απελευθερώνουν υδρατμούς, CO2, CH4, N2O και αερολύματα, επηρεάζοντας το ισοζύγιο ακτινοβολίας

1.3 Γενικές πληροφορίες για την δασική πυρκαγιά

Παρακάτω θα αναφέρουμε τα χαρακτηριστικά εκείνα που επηρεάζουν τη συμπεριφορά της δασικής πυρκαγιάς

a) **Βλάστηση**

Οι νεκροί ή ζωντανοί φυτικοί ιστοί αποτελούν την καύσιμη ύλη στις δασικές πυρκαγιές και διακρίνονται σε δυο κατηγορίες με βάση την θέση τους, σε εδάφους και αέρος. Η καύσιμη ύλη εδάφους περιλαμβάνει υλικά όπως

- τύρφη, δηλαδή η νεκρή βιομάζα (βελόνες, φύλλα, κλαδάκια) που η προέλευση της δεν μπορεί να αναγνωριστεί καθώς έχει αποσυντεθεί σε μεγάλο βαθμό.
- ξηροφυλλοτάπητας, δηλαδή βιομάζα που μπορούμε να αναγνωρίσουμε την προέλευση της καθώς δεν έχει αποσυντεθεί σε μεγάλο βαθμό.
- Ρίζες δέντρων
- νεαρά δέντρα
- θάμνους
- κλαδιά

Η πυρκαγιά που έχει ως κύριο καύσιμο την συγκεκριμένη κατηγορία διαδίδεται ταχύτατα αλλά σβήνει και σχετικά ευκολά.

Αντίθετα η εναέρια καύσιμη ύλη, είναι ο συνολικός αριθμός καύσιμης ύλης με ύψος από 1,5 μέτρα και υψηλοτέρα, όπως για παράδειγμα τα φύλλα πάνω στο δέντρο, τα κλαδιά στην κόμη των δένδρων τα αναρριχώμενα φυτά κ.α. Η συγκεκριμένη καύσιμη ύλη είναι υπεύθυνη για την αύξηση της έντασης της φωτιάς και την δημιουργία νέων εστιών πυρκαγιών σε μεγάλες αποστάσεις καθώς με τον άνεμο μεταφέρονται καύτρες η σε περίπτωση που υπάρχουν πεύκα τα κουκουνάρια. Οι πυρκαγιές που δημιουργούνται σε αυτού του είδους την καύσιμη ύλη είναι πολύ μεγάλης ταχύτητας και συνήθως μεγάλης έντασης.

b) Θερμοκρασία αέρα

Η συμπεριφορά της Φωτιάς επηρεάζεται σε μεγάλο βαθμό από την θερμοκρασία του αέρα λόγω του κατωτάτου ποσού θερμότητας που χρειάζεται η ανάφλεξη και η συνέχισης της διαδικασίας της καύσης. Μέσω της αντανάκλασης της ηλιακής ακτινοβολίας από την γη, αυτή απορροφάτε και θερμαίνεται η επιφάνεια της γης και κατά συνέπεια επηρεάζεται η θερμοκρασία των καυσίμων. Αυτή η μεταβολή επηρεάζει την ευαισθησία της καύσιμης ύλης στην ανάφλεξη με αποτέλεσμα σε υψηλές θερμοκρασίες να έχουμε ευκολότερη ανάφλεξη σε σχέση με τις χαμηλές. Για το

συγκεκριμένο ζήτημα οι (Dimitrakopoulos, Gogi, Stamatelos, & Mitsopoulos, 2011) αναφέρουν πως ο συνδυασμός μέτριοι ή/και δυνατοί άνεμοι, περίοδος καύσωνα και θερμοκρασίες πάνω από 30 °C στην χωρά μας οδηγεί σε μεγάλες δασικές πυρκαγιές. Επίσης μια άλλη ερευνά των (Jesús, Jose, & Camia, 2013) κατέληγαν στο συμπέρασμα ότι το μέγεθος της καμένης έκτασης επηρεάζεται σε σημαντικό βαθμό από τις υψηλές θερμοκρασίες.

c) Άνεμος

Ο αέρας σε κίνηση αναφέρεται ως άνεμος και χαρακτηρίζεται από την ταχύτητά του σε ένα συγκεκριμένο ύψος πάνω από το έδαφος (ή την κάλυψη της βλάστησης) και την κατεύθυνσή του σε σχέση με τον Βορρά (ή την ανοδική κατεύθυνση).
Η ταχύτητα του ανέμου υφίσταται συνεχείς μεταβολές με την πάροδο του χρόνου και, ως εκ τούτου, υπολογίζεται κατά μέσο όρο για μια συγκεκριμένη περίοδο για σκοπούς αναφοράς. Η ταχύτητα ριπής ανέμου, που συχνά περιγράφεται ως ταχύτητα αιχμής του ανέμου. (Scott, 2012) Επίσης μια τοποθεσία με συγκεκριμένη τοπογραφική διαμόρφωση σε συνδυασμό με κάποιες καιρικές ιδιαιτερότητές μπορεί να έχει τοπικούς ανέμους πχ μελτέμια (Ταμπάκη & Καρανικόλα, 2015). Για την σημασία της έντασης και την κατεύθυνση του ανέμου στην εξάπλωση μιας πυρκαγιάς έχουν γίνει αρκετές έρευνες. Οι (Dimitrakopoulos, Gogi, Stamatelos, & Mitsopoulos, 2011) έδειξαν ότι υπάρχει σχέση αναμεσά στο μέγεθος της καμένης έκτασης και την ένταση του ανέμου.

1.4 Πρόληψη

Η Δασική υπηρεσία στην χωρά μας είναι ο υπεύθυνος φορέας για την πρόληψη των δασικών πυρκαγιών. Με τον ορό πρόληψη εννοούμαι τις κατάλληλες πολίτικες και μετρά που συνίστανται ώστε να μειωθεί όχι μόνο η πιθανότητα εκδήλωσης και εξάπλωσης μιας πυρκαγιάς αλλά και στην μείωση των καταστροφών που προκαλούν. Τέλος η αναγκαιότητα ύπαρξης ενός μηχανισμού ικανού να εντοπίζει άμεσα κάθε πυρκαγιά στέλνοντας τις απαραίτητες δυνάμεις για την έγκαιρη κατάσβεσή της είναι αναγκαία για την σωστή πρόληψη (Ξανθόπουλος, 2016).

Εν αντιθέσει με την κοινή άποψη ότι η πρόληψη των πυρκαγιών είναι κάτι εύκολο και περιορίζεται μόνο στον καθαρισμό των δασών και την ευαισθητοποίηση της κοινωνίας, εν τούτης οι ενέργειες προλήψεις είναι πολύ περισσότερες όπως:

a) Ευαισθητοποίηση και ενημέρωση του κοινού

Αυτή η διαδικασία είναι εξαιρετικά πολύπλοκη, καθώς η καμπάνια ενημέρωσης των πολίτων πρέπει να αρχίζει από τους μαθητές και το σχολείο που είναι πιο δεχτικοί σε νέες ιδέες και η αποτελεσματικότητα θα είναι πολύ καλύτερη και φυσικά οι υπόλοιποι πολίτες να ενημερώνονται μέσω διάφορων διαθέσιμων μέσων επικοινωνίας (Παπαγεωργίου, Καρέτσος, & Κατσαδωράκης, 2012).

b) Τεχνολογικά μέτρα

Ορισμένες αιτίες των πυρκαγιών μπορούν να αποτραπούν με την εφαρμογή τεχνολογικών βελτιώσεων στα χρησιμοποιούμενα μέσα και τις πρακτικές που συχνά προκαλούν πυρκαγιές. Ορισμένα παραδείγματα είναι οι καταλύτες των οχημάτων, τα καπνοδοχεία που χρησιμοποιούνται από τους μελισσοκόμους για τον κάπνισμα των μελισσών κ.λπ. (Ξανθόπουλος, 2016).
Ειδικά τα τελευταία είναι υπεύθυνα για πολλές πυρκαγιές. Η χρήση ενός ψεκαστήρα νερού αποτελεί μια εξίσου αποτελεσματική εναλλακτική λύση που εφαρμόζεται από πολλούς μελισσοκόμους. Αντί να χρησιμοποιούν το καπνιστήρι, οι μελισσοκόμοι φέρουν έναν ψεκαστήρα που εκτοξεύει νερό και ακινητοποιεί τις μέλισσες, αποτρέποντάς τις από το να πετάξουν λόγω του νερού που είναι πάνω στα φτερά τους. Ο ψεκαστήρας νερού χρησιμοποιείται με τον ίδιο τρόπο με το καπνιστήρι.

c) Νομοθετικά μέτρα

Η παρουσία στρεβλώσεων, αντιθέσεων, ή ύπαρξη κενών και ανεπαρκειών στην νομοθεσία και στο ποινολόγιο δεν αποθαρρύνει τους επίδοξους εμπρηστές ώστε να πραγματοποιήσουν το έργο τους. Για παράδειγμα αποτελεί η δημιουργία νομοθεσίας για τη δημιουργία δασολογίου και δασικών χαρτών, αυστηρότερες ποινές στους εμπρηστές

και η σωστή εφαρμογή κείμενης δασικής και περιβαλλοντικής νομοθεσίας. (Ξανθόπουλος, 2016)

Στην χωρά μας, παρατηρείται το φαινόμενο ότι οι υπαίτιοι των πυρκαγιών (αστικών και αγροτοδασικών), δεν παραπέμπονται στην δικαιοσύνη καθώς είτε επειδή δεν μπορούν να τους εντοπίσουν, είτε τα αίτια που προκάλεσαν την φωτιά δεν εξακριβωθήκαν οπότε οι δικογραφίες πηγαίνουν στο αρχείο. Στις περιπτώσεις όμως που εξακριβωθήκαν τα αίτια, πάλι δεν παραπέμφθηκαν όλοι οι υπεύθυνοι για τα συγκεκριμένα εγκλήματα καθώς: α) ήταν αδύνατος ο εντοπισμός των δραστών καθώς δεν εξιχνιάσθηκαν όλες οι υποθέσεις και β) δεν περιείχαν επαρκή αποδεικτικά στοιχεία για την ποινική δίωξη των δραστών η παραπομπή στο ακροατήριο του αρμοδίου δικαστηρίου λόγω ανεπαρκούς άσκησης των ειδικών ανακριτικών καθηκόντων από το προσωπικό (Γκουρμπάτσης Α. , 2014).

d) Αποτελεσματική και κατάλληλη διαχείριση των δασών

Η πρόληψη των δασικών πυρκαγιών και η ορθή διαχείριση των δασών είναι κρίσιμης σημασίας παράγοντας για τον έλεγχο του προβλήματος και τον περιορισμό των δαπανών. Όταν λεμέ ορθή διαχείριση των δασών εννοούμε την απομάκρυνση της παραγόμενης δασικής βιομάζας, την φροντίδα των δέντρων με αραίωση και κλάδεμα των κλαδιών τους και απομάκρυνση των νεκρών δέντρων και όσων προσβλήθηκαν από παθογόνους οργανισμούς και έντομα κ.α (Ξανθόπουλος, 2016), (Τσαγκάρη, Καρέτσος, & Προύτσος, 2011). Επίσης με σκοπό την διάσπαση της συνέχειας της καύσιμης ύλης αλλά και την διευκόλυνση των πυροσβεστικών οχημάτων, είναι σύνηθες το φαινόμενο να δημιουργούνται δρόμοι και αντιπυρικές ζώνες μέσα στα δάση. Δυστυχώς όμως στην χωρά μας η μη συντήρηση αυτών έχει τα αντίθετα αποτελέσματα.

e) Προληπτικός (αντιπυρικός) σχεδιασμός

Η τεχνολογία και πιο συγκεκριμένα η ανάλυση των δεδομένων σε μια περιοχή όπως για παράδειγμα την πιθανότητα εκδήλωσης πυρκαγιάς, την έντασή της, την ικανότητα αντιμετώπισης, την προϊστορία της και αλλά δεδομένα, μπορεί να δώσει πληροφορίες που πάνω σε αυτές μπορούν οι σχετικοί φορείς να δημιουργήσουν έναν αντιπυρικό

σχεδιασμό που θα προβλέπει από τον τρόπο συνεργασίας τους μέχρι και το πως θα γίνει με τον αποδοτικότερο τρόπο η κατανομή των δυνάμεων

f) Προληπτικά έργα (οδοί, αποθήκες νερού, αεροδρόμια, πυροπροστατευτικές ζώνες κ.λπ.)

Απόρροια των προηγουμένων είναι ο καθορισμός όχι μόνο των έργων που θα γίνουν αλλά και η προτεραιότητα τους. Φυσικά στα παραπάνω λαμβάνεται υπόψιν και ο διαθέσιμος προϋπολογισμός.

g) Ετοιμότητα - σύστημα αξιολόγησης κινδύνου

Στην χωρά μας, η Γενική Γραμματεία Πολιτικής Προστασίας (ΓΓΠΠ) δημοσιεύει καθημερινά στους αρμοδίους φορείς ένα χάρτη με την πρόγνωση του κίνδυνου πυρκαγιάς (https://civilprotection.gov.gr/). Αποτελεί βασικό εργαλείο του σχεδιασμού για την πρόληψη της δασικής πυρκαγιάς και έχει ως σκοπό σε κρίσιμες ημέρες, την επιφυλακή των δυνάμεων ενώ σε ημέρες χαμηλού κίνδυνου μπορούν να εξοικονομηθούν πόροι. Επίσης χρησιμεύει και για την ενημέρωση των πολίτων ώστε και αυτοί με την σειρά τους να είναι προσεκτικοί και προληπτικοί. Παρακάτω παρατίθεται ένας τέτοιος χάρτης.

Εικόνα 1. 1 Παράδειγμα Χάρτη Πρόβλεψης Κίνδυνου

h) Επίγειες περιπολίες στο δάσος

Είναι μια πολύ σημαντική δράση καθώς με τις επίγειες περιπολίες στα δάση από πυροσβέστες η εθελοντές όχι μόνο ευαισθητοποιούν τους πολίτες αλλά και αποτρέπουν την έναρξη πυρκαγιάς από αμέλεια η κάποιον εμπρησμό. Η συχνότητα των περιπολίων και η διαδρομή τους καθορίζονται από το αντιπυρικό σχέδιο και την επικινδυνότητα της μέρας.

i) Εντοπισμός των πυρκαγιών από το έδαφος, τον αέρα και το διάστημα

Τέλος μια από τις σημαντικότερες δράσεις είναι ο γρήγορος και άμεσος εντοπισμός της πυρκαγιάς που είναι ζωτικής σημασίας για τον περιορισμό των καταστροφών που θα προκληθούν από αυτή. Ο συνδυασμός πολλαπλών μεθόδων εντοπισμού βοηθούν στην ορθή κατανόηση της κατάστασης και την γρήγορη κινητοποίηση των πυροσβεστικών δυνάμεων.

Πιο συγκεκριμένα ένα δίκτυο πυροφυλακίων που είναι επανδρωμένα με εθελοντές και πυροσβέστες, οι περιπολίες στα δάση καθώς και η αναφορά νέων πυρκαγιών από όλους, αποτελούν τον επίγειο εντοπισμό πυρκαγιών. Η επισήμανση και παρατήρηση των πυρκαγιών κατά την διάρκεια των νυχτερινών ωρών γίνεται μέσω των φλογών της

φωτιάς ενώ την ημέρα από την αναδυομένη στήλη καπνού που παράγεται (Τσαγκάρη, Καρέτσος, & Προύτσος, 2011).

Τα τελευταία χρονιά έχει αυξηθεί και η χρήση καμερών για την επίγεια παρακολούθηση των δασών που προωθούν τις εικόνες σε ένα κέντρο επιτήρησης οπού εκεί υπάρχει η δυνατότητα να εφαρμοστούν προηγμένες τεχνολογίες που ανιχνεύουν αυτόματα πιθανές εστίες πυρκαγιάς.

Από τον εναέριο χώρο οι πιλότοι (είτε της πολιτικής είτε της στρατιωτικής αεροπορίας) είναι υποχρεωμένα να αναφέρουν πυρκαγιές που παρατηρούν. Επιπρόσθετα τα μη επανδρωμένα αεροσκάφη (UAVs) μπορούν να πραγματοποιούν περιπολίες και να εντοπίσουν κάποιο γεγονός (Παπαγεωργίου, Καρέτσος, & Κατσαδωράκης, 2012).

Από το διάστημα, υπάρχει η δυνατότητα μέσω των δορυφόρων να εντοπίζονται οι φωτιές και να παρέχουν πληροφορίες για την θέση τους και την εξάπλωση της και να δώσουν άμεσα χρήσιμες γεωγραφικές πληροφορίες (Παπαγεωργίου, Καρέτσος, & Κατσαδωράκης, 2012).

Από όλα τα παραπάνω καταλαβαίνουμε ποσό σημαντική και πολύπλοκη είναι η πρόληψη των πυρκαγιών.

1.5 Καταστολή

Το μέγεθος της καμένης δασικής έκτασής εξαρτάται από ένα πολύ σημαντικό παράγοντα αυτόν της καταστολής των πυρκαγιών. Όλες οι διάφορες τεχνικές που υπάρχουν σήμερα έχουν ως αρχή την απομάκρυνση ενός η περισσοτέρων στοιχείων που συνθέτουν το «τρίγωνο της φωτιάς» που αναλύθηκε προηγούμενος. Θεωρητικά, η επιλογή της κατάλληλης μεθόδου κατάσβεσης δεν είναι τυχαία επιλογή, εξαρτάται από το είδος της καύσιμης βιομάζας, το ποσό δύσβατη είναι η όχι η περιοχή, τις καιρικές συνθήκες που επικρατούν, τον τύπο πυρκαγιάς και αλλά. Η πραγματικότητα όμως είναι διαφορετική καθώς επηρεάζεται από το διαθέσιμο προσωπικό, την διαθεσιμότητα των πόρων καθώς και την προσβασιμότητα στην περιοχή (Ταμπάκη & Καρανικόλα, 2015). Επίσης πολύ σημαντική είναι και η σωστή πρώτη εκτίμηση και αναφορά των πρώτων δασοπυροσβεστών που καταφθάνουν σε μια πυρκαγιά βάση της οποίας γίνεται η κινητοποίηση επικουρικών δυνάμεων.

Στην χωρά μας από το 1998 μέχρι και σήμερα, την ευθύνη για την καταστολή των πυρκαγιών την έχει η Πυροσβεστική Υπηρεσία. (Γκουρμπάτσης Α. , 2015)
Ο διαχωρισμός των μέσων κατάσβεσης είναι αναμεσά σε επίγεια και εναερία αλλά την τελική κατάσβεση και τον κύριο όγκο των προσπαθειών κατασβέσεις τον αναλαμβάνουν οι επίγειες δυνάμεις ενώ επικουρικά λειτουργούν τα εναέρια μέσα με πολύ σημαντική συνεισφορά καθώς ρίχνουν νερό ή και επιβραδυντικά υλικά στην καύσιμη ύλη (Ταμπάκη & Καρανικόλα, 2015).

1.6 Η μεταπυρική αποκατάσταση

Η σειρά μέτρων που έχουν ως στόχο να αντιμετωπίσουν τις ζημίες που προκλήθηκαν από τις πυρκαγιές, να αποκαταστήσουν τις καμένες περιοχές και να αποσοβήσουν δευτερεύουσες μελλοντικές καταστροφές (διάβρωση του εδάφους, πλημμύρες) ονομάζεται μεταπυρικη αποκατάσταση (Παπαγεωργίου, Καρέτσος, & Κατσαδωράκης, 2012). Τα μέτρα αυτά καλύπτουν τα έξης:

a. Την διαχείριση των καμένων κορμών δένδρων και πιο συγκεκριμένα την ρίψη τους στο έδαφος ώστε και για λόγους ασφάλειας και επιτάχυνσης της αποσύνθεση τους. Επίσης πολύ σημαντικό ρόλο παίζουν και στην προστασία του εδάφους από την διάβρωση και την απομάκρυνση του από την καμένη περιοχή (Παπαγεωργίου, Καρέτσος, & Κατσαδωράκης, 2012).
b. Πιο συγκεκριμένα για την διάβρωση του εδάφους και την προστασία του ξεγυμνωμένου εδάφους σύνηθες μετρό είναι η δημιουργία κορμοδεματων,κλαδοπλεγματων και κορμοφραγματων από οπλισμένο σκυρόδεμα σε πλάγιες και ρέματα με σκοπό να παραμείνει το χώμα στην περιοχή και να αρχίσει η αναγέννηση (Παπαγεωργίου, Καρέτσος, & Κατσαδωράκης, 2012).

   Για παράδειγμα στην Εύβοια το 2021, η Δασική Υπηρεσία και ο αρμόδιος υφυπουργός Γ. Αμυράς εξέδωσαν την εγκύκλιο «ΥΠΕΝ/ΔΔΔ/117627/3873 «Γενικές Οδηγίες – Κατευθύνσεις για τη διαχείριση και αξιοποίηση της ξυλείας των καμένων δασών».» όπου σε αυτή αναλύονται όχι μονό τα μετρά κατά της διάβρωσης του εδάφους και της εκμετάλλευσης του ξύλου ως υλικό άλλα βασιζόμενη στις ειδικές

γνώσεις και την εμπειρία του προσωπικού της Δασικής υπηρεσίας θέτει συγκεκριμένους άξονες της προστασίας των υδάτινων πόρων, της βιοποικιλιοτηας και την προστασία του εδαφους. Από τα πραπανω συμπεραίνουμε ότι είναι πολύ σημαντική η συγκεκριμένη διαδικασία και αρκετά απαιτητική.

c. Την αναγέννηση της βλάστησης στις συγκεκριμένες περιοχές μέσω της σποράς και της αναδάσωσης οπού αυτό είναι απαραίτητο για την αποκατάσταση της περιοχής (Παπαγεωργίου, Καρέτσος, & Κατσαδωράκης, 2012).
d. Τέλος τα τελευταία χρόνια υπολογίζονται στις ενέργειες μεταπυρικης αποκατάστασης και τα οικονομικά και κοινωνικά προβλήματα που προκύπτουν από τις πυρκαγιές στην τοπική κοινωνία (Παπαγεωργίου, Καρέτσος, & Κατσαδωράκης, 2012). Για παραδούμε μια τέτοια ενέργεια είναι το επίδομα 800 ευρώ το μήνα για τους ρητινοπαραγωγους που πλέον αντιμετωπίζουν μεγάλο πρόβλημα καθώς το δάσος τους κάηκε και έχασαν το εισόδημα τους.

2 Περιγραφική Στατιστική των πυρκαγιών από το 2011-2022

Στην ενότητα αυτή θα αναλύσουμε εν συντομία ορισμένα στατιστικά και θα προσπαθήσουμε να εξάγουμε κάποια συμπεράσματα από τα δεδομένα πυρκαγιών που υπάρχουν στον ιστοτόπο («https://www.fireservice.gr/el/synola-dedomenon»). Στην συγκεκριμένη έρευνα χρησιμοποιήσαμε τα δεδομένα από το 2011-2022 καθώς στα προηγούμενα χρόνια δεν συμπλήρωναν τις πυροσβεστικές δυνάμεις που συμμετείχαν στην κατάσβεση της πυρκαγιάς. Το Excel έχει 116256 τιμές και 16 στήλες. Πιο συγκεκριμένα οι στήλες είναι Νομός, Ημερ/νία Έναρξης πυρκαγιάς, Δάση, Δασική Έκταση, Άλση, Χορτ/κές Εκτάσεις, Καλάμια_Βάλτοι, Γεωργικές Εκτάσεις, Υπολείμματα Καλλιεργειών, Σκουπι_δότοποι, Πυρο/σωμα, Πεζοπόρα Τμήματα, Εθελοντές, Στρατός, Άλλες Δυνάμεις και αναλύονται στον παρακάτω πίνακα.

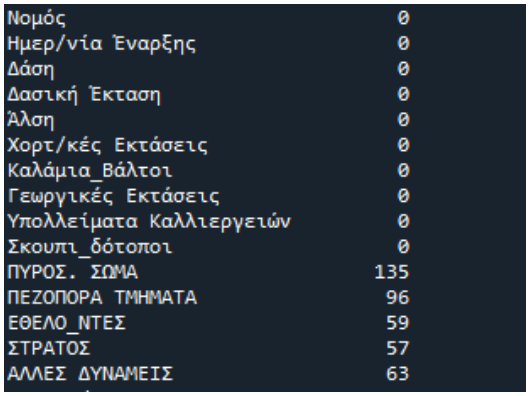
Νομός	Αναφέρουμε την περιοχή που ξεκίνησε η πυρκαγιά. Στην συγκεκριμένη ερευνά εξαιρέσαμε την στήλη δήμος καθώς δεν ήταν σε όλα τα πεδία συμπληρωμένη.
Ημερ/νία Έναρξης πυρκαγιάς	Η ημερομηνία που ξεκίνησε η φωτιά.
Δάση	Σύμφωνα με τη παρ. 1 του άρθρου 3 ν. 998/79, δάσος είναι ο χώρος εκείνος που αποτελείται από πυκνή βλάστηση, δέντρα και άλλα φυτά. Επίσης συμβάλλει στην δημιουργία μιας ειδικής οικολογικής κοινότητας και αποτελεί ένα ξεχωριστό φυσικό περιβάλλον.
Δασική έκταση	Όταν η άγρια ξυλώδης βλάστηση, υψηλή η θαμνώδης είναι αραιή τότε το δάσος θεωρείται Δασική έκταση (Παρ 2 του ιδίου νόμου)
Χορτολιβαδικές εκτάσεις	Το χαρακτηριστικό των Χορτολιβαδικων εκτάσεων είναι η μη συστηματική καλλιέργεια και η ύπαρξη μικρού ποσοστού κάλυψης από ξυλώδη φυτά ικανού να μην πληρούν τα κριτήρια ένταξης στην δασική έκταση.
Υπολείμματα καλλιεργειών	Μια σύνηθες πρακτική των γεωργών είναι τα υπολείμματα των σοδιών τους η των γεωργικών εργασιών τους να τα καινέ με αποτέλεσμα οι φωτιές να επεκτείνονται και στις γειτονικές εκτάσεις. Η καύση των γεωργικών υπολειμμάτων κατά τους χειμερινούς μήνες ευθύνεται για το μεγαλύτερο ποσοστό των πυρκαγιών.
Βάλτοι	Η πλημμυρισμένοι έκταση γης με στατικό γλυκό νερό και συγκεκριμένη βλάστηση όπως καλαμιώνες, χαρακτηρίζεται ως Βάλτοι.
Αλσος	Το μικρό σε μέγεθος δάσος που βρίσκεται μέσα η κοντά σε κατοικημένες περιοχές, ονομάζεται Άλσος
Σκουπιδότοποι	Ο χώρος που γίνεται η ρίψη των απορριμμάτων χαρακτηρίζεται ως Σκουπιδότοπος.
Γεωργικές εκτάσεις	Η χρήση γης αποκλειστικά για γεωργική παραγωγή χαρακτηρίζεται ως Γεωργική έκταση.
Πυρο/σωμα	Το πυροσβεστικό σώμα είναι μια επαγγελματική υπηρεσία με άρτια οργάνωση και συγκεκριμένους ρόλους οπού αναλαμβάνει την πυρόσβεση και την διάσωση των ανθρώπων

	και των περιουσιακών στοιχείων κατά τη διάρκεια μιας πυρκαγιάς η άλλων έκτακτων αναγκών.
Πεζοπόρα Τμήματα	Τα πεζοπόρα τμήματα είναι ομάδες εκπαιδευμένων ατόμων που ειδικεύονται στον περιορισμό και στην κατάσβεση των πυρκαγιών σε περιοχές που δεν είναι προσβάσιμες με τα πυροσβεστικά οχήματα. Είναι εξοπλισμένοι με πυροσβεστικές αντλίες, χειροκίνητα εργαλεία και προστατευτικό εξοπλισμό που χρησιμοποιείται για την ασφάλεια τους.
Εθελοντές	Είναι τα άτομα εκείνα που προσφέρουν τον ελεύθερο χρόνο τους και τις δεξιότητές τους, χωρίς αμοιβή για το έργο τους και βοηθούν σε μια συγκεκριμένη δραστηριότητα ή πρόγραμμα, συνήθως για το κοινό καλό.
Στρατός	Οι στρατιωτικές δυνάμεις που συμμετέχουν στην κατάσβεση μιας πυρκαγιάς.
Άλλες Δυνάμεις	Διαφορετικοί τύποι πυροσβεστικών δυνάμεων που δεν συμπεριλαμβάνονται στις παραπάνω

Πίνακας 2. 1 Στήλες Excel με στατιστικά πυρκαγιών 2011-2022

2.1 Προετοιμασία Δεδομένων

Από το site κατεβάσαμε 12 excel και πήραμε τις στήλες που περιγράψαμε παραπάνω. Στην συνέχεια καθαρίσαμε τα δεδομένα μας από τα κενά και διορθώσαμε όσα δεδομένα δεν είχαν ορθή μορφή, το μετατρέψαμε σε csv και μέσω python code πραγματοποιήσαμε στατιστική ανάλυση των πεδίων καθώς και δημιουργήσαμε και ορισμένα διαγράμματα. Τέλος τις κενές τιμές θα τις αντικαταστήσουμε μέσω της Python με 0 ώστε να μην επηρεαστούν τα σχήματα. (#na_count = df.isna(). sum(), #print(na_count), #df_cleaned = df.fillna(0))



Εικόνα 2. 1 NA τιμές ανά στήλη

```
 0   Νομός                     116256 non-null  object
 1   Ημερ/νία Έναρξης          116256 non-null  datetime64[ns]
 2   Δάση                      116256 non-null  float64
 3   Δασική Έκταση             116256 non-null  float64
 4   Άλση                      116256 non-null  float64
 5   Χορτ/κές Εκτάσεις         116256 non-null  float64
 6   Καλάμια_Βάλτοι            116256 non-null  float64
 7   Γεωργικές Εκτάσεις        116256 non-null  float64
 8   Υπολλείματα Καλλιεργειών  116256 non-null  float64
 9   Σκουπι_δότοποι            116256 non-null  float64
 10  ΠΥΡΟΣ. ΣΩΜΑ               116121 non-null  float64
 11  ΠΕΖΟΠΟΡΑ ΤΜΗΜΑΤΑ           116160 non-null  float64
 12  ΕΘΕΛΟ_ΝΤΕΣ                116197 non-null  float64
 13  ΣΤΡΑΤΟΣ                   116199 non-null  float64
 14  ΑΛΛΕΣ ΔΥΝΑΜΕΙΣ            116193 non-null  float64
```

Εικόνα 2. 2 Τα δεδομένα ανά στήλη

2.2 Five Number Summary

Η ανάλυση μας θα ξεκινήσει με την σύνοψη των πέντε αριθμών (five number summary) η οποία θα εφαρμοστεί στο csv που δημιουργήσαμε από τα δεδομένα των δασικών πυρκαγιών. Αυτοί οι αριθμοί μας προσφέρουν μια εποπτική εικόνα των δεδομένων που είναι πολύ σημαντική ώστε να διακρίνουμε από την αρχή και χωρίς ιδιαίτερη επεξεργασία το εύρος των δεδομένων, την μεγαλύτερη και την μικρότερη τιμή, την θέση της διάμεσου και αλλά. Επιπλέον μας δίνεται η δυνατότητα να πραγματοποιήσουμε συγκρίσεις μεταξύ διαφορετικών συνόλων και να εξάγουμε χρήσιμα συμπεράσματα.
Τέλος τα μέτρα για τα οποία μιλάμε είναι τα παρακάτω: η ελάχιστη τιμή, το κατώτερο τεταρτημόριο (Q1), η διάμεσος (Q2) το 3 τεταρτημόριο (Q3) και η μέγιστη τιμή. Επίσης στον παρακάτω πίνακα προσθέσαμε και το std (η τυπική απόκλιση) που είναι ένα μέτρο, το οποίο φανερώνει πόσο διασκορπισμένα είναι τα δεδομένα γύρω από το κέντρο της κατανομής (τη μέση τιμή).
Μέσω της εντολής describe δημιουργήθηκε ο παρακάτω πίνακας που περιέχει συγκεντρωτικά όλες τις στατιστικές μετρήσεις των στηλών με τις καμένες εκτάσεις:

	Statistic_measures	Δάση	Δασική Έκταση	Άλση	Χορτ/κές Εκτάσεις	Καλάμια_βάλτοι	Γεωργικές Εκτάσεις	Υπολείματα Καλλιεργειών	Σκουπι_δότοποι
2	count	116256	116256	116256	116256	116256	116256	116256	116256
3	mean	8,129502993	11,88900547	0,008741312	5,225621473	0,872438068	7,091414207	3,996335243	0,056808337
4	std	931,7662701	442,055791	0,771057526	123,1927702	18,43561204	556,6593871	84,02934511	6,01685142
5	min	0	0	0	0	0	0	0	0
6	25%or Q1	0	0	0	0	0	0	0	0
7	50% or Q2	0	0	0	0	0	0	0	0
8	75% or Q3	0	0	0	0,1	0	0,01	0	0
9	max	288428,47	65650	200	16000,56	4990	170224,27	23960	2000
10									

Πίνακας 2. 2 Στατιστικές μετρήσεις διαφορετικών ειδών καμένων εκτάσεων

Για παράδειγμα η στήλη Δασική Έκταση μας λέει ότι:

- count: Το πλήθος των τιμών είναι 116256
- mean: Η μέση τιμή των καμένων εκτάσεων (σε στρέμματα) είναι περίπου 11.889.
- std: Είναι η διακύμανση των τιμών από την μέση τιμή και είναι 442,056. Ο συγκεκριμένος αριθμός φανερώνει μια σημαντική διακύμανση στα δεδομένα μας
- min: Η ελάχιστη τιμή στη στήλη. Σε αυτήν την περίπτωση, είναι 0
- 25%: Το ποσοστό των τιμών που είναι μικρότερες από ή ίσες με το 25% του δείγματος. Σε αυτήν την περίπτωση, το 25% των τιμών είναι μηδέν, που μας φανερώνει ότι ένα μεγάλο ποσοστό των δειγμάτων έχει την τιμή μηδέν.
- 50%: Το ποσοστό των τιμών που είναι μικρότερες από ή ίσες με το 50% του δείγματος. Σε αυτήν την περίπτωση, το 50% των τιμών είναι επίσης μηδέν.
- 75%: Το ποσοστό των τιμών που είναι μικρότερες από ή ίσες με το 75% του δείγματος. Σε αυτήν την περίπτωση, το 75% των τιμών είναι επίσης μηδέν.
- max: Η μέγιστη τιμή στη στήλη. Η μέγιστη τιμή είναι 65,650.

Η σύγκριση αυτών των στατιστικών μπορεί να δώσει μια γενική εικόνα της κατανομής των δεδομένων και της διακύμανσης τους. Βάσει αυτών των στατιστικών, μπορούμε να παρατηρήσουμε ότι οι τιμές κάθε μεταβλητής έχουν μεγάλη διακύμανση. Την μεγαλύτερη διακύμανση την έχουν τα δάση με 931,76. Επίσης, ορισμένες μεταβλητές έχουν υψηλές μέσες τιμές όπως το 11,89 (δασικές εκτάσεις), ενώ άλλες έχουν χαμηλές. Τέλος οι τιμές του 25%, 50% και 75% που είναι μηδέν υποδηλώνουν ότι η πλειονότητα των τιμών είναι μηδέν. Αυτό συμβαίνει καθώς οι μηδενικές τιμές αντιπροσωπεύουν απουσία δεδομένων ή έλλειψη πληροφοριών για τις εκτάσεις καμένης γης. Για

παράδειγμα στο excel που έχουμε δημιουργήσει 1377 συμβάντα πυρκαγιών δεν έχουν συμπληρωμένα τα πεδία με την καμένη γη. Εικάζουμε ότι οι άνθρωποι που καταχωρούν τα δεδομένα, έχουν ως υποχρεωτικά πεδία την ημερομηνία έναρξης της πυρκαγιάς, την ώρα και τον νομό (είναι πεδία που είναι πάντα συμπληρωμένα) και δεν δίνουν τόση μεγάλη σημασία στην καταχώρηση της καμένης έκτασης ούτε στα σώματα της πυροσβεστικής που συμμετείχαν.
Για να αντιμετωπίσουμε το συγκεκριμένο ζήτημα δημιουργήσαμε και έναν δεύτερο πίνακα όπου για να υπολογιστεί ως πυρκαγιά πρέπει να είναι πάνω από 0,1 στρέμματα η καμένη έκταση.

Statistic_measure	Δάση	Δασική Έκταση	Άλση	Χορτ/κές Εκτάσεις	Καλάμια_Βάλτοι	Γεωργικές Εκτάσεις	Υπολλείματα Καλλιεργειών	Σκουπι_δότοποι
count	4406	20725	365	32283	24383	29970	18580	3252
mean	214,5037449	66,69086707	2,784191781	18,81825884	4,159707993	27,50815649	25,0052718	2,030845633
std	4782,112797	1045,251725	13,49568056	233,2341403	40,08561707	1096,119104	208,943143	35,92468197
min	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01
25%or Q1	0,1	0,3	0,1	0,2	0,2	0,2	0,3	0,1
50% or Q2	1	1,5	0,2	1	0,9	1	2	0,2
75% or Q3	5	6	1	4	2	3	10	1
max	288428,47	65650	200	16000,56	4990	170224,27	23960	2000

Πίνακας 2. 3 Στατιστικές μετρήσεις διαφορετικών ειδών καμένων εκτάσεων!=0

Κώδικας σε python

```
columns = ['Δάση', 'Δασική Έκταση', 'Άλση', 'Χορτ/κές Εκτάσεις', 'Καλάμια_Βάλτοι',
'Γεωργικές Εκτάσεις', 'Υπολείμματα Καλλιεργειών', 'Σκουπι_δότοποι']
df_selected = df_cleaned[columns]
df_selected_no_miden = df_selected[df_selected != 0]
df_sum_desc = df_selected_no_miden.describe()
df_sum_desc.to_excel("aaa.xlsx", index=False)
```

Όπως ήταν αναμενόμενο από αυτή την διάκριση επηρεάστηκε πάρα πολύ η μέση τιμή και η τυπική απόκλιση. Πιο συγκεκριμένα από το παραπάνω σχήμα φαίνεται ότι τα δάση έχουν τη μεγαλύτερη μέση τιμή, διακύμανση και το max ενώ οι δασικές εκτάσεις έχουν το μεγαλύτερο Q2 και Q3

Στην συνέχεια ενοποιήσαμε όλα τα είδη Α/Φ Μέσων και οχημάτων και δημιουργήσαμε την στήλη All_vehi, ενοποιήσαμε όλα τα διαφορετικά σώματα της πυροσβεστικής στην στήλη Puro_Men και προσθέσαμε όλα τα είδη καμένης γης και δημιουργήσαμε την Burn_Land και πραγματοποιήσαμε την παραπάνω ανάλυση και τα αποτελέσματα συνοψίζονται στον παρακάτω πίνακα:

Στήλη1	**Puro_Men**	**All_vehi**	**Burn_Land**
count	112744	112206	106445
mean	6,742451926	2,781651605	36,68043196
std	19,22843531	5,523701893	1792,46826
min	1	1	0,01
25%	2	1	0,2
50%	4	2	1
75%	6	3	5
max	2661	559	511854,14
Kurt	4253,79	1430,67	62933,4
skew	45,14	24,5	228,29

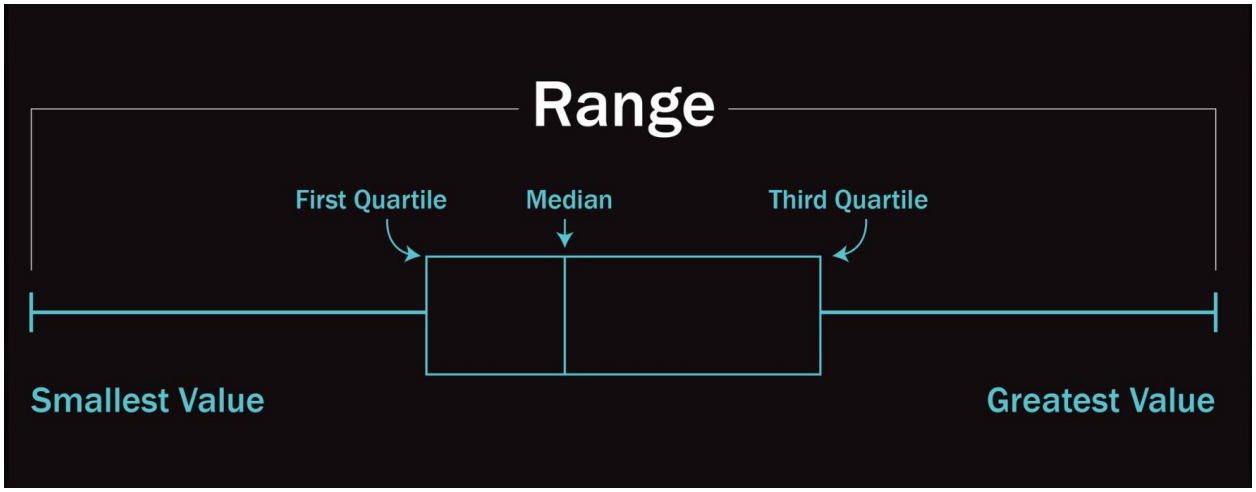
Πίνακας 2. 4 Στατιστικές μετρήσεις συνολικών Μηχ/των μέσων, πυροσβεστικών μέσων και οχημάτων.

Εδώ προσθέσαμε και 2 παραπάνω μέτρα διασποράς που μας λένε πόσο ευρύ είναι το σύνολο των δεδομένων μας:

- Skewness: Αυτό είναι ένα σημαντικό μέτρο που μας δείχνει το σχήμα μιας κατανομής. Η λοξότητα υπολογίζετε χρησιμοποιώντας την εντολή της python.skew(). Επιπλέον, εάν οι τιμές είναι κάτω από -1, έχουμε αρνητική ή αριστερή λοξότητα και αν η λοξότητα είναι πάνω από +1, έχουμε θετική ή δεξιά λοξότητα. Σε κάθε περίπτωση μας φανερώνει ότι έχουμε ακραίες τιμές. Στην περίπτωσή μας όλες οι τιμές είναι θετικές οπότε έχουμε ιδιαίτερα δεξιά λοξότητα
- Kurtosis: Είναι ο βαθμός αιχμής μιας κατανομής που λαμβάνεται συνήθως σε σχέση με μια κανονική κατανομή. Εάν τα δεδομένα έχουν μεγαλύτερη κορυφή από ό,τι στην κανονική κατανομή, τότε η Kurt >0 και ονομάζεται λεπτόκυρτη, ενώ μια χαμηλότερη κορυφή έχει Kurt<0 και ονομάζεται πλατύκυρτη κατανομή.

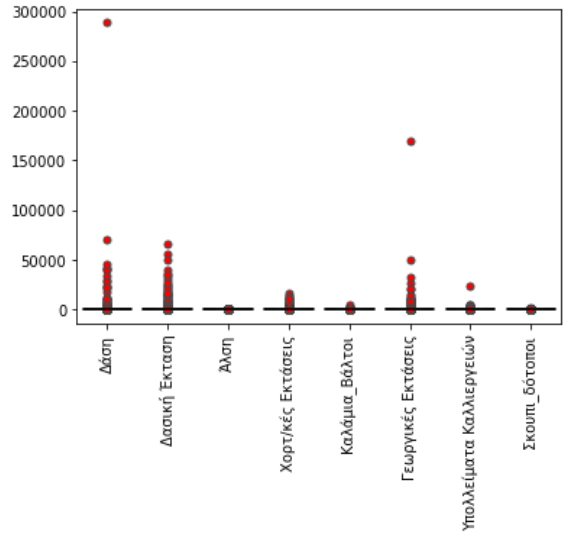
Στην περίπτωσή μας είναι θετικές και η κατανομή μας είναι λεπτόκυρτη και μας προϊδεάζει ότι υπάρχουν ακραίες τιμές.

Μια σύνοψη πέντε αριθμών μπορεί να αναπαρασταθεί σε ένα διάγραμμα γνωστό ως διάγραμμα box and whisker. Παρακάτω δίνεται ο ορισμός "*Box-and-Whisker Plot a plot that shows the center, spread, and skewness of a data set. It is constructed by drawing a box and two whiskers that use the median, the first quartile, the third quartile, and the smallest and the largest values in the data set between the lower and the upper inner fences"* (Mann, 2011).

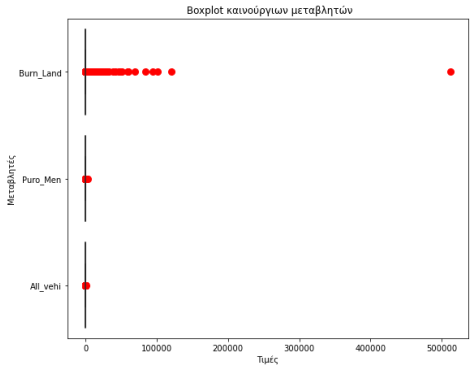


Εικόνα 2. 3 Παράδειγμα ενός Boxplot

Εδώ βλέπουμε και σχηματικά τις ακραίες τιμές (κόκκινο χρώμα) που υπάρχουν στο δείγμα μας.



Εικόνα 2. 4 Boxplot ειδών καμένης Γης



Εικόνα 2. 5 Boxplot καινούργιων μεταβλητών

2.3 Προσωπικό

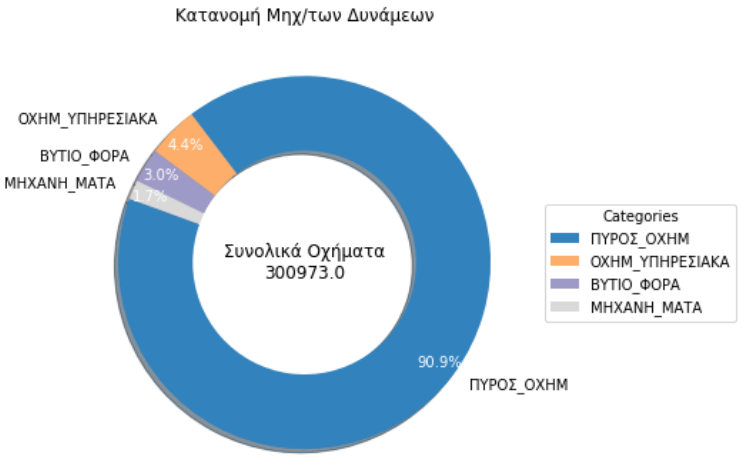
Η κατανομή των δυνάμεων για τις πυρκαγιές από το 2011-2022 φαίνεται στο παρακάτω σχήμα.

Εικόνα 2. 6 Κατανομή πυροσβεστικών δυνάμεων

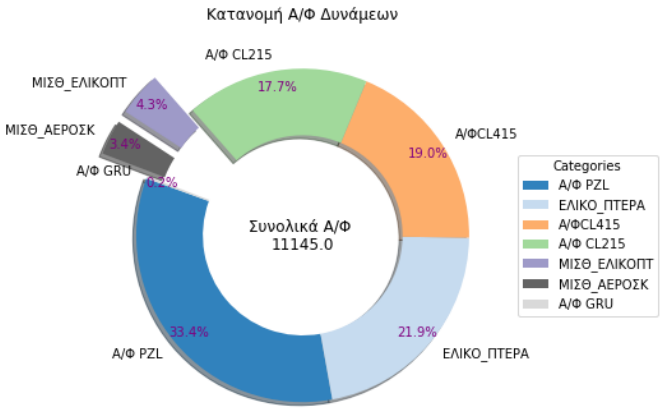
Το μεγαλύτερο ποσοστό ανήκει στο πυροσβεστικό σώμα ενώ εντύπωση μας προκαλεί το ποσοστό των εθελοντών το οποίο είναι μόλις στο 3,2%.

2.4 Μηχανοκίνητα Μέσα

Τα παρακάτω σχήματα μας δείχνουν ποια μηχανοκίνητα μέσα και Α/Φ χρησιμοποιήθηκαν συνολικά την περίοδο 2011-2022. Το μεγαλύτερο ποσοστό καταλαμβάνουν τα πυροσβεστικά οχήματα.



Εικόνα 2. 7 Κατανομή Μηχ/των Δυνάμεων



Εικόνα 2. 8 Κατανομή Α/Φ Δυνάμεων

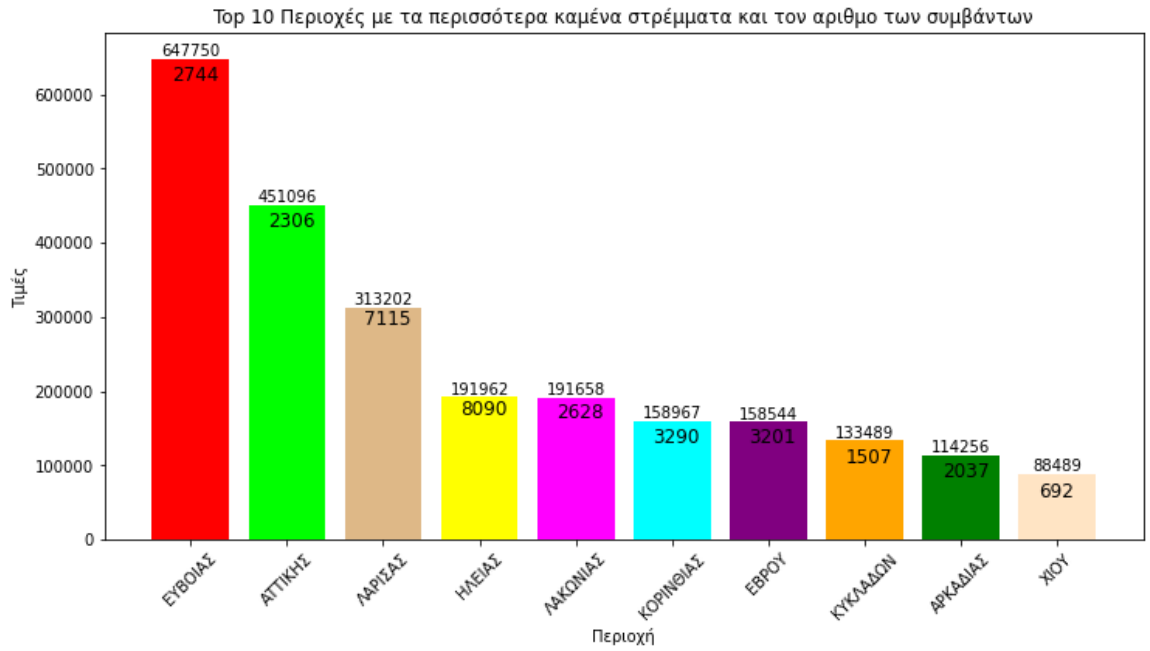
Εδώ βλέπουμε ότι περισσότερο χρησιμοποιείται το PZL και τα Ελικόπτερα καθώς τα συγκεκριμένα μέσα μπορούν να προσεγγίσουν δύσβατες περιοχές πιο εύκολα από τα Canadair.

2.5 Καμένη έκταση (σε στρέμματα)

Παρακάτω απεικονίζουμε διαγραμματικά την καμένη Έκταση

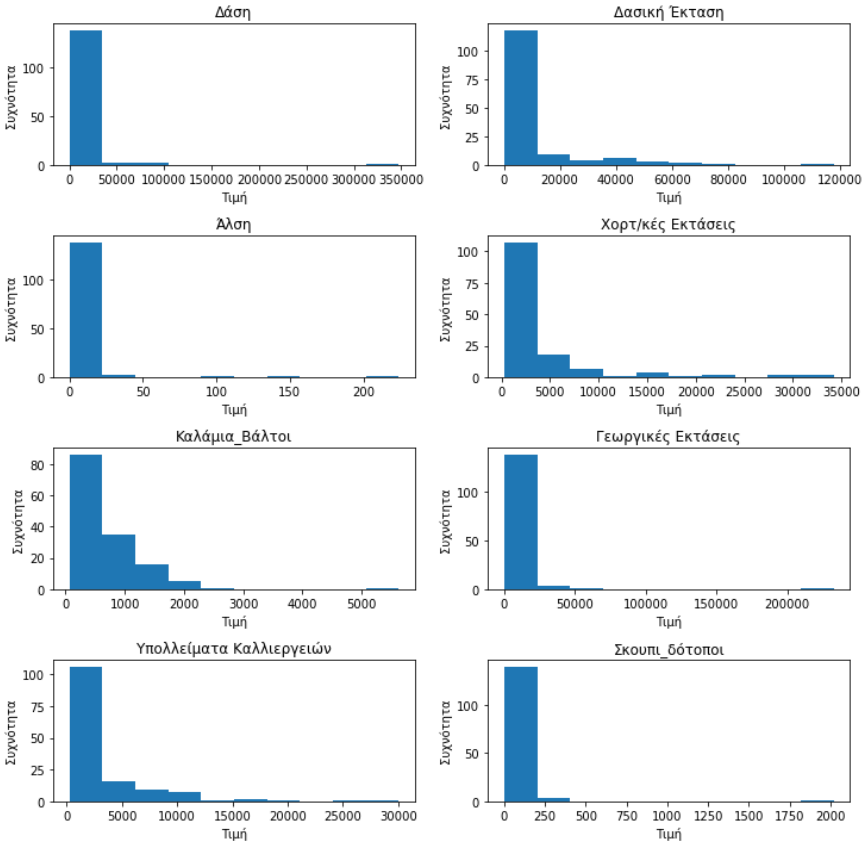
Εικόνα 2. 9 κατανομή καμένων εκτάσεων

Παρατηρούμε ότι Δάση και Δασικές έκτασης έχουν το μεγαλύτερο ποσοστό σε σχέση με τα υπόλοιπα είδη.



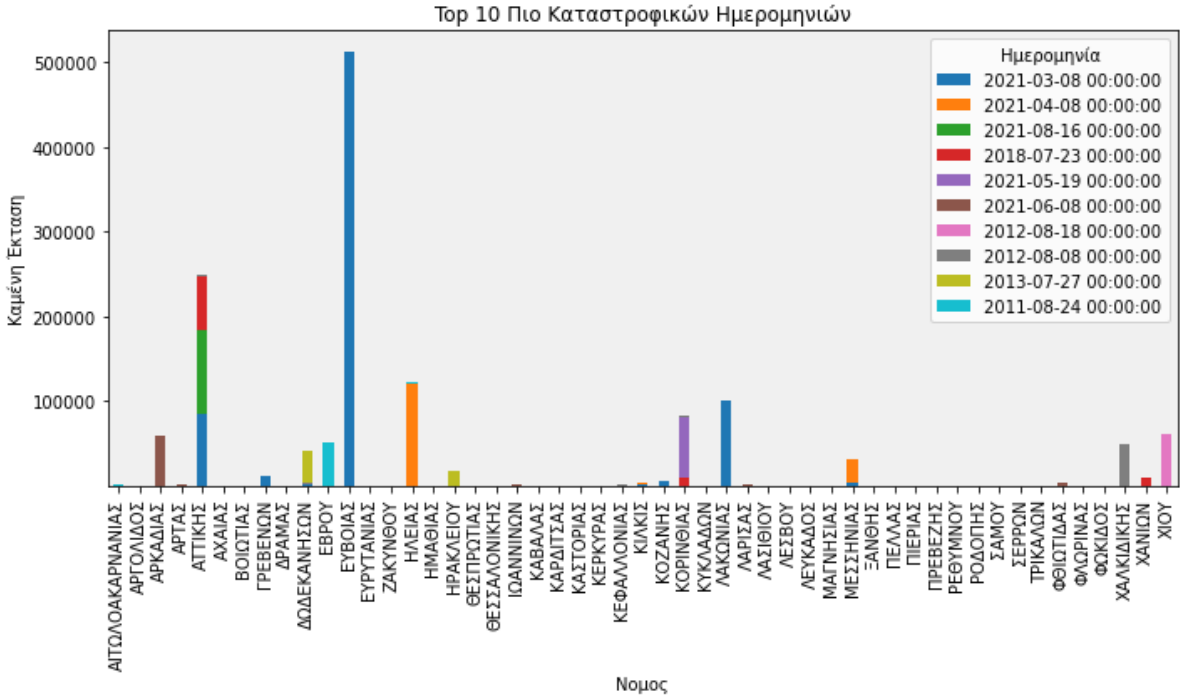
Εικόνα 2. 10 Top 10 Περιοχές με τα περισσότερα καμένα στρέμματα και τον αριθμό των συμβάντων.

Από τα παραπάνω προκύπτει ότι στην Εύβοια και την Αττική έχουμε τις πιο καταστροφικές πυρκαγιές (σε αριθμό στρεμμάτων) ενώ όπως φαίνεται ο αριθμός των πυρκαγιών είναι μικρότερος από των υπολοίπων. Αυτό μας δείχνει ότι στην Ηλεία μπορεί να έχουμε τις περισσότερες πυρκαγιές, αλλά αυτές καταστέλλονται αμέσως ενώ στην Εύβοια και στην Αττική είναι λιγότερες αλλά πιο καταστροφικές.
Εδώ παρουσιάζουμε μια εποπτική εικόνα της κατανομής των καμένων εκτάσεων.



Εικόνα 2. 11 Κατανομή των καμένων εκτάσεων

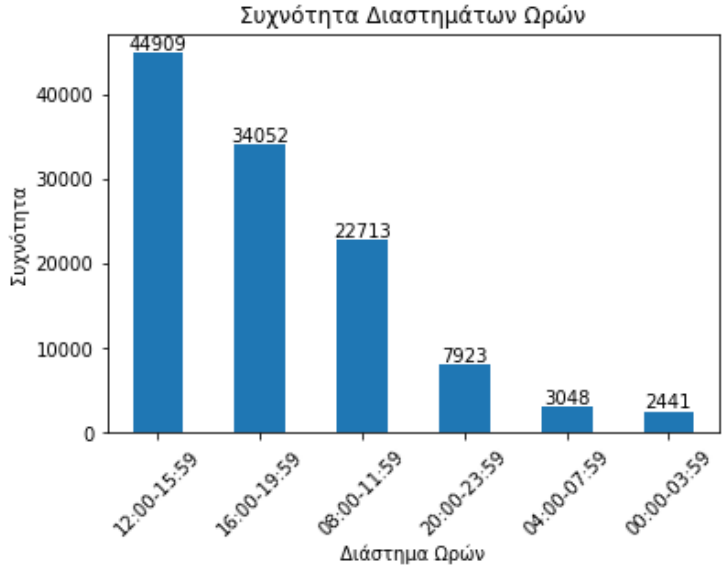
Τέλος παρουσιάζουμε τις 10 πιο καταστροφικές ημερομηνίες καμένης γης και σε ποιους νομούς έλαβαν χωρά.



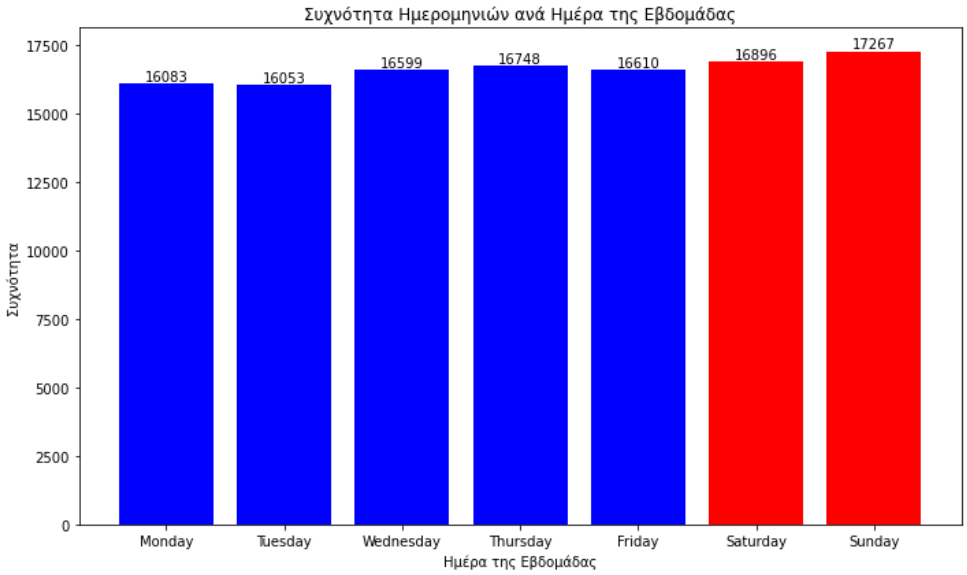
Εικόνα 2. 12 Τοπ 10 πιο Καταστροφικών Ημερομηνιών

2.6 Ώρες και Ημέρες Βδομάδας

Εδώ παρουσιάζεται η πιο κοινή ώρα που ξεκινούν οι πυρκαγιές. Η πλειοψηφία λαμβάνει χώρα από τις 12 μέχρι τις 4 ενώ τα ξημερώματα δεν ξεκίνησαν πολλές πυρκαγιές.



Εικόνα 2. 13 Συχνότητα Διαστημάτων Ωρών



Εικόνα 2. 14 Συχνότητα Ημερομηνιών ανά Ημέρα της Εβδομάδας

Στο παραπάνω σχήμα έχει γίνει μια ομαδοποίηση όπου με μπλε χρώμα παρουσιάζονται οι εργάσιμες μέρες και με κόκκινο η μη εργάσιμες ημέρες. Παρατηρούμε αυξημένες τιμές των συμβάντων στις μη εργάσιμες ημέρες. Επίσης διαπιστώνουμε ότι η Κυριακή είναι η πιο κοινή μέρα που ξεκινάει μια φωτιά ενώ η Τρίτη είναι η πιο ασυνήθιστη.

3 Weka και Τεχνικές Data Mining

3.1 Τι είναι το Weka

Το Weka είναι ένα μικρό πτηνό το οποίο βρίσκεται στην Νέα Ζηλανδία, δεν μπορεί να πετάξει και έχει το μέγεθος πάπιας. Στην περίπτωση μας όμως το Weka είναι μια εργαλειοθήκη Data Mining και σημαίνει Waikato Environment for Knowledge Analysis (WEKA), είναι γραμμένο σε Java και τρέχει σε όλα τα υπολογιστικά περιβάλλοντα. Ο αρθρωτός του σχεδιασμός δίνει την δυνατότητα να πραγματοποιηθούν ενέργειες όπως η οπτικοποίηση και η προ επεξεργασία των δεδομένων καθώς και χρήση αλγόριθμων μηχανικής μάθησης, όλα με ένα εργαλείο. Τέλος υποστηρίζει διάφορους τύπους αρχείων όπως.csv κ.α. αλλά για ανάλυση δεδομένων ο πιο κοινός τύπος είναι το .arff .

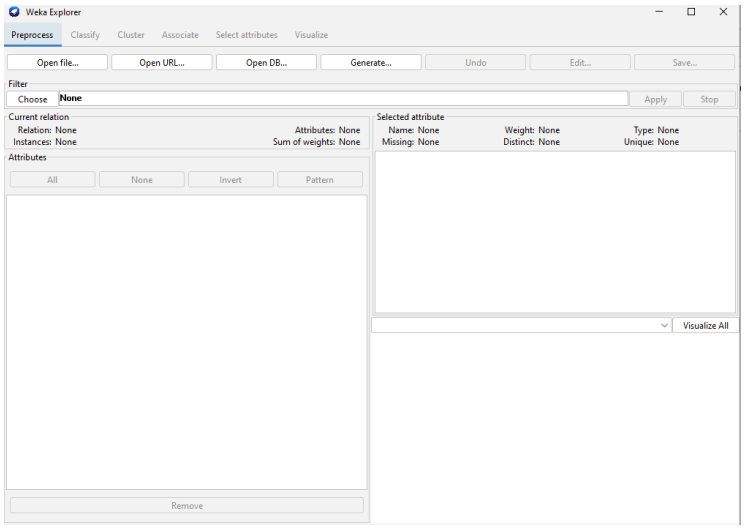
Εικόνα 3. 1 Weka Interface

Το weka έχει 4 interface, τα οποία είναι :

- Experimenter: Εδώ μπορούμε να διεξάγουμε πειράματα και στατιστικές δοκιμές μεγάλης κλίμακας σε διαφορετικά datasets με μεθόδους μηχανικής μάθησης.
- Knowledge Flow: Εδώ είναι ένα graphical interface το οποίο υποστηρίζει παρόμοιες λειτουργίες με τον Explorer αλλά με τηνδιαφορά ότι υπάρχει η

δυνατότητα drag and drop των πεδίων και αυτό είναι πολύ σημαντικό στην σταδιακή μάθηση.

- Simple CLI: Είναι ένα περιβάλλον το οποίο γράφεις της εντολές σε στυλ command-line.
- Explorer: Είναι το περιβάλλον που χρησιμοποιούμε για να επεξεργαστούμε τα δεδομένα. Το interface φαίνεται στην παρακάτω εικόνα



Εικόνα 3. 2 Weka Explorer

Στο επάνω μέρος, διακρίνουμε διάφορες καρτέλες στις οποίες βρίσκονται διάφοροι αλγόριθμοι μηχανικής μάθησης. Πιο συγκεκριμένα οι καρτέλες είναι:

- Προ επεξεργασία
- Ταξινόμηση
- Συστάδα
- Συσχέτιση
- Επιλογή χαρακτηριστικών
- Οπτικοποίηση

a) Προ επεξεργασία (Preprocess Tab)

Αρχικά μόνο αυτό το tab είναι διαθέσιμο. Η προ επεξεργασία των δεδομένων είναι το αρχικό βήμα της μηχανικής μάθησης και μέσω αυτού μπορούμε να εφαρμόσουμε

διάφορους αλγορίθμους πάνω στα δεδομένα που θα εισάγουμε για παράδειγμα υπάρχει το Discretize το οποίο μετατρέπει τα numeric πεδία σε nominal κ.α

b) Καρτέλα Ταξινόμηση (Classify Tab)

Η συγκεκριμένη καρτέλα περιλαμβάνει αλγόριθμους μηχανικής μάθησης οι οποίοι χρησιμοποιούνται στην ταξινόμηση των δεδομένων όπως Regressions (Linear,Logistic), Support Vector Machines, Decision Trees κ.α. Επίσης χωρίζονται σε εποπτευομένους και μη.

c) Καρτέλα συστάδων(Cluster Tab)

Στην καρτέλα περιλαμβάνει αλγόριθμους όπως οι SimpleKMeans, FilteredClusterer, HierarchicalClusterer, οι οποίοι ομαδοποιούν τα δεδομένα.

d) Καρτέλα Associate

Εδώ υπάρχουν αλγόριθμοι όπως οι Apriori, FilteredAssociator και FPGrowth οι οποίοι με το κατάλληλο διάστημα εμπιστοσύνης σου δημιουργούν κανόνες που είναι κρυμμένοι μέσα στο δείγμα.

e) Καρτέλα Select Attributes

Η καρτέλα Select Attributes σας επιτρέπει την επιλογή χαρακτηριστικών με βάση διάφορους αλγορίθμους, όπως ClassifierSubsetEval, PrinicipalComponents κ.λπ.

f) Καρτέλα Visualize

Τέλος η οπτικοποιηση των δεδομένων μας γίνεται στην καρτέλας Visualize.

Το Weka επεξεργάζεται πολλούς τύπους αρχείου όπως το csv, το json και το arff. Στην εργασία αυτή χρησιμοποιήσαμε το arff αρχείο και το csv.
Στην συνέχεια θα δούμε τις βασικές τεχνικές data mining που θα χρησιμοποιήσουμε στην εργασία μας.

3.2 Τι είναι το Data Mining

Η διαδικασία ανακάλυψης ενδιαφερόντων μοτίβων από διαφορετικές πήγες όπως βάσεις δεδομένων, excel κ.α. ονομάζεται εξόρυξη δεδομένων (Data Mining). Η συγκεκριμένη διαδικασία απαιτεί πολλούς γύρους ώστε τα αποτελέσματα να είναι ικανοποιητικά, για αυτό και θεωρείται μια επαναληπτική διαδικασία, συνδυάζει γνώσεις από πολλά

διαφορετικά πεδία όπως στατιστική, μηχανική μάθηση κ.α. και περιλαμβάνει τα παρακάτω βήματα:

- η προ επεξεργασία των δεδομένων δηλαδή ο καθαρισμός των δεδομένων και η μετατροπή τους από πρωτογενή δεδομένα σε επεξεργάσιμα.
- οι αλγόριθμοι εξόρυξης δεδομένων στους οποίους τροφοδοτούμε τα επεξεργασμένα δεδομένα και μέσω αυτών θα παραχθεί το μοτίβο.
- η ερμηνεία των αποτελεσμάτων από τους αλγορίθμους εξόρυξης.

Οι εργασίες εξόρυξης περιλαμβάνουν την εποπτευομένη μάθηση (classification), την μάθηση χωρίς επίβλεψη (clustering), την εξόρυξη κανόνων συσχετίσεις (association rule mining) και την εξόρυξή διαδοχικών προτύπων (sequential pattern mining) (Liu, 2011)
Οι πρωταρχικοί στόχοι της διαδικασίας της εξόρυξης δεδομένων είναι η πρόβλεψη και η περιγραφή.
Η πρόβλεψη αγνώστων η μελλοντικών τιμών η κλάσεων, μπορεί να επιτευχθεί με την χρήση ορισμένων μεταβλητών που μας ενδιαφέρουν και τις ορίσουμε εμείς και η περιγραφή, η οποία επικεντρώνεται στην περιγραφή των δεδομένων μέσω προτύπων που μπορούν να ερμηνευτούν από τον άνθρωπο. (Fayyad, Piatetsky-Shapiro, & Smyth, 1996)

3.3 Πρόβλεψη αποτελεσμάτων

a) Regression

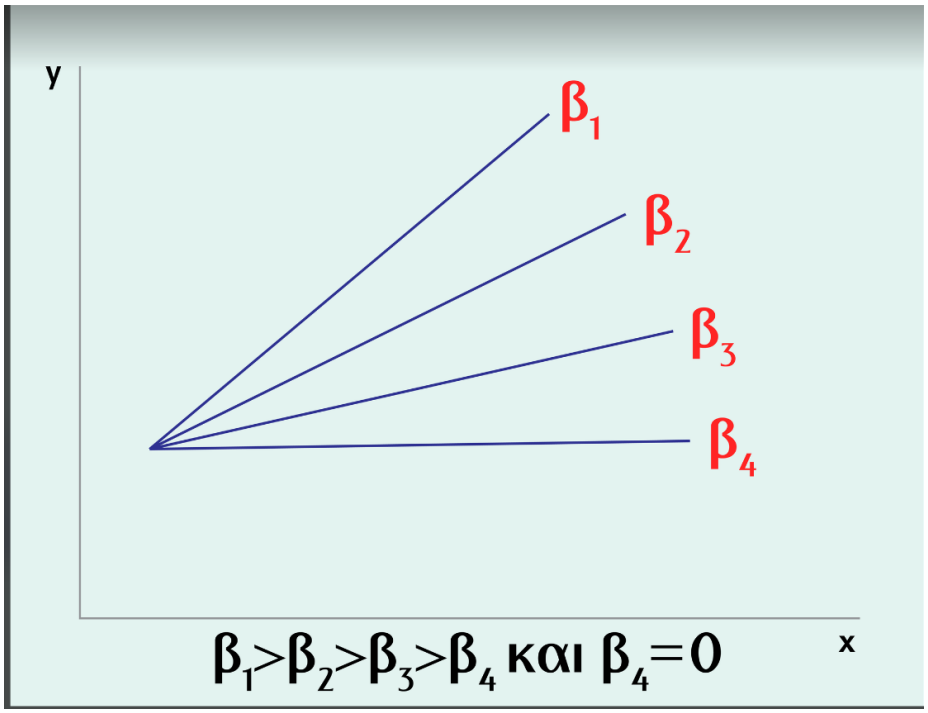
Η διαδικασία της γραμμικής παλινδρόμησης (Linear Regression) αποτελεί την εξέταση της σχέσης αναμεσά σε μια εξαρτημένη μεταβλητή και μιας η περισσότερων ανεξάρτητων μεταβλητών χρησιμοποιώντας κάποιον αλγόριθμο παλινδρόμησης. Η γενική ιδέα της διαδικασίας αυτής είναι να εξεταστούν τα παρακάτω:

- Ποσό καλά αποτελέσματα μπορώ να έχω στην πρόβλεψη μιας μεταβλητής (εξαρτημένης μεταβλητής) από ένα σύνολο μεταβλητών που έχω (ανεξάρτητες μεταβλητες)

- Από το παραπάνω σύνολο, ποιες από αυτές είναι σημαντικοί προγνωστικοί παράγοντες της εξαρτημένης μεταβλητής και με ποιον τρόπο την επηρεάζουν. (Guan, 2023)

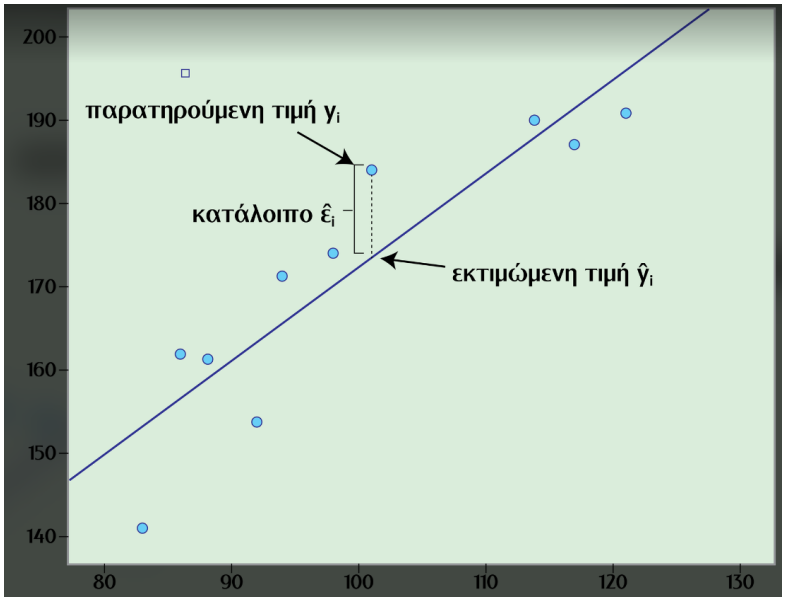
Η εξίσωση της παλινδρόμησης είναι της μορφής: Y= A+ BX+ E οπού:

- Y: η μεταβλητή που θέλω να προβλέψω
- B: η κλίση της γραμμής (slope) που σημαίνει ότι εάν αυξηθεί το X κατά μια μονάδα, το Y θα αυξηθεί κατά B.



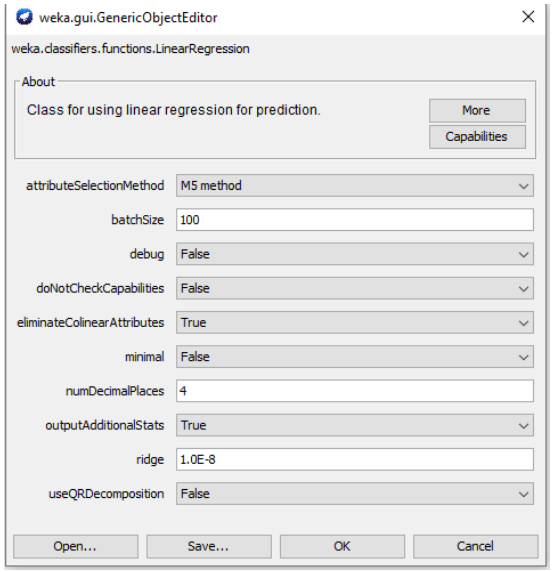
Εικόνα 3. 3 Παράδειγμα Εξίσωσης παλινδρόμησης

- A: Είναι η σταθερά (intercept/constant) και είναι το σημείο που τέμνει τον οριζόντιο άξονα η ευθεία παλίνδρομης
- E: Το κατάλοιπο (residual) οπού είναι η απόσταση της πραγματικής τιμής με την προβλεπόμενη.



Εικόνα 3. 4 Παράδειγμα μεταβλητών παλινδρόμησης

Στο εργαλείο Weka η παλινδρόμηση βρίσκεται στο classify ->functions -> Linear Regression και το μενού φαίνεται στην παρακάτω εικόνα:



Εικόνα 3. 5 Παλινδρόμηση Weka

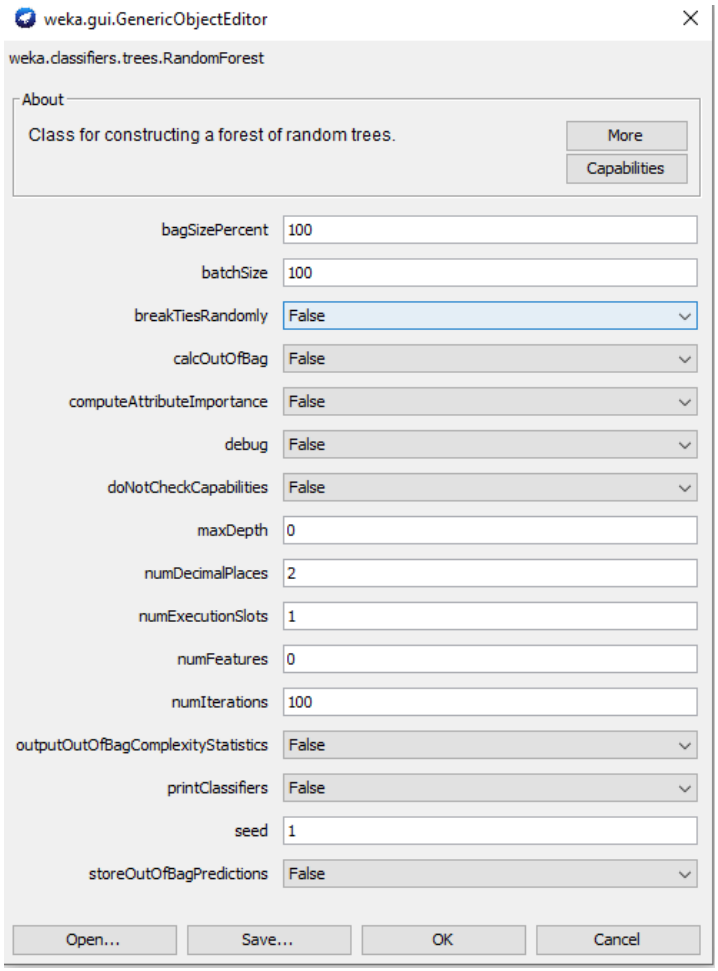
b) Random Forest

Η μέθοδος Random Forest χρησιμοποιείται ευρέως και σε πολλούς διαφορετικούς τομείς όπως το μάρκετινγκ, την ιατρική ασφάλιση κ.α. Είναι μια ολοκληρωμένη μέθοδος που

χρησιμοποιείται τόσο από προβλήματα ταξινόμησης όσο και παλινδρόμησης και ανήκει στην κατηγορία μεθόδων συνόλου (ensemble methods) με την βασική μονάδα της να ονομάζεται δέντρο απόφασης. Είναι μια απλή μέθοδος όπου ουσιαστικά εισάγουμε τα δεδομένα μας, δημιουργεί δέντρα απόφασης, παίρνει τις πληροφορίες από αυτά και μετά ενοποιεί τα αποτελέσματα και δημιουργεί ένα ενιαίο. Η τελική πρόβλεψη δημιουργείται από τον συνδυασμό των προβλέψεων που πραγματοποιούν τα μεμονωμένα δέντρα για κάθε δείγμα που επεξεργάζονται. Τα στάδια είναι τα εξής: (Chandel, Sarwat, Najah, Dhanagare, & Agarwala, 2022)

- Δημιουργία Δέντρων, σε αυτό το στάδιο κάθε δέντρο εκπαιδεύεται σε τυχαίο διαφορετικό υποσύνολο των δεδομένων που κάναμε εισαγωγή.
- Πρόβλεψη, κάθε δέντρο που δημιουργήθηκε στο παραπάνω βήμα κάνει την δική του πρόβλεψη και το τελικό αποτέλεσμα είναι ο μέσος ορός όλων των προβλέψεων των δέντρων. Στην περίπτωση της ταξινόμησης όμως, ο αλγόριθμος διαμορφώνει το τελικό αποτέλεσμα βάση την πλειοψηφία των προβλέψεων.
- Συνδυασμός Αποτελεσμάτων: Όλες οι προβλέψεις σε αυτό το στάδιο συνδυάζονται για να δημιουργηθεί το τελικό προβλεπτικό μοντέλο του Random Forest (Guan, 2023).

Στο weka το RF είναι στα trees Random Forest



Εικόνα 3. 6 Weka Random Forest

Όπου το κάθε πεδίο σημαίνει:

1. Seed: Ένας τυχαίος αριθμός seed που χρησιμοποιείται για την επιλογή των χαρακτηριστικών (attributes).
2. NumFolds : Εδώ κάνουμε εισαγωγή των δεδομένων που θα χρησιμοποιήσουμε για την διαδικασία backfitting (αφαιρεί από το τελικό μοντέλο όσα χαρακτηριστικά δεν είναι σημαντικά). Το μηδέν σημαίνει ότι δεν γίνεται backfitting
3. NumDecimalPlaces: Ο αριθμός των δεκαδικών ψηφίων που χρησιμοποιούνται για την εμφάνιση των αριθμών στο μοντέλο.
4. MinVarianceProp: Ορίζουμε το ελάχιστο ποσοστό της διακύμανσης σε όλα τα δεδομένα σε έναν κόμβο (Node) ώστε να γίνει ο διαχωρισμός (splitting)
5. minNum: Το ελάχιστο βάρος των περιπτώσεων σε ένα φύλλο (leaf)

6. maxDepth: Το μέγιστο ύψος ενός δέντρου, το μηδέν σημαίνει ότι δεν έχει όρια.
7. doNotCheckCapabilities : Εδώ εάν ενεργοποιηθεί, οι δυνατότητες του ταξινομητή δεν ελέγχονται πριν από την κατασκευή του (απαιτείται μεγάλη προσοχή στην χρήση του και μόνο για την μείωση του χρόνου εκτέλεσης).
8. debug: Εμφανίζει επιπλέον πληροφορίες εάν είναι true
9. breakTiesRandomly: τυχαίος διαχωρισμός όταν αρκετά χαρακτηριστικά (attributes) δείχνουν εξίσου καλά
10. batchSize: ορίζουμε το μέγεθος των περιπτώσεων που θα επεξεργαστούν εάν η πρόβλεψη (batch prediction) έχει καλή απόδοση.
11. allowUnclassifiedInstances: Εάν θα συμπεριλάβουμε και μη ταξινομημένες περιπτώσεις
12. KValue: Εισάγουμε τον αριθμό των τυχαία επιλεγμένων χαρακτηριστικών, στην περίπτωση του μηδέν χρησιμοποιείται το int(log_2(#predictors) + 1).

c) C4.5 algorithm/J48

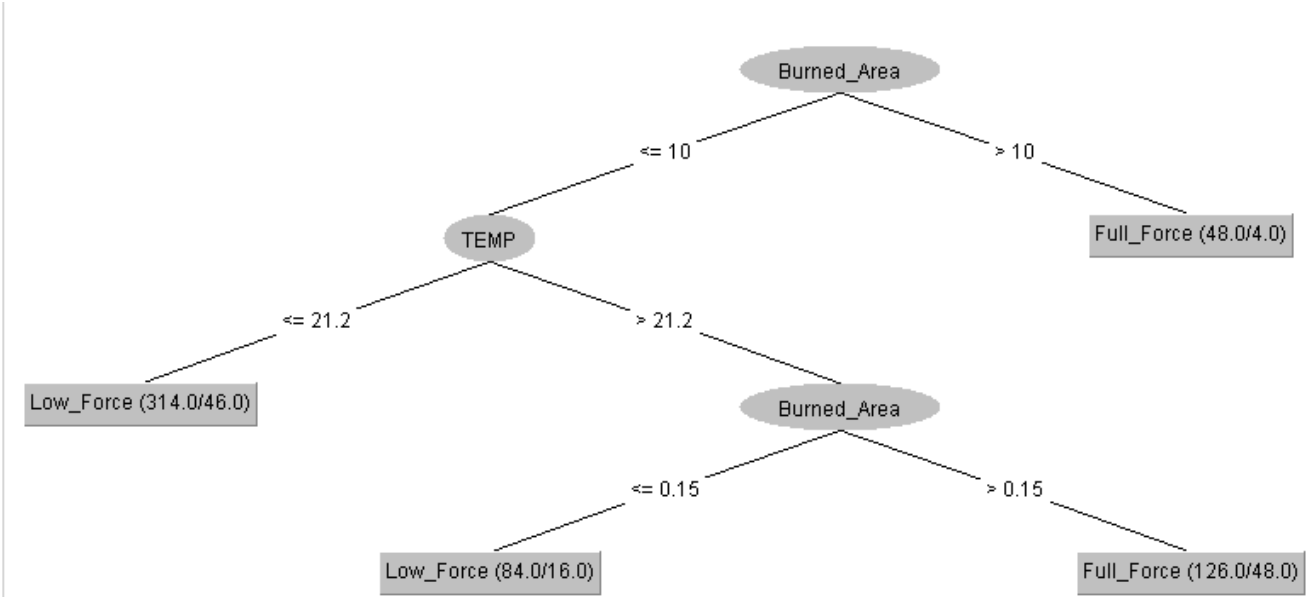
Ο αλγόριθμος C4.5 είναι μια επέκταση ενός αλλού αλγορίθμου, του ID3 του Ross Quinlan, και στο weka υπάρχει ως j48 (το J από την Java). Επιλεγεί μέσα από ένα κεντροποιημένο (standardized) σύνολο δεδομένων και διαχωρίζει την πληροφορία επιλέγοντας ένα χαρακτηριστικό (attribute). Στην συνέχεια μέσα από τα ομαδοποιημένα δεδομένα επιλέγει εκ νέου το καλύτερο χαρακτηριστικό και επιστρέφει τα μικρότερα υποσύνολα. Ο αλγόριθμος τελειώνει μόλις ένα υποσύνολο ανήκει στην ιδιά κλάση σε όλες τις περιπτώσεις. Το j48 έχει αρκετά επιπρόσθετα χαρακτηριστικά όπως η διαχείριση των ελλειπουσών τιμών, εξαγωγή κανόνων κ.α. και η ακρίβεια του μπορεί να αυξηθεί μέσω της περικοπής (pruning) κόστος των χαρακτηριστικών. Εδώ η ακρίβεια μπορεί να αναβαθμιστεί με το κλάδεμα (Venkatesan, 2015).

Ο αλγόριθμος:

- Στάδιο 1:Στην περίπτωση που οι τιμές ανήκουν στην ιδιά κλάση, το φύλλο (leaf) χαρακτηρίζεται (labeled) με την ιδιά κλάση.
- Στάδιο 2: Οι πιθανές πληροφορίες υπολογίζονται για κάθε χαρακτηριστικό και το κέρδος της πληροφορίας που προκύπτει από το τεστ, το προσθέτει στο χαρακτηριστικό.

- Στάδιο 3: Στο τελευταίο στάδιο, το καλύτερο χαρακτηριστικό θα επιλεγεί ανάλογα με την τρέχουσα παράμετρο επιλογής (Gayathri, 2018).

Εδώ παρουσιάζεται ένα δέντρο απόφασης που παράγει ο αλγόριθμος J48 το οποίο είναι της μορφής if - then. Πιο συγκεκριμένα ο κάθε κόμβος συμβολίζει ένα τεστ που ανάλογα το αποτελέσματα (αριθμός που φαίνεται πχ το >0,89) οδηγεί σε διαφορετικό φύλο (leaf) που φανερώνει το αποτέλεσμα. Ο αριθμός στα φύλα είναι οι παρατηρήσεις που ταξινομήθηκαν εκεί.



Εικόνα 3. 7 Weka Decision tree

Επίσης το weka μας παρέχει και κάποια στατιστικά ώστε να αξιολογήσουμε το μοντέλο μας.

```
Number of Leaves  :     4

Size of the tree :      7


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         453               79.1958 %
Incorrectly Classified Instances       119               20.8042 %
Kappa statistic                          0.5085
Mean absolute error                      0.3047
Root mean squared error                  0.3971
Relative absolute error                 69.7592 %
Root relative squared error             85.0171 %
Total Number of Instances              572

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,874     0,380     0,829     0,874     0,851     0,510     0,740     0,793     Low_Force
               0,620     0,126     0,699     0,620     0,657     0,510     0,740     0,646     Full_Force
Weighted Avg.  0,792     0,299     0,787     0,792     0,788     0,510     0,740     0,746

=== Confusion Matrix ===

   a   b   <-- classified as
 339  49 |   a = Low_Force
  70 114 |   b = Full_Force
```

Εικόνα 3. 8 Weka Output

Όπου το κάθε ένα σημαίνει:

1. Number of Leaves: ο αριθμός των φύλων

2. Size of the tree: Ο συνολικός αριθμός node και Leaf
3. Correctly Classified Instances: Το σύνολο των σωστά ταξινομημένων παρατηρήσεων
4. Incorrectly Classified Instances: Το σύνολο των λανθασμένων ταξινομημένων παρατηρήσεων
5. Kappa statistic: Ο συγκεκριμένος αριθμός μας αποκαλύπτει σε τι βαθμό συμφωνούν οι πραγματικές παρατηρήσεις με αυτή που υπολόγισε το μοντέλο. Οι τιμές που μπορείς να πάρει είναι -1 μέχρι 1 και όσο πιο κοντά στο 1 τόσο το καλύτερο.
6. Mean absolute error: Μετράει την απόσταση της πραγματικής τιμής με αυτή που υπολόγισε το μοντέλο, πχ εάν το μοντέλο δείξει 20 βαθμούς κελσίου και η πραγματική είναι 25, τότε το MAE είναι 5. Γενικά όσο πιο χαμηλό, τόσο το καλύτερο για το μοντέλο μας.
7. Root mean squared error (RMSE): Ο μαθηματικός τύπος του συγκεκριμένου δείκτη είναι math.sqrt((Σ(πραγματική τιμή - πρόβλεψη)^2/ τον συνολικό αριθμό παραδειγμάτων)) και όσο μικρότερο τόσο καλύτερη είναι η προσαρμογή του μοντέλου μας στα δεδομένα.
8. Relative absolute error (RAE): Ο μαθηματικός τύπος είναι το Σ(|πραγματική τιμή - πρόβλεψη του μοντέλου|)/Σ(|μέση τιμή (των πραγματικών τιμών)- πραγματική τιμή|) και εδώ όσο μικρότερο τόσο το καλύτερο.
9. Root relative squared error (RRSE): Υπολογίζει την τετραγωνική ρίζα τυπική απόκλιση των προβλέψεων από τις πραγματικές τιμές και διαιρεί με την τυπική απόκλιση των πραγματικών. Όπως και τα παραπάνω, όσο χαμηλότερο είναι τόσο πιο ακριβείς είναι οι προβλέψεις του μοντέλου μας.
10. Total Number of Instances: Ο συνολικός αριθμός των παραδειγμάτων.
11. TP Rate (True Positive Rate): Είναι το ποσοστό των πραγματικών θετικών παρατηρήσεων που προέβλεψε σωστά το μοντέλο μας σε σχέση με όλες τις θετικές τιμές. Όσο μεγαλύτερο τόσο το καλύτερο.
12. FP Rate (False Positive Rate): Είναι το ποσοστό των πραγματικών αρνητικών παρατηρήσεων που το μοντέλο μας προβλέπει ως θετικά σε σχέση με όλα τα αρνητικά παραδείγματα

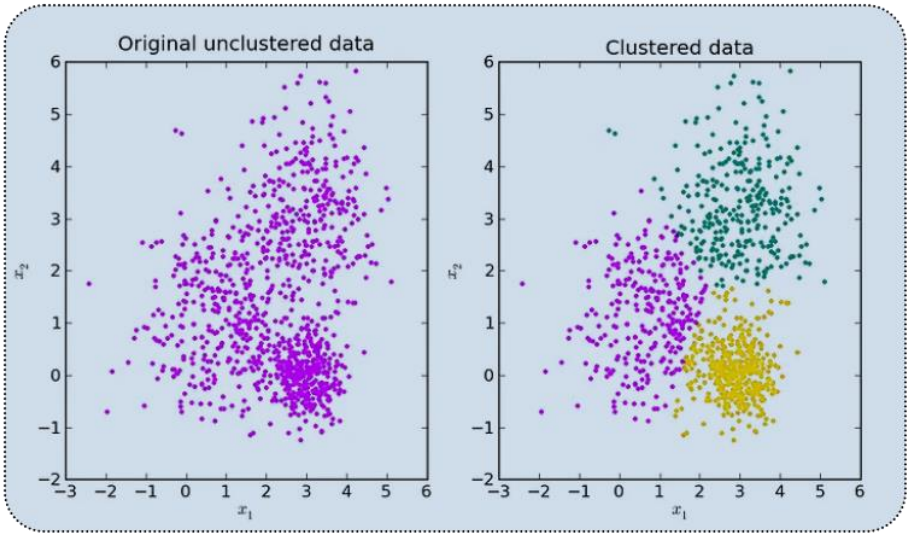
13. Precision: Την τιμή του συγκεκριμένου δείκτη επιθυμούμε να είναι υψηλή καθώς μας φανερώνει ότι για θετικές παρατηρήσεις, το μοντέλο μας προβλέπει σωστά και πιο συγκεκριμένα είναι το ποσοστό των πραγματικά θετικών τιμών με τις θετικές τιμές που προέβλεψε το μοντέλο μας.
14. Recall: Ίδιο με το True Positive Rate
15. F-Measure: Ο συνδυασμός του precision και του recall και μετράει την ισορροπία ανάμεσα στην ακρίβεια και την ανάκληση. Ιδανικά όσο πιο υψηλό γίνεται.
16. Mcc (Matthews Correlation Coefficient): Μετράει την συσχέτιση της πραγματικής με την προβλεπόμενη τιμή. Οι τιμές που λαμβάνει είναι 1 (τελειά συσχέτιση) και -1 (αντίστροφη συσχέτιση)
17. ROC Area: Αναπαριστά την σχέση μεταξύ του TP και του FT και το θέλουμε υψηλό.
18. PRC Area: Αναπαριστά την σχέση μεταξύ του Precision και Recall και αυτό όσο πιο υψηλό τόσο το καλύτερο.

3.4 Περιγραφή αποτελεσμάτων

a) Clustering

Η συσταδοποίηση (clustering) είναι η τεχνική με την οποία γίνεται κατανοητή η δομή των δεδομένων που θέλουμε να επεξεργαστούμε με βάση ένα μέτρο σύγκρισης π.χ. την ευκλείδεια απόσταση. Σε αυτή περιλαμβάνεται ο εντοπισμός συστάδων (cluster) οπού το χαρακτηριστικό τους είναι ότι τα στοιχεία της κάθε συστάδας, να είναι παρόμοια μεταξύ τους ενώ ταυτόχρονα να διαφέρουν αρκετά όταν ανήκουν σε διαφορετικές συστάδες. (Kusak, Unel, Alptekin, Celik, & Yakar, 2021). Ένας χαρακτηριστικός τομέας που χρησιμοποιείται είναι στην τμηματοποιηση της αγοράς, όπου παρόμοιοι πελάτες ομαδοποιούνται με βάση κάποια ιδιαίτερα χαρακτηριστικά κ.α. Τέλος κατηγοριοποιείται ως μέθοδος μάθησης χωρίς επίβλεψη καθώς προσπαθεί να ανακαλύψει μοτίβα, ομοιότητες και σχέσεις μεταξύ των δεδομένων χωρίς η σωστή απάντηση να είναι προκαθορισμένη από εμάς (ετικέτες βασικής αλήθειας).

Στο weka υπάρχουν αρκετοί αλγόριθμοι που πραγματοποιούν συσταδοποίηση όπως ο K-Mean Algorithm, ο COMWEB Algorithm, Canopy Algorithm και ο HierarchicalClusterer όπου αναλύονται παρακάτω συνοπτικά.



Εικόνα 3. 9 Παράδειγμα Clustered data vs Unclustered data

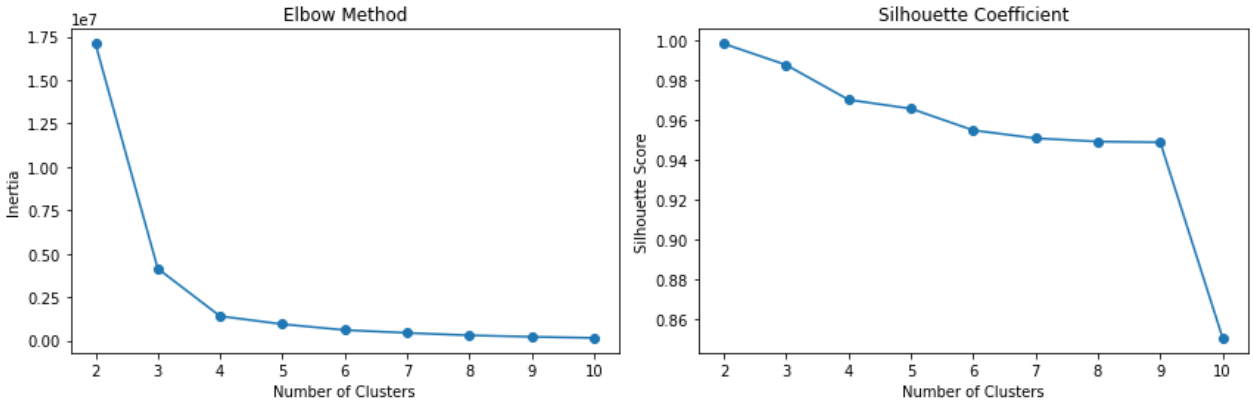
b) K-Mean Algorithm

Το 1967 ο MacQueen εισηγήθηκε τον Αλγόριθμό K-mean όπου ο σκοπός του ήταν η ταξινόμηση δεδομένων σε διαφορετικές κ συστάδες. Είναι μια απλή σχετικά μέθοδος επανάληψης για τον χωρισμό των δεδομένων σε ένα αριθμό συστάδων που τον παρέχει ο χρήστης. Επίσης είναι μη ιεραρχική μέθοδος και τα δεδομένα σε μια ομάδα έχουν ίδια χαρακτηριστικά, αλλά διαφέρουν σε μεγάλο βαθμό από τα στοιχεία των άλλων συστάδων καθώς ο διαχωρισμός γίνεται με βάση την ευκλείδεια απόσταση. Αυτό έχει ως αποτέλεσμα οι συστάδες να είναι πυκνές και ανεξάρτητες μεταξύ τους.
Πιο συγκεκριμένα θα δώσουμε ένα απλό παράδειγμα πως λειτουργεί ο αλγόριθμος αυτός.

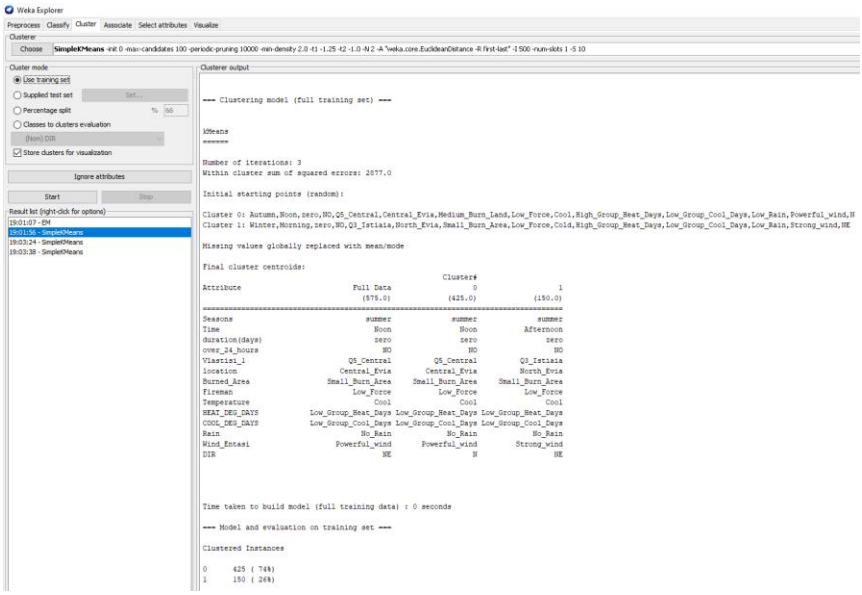
1. Επιλέγουμε τον αριθμό Cluster (K) που θέλουμε να δημιουργήσουμε.
2. Το κέντρο Cluster αρχικοποιείται τυχαία

3. Γίνεται ο υπολογισμός της απόστασης του κάθε σημείου από το κέντρο της συστάδας και βάση αυτής, γίνεται η ταξινόμηση στις διαφορετικές συστάδες (Khairani & Sutoyo, 2020)

Στην παρούσα εργασία για να βρούμε τον αριθμό των cluster χρησιμοποιήσαμε από python τον αλγόριθμο Sihouette για την αξιολόγηση των Cluster και τον αλγόριθμο Elbow για να βρούμε τον ιδανικό αριθμό. (Hidayati, Nalaratih, Shabrina, Wahyuni, & Latifah, 2020)



Εικόνα 3. 10 Παράδειγμα εύρεσης των αριθμών cluster και αξιολόγηση τους



Εικόνα 3. 11 Weka Simple k-means Output

c) Αλγόριθμος COBWEB

Το 1980 ο καθηγητής Douglas H. Fisher ο οποίος σήμερα εργάζεται στο Πανεπιστήμιο Vanderbilt, ανέπτυξε έναν αλγόριθμο μηχανικής μάθησης που είχε ως σκοπό την ομαδοποίηση αντικειμένων σε ένα μεγάλο σύνολο δεδομένων. Αυτός ο αλγόριθμος λέγεται COBWEB και παράγει ένα δενδρόγραμμα (dendrogram) που ονομάζεται δέντρο ταξινόμησης. Εδώ κάθε συστάδα διαθέτει μια πιθανολογική περιγραφή και η ποιότητα της διαμορφώνεται με βάση την χρησιμότητας της. Επιπλέον υπάρχει η δυνατότητα να προβλεφθούν τα χαρακτηριστικά που λείπουν η μια νέα κλάση χάρης στο δέντρο ταξινόμησης (Dafallah, Elhassan, & Ahamed, 2020).

Η λειτουργία του βασίζεται στον διαχωρισμό και την σύζευξη μιας συστάδας.

1. Ο αλγόριθμος ξεκινά με έναν κενό κόμβο ρίζας και οι τιμές προστίθενται μία προς μία
2. Μόλις δεχτεί μια τιμή, υπολογίζει την πιθανότητα να ανήκει σε ήδη δημιουργημένες κατηγορίες η να παραχθεί μια νέα
3. Βάση της πιθανότητας που υπολογίστηκε στο προηγούμενο βήμα κατηγοριοποιείτε σε υπάρχουσα κατηγορία η σε νέα
3. Σε αυτό το στάδιο, ανάλογα με το επίπεδο ομοιότητας των κατηγοριών που έχουν δημιουργηθεί, υπολογίζει την πιθανότητά για διαχωρισμό η συγχώνευση αυτών.
4.Η ίδια διαδικασία γίνεται σε κάθε νέο όρισμα και βελτιώνει τις κατηγορίες του. (Kanageswari & Pethalakshmi, 2017)

Στην παρακάτω εικόνα αναλύεται σε ψευδογλώσσα ο αλγόριθμος COBWEB (Panda & Patra, 2009)

COBWEB ALGORITHM

COBWEB (Node, Instance)

Begin

- If Node is a leaf then begin

Create two children of Node- L_1 and L_2;

Set the probabilities of L_1 to those of Node;

Set the probabilities of L_2 to those of Instance;

Add Instance to Node, updating Node's probabilities,

End.

- Else begin

Add Instance to Node, updating Node's probabilities; for each child C of Node, compute the category utility of clustering achieved by placing Instance in C;

Calculate:

S_1= the score for the best categorization (Instance is placed in C_1);

S_2=the score for the second best categorization (Instance is placed in C_2);

S_3=the score for placing Instance in a new category;

S_4=the score for merging C_1 and C_2 into one category;

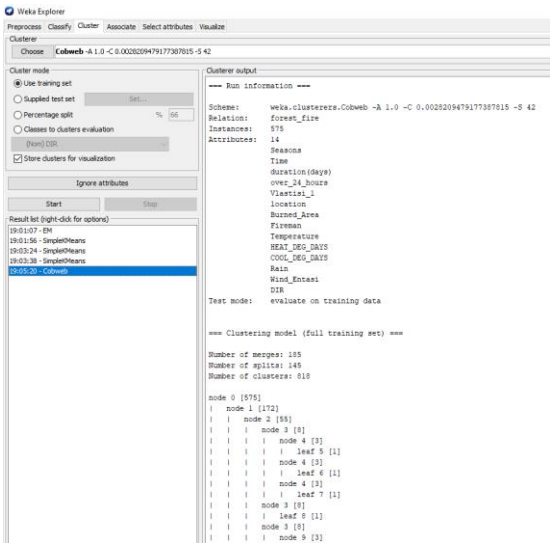
S_5= the score for splitting C_1 (replacing it with its child categories).

End

- If S_1 is the best score then call COBWEB (C_1, Instance).
- If S_3 is the best score then set the new category's probabilities to those of Instance.
- If S_4 is the best score then call COBWEB (C_m, Instance), where C_m is the result of merging C_1 and C_2.
- If S_5 is the best score then split C_1 and call COBWEB (Node, Instance).

end

Εικόνα 3. 12 Αλγόριθμος COBWEB σε ψευδογλώσσα



Εικόνα 3. 13 Weka COBWEB

d) Αλγόριθμος Canopy

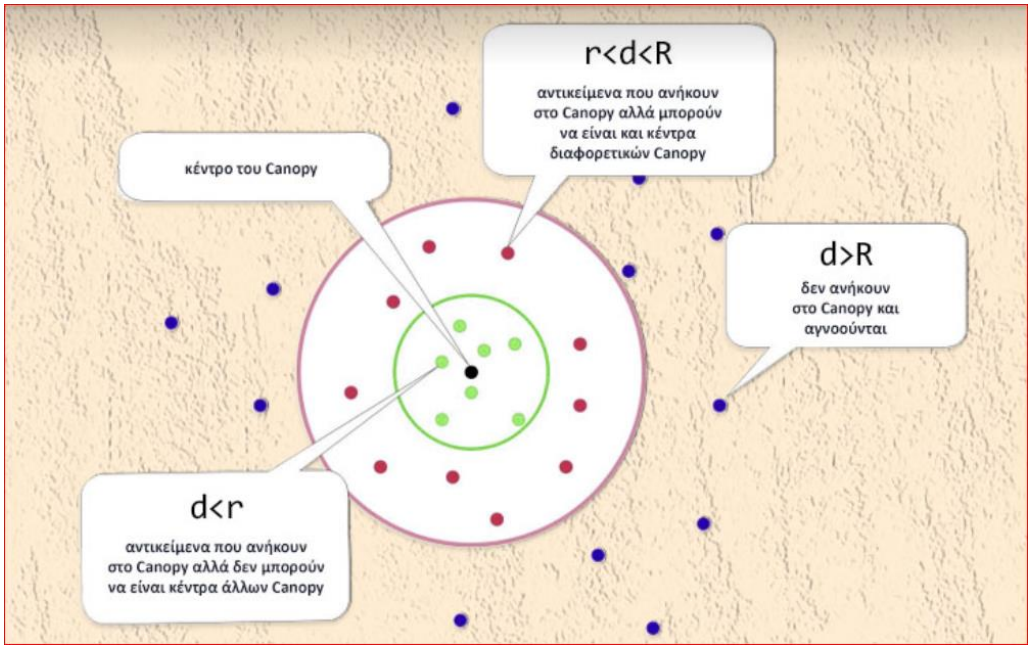
Οι Andrew McCallum, Kamal Nigam and Lyle Ungar το 2000 προτείναν έναν unsupervised αλγόριθμο που χρησιμοποιείται ως βήμα προετοιμασίας για τον K-mean Algorithm και τον Hierarchical Algorithm

Σε πολλές περιπτώσεις η άμεσή χρήση αλγόριθμο ομαδοποίησης δεν είναι δυνατή καθώς το μέγεθος των δεδομένων είναι πολύ μεγάλο. Η χρήση του συγκεκριμένου αλγόριθμου επιταχύνει τις εργασίες ομαδοποίησης και μειώνει το υπολογιστικό κόστος.

Ο αλγόριθμος χρησιμοποιεί δυο παραμέτρους, έστω r το μικρότερο κατώφλι και R το μεγαλύτερο κατώφλι οπού ισχύει η σχέση R>r. Τα βήματα είναι τα εξής:

1. Εισάγουμε τα δεδομένα που θέλουμε να ομαδοποιηθούν
2. Ορίζουμε τα r και R και υπολογίζουμε την απόσταση των δεδομένων από τα συγκεκριμένα σημεία
3. Εάν η απόσταση είναι μικρότερη του r τότε το αντικείμενο προστίθεται στο canopy με κέντρο r ως ένα σημείο κοντά στο κέντρο (3.14 εικόνα)
4. Εάν η απόσταση είναι αναμεσά στο r και R, τότε το αντικείμενο προστίθεται ως ένα μέλος του canopy
5. Εάν η απόσταση του σημείου είναι μεγαλύτερη από το R τότε διαγράφουμε το αντικείμενο

6. Όταν έχουν ομαδοποιηθεί όλα τα δεδομένα ενώνουμε τα canopy που έχουν κοινά μέλη η κέντρα και η διαδικασία σταματά μόλις δεν υπάρχουν άλλα canopy να προστεθούν (McCallum, Nigam, & Ungar, 2000).



Εικόνα 3. 14 Παράδειγμα Αλγορίθμου Canopy

Στο Weka ο συγκεκριμένος αλγόριθμος δεν απαιτεί από τον χρήστη να δώσει τις παράμερους r και R και χρησιμοποιεί αυτόματα προκαθορισμένες τιμές και το αποτέλεσμα φαίνεται από την παρακάτω εικόνα.

Εικόνα 3. 15 Weka Canopy

e) HierarchicalClusterer

Η ιεραρχική ομαδοποίηση λειτουργεί ομαδοποιώντας τα δεδομένα σε δέντρα συστάδων. Αρχικά αντιμετωπίζει ως ξεχωριστή συστάδα το κάθε σημείο από τα εισαγόμενα δεδομένα. Στην συνέχεια επαναλαμβάνει τα ακόλουθα βήματα:

1. Εντοπίζει τις 2 πιο κοντινές συστάδες
2. Συγχωνεύει τις 2 πιο κοινές συστάδες. Η συγκεκριμένη διαδικασία ολοκληρώνεται έως ότου συγχωνευτούν όλες οι συστάδες.

Ο στόχος εδώ είναι η δημιουργία μιας ιεραρχικής σειράς εμφωλευμένων συστάδων που ονομάζεται Δενδρόγραμμα (Dendrogram) το οποίο είναι ένα διάγραμμα που αναπαριστά γραφικά την στατιστικοποίηση των συγχωνευμένων συστάδων. Έχει την μορφή ανεστραμμένου δέντρου που μας δείχνει την σειρά με την οποία συγχωνεύονται (από κάτω προς τα πάνω) ή διασπώνται (από πάνω προς τα κάτω) οι συστάδες. Υπάρχουν δυο προσεγγίσεις για τον συγκεκριμένο αλγόριθμο: (Dafallah, Elhassan, & Ahamed, 2020)

1.Συσσωρευτική (Agglomerative)

Σε αυτή την προσέγγισή ο αλγόριθμος θεωρεί κάθε μεμονωμένο δεδομένα ως ξεχωριστή συστάδα και σε κάθε βήμα, συγχωνεύει τα πιο κοντινά (παρόμοια) ζευγάρια και δημιουργείτε μια μεγαλύτερη συστάδα (από κάτω προς τα πάνω μέθοδο)

Πιο συγκεκριμένα:

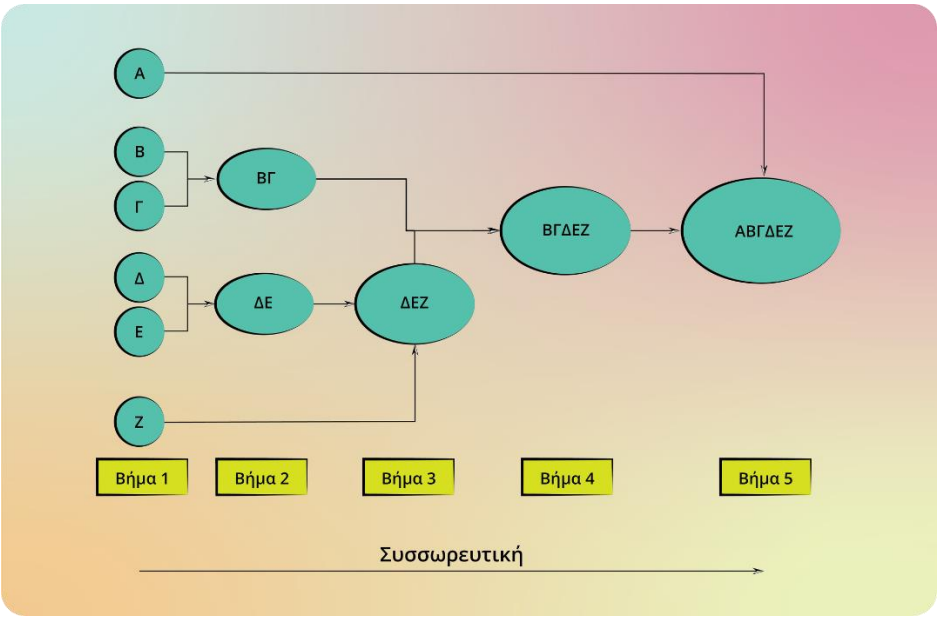
1. Υπολογίζει την ομοιότητα μιας συστάδας με τις υπόλοιπες. (proximity matrix)

2.Συγχωνευει τις πιο παρόμοιες.

3.Επαναυπολογιζει τον πίνακα εγγύτητας (proximity matrix) για κάθε συστάδα

4. Επανάληψη των βημάτων 2 και 3 μέχρι να μείνει μια τελική συστάδα.

Η παραπάνω διαδικασία φαίνεται από το σχήμα το οποίο και σχολιάζεται παρακάτω. (Kumar, χ.χ.)



Εικόνα 3. 16 Συσσωρευτική ιεραρχική ομαδοποίηση

Πιο αναλυτικά:

Βήμα 1: Έστω ότι κάθε γράμμα παριστάνει μια συστάδα και υπολογίζεται από τον αλγόριθμο η ομοιότητα της σε σχέση με τις υπόλοιπες.

Βήμα 2: Εδώ συγκρίνει τις συστάδες και συγχωνεύει τις πιο κοντινές. στο παράδειγμα μας η Β και η Γ είναι κοντά και δημιουργείτε η ΒΓ ενώ η Α δεν έχουν κάποια ομοιότητα και θα μείνει ανεξάρτητη για την ώρα.

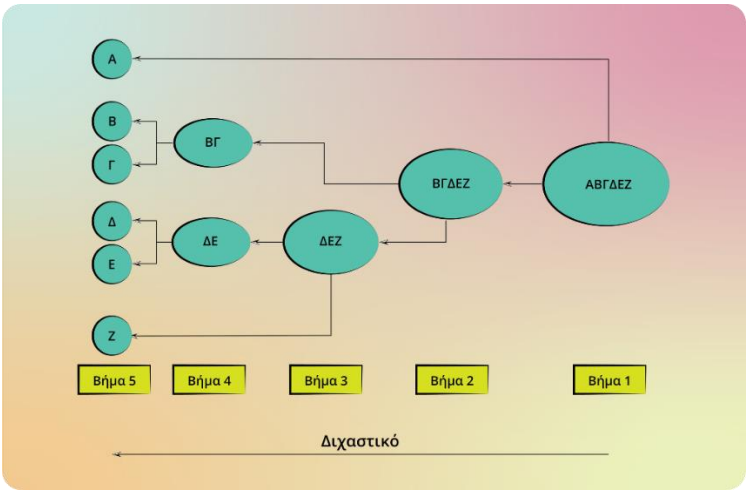
Βήμα 3: Υπολογίζει εκ νέου την ομοιότητα και συγχωνεύει τις κοινές συστάδες και δημιουργούνται η DEF. Οπότε έχουμε τις Α,ΒΓ,ΔΕΖ

Βήμα 4: τώρα συγκρίνει τις συστάδες ΒΓ ΚΑΙ ΔΕΖ και αφού είναι κοινές τότε τις ενώνει και έχουμε Α, ΒΓΔΕΖ

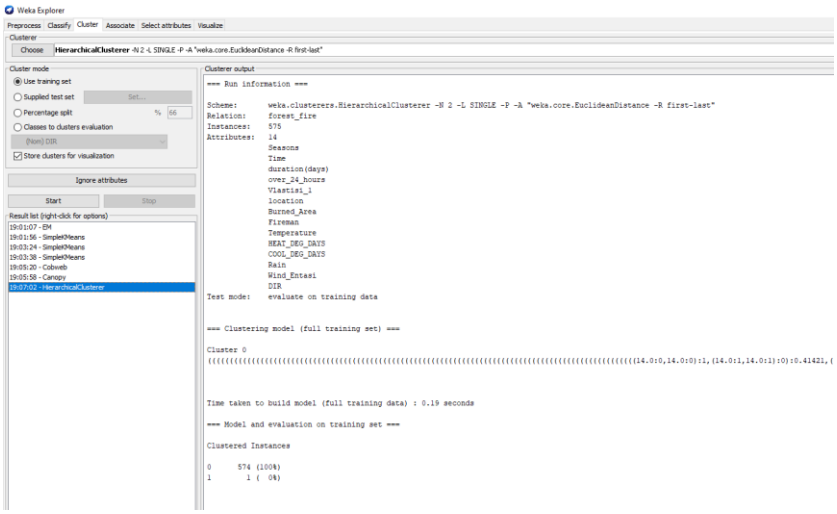
Βήμα 5: Τέλος, οι δύο εναπομείνασες συστάδες συγχωνεύονται μεταξύ τους για να σχηματίσουν μια ενιαία συστάδα ΑΒΓΔΕΖ.

2.Διαχωριστικό (Divisive)

Μια διαφορετική προσέγγιση είναι η διαχωριστική ιεραρχική ομαδοποίηση (Divisive) η οποία είναι το αντίθετο της συσσωρευτικής. Πιο συγκεκριμένα λαμβάνει όλα τα δεδομένα ως μια ενιαία συστάδα και σε κάθε επανάληψή διαχωρίζει τα δεδομένα σε κοινές συστάδες. Στο τέλος, απομένουν Ν συστάδες.



Εικόνα 3. 17 Διχαστική ιεραρχική ομαδοποίηση



Εικόνα 3. 18 Weka ιεραρχική ομαδοποίηση

Στον παρακάτω πίνακα αναφέρονται συνοπτικά τα θετικά και αρνητικά των αλγορίθμων ομαδοποίησης. (Ahmad & Dang, 2015)

Αλγόριθμοι	Πλεονέκτημα	Μειονεκτήματα
K-Means	Λειτουργεί καλά για μεγάλα σύνολα δεδομένων καθώς είναι απλός και γρήγορος.	Πρέπει να γνωρίζουμε τον αριθμό των κλάσεων καθώς τα αποτελέσματα του εξαρτώνται από αυτό
COBWEB	Χρησιμοποιείται για την πρόβλεψη των τιμών των υπολοίπων χαρακτηριστικών η την κλάση ενός νέου αντικειμένου καθώς οργανώνει τα δεδομένα σε ένα δένδρο κατηγοριών	Απαιτητικός για μεγάλα σύνολα δεδομένων, ειδικά όταν το δέντρο κατηγοριών γίνεται πολύ μεγάλο. Για μεγάλα σύνολα δεδομένων απαιτεί αρκετά μεγάλη υπολογιστική ισχύει και το δέντρο κατηγοριών γίνεται αρκετά μεγάλο

Canopy	Ο αλγόριθμος Canopy είναι πολύ απλός, γρήγορος και ακριβής για την ομαδοποίηση αντικειμένων σε Cluster	Δεν παρέχει την ακρίβεια και την ευελιξία άλλων πιο περίπλοκων αλγορίθμων ομαδοποίησης.
Ιεραρχικής Ομαδοποίησης	Οι ιεραρχικοί αλγόριθμοι έχουν την δυνατότητα να οργανώσουν τα δεδομένα σε ένα δένδρο κατηγοριών και παρέχουν πολύτιμες πληροφορίες για την συσχέτιση και την ιεραρχία των Cluster μεταξύ τους.	Είναι υπολογιστικά απαιτητικοί για μεγάλα σύνολα δεδομένων, και η επιλογή του επιθυμητού επιπέδου κοπής του δένδρου μπορεί να είναι δύσκολη. Είναι δύσκολη η επιλογή κοπής του δένδρου καθώς αυτό επηρεάζει την δημιουργία Cluster και την ποιότητα της ομαδοποίησης. Επίσης για μεγάλα δεδομένα είναι υπολογιστικά απαιτητικοί.

Πίνακας 3. 1 Σύγκριση περιγραφικών Αλγορίθμων

3.5 Association rules & Sequence Discovery

Για την εξόρυξη δεδομένων η χρησιμότητα των κανόνων συσχετίσεων (Association rules) και των διαδοχικών συσχετίσεων (Sequence Discovery) είναι αναμφισβήτητη. Κατά την διάρκεια των κανόνων συσχετίσεων ανακαλύπτονται σχέσεις συνύπαρξης (συσχετίσεις) αναμεσά στα δεδομένα με χαρακτηριστικότερο παράδειγμα, την ανάλυση δεδομένων του καλαθιού αγοράς και τον εντοπισμό προϊόντων που αγοράζονται μαζί για παράδειγμα Ξηροί Καρποί και Μπύρα. Αυτός ο κανόνας θα είχε την μορφή «Ξηροί Καρποί => Μπύρα [support = 10%, confidence = 80%]» υποδηλώνει ότι το 10% των πελατών αγοράζουν Ξηρούς Καρπούς και Μπύρα μαζί, και όσοι αγοράζουν Ξηρούς Καρπούς αγοράζουν και μπύρα στο 80% των περιπτώσεων.
Αντίθετά στην δημιουργία διαδοχικών προτύπων εξετάζεται η σειρά με την οποία αγοράζονται τα αντικείμενα. Για παράδειγμα το 10% των πελατών πρώτα αγοράζει τυρί, μετά γαλοπούλα και τέλος ψωμί. Τα συγκεκριμένα μοτίβα έχουν μεγάλη σημασία για την ανάλυση των clickstreams στα logs των σέρβερ κ.α (Liu, 2011)

a) Assosiacion rules

Η συγκεκριμένη μεθοδολογία αναπτύχθηκε από τους Aggarwal και Srikant το 1994 και μέσω αυτής, υπάρχει η δυνατότητα να ανακαλύπτουμε συσχετίσεις αναμεσά στα

δεδομένα. Η μορφή τους είναι απλή και της μορφής «εάν-τότε» που αυτό βοηθάει στην καλύτερη κατανόηση των σχέσεων μεταξύ των δεδομένων. Για παράδειγμα, ένας τέτοιος κανόνας είναι «εάν ένας πελάτης αγοράσει βάση για πίτσα με πιθανότητα 80% θα αγοράσεις και τυρί μοτσαρελα» (Kusak, Unel, Alptekin, Celik, & Yakar, 2021)
Η υποστήριξη και η εμπιστοσύνη είναι τα σημαντικότερα μετρά ισχύος ενός κανόνα. Πιο συγκεκριμένα η υποστήριξη (support) αποκαλύπτει την συχνότητα ενός κανόνα σε σχέση με τα συνολικά δεδομένα, ενώ η εμπιστοσύνη (confidence) από την άλλη πλευρά, μετρά όλες τις περιπτώσεις που περιέχουν στοιχεία A και B και το διαιρεί με την υποστήριξη του A.
Επίσης υπάρχει και το μέγεθος Lift το οποίο θα δούμε και στο weka και είναι ο λόγος της πιθανότητας L και R ως προς το γινόμενο της πιθανότητας L και της πιθανότητας R
lift = Pr(L, R) / Pr(L).Pr(R).
Σε αυτό το μέτρο, εάν το αποτέλεσμα είναι 1 τότε το L και R είναι ανεξάρτητα. Για το συγκεκριμένο μέτρο ισχύει ότι όσο ψηλότερη είναι η τιμή τόσο μεγαλύτερη είναι και η πιθανότητα να υπάρχει σχέση μεταξύ τους.
Σημαντικό μειονεκτικά στην περίπτωση των αλγορίθμων κανόνων συσχέτισης είναι η παραγωγή μεγάλου πλήθους κανόνων που δεν έχουν αξία για την ερευνά μας και ας έχουν πολύ μεγάλο βαθμό conference και support.
Παρακάτω δίνουμε ένα παράδειγμα τι είναι το support και το confidence.
Στην εργασία αυτή θα χρησιμοποιήσουμε τον πιο γνωστό αλγόριθμο για την εξόρυξη κανόνων τον Apriori στον οποίο ορίζουμε εμείς το κατώτατο και το ανώτερο όριο της υποστήριξης και της εμπιστοσύνης.
Ο αλγόριθμος Apriori ουσιαστικά αποτελείται από 2 βήματα, την ένωση (join step) οπού σε αυτό το βήμα δημιουργούμε (k+1) καλαθιά (itemsets), οπού κάθε στοιχείο ενώνεται με τον εαυτό του και 2 η περικοπή (Prune Step) οπού αρχικά γίνεται η σάρωση της συχνότητας κάθε στοιχείου στο σύνολο των δεδομένων και ανάλογα με το ελάχιστο support που του έχουμε θέσει αφαιρεί όσα δεδομένά δεν εμφανίζονται συχνά.
Για παράδειγμα έστω ότι έχουμε τις ακόλουθες περιπτώσεις και θέλουμε να βρούμε ποια προϊόντα αγοράζονται μαζί συχνά ώστε να τα τοποθετήσουμε κοντά στο ράφι.

Περίπτωση: {Αρτοποιείο, Γάλα, Καφές}

Περίπτωση: {Αρτοποιείο, Γάλα, Τυρί}
Περίπτωση: {Αρτοποιείο, Γάλα}
Περίπτωση: {Αρτοποιείο, Καφές}
Περίπτωση: {Αρτοποιείο, Γάλα, Καφές, Τυρί}

Αρχικά, κάθε αντικείμενο θεωρείται ως 1-καλαθι. Ο αλγόριθμος μετρά την συχνότητα που εμφανίζεται το κάθε αντικείμενο στην βάση δεδομένων μας

Αρτοποιείο: 5 φορές
Γάλα: 4 φορές
Καφές: 2 φορές
Τυρί: 2 φορές

Στην συνέχεια ορίζουμε το ελάχιστο κατώφλι υποστήριξης πχ min_sup=3 και βρίσκει τα 1-καλαθι που εμφανίζονται τουλάχιστον 3 φορές. Στο συγκεκριμένο παράδειγμα είναι τα παρακάτω:
Αρτοποιείο: 5 φορές
Γάλα: 4 φορές
Ο Καφές και το Τυρί δεν έχουν αρκετά συχνές εμφανίσεις και αφαιρούνται.
Στην συνέχεια δημιουργούμε 2-καλαθια συνδυάζοντας τα αντικείμενα με τον εαυτό τους δηλαδή:

2-καλαθια: { Αρτοποιείο, Γάλα },
{ Αρτοποιείο, Καφές },
{ Αρτοποιείο, Τυρί },
{ Γάλα, Καφές },
{ Γάλα, Τυρί },
{Καφές, Τυρί }

Στο επόμενο βήμα αφαιρούμε τα παραπάνω 2-καλαθια που δεν εμφανίζονται συχνότερα από 3 φορές στο αρχικό μας δείγμα και επιλέγουμε το {Αρτοποιείο, Γάλα}: 4 φορές
Στην συνέχεια δημιουργήσαμε τα παρακάτω καλάθια:

3-καλαθια: {Αρτοποιείο, Γάλα, Καφές},

{ Αρτοποιείο, Γάλα, Τυρί }

Εάν ελέγξουμε τα παραπάνω καλάθια με τις περιπτώσεις που έχουμε θα διαπιστώσουμε ότι το μόνο καλάθι που εμφανίζεται 2 φορές είναι το {Αρτοποιείο, Γάλα, Καφές} και εδώ τελειώνει ο αλγόριθμος καθώς δεν υπάρχουν καλάθια που να εμφανίζονται περισσότερες από 3 φορές (Min_Sup).Αρά τα πιο συχνά καλάθια είναι το { Αρτοποιείο, Γάλα } και το {Αρτοποιειο,Γαλα,Καφες}

b) Sequence Discovery

Ο εντοπισμός συχνά εμφανιζομένων ακολουθιών η υπό ακολουθιών σε ένα σύνολο δεδομένων ονομάζεται ανίχνευση ακολουθιών μοτίβων (Sequence Discovery).

Στην περίπτωση αυτή όμως δίνεται πολύ μεγάλη σημασία, στην σειρά η στον χρόνο που συμβαίνει το κάθε συμβάν, π.χ. οι χρήστες που πρώτα αγοράζουν μια τηλεόραση, στη συνέχεια θα αγοράζουν και μια παιχνιδομηχανή, εντός ενός μήνα.

Για παράδειγμα ένα δεδομένο το οποίο εμφανίζεται μόνο μία φορά σε ένα γεγονός μιας ακολουθίας, μπορεί να εμφανιστεί σε διαφορετικά γεγονότα της ίδιας ακολουθίας και αυτό το ανακαλύπτει η ανίχνευση ακολουθιών μοτίβων.

Αξίζει να σημειωθεί πως ενώ η συσχέτιση κανόνων (Association Rules) υποδεικνύει τις σχέσεις εντός της συναλλαγής, τα διαδοχικά μοτίβα αντιπροσωπεύουν τη συσχέτιση μεταξύ συναλλαγών

Επίσης ο χρήστης και εδώ έχει την δυνατότητα να ορίσει το κατώτατο κατώφλι υποστήριξης (min-sup), ανακαλύπτοντας όλες τι συχνές ακολουθίες με βάση το min_sup που έχει ορίσει. Τέλος έχουν εφαρμογή σε τομείς όπως το μάρκετινγκ, η τοποθέτηση και προώθηση προϊόντων, η πρόβλεψη καιρού, στο cybersecurity (η ανίχνευση εισβολών στο Web) κ.α. (Liu, 2011).

Ο πιο γνωστός αλγόριθμος είναι ο GSP (Generalized Sequential Pattern). Προτάθηκε από τους Jiawei Han, Jian Pei και Yiwen Yin το 2000 και η λειτουργία του είναι παρόμοια τον αλγόριθμο Apriori, με την διαφορά ότι η εύρεση όλων των συχνών itemsets δεν είναι υποχρεωτική. Πιο συγκεκριμένα δίνεται η δυνατότητα στον χρήστη να α) καθορίσει τα χρονικά όρια μεταξύ γειτονικών στοιχείων σε ένα μοτίβο, β) το σύνολο των στοιχείων που καλύπτουν ένα σύνολο συναλλαγών εντός ενός χρονικού παραθύρου γ) το πρότυπο

να ανακαλυφθεί σε διαφορετικό επίπεδο ταξινόμησης από αυτό που έχει αρχικά ορίσει ο χρήστης.

Ο αλγόριθμος GSP πραγματοποιεί πολλαπλά περάσματα μέσα στο σύνολο των δεδομένων που εξετάζουμε καθώς α) στο αρχικό πέρασμα αποκαλύπτονται οι συχνές ακολουθίες που έχουν την ελαχίστη υποστήριξη.

και β) στην συνέχεια με κάθε επανάληψή ελέγχει την κάθε ακολουθία δεδομένων και ανανεώνει την συχνότητα εμφάνισης του αριθμού των στοιχείων που περιέχονται σε αυτήν (Slimani & Lazzez, 2013).

Παρακάτω παραθέτουμε ένα απλό παράδειγμα ανίχνευσης μοτίβου (Sequence Discovery).

Έστω ότι έχουμε τα παρακάτω δεδομένα από ένα βιβλιοπωλείο και θέλουμε να βρούμε ποια αντικείμενο αγοράζονται μαζί ώστε να τα τοποθετήσουμε σε κοντινή απόσταση για να αυξήσουμε τις πωλήσεις μας.

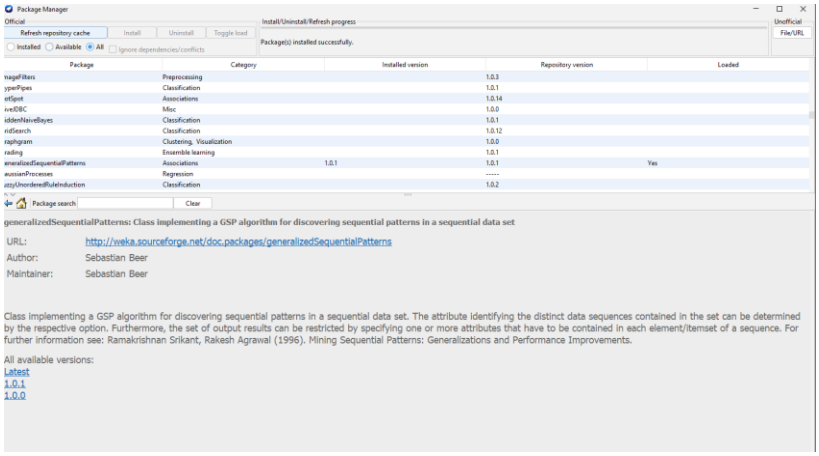
Χαρτί, Μολύβι, Στυλό

Χαρτί, Στυλό

Χαρτί, Μολύβι

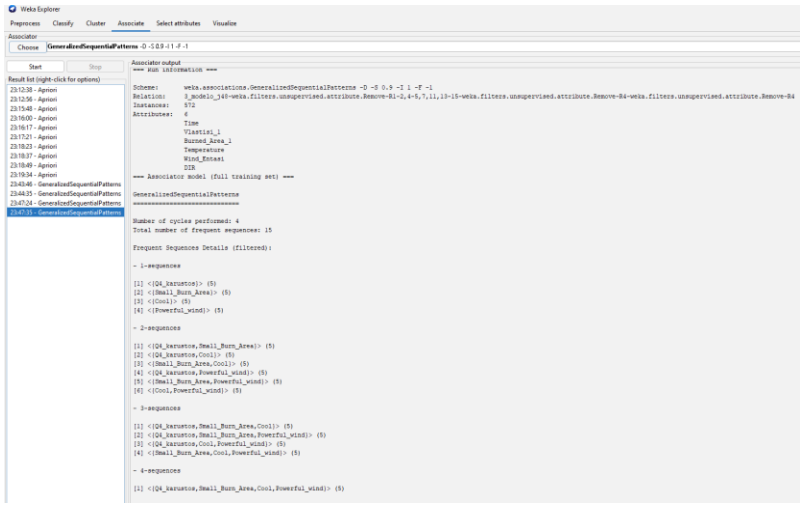
Χαρτί, Κόλλα

Εάν ορίσουμε επίπεδο συχνότητας 50% δηλαδή 2 από τις 4 επιλογές, βλέπουμε ότι το Χαρτί και το Στυλό θα αποτελέσουν το πρώτο επίπεδο προτύπων. Στην συνέχεια ο αλγόριθμος θα ελέγξει ξανά όλα τα δεδομένα και με βάση το Χαρτί και το Στυλό που βρήκε στο προηγούμενο βήμα θα προσπαθήσει να βρει ποια αλλά αντικείμενα αγοράζονται συχνά μαζί με αυτά. Το Weka δεν έχει στην προ επιλογή τον συγκεκριμένο αλγόριθμο και τον εγκαταστήσαμε από το Package Manager όπως δείχνει και η παρακάτω εικόνα.



Εικόνα 3. 19 Εγκατάσταση GSP στο Weka

Ένα παράδειγμα από την χρήση του αλγορίθμου στην συγκεκριμένη ερευνά



Εικόνα 3. 20 GSP στο Weka

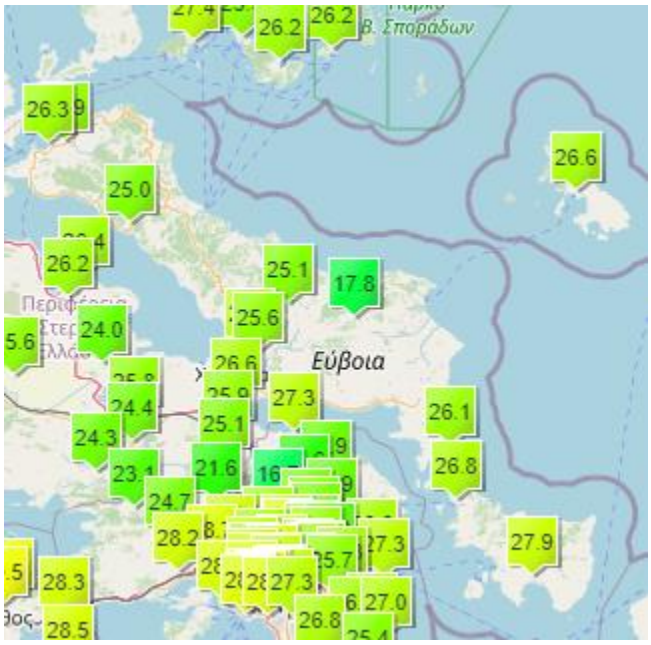
4 Στατιστική Ανάλυση Εύβοια

Το νησί της Ευβοίας κάθε χρόνο δοκιμάζεται από πυρκαγιές. Από το 2011 μέχρι το 2022 έχουν καεί περίπου 696374 στρέμματα δάσους. Στο κεφάλαιο αυτό θα προσπαθήσουμε μέσα από τα δεδομένα από τον σχετικό ιστοτόπο του Πυροσβεστικού Σώματος «https://www.fireservice.gr/el/synola-dedomenon») και τα μετεωρολογικά δεδομένα που αποκτήσαμε, να εφαρμόσουμε τεχνικές Data Mining και να εξάγουμε χρήσιμες πληροφορίες και κυρίως εάν έχει σημαντικό ρολό η κατεύθυνση του ανέμου σε μια πυρκαγιά. Αρχικά θα περιγράψουμε την διαδικασία συλλογής δεδομένών και στην συνέχεια θα πραγματοποιήσουμε Regression με τους αλγόριθμους Linear Regression και Random Forest και θα κατασκευάσουμε ένα δέντρο απόφασης με τον αλγόριθμο J48. Στην συνέχεια θα εφαρμόσουμε τον αλγόριθμο Apriori και GSP για τα δεδομένα που δημιουργήσαμε το δέντρο απόφασης. Επιπρόσθετα θα πραγματοποιήσουμε και μια ανάλυση με τον αλγόριθμο k-mean ώστε να ομαδοποιήσουμε τα δεδομένα και τέλος για κάθε ξεχωριστό cluster θα εφαρμόσουμε τον αλγόριθμο apriori rule ώστε να ανακαλύψουμε pattern αναμεσά στις τιμές των ξεχωριστών ομάδων.

4.1 Συλλογή δεδομένων

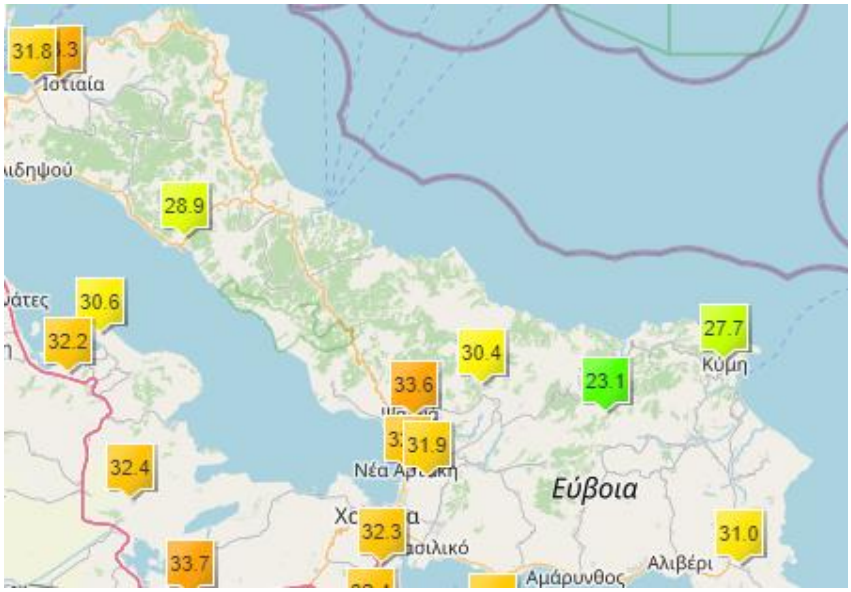
Αρχικά από το 2011 μέχρι και το 2022 έχουμε 2744 συμβάντα πυρκαγιών. Προχωρήσαμε όμως σε αφαίρεση 142 γραμμών καθώς δεν είχαν συμπληρωμένη την έκταση καμένης γης. Επίσης αφαιρέθηκαν και 12 πυρκαγιές καθώς δεν είχαν συμπληρωμένα τα στοιχεία των πυροσβεστικών δυνάμεων που συμμετείχαν στην κατάσβεση (34 στρέμματα καμένης γης). Συνεπώς το δείγμα μας έχει 2590 συμβάντα πυρκαγιών για την Εύβοια. Στην συνέχεια για κάθε ημερομηνία προσπαθήσαμε να συλλέξουμε τα περιβαλλοντικά δεδομένα που επικρατούσαν στην περιοχή μόλις ξεκίνησε η φωτιά. Προχωρήσαμε σε επικοινωνία με την Εθνική Μετεωρολογική Υπηρεσία ώστε να μας προωθήσουν τα μετεωρολογικά δεδομένα για την περίοδο 2011-2022 αλλά αυτό δεν κατέστη δυνατό και τα κατεβάσαμε από το site https://meteosearch.meteo.gr/data/index.cfm μόνο για την περίοδο 2020-2022. Για τα προηγούμενα χρόνια τα δεδομένα είτε δεν υπήρχαν είτε ήταν ελλιπέστατα και για αυτό περιορίσαμε την ερευνά μας στην συγκεκριμένη περίοδο.

Η Εύβοια σε αντίθεση με άλλες περιοχές όπως η Ζάκυνθος έχει αρκετούς διαφορετικούς σταθμούς. Πιο συγκεκριμένα όπως δείχνει και η παρακάτω εικόνα υπάρχουν 15 διαφορετικοί σταθμοί (Λίμνη Ευβοίας, Ζαρακες, Ιστιαία, Κάρυστος, Κριεζά, Κύμη, Μακρυκαπα, Νέα Αρτάκη, Παξιμάδα, Σέττα, Στενή, Στύρα, Χαλκίδα, Ψαχνά και Ωρεοι.)



Εικόνα 4. 1 Μετεωρολογικοί σταθμοί Ευβοίας

Όσα περιβαλλοντικά δεδομένα δεν υπήρχαν, τα συμπληρώσαμε με δεδομένα από κοντινούς μετεωρολογικής σταθμούς όπως για παράδειγμα το αρχείο για τους μήνες Σεπτέμβριο, Οκτώβριο και Νοέμβριο του μετεωρολογικό σταθμού της Κύμης ήταν κενό, το αντικαταστήσαμε με το αρχείο από τον σταθμό της Σεττας που είναι ο κοντινότερος όπως δείχνει και η παρακάτω εικόνα.



Εικόνα 4. 2 Παράδειγμα γειτονικού μετεωρολογικού σταθμού

Από το αρχικό Excel και με βάση την στήλη Δήμος κατηγοριοποιήσαμε της περιοχές που εμφανίστηκαν πυρκαγιές σε 6 μεγάλες κατηγορίες και από αυτές τις περιοχές, συλλέξαμε τα μετεωρολογικά δεδομένα και πιο συγκεκριμένα είναι οι Ιστιαία, Κάρυστος, Κύμη, Χαλκίδα, Ψαχνά και Ωρεοι. Όσες τιμές ήταν κενές βρήκαμε την περιοχή είτε από την διεύθυνση της περιοχής που μπορεί να ήταν συμπληρωμένη είτε από το διαδίκτυο για την συγκεκριμένη ημερομηνία στην Εύβοια και για αυτό το excel μας δεν έχει κάποιο missing value ως προς τα συγκεκριμένα πεδία.

Υπηρεσία	Νομός	Ημερ/νία Έναρξης	Ώρα Έναρξης	Ημερ/νία Κατασβεσης	Ώρα Κατάσβεσης	Δασαρχείο	Δήμος	Περιοχή	Διεύθυνση
Π.Υ. ΧΑΛΚΙΔΑΣ	ΕΥΒΟΙΑΣ	3/1/2022	15:56	3/1/2022	18:35		Δ. ΔΙΡΦΥΩΝ - ΜΕΣΣΑΠΙΩΝ	ΜΙΣΤΡΟΣ	"ΚΟΡΟΜΗΛΙΑ", ΤΚ ΜΙΣΤΡΟΥ
Π.Υ. ΧΑΛΚΙΔΑΣ	ΕΥΒΟΙΑΣ	12/2/2022	15:20	12/2/2022	16:10		Δ. ΔΙΡΦΥΩΝ - ΜΕΣΣΑΠΙΩΝ	ΠΟΛΙΤΙΚΑ	"ΚΑΚΑΟΠΕΡΑΤΟ" ΠΟΛΙΤΙΚΑ
Π.Υ. ΧΑΛΚΙΔΑΣ	ΕΥΒΟΙΑΣ	21/2/2022	15:40	21/2/2022	21:01		Δ. ΔΙΡΦΥΩΝ - ΜΕΣΣΑΠΙΩΝ	ΚΑΜΠΙΑ	"ΚΟΥΚΟΣ" ΚΑΜΠΙΩΝ
Π.Υ. ΧΑΛΚΙΔΑΣ	ΕΥΒΟΙΑΣ	28/3/2022	16:30	28/3/2022	19:33		Δ. ΔΙΡΦΥΩΝ - ΜΕΣΣΑΠΙΩΝ	ΨΑΧΝΑ	ΛΙΒΑΔΙ
Π.Υ. ΧΑΛΚΙΔΑΣ	ΕΥΒΟΙΑΣ	31/3/2022	14:24	31/3/2022	16:19		Δ. ΔΙΡΦΥΩΝ - ΜΕΣΣΑΠΙΩΝ		"ΣΠΗΛΙΑ", ΤΚ ΘΕΟΛΟΓΟΥ
Π.Υ. ΧΑΛΚΙΔΑΣ	ΕΥΒΟΙΑΣ	6/4/2022	16:56	6/4/2022	17:53		Δ. ΔΙΡΦΥΩΝ - ΜΕΣΣΑΠΙΩΝ		ΠΑΛΙΟΥΡΑΣ
Π.Υ. ΧΑΛΚΙΔΑΣ	ΕΥΒΟΙΑΣ	15/4/2022	10:42				Δ. ΧΑΛΚΙΔΕΩΝ	ΛΟΥΚΙΣΙΑ	ΤΚ ΛΟΥΚΙΣΙΩΝ
Π.Υ. ΧΑΛΚΙΔΑΣ	ΕΥΒΟΙΑΣ	4/5/2022	15:45	4/5/2022			Δ. ΧΑΛΚΙΔΕΩΝ	ΒΑΘΥ	ΜΟΝΟΠΑΤΙ-ΑΓ. ΡΑΦΑΗΛ
Π.Υ. ΧΑΛΚΙΔΑΣ	ΕΥΒΟΙΑΣ	13/5/2022	10:30	13/5/2022	11:56		Δ. ΧΑΛΚΙΔΕΩΝ	ΧΑΛΚΙΣ	ΚΑΣΤΡΟ ΚΑΡΑΜΠΑΜΠΑ, ΔΚ ΧΑΛΚΙΔΑΣ
Π.Υ. ΧΑΛΚΙΔΑΣ	ΕΥΒΟΙΑΣ	17/5/2022	21:05	17/5/2022	21:48		Δ. ΔΙΡΦΥΩΝ - ΜΕΣΣΑΠΙΩΝ	ΨΑΧΝΑ	"ΛΙΒΑΔΙ", ΔΚ ΨΑΧΝΩΝ
Π.Υ. ΧΑΛΚΙΔΑΣ	ΕΥΒΟΙΑΣ	18/5/2022	20:24	18/5/2022	21:04		Δ. ΔΙΡΦΥΩΝ - ΜΕΣΣΑΠΙΩΝ		ΤΚ ΠΙΣΣΩΝΑ
Π.Υ. ΧΑΛΚΙΔΑΣ	ΕΥΒΟΙΑΣ	19/5/2022	16:33	19/5/2022	23:33		Δ. ΔΙΡΦΥΩΝ - ΜΕΣΣΑΠΙΩΝ	ΑΜΦΙΘΕΑ	ΤΚ ΑΜΦΙΘΕΑΣ
Π.Υ. ΧΑΛΚΙΔΑΣ	ΕΥΒΟΙΑΣ	20/5/2022	17:23	20/5/2022	19:47		Δ. ΔΙΡΦΥΩΝ - ΜΕΣΣΑΠΙΩΝ	ΝΕΡΟΤΡΙΒΙΑ	ΑΓ. ΜΗΝΑΣ- ΝΕΡΟΤΡΙΒΙΑΣ

Εικόνα 4. 3 Παράδειγμα κατηγοριοποίησης Περιοχών Εύβοια

Τα μετεωρολογικά δεδομένα για την Εύβοια, από την μορφή txt που κατέβηκαν από το site του Meteo, τα ενοποιήσαμε και δημιουργήσαμε το παρακάτω Excel το οποίο περιέχει

31 στήλες όπως δείχνει και η παρακάτω εικόνα και πλέον για κάθε ημερομηνία έναρξης έχουμε την αντίστοιχη θερμοκρασία, ταχύτητα του ανέμου κ.α

MONTHLY CLIMATOLOGICAL SUMMARY for SEP. 2022

NAME: chalkida CITY: STATE:
ELEV: 0 m LAT: LONG:

TEMPERATURE (°C), RAIN (mm), WIND SPEED (km/hr)

DAY	MEAN TEMP	HIGH	TIME	LOW	TIME	HEAT DEG DAYS	COOL DEG DAYS	RAIN	AVG WIND SPEED	HIGH	TIME	DOM DIR
1	27.3	33.2	17:00	23.1	6:30	0.0	8.9	0.0	6.8	25.7	17:30	W
2	26.8	32.9	13:30	22.8	7:50	0.0	8.5	0.0	6.0	30.6	18:00	E
3	26.3	31.8	16:30	22.7	5:10	0.0	7.9	0.0	6.0	25.7	17:20	N
4	26.4	32.2	15:40	23.1	6:50	0.0	8.1	0.0	7.6	29.0	12:40	WSW
5	22.5	24.6	15:50	18.8	8:20	0.0	4.2	9.0	11.4	54.7	11:40	NE
6	22.1	25.6	14:00	19.7	7:20	0.0	3.8	0.0	12.4	53.1	15:40	NE
7	22.2	26.2	16:10	19.8	0:50	0.0	3.9	0.0	11.7	53.1	13:10	NW
8	23.7	28.5	15:50	20.0	2:40	0.0	5.4	0.0	12.1	46.7	10:20	NNW
9	25.9	33.1	15:50	20.3	5:30	0.0	7.6	0.0	5.6	20.9	20:40	SSW
10	29.0	36.6	15:00	23.5	6:20	0.0	10.7	0.0	4.2	17.7	15:00	ESE
11	28.6	34.7	16:00	24.9	7:00	0.0	10.2	0.0	9.7	37.0	20:00	WNW
12	24.4	27.7	13:30	21.3	23:40	0.0	6.1	0.0	15.9	54.7	12:50	NW
13	23.2	27.2	15:20	20.8	3:50	0.0	4.8	0.0	12.4	46.7	14:10	NW

Εικόνα 4. 4 Αρχείο μετεωρολογικών δεδομένων Meteo

Ημερ/νία Έναρξης	Ώρα Έναρξης	Ημερ/νία Κατάσβεσης	Ώρα Κατάσβεσης	Δασαρχείο	Δήμος	meteo_stathmoi	Περιοχή	Διεύθυνση	Δάση	Δασική Έκταση	Άλση	Χορτ/κές Εκτάσεις	Καλάμια - Βάλτοι	Γεωργικές Εκτάσεις	Υπολείματα Καλλιεργειών	Εκσυμ-βάτοποι	ΠΥΡΟΣ. ΣΩΜΑ	ΠΕΖΟΠΟΡΑ ΤΜΗΜΑΤΑ	ΕΘΕΛΟ-ΝΤΕΣ	ΣΤΡΑΤΟΣ	ΑΛΛΕΣ ΔΥΝΑΜΕΙΣ	TEMP	HIGH	LOW	HEAT DEG DAYS	COOL DEG DAYS	RAIN	SPEED	HIGH	DIR
14/2/2022	17:50	14/2/2022	20:20		Δ. ΙΣΤΙΑΙΑΣ - Ιστιαία Ευβοίας		ΑΓΙΟΣ ΓΕΩΡ		0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	7	0	0	0	0	8.1	12.7	4.0	10.3	0.0	0.0	0.0	6.4	ESE
26/3/2022	10:39	26/3/2022	12:27		Δ. ΙΣΤΙΑΙΑΣ - Ιστιαία Ευβοίας	ΛΙΧΑΣ	ΑΓΙΟΣ ΠΑΝΤ		0,00	0,05	0,00	0,00	0,00	0,00	0,00	0,00	7	0	0	0	0	11.9	17.6	5.5	6.4	0.0	0.0	0.2	12.9	ESE
26/3/2022	12:55	26/3/2022	15:10		Δ. ΙΣΤΙΑΙΑΣ - Ιστιαία Ευβοίας	ΚΑΜΑΡΙΑ	ΚΑΜΑΡΙΑ-Π		0,00	0,00	0,00	0,00	0,00	0,10	0,00	0,00	8	0	0	0	0	11.9	17.6	5.5	6.4	0.0	0.0	0.2	12.9	ESE
27/3/2022	10:30	27/3/2022	14:29		Δ. ΙΣΤΙΑΙΑΣ - Ιστιαία Ευβοίας	ΛΟΥΤΡΑ ΑΙΔ	ΑΙΔΗΨΟΣ Π		0,00	0,00	0,00	0,00	0,00	0,00	0,20	0,00	8	0	0	0	0	11.3	19.5	4.0	6.7	0.1	0.0	0.0	4.8	ESE
29/3/2022	18:52	29/3/2022	20:58		Δ. ΙΣΤΙΑΙΑΣ - Ιστιαία Ευβοίας	ΝΕΟΣ ΠΥΡΓΟ	ΝΕΟΣ ΠΥΡΓΟ		0,00	0,00	0,00	0,00	0,50	0,00	0,00	0,00	6	0	0	0	0	10.8	18.3	4.6	7.5	0.0	0.0	0.0	4.8	SE
29/3/2022	18:54	29/3/2022	21:02		Δ. ΙΣΤΙΑΙΑΣ - Ιστιαία Ευβοίας	ΑΣΜΗΝΙΟ	ΑΓΙΑ ΕΛΕΝΗ		0,00	3,00	0,00	0,00	0,00	0,00	0,00	0,00	6	0	0	0	0	10.8	18.3	4.6	7.5	0.0	0.0	0.0	4.8	SE
30/3/2022	18:22	30/3/2022	23:35		Δ. ΙΣΤΙΑΙΑΣ - Ιστιαία Ευβοίας	ΚΑΣΤΑΝΙΩΤ	ΑΓΙΟΙ ΘΕΟΔ		0,00	0,40	0,00	0,00	0,00	0,00	0,00	0,00	6	0	0	0	0	12.9	19.3	4.6	5.4	0.1	0.0	0.0	6.4	SE
30/3/2022	20:27	31/3/2022	00:55		Δ. ΙΣΤΙΑΙΑΣ - Ιστιαία Ευβοίας	ΛΙΧΑΣ	ΛΙΧΑΔΟΝΗΣ		0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	15	0	0	0	0	12.9	19.3	4.6	5.4	0.1	0.0	0.0	6.4	SE
5/4/2022	18:15	5/4/2022	18:35		Δ. ΙΣΤΙΑΙΑΣ - Ιστιαία Ευβοίας	ΩΡΕΟΙ	ΩΡΕΟΙ		0,00	0,00	0,00	0,00	0,04	0,00	0,00	0,00	6	0	0	0	0	11.9	14.7	10.1	6.4	0.0	0.0	0.5	9.7	NE
6/4/2022	15:15	6/4/2022			Δ. ΙΣΤΙΑΙΑΣ - Ιστιαία Ευβοίας	ΑΣΜΗΝΙΟ	ΚΑΡΥΕΣ ΑΣΜ		0,00	1,00	0,00	0,00	1,00	0,00	0,00	0,00	10	0	0	0	0	12.5	18.2	6.5	5.8	0.0	0.0	0.0	4.8	ESE
14/4/2022	22:16	14/4/2022	00:25		Δ. ΙΣΤΙΑΙΑΣ - Ιστιαία Ευβοίας		ΝΕΟΧΩΡΙ Θ		0,00	0,40	0,00	0,00	0,00	0,00	0,00	0,00	7	0	0	0	0	11.9	19.6	4.6	6.4	0.1	0.0	0.2	12.9	NNE
15/4/2022	16:00	15/4/2022	18:44		Δ. ΙΣΤΙΑΙΑΣ - Ιστιαία Ευβοίας	ΚΑΜΑΡΙΑ	ΚΑΜΑΡΙΑ		0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	4	0	0	0	0	12.4	20.7	4.1	6.3	0.3	0.0	0.0	4.8	NNE

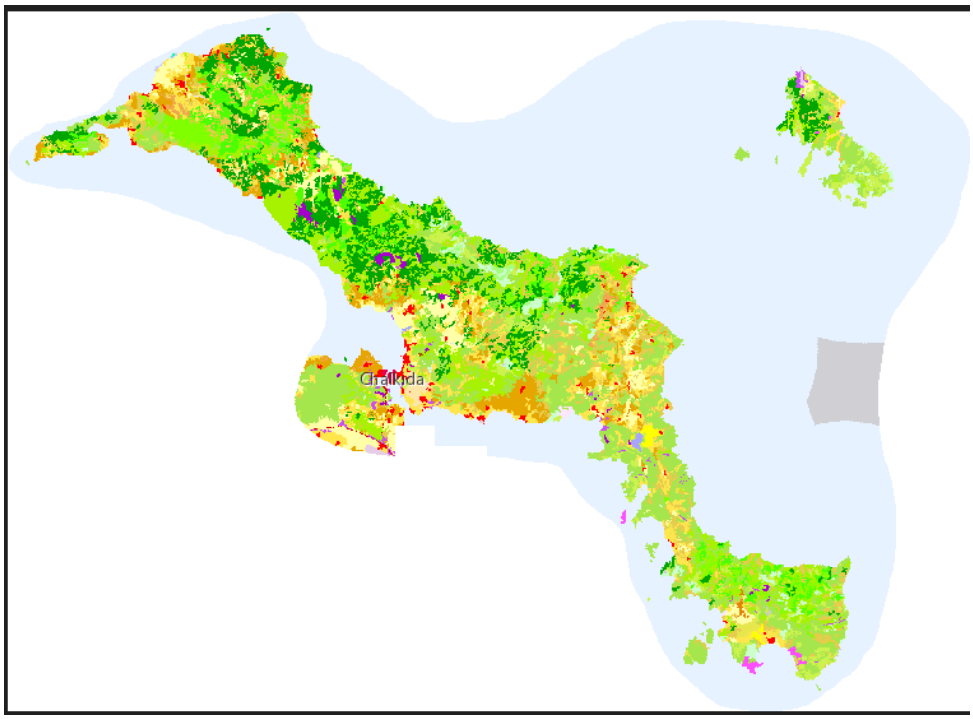
Εικόνα 4. 5 Τελικό Excel με μετεωρολογικά δεδομένα.

Δυστυχώς οι μετεωρολογικοί σταθμοί στην Εύβοια όπως προ είπαμε δεν έχουν καταγεγραμμένη την υγρασία όπως άλλοι σταθμοί στην Ελλάδα (π.χ. η Ζάκυνθος). Αλλά μας παρέχουν τους Cooling and Heating Degree Days. Είναι δείκτες ενέργειας που μας παρουσιάζουν, την ενέργεια που απαιτείται για την ψύξη και την θέρμανσή ενός κτηρίου. Εκφράζουν την σχέση (σε ημέρες) αναμεσά στην ενέργεια για την θέρμανσή η την ψύξη των κτηρίων (για την άνεση των ανθρώπων) με την θερμοκρασία και πιο συγκεκριμένα είναι το μηνιαίο άθροισμα της διαφοράς μεταξύ ενός κατώτατου ορίου θερμοκρασίας (Tr) και μιας μέσης ημερήσιάς θερμοκρασίας αέρα (T). Στην περίπτωση του CDD είναι

υψηλότερη ενώ του HDD είναι χαμηλότερη. (Corrales-Suastegui, Ruiz-Alvarez, Torres-Alavez, & Pavia, 2021)

Στην συνέχεια για κάθε ημερομηνία πυρκαγιάς και ανάλογα με την περιοχή που την ταξινομήσαμε, κατηγοριοποιήσαμε την βλάστηση που υπήρχε στην περιοχή.

Η συγκεκριμένη κατηγοριοποίηση βασίστηκε στις εικόνες που αποκτήσαμε από το CLC 2018 (https://land.copernicus.eu/pan-european/corine-land-cover/clc2018?tab=mapview). Η υπηρεσία παρακολούθησής γης Copernicus Land Monitoring Service δημιούργησε το CLC2018 το οποίο είναι η κάλυψη της βλάστησης της γης του έτους 2018 με ανάλυση 100μ. Κάθε χρώμα αντιπροσωπεύει και διαφορετικό είδος Βλάστησης όπως #a6e64d 323 - Sclerophyllous vegetation και το λεξικό των χρωμάτων είναι στο site https://collections.sentinel-hub.com/corine-land-cover/readme.html. Παρακάτω παρουσιάζεται η Εύβοια από το CLC 2018.



Εικόνα 4. 6 Η Εύβοια μέσω του CLC 2018

Για τη συγκεκριμένη εργασία, χωρίσαμε την Εύβοια σε 4 σημεία, Βορειά Εύβοια (Ωρεοί, Ιστιαία), Κεντρική Εύβοια (Χαλκίδα, Κύμη, ψαχνά), Νότια Εύβοια (Κάρυστος) και Σκύρος. Δυστυχώς τα δεδομένα δεν επιτρέπουν μεγαλύτερη ανάλυση όσον αφορά την

βλάστηση καθώς σε πολλές περιπτώσεις ο δήμος που αναγράφεται ότι ξεκίνησε η πυρκαγιά δεν συμφωνεί με την διεύθυνση που αναγράφεται στο Excel από την πυροσβεστική με αποτέλεσμα να μην μπορεί να γίνει ακριβής αποτύπωση της βλάστησης στο σημείο που ξεκίνησε η φωτιά. Για να αντιμετωπίσουμε αυτό το ζήτημα ομαδοποιήσαμε την βλάστηση σε 5 μεγάλες κατηγορίες που θα αναλυθούν παρακάτω. Ο εντοπισμός της βλάστησης έγινε μέσω ενός προγράμματος σε Python που δημιουργήσαμε για τις ανάγκες της εργασίας. Για παράδειγμα από την παρακάτω εικόνα, για να καταγράψουμε τα είδη βλάστησης, κάνουμε εισαγωγή της εικόνας στο πρόγραμμα μας (έχοντας πρώτα δημιουργήσει ένα λεξικό με τα χρώματα που επιθυμούμε να αναγνωρίσει) και αυτόματα εξάγει την βλάστηση και τα ποσοστά τους. Όπως για παράδειγμα τα παρακάτω αποτελέσματα.

223 - Olive groves: 6.90%

243 - Land principally occupied by agriculture with significant areas of natural vegetation: 13.79%

324 - Transitional woodland-shrub: 15.52%

323 - Sclerophyllous vegetation: 8.62%

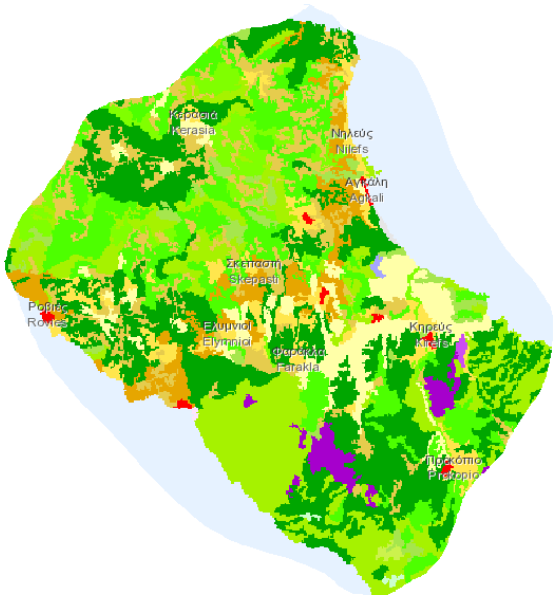
311 - Broad-leaved forest: 5.17%

312 - Coniferous forest: 13.79%

313 - Mixed forest: 17.24%

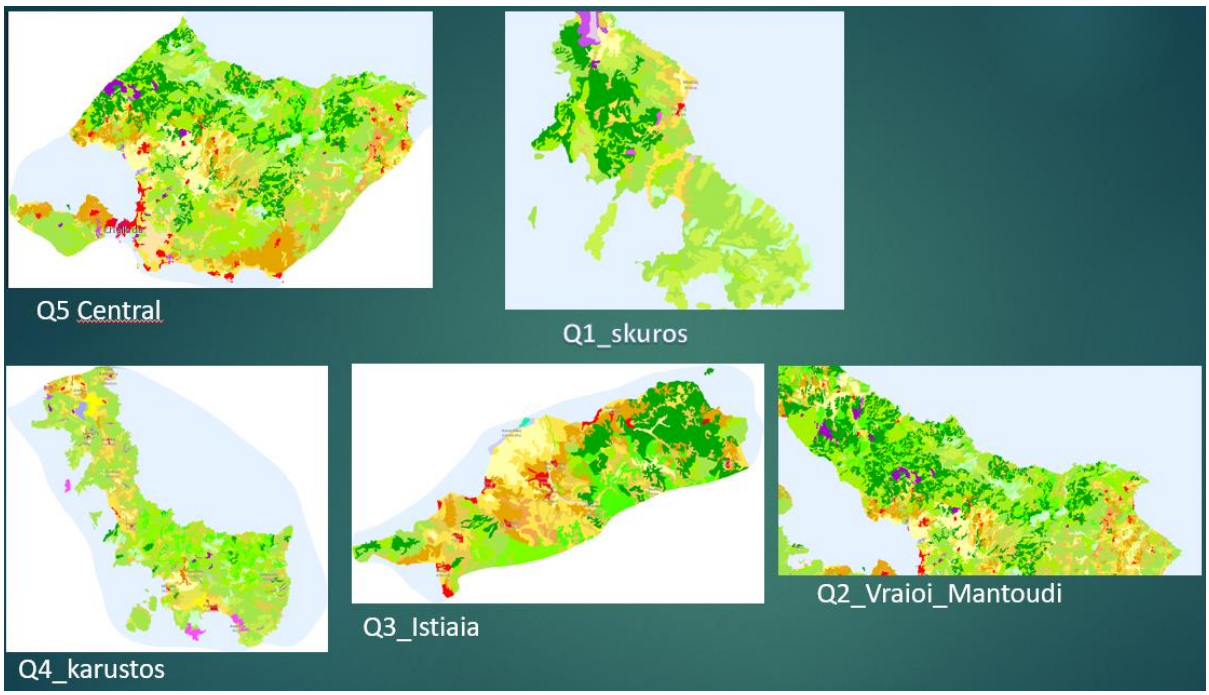
242 - Complex cultivation patterns: 6.90%

211 - Non-irrigated arable land: 5.17%



Εικόνα 4. 7 Παράδειγμα εικόνας εφαρμογής εργαλείου Βλάστησης

Με παρόμοιο τρόπο εργαστήκαμε και για τις παρακάτω εικόνες και προέκυψε ο πίνακας με την βλάστηση.



Εικόνα 4. 8 Εικόνες που χρησιμοποιήσαμε για την εύρεση της βλάστησης.

Βλάστηση	**Είδος Βλάστησης (**CLC 2018)	**Είδος Βλάστησης (Ελληνικά)**
Q2_Vraioi_Mantoudi (Βόρεια Εύβοια)	324 - Transitional woodland-shrub: 9.95% 311 - Broad-leaved forest: 0.96% 313 - Mixed forest: 2.64% 243 - Land principally occupied by agriculture with significant areas of natural vegetation: 4.49% 323 - Sclerophyllous vegetation: 7.79% 312 - Coniferous forest: 9.64% 242 - Complex cultivation patterns: 3.62% 223 - Olive groves: 2.53% 211 - Non-irrigated arable land: 3.35% 121 - Industrial or commercial units: 0.06% 131 - Mineral extraction sites: 0.60% 333 - Sparsely vegetated areas: 1.09% 321 - Natural grasslands: 1.66% 322 - Moors and heathland: 0.27% 222 - Fruit trees and berry plantation: 0.33% 231 - Pastures: 0.11%	324 - Μεταβατικό δάσος-θάμνος: 9,95% 311 - Δάσος πλατύφυλλων: 0.96% 313 - Μικτό δάσος: 2,64% 243 - Γη που καταλαμβάνεται κυρίως από γεωργία με σημαντικές εκτάσεις φυσικής βλάστησης: 4.49% 323 - Σκληρόφυλλη βλάστηση: 7.79% 312 - Δάσος κωνοφόρων: 9.64% 242 - Πολύπλοκα πρότυπα καλλιέργειας: 3.62% 223 - Ελαιώνες: 2,53% 211 - Μη αρδευόμενες καλλιεργήσιμες εκτάσεις: 3.35% 121 - Βιομηχανικές ή εμπορικές μμονάδες: 0,06% 131 - Χώροι εξόρυξης ορυκτών: 0,60% 333 - Περιοχές με αραιή βλάστηση: 1.09% 321 - Φυσικά λιβάδια: 1.66% 322 - Βάλτοι και έλη: 0,27% 222 - Φυτείες οπωροφόρων δένδρων και μούρων: 0.33% 231 - Βοσκότοποι: 0.11%
Q3_Istiaia (Βόρεια Εύβοια)	312 - Coniferous forest: 7.11% 243 - Land principally occupied by agriculture with significant areas of natural vegetation: 4.97% 223 - Olive groves: 4.70% 242 - Complex cultivation patterns: 5.66% 211 - Non-irrigated arable land: 2.59% 311 - Broad-leaved forest: 1.25% 324 - Transitional woodland-shrub: 1.91% 323 - Sclerophyllous vegetation: 2.35% 313 - Mixed forest: 2.73% 222 - Fruit trees and berry plantation: 0.76% 321 - Natural grasslands: 0.12%	312 - Δάσος κωνοφόρων: 7.11% 243 - Γη που καταλαμβάνεται κυρίως από γεωργία με σημαντικές εκτάσεις φυσικής βλάστησης: 4.97% 223 - Ελαιώνες: 4.70% 242 - Πολύπλοκα πρότυπα καλλιέργειας: 5.66% 211 - Μη αρδευόμενες καλλιεργήσιμες εκτάσεις: 2,59% 311 - Δάση πλατύφυλλων: 1.25% 324 - Μεταβατικά δάση-θάμνοι: 1,91% 323 - Σκληρόφυλλη βλάστηση: 2,35% 313 - Μικτό δάσος: 2,73% 222 - Φυτείες οπωροφόρων δένδρων και μούρων: 0.76% 321 - Φυσικά λιβάδια: 0.12%
Q5 Central (Κεντρική Εύβοια)	324 - Transitional woodland-shrub: 8.54% 323 - Sclerophyllous vegetation: 9.65% 333 - Sparsely vegetated areas: 1.31% 312 - Coniferous forest: 6.23% 321 - Natural grasslands: 2.06% 313 - Mixed forest: 1.15% 322 - Moors and heathland: 0.32% 243 - Land principally occupied by agriculture with significant areas of natural vegetation: 4.11% 311 - Broad-leaved forest: 0.82% 131 - Mineral extraction sites: 0.36% 211 - Non-irrigated arable land: 2.53% 242 - Complex cultivation patterns: 4.34% 222 - Fruit trees and berry plantation: 0.41% 121 - Industrial or commercial units: 0.16% 223 - Olive groves: 4.01% 231 - Pastures: 0.13%	324 - Μεταβατικό δάσος-θάμνος: 8,54% 323 - Σκληρόφυλλη βλάστηση: 9,65% 333 - Περιοχές με αραιή βλάστηση: 1.31% 312 - Δάση κωνοφόρων: 6.23% 321 - Φυσικά λιβάδια: 2,06% 313 - Μικτά δάση: 1.15% 322 - Βάλτοι και ερείπια: 0,32% 243 - Γη που καταλαμβάνεται κυρίως από γεωργία με σημαντικές εκτάσεις φυσικής βλάστησης: 4.11% 311 - Δάση πλατύφυλλων: 0.82% 131 - Χώροι εξόρυξης ορυκτών: 0,36% 211 - Μη αρδευόμενη αρόσιμη γη: 2,53% 242 - Πολύπλοκα πρότυπα καλλιέργειας: 4.34% 222 - Φυτείες οπωροφόρων δένδρων και μούρων: 0.41% 121 - Βιομηχανικές ή εμπορικές μονάδες: 0,16%

	241 - Annual crops associated with permanent crops: 0.52%	223 - Ελαιώνες: 4.01% 231 - Βοσκότοποι: 0.13% 241 - Ετήσιες καλλιέργειες που συνδέονται με μόνιμες καλλιέργειες: 0,52%
Q1_skuros (Κεντρική Ευβοια-Σκυρος)	121 - Industrial or commercial units: 0.48% 333 - Sparsely vegetated areas: 2.29% 321 - Natural grasslands: 7.73% 243 - Land principally occupied by agriculture with significant areas of natural vegetation: 1.45% 231 - Pastures: 0.34% 323 - Sclerophyllous vegetation: 10.97% 211 - Non-irrigated arable land: 0.70% 312 - Coniferous forest: 5.30% 324 - Transitional woodland-shrub: 2.86% 242 - Complex cultivation patterns: 1.92% 131 - Mineral extraction sites: 0.04%	121 - Βιομηχανικές ή εμπορικές μονάδες: 0,48% 333 - Περιοχές με αραιή βλάστηση: 2.29% 321 - Φυσικά λιβάδια: 7.73% 243 - Εκτάσεις που καταλαμβάνονται κυρίως από τη γεωργία με σημαντικές εκτάσεις φυσικής βλάστησης: 1.45% 231 - Βοσκότοποι: 0.34% 323 - Σκληροφυλλική βλάστηση: 10.97% 211 - Μη αρδευόμενη αρόσιμη γη: 0,70% 312 - Δάση κωνοφόρων: 5.30% 324 - Μεταβατικό δάσος-θάμνος: 2,86% 242 - Πολύπλοκα πρότυπα καλλιέργειας: 1.92% 131 - Χώροι εξόρυξης ορυκτών: 0,04%
Q4_karustos (Νοτια Εύβοια)	243 - Land principally occupied by agriculture with significant areas of natural vegetation: 2.76% 323 - Sclerophyllous vegetation: 9.30% 242 - Complex cultivation patterns: 1.94% 211 - Non-irrigated arable land: 0.69% 223 - Olive groves: 0.22% 221 - Vineyards : 0.12% 121 - Industrial or commercial units: 0.18% 212 - Permanently irrigated land: 0.38% 321 - Natural grasslands: 2.11% 131 - Mineral extraction sites: 0.07% 324 - Transitional woodland-shrub: 3.22% 231 - Pastures: 0.50% 311 - Broad-leaved forest: 0.97% 333 - Sparsely vegetated areas: 0.80% 313 - Mixed forest: 1.22% 312 - Coniferous forest: 0.38%	243 - Γη που καταλαμβάνεται κυρίως από γεωργία με σημαντικές εκτάσεις φυσικής βλάστησης: 2,76% 323 - Σκληροφυλλική βλάστηση: 9,30% 242 - Πολύπλοκα πρότυπα καλλιέργειας: 1.94% 211 - Μη αρδευόμενη καλλιεργήσιμη γη: 0,69% 223 - Ελαιώνες: 0.22% 221 - Αμπελώνες: 0,12% 121 - Βιομηχανικές ή εμπορικές μονάδες: 0,18% 212 - Μόνιμα αρδευόμενες εκτάσεις: 0,38% 321 - Φυσικά λιβάδια: 2,11% 131 - Χώροι εξόρυξης ορυκτών: 0,07% 324 - Μεταβατικές δασικές εκτάσεις-θάμνοι: 3,22% 231 - Βοσκότοποι: 0.50% 311 - Δάση πλατύφυλλων: 0.97% 333 - Περιοχές με αραιή βλάστηση: 0.80% 313 - Μικτά δάση: 1.22% 312 - Δάση κωνοφόρων: 0.38%

Πίνακας 4. 1 Πίνακας Βλάστησης

Το εργαλείο που χρησιμοποιήσαμε για τον εντοπισμό της βλάστησης έχει την παρακάτω μορφή:

```python
import pandas as pd
import cv2
import numpy as np


#Αυτή είναι η συνάρτηση get_color_name() που χρησιμοποιείται για να βρεθεί το όνομα του πλησιέστερου χρώματος βάσει της ευκλείδιας απόστασης
#ανάμεσα στα δοθέντα RGB της εικόνας και των χρωμάτων στο DataFrame.
def get_color_name(R, G, B):
    minimum = 1000
    cname = ""
    for i in range(len(df)):
        d = abs(R - df['R'][i]) + abs(G - df['G'][i]) + abs(B - df['B'][i])
        if d <= minimum:
            minimum = d
            cname = df['color_name'][i]
    return cname

# ----------------------------------------------------------------------------

#εισαγωγη δεδομενων
img_path = r"Q4_karustos.png"
csv_path = r"colors.csv"

# Δημιουργία DataFrame από το CSV αρχείο με τις παρακάτω στήλες
df = pd.read_csv(csv_path, names=['hex', 'color_name', 'hex_value', 'R', 'G', 'B'], header=None)

# Φόρτωση εικόνας ως numpy array
img = cv2.imread(img_path)

color_freq = {}  # Λεξικό για αποθήκευση συχνοτήτων χρωμάτων


#Δημιουργεί έναν πίνακα που περιέχει τις τιμές R, G και B των χρωμάτων από το DataFrame.
available_colors = np.array(df[['R', 'G', 'B']])

# Αυτό το τμήμα επεξεργάζεται κάθε pixel της εικόνας.
#Για κάθε pixel, υπολογίζεται η ευκλείδια απόσταση ανάμεσα στις τιμές RGB του pixel και των τιμών RGB των διαθέσιμων χρωμάτων.
#Το όνομα του πλησιέστερου χρώματος βρίσκεται από το DataFrame και το λεξικό color_freq αυξάνει την συχνότητα αυτού του χρώματος.
for y in range(img.shape[0]):
    for x in range(img.shape[1]):
        b, g, r = img[y, x]
        b = int(b)
        g = int(g)
        r = int(r)

        # Υπολογίζουμε την ευκλείδια απόσταση για κάθε διαθέσιμο χρώμα
        distances = np.sqrt(np.sum((available_colors - [r, g, b]) ** 2, axis=1))

        # Βρίσκουμε τον ελάχιστο δείκτη απόστασης
        min_distance_index = np.argmin(distances)

        color_name = df['color_name'][min_distance_index]
        color_freq[color_name] = color_freq.get(color_name, 0) + 1

# Ορίζουμε το όριο ποσοστού για να εμφανιστεί ένα χρώμα (σε ποσοστό)
min_percentage = 0.01

# Εμφανίζει το όνομα και το ποσοστό εμφάνισης των χρωμάτων, μόνο εάν το ποσοστό είναι μεγαλύτερο από το επιθυμητό όριο.
#Το ποσοστό εμφάνισης υπολογίζεται ως το πλήθος των εμφανίσεων του χρώματος διαιρεμένο με το συνολικό πλήθος των χρωμάτων.
total_colors = sum(color_freq.values())

for color_name, freq in color_freq.items():
    percentage = (freq / total_colors) * 100
    if percentage > min_percentage:
        print(f"{color_name}: {percentage:.2f}%")
```

Εικόνα 4. 9 Ο κώδικας του εργαλείου Βλάστησης

Ο συγκεκριμένος κώδικας διαβάζει από το img_path την εικόνα που θέλουμε να εξετάσουμε και βάση το λεξικό χρωμάτων που δημιουργήσαμε (csv_pth) από το σαιτ https://www.colorhexa.com που περιέχει τους κωδικούς από το CLC 2018 μεταφρασμένους σε RGB values, εμφανίζεται η βλάστηση.

```
#ffffa8,"211 - Non-irrigated arable land",#ffffa8,255,255,168
#ffff00,"212 - Permanently irrigated land",#ffff00,255,255,0
#e6e600,"213 - Rice fields",#e6e600,230,230,0
#e68000,"221 - Vineyards  ",#e68000,230,128,0
#f2a64d,"222 - Fruit trees and berry plantation",#f2a64d,242,166,77
#e6a600,"223 - Olive groves",#e6a600,230,166,0
#e6e64d,"231 - Pastures",#e6e64d,230,230,77
#ffe6a6,"241 - Annual crops associated with permanent crops",#ffe6a6,255,230,166
#ffe64d,"242 - Complex cultivation patterns",#ffe64d,255,230,77
#e6cc4d,"243 - Land principally occupied by agriculture with significant areas of natural vegetation",#e6cc4d,230,204,77
#f2cca6,"244 - Agro-forestry areas",#f2cca6,242,204,166
```

Εικόνα 4. 10 Μορφή λεξικού χρωμάτων

1	#e6004d	111 - Continuous urban fabric
2	#ff0000	112 - Discontinuous urban fabric
3	#cc4df2	121 - Industrial or commercial units
4	#cc0000	122 - Road and rail networks and associated land
5	#e6cccc	123 - Port areas
6	#e6cce6	124 - Airports
7	#a600cc	131 - Mineral extraction sites
8	#a64d00	132 - Dump sites
9	#ff4dff	133 - Construction sites
10	#ffa6ff	141 - Green urban areas
11	#ffe6ff	142 - Sport and leisure facilities
12	#ffffa8	211 - Non-irrigated arable land
13	#ffff00	212 - Permanently irrigated land
14	#e6e600	213 - Rice fields
15	#e68000	221 - Vineyards
16	#f2a64d	222 - Fruit trees and berry plantations
17	#e6a600	223 - Olive groves
18	#e6e64d	231 - Pastures
19	#ffe6a6	241 - Annual crops associated with permanent crops
20	#ffe64d	242 - Complex cultivation patterns
21	#e6cc4d	243 - Land principally occupied by agriculture with significant areas of natural vegetation
22	#f2cca6	244 - Agro-forestry areas
23	#80ff00	311 - Broad-leaved forest
24	#00a600	312 - Coniferous forest
25	#4dff00	313 - Mixed forest
26	#ccf24d	321 - Natural grasslands
27	#a6ff80	322 - Moors and heathland
28	#a6e64d	323 - Sclerophyllous vegetation
29	#a6f200	324 - Transitional woodland-shrub

Εικόνα 4. 11 Λίστα Χρωμάτων CLC 2018

4.2 Σύγκριση απόδοσης και ακρίβειας

a) Linear Regression

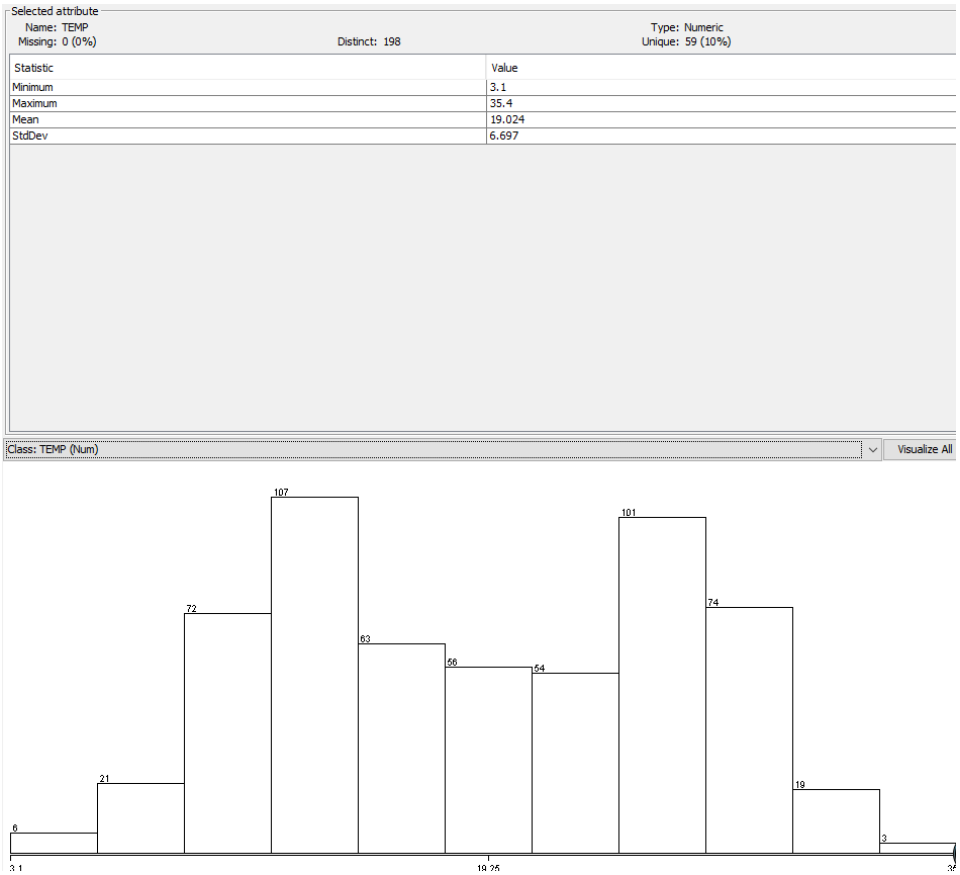
Ο αλγόριθμος Linear Regression για να λειτουργήσει, θα πρέπει όλα τα πεδία να είναι αριθμητικά όπως για παράδειγμα ο Βόρειος άνεμος θα δηλωθεί ως εξής Βόρειος : 1 Νότος: 0 Ανατολικά: 0 Δυτικά: 0. Τροποποιήσαμε το excel μας και πλέον έχει 52 στήλες. Επίσης προσθέσαμε όλους τους διαφορετικούς τύπους καμένης γης (Δάσος, Άλσος κ.α) σε μια στήλη την Burned_Area όπως επίσης και τις πυροσβεστικές δυνάμεις στην fireman (πυροσβεστικό σώμα, πεζοπόρο τμήμα, εθελοντές, άλλες δυνάμεις και στρατός). Σε αυτό το σημείο αξίζει να αναφέρουμε ότι στις πυροσβεστικές δυνάμεις δεν

υπολογίσαμε τα μηχανοκίνητα μέσα καθώς δεν ήταν συμπληρωμένα σε όλες τις περιπτώσεις. Στην συνέχεια, με σκοπό την καλύτερη μελέτη των συσχετίσεων μεταξύ της εξαρτημένης μεταβλητής (Burned_Area) και των υπολοίπων ανεξάρτητων μεταβλητών, χωρίσαμε την Burned_Area (καμένη έκταση) σε δυο μεταβλητές, την Low_Burned_Area που περιλαμβάνει τις φωτιές που έκαψαν από 9 στρέμματα και κάτω και την Large_Burned_Area που περιλαμβάνει τις υπόλοιπες. Μετατρέψαμε το Excel σε arff και κάναμε εισαγωγή στο Weka για περαιτέρω ανάλυση.

```
@relation Burned_Area_Prediction

@attribute Mon numeric
@attribute Tue numeric
@attribute wed numeric
@attribute Thu numeric
@attribute Friday numeric
@attribute Sat numeric
@attribute Sun numeric
@attribute Winter numeric
@attribute spring numeric
@attribute summer numeric
@attribute Autumn numeric
@attribute Early numeric
@attribute Morning numeric
@attribute Noon numeric
@attribute Afternoon numeric
@attribute MidNight numeric
@attribute Istiaia numeric
@attribute Chalkida numeric
@attribute Karistos numeric
@attribute Kimi numeric
@attribute Oreoi numeric
@attribute Psaxna numeric
@attribute Q1_skuros numeric
@attribute Q2_Vraioi_Mantoudi numeric
@attribute Q3_Istiaia numeric
@attribute Q4_karustos numeric
@attribute Q5_Central numeric
@attribute forest numeric
@attribute forest_ektash numeric
@attribute alsos numeric
@attribute xorta_ektaseis numeric
@attribute kalamia_valtoi numeric
@attribute georgikes_ektaseis numeric
@attribute upolleimata_kalliergeivn numeric
@attribute rubish numeric
@attribute Burned_Area numeric
@attribute Low_Burned_Area numeric
@attribute Large_Burned_Area numeric
@attribute Purosvestiko_svma numeric
@attribute footer numeric
@attribute volunteer numeric
@attribute soliders numeric
@attribute other_forces numeric
@attribute Fireman numeric
@attribute TEMP numeric
@attribute HEAT_DEG_DAYS numeric
@attribute COOL_DEG_DAYS numeric
@attribute RAIN numeric
@attribute SPEED numeric
@attribute ESE numeric
@attribute SE numeric
@attribute NE numeric
@attribute NNE numeric
@attribute SSE numeric
@attribute NNW numeric
@attribute W numeric

@data
1,0,0,0,0,0,1,0,0,0,0,0,1,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0.01,0,0,0.01,1,0,7,0,0,0,0,7,8.1,10.3,0.0,0.0,0.0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0
0,0,0,0,0,1,0,0,1,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0.05,0,0,0,0,0,0,0.05,1,0,7,0,0,0,0,7,11.9,6.4,0.0,0.0,0.2,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0
0,0,0,0,0,1,0,0,1,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0.1,0,0,0.1,1,0,8,0,0,0,0,8,11.9,6.4,0.0,0.0,0.2,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0
0,0,0,0,0,0,1,0,1,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0.2,0,0.2,1,0,8,0,0,0,0,0,8,11.3,6.7,0.1,0.0,0.0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0
```

Εικόνα 4. 10 Παράδειγμα arff αρχείου για παλινδρόμηση

Αρχικά το Weka μας δίνει κάποια περιγραφικά στοιχεία για την κάθε στήλη όπως και το ιστόγραμμα τους. Για παράδειγμα η στήλη θερμοκρασία έχει μηδέν missing value, είναι αριθμητικό, η ελάχιστη τιμή είναι 3,1, η μέγιστη 35,4 και η μέση τιμή είναι 19,024.



Εικόνα 4. 11 Περιγραφικά στοιχεία θερμοκρασίας Weka

Αρχικά, θα δημιουργήσουμε τον πίνακα συσχετίσεων της καμένης Γης και της κατηγοριοποίησης που πραγματοποιήσαμε ώστε να ελέγξουμε εάν υπάρχει κάποια ισχυρή συσχέτιση με κάποια μεταβλητή.

Στήλη1	Burned_Area	Low_Burned_Area	Large_Burned_Area
Mon	-0.02	0.02	-0.02
Tue	0.11	-0.01	0.01
wed	-0.02	0.04	-0.04
Thu	-0.02	0	0
Friday	-0.02	-0.03	0.03
Sat	-0.02	-0.01	0.01
Sun	-0.02	-0.01	0.01
Winter	-0.02	-0.03	0.03
spring	-0.02	0.13	-0.13
summer	0.06	-0.12	0.12
Autumn	-0.02	0.03	-0.03
Early	-0.01	0.03	-0.03
Morning	-0.02	0.04	-0.04
Noon	-0.03	-0.06	0.06
Afternoon	0.05	0.05	-0.05
MidNight	-0.01	-0.07	0.07
Istiaia	-0.02	0.08	-0.08
Chalkida	-0.01	0.02	-0.02
Karistos	-0.02	-0.28	0.28
Kimi	-0.03	0.06	-0.06
Oreoi	0.12	0.04	-0.04
Psaxna	-0.02	0.07	-0.07
Q1_skuros	-0.01	-0.06	0.06
Q2_Vraioi_Man	0.12	0.04	-0.04
Q3_Istiaia	-0.02	0.08	-0.08
Q4_karustos	-0.02	-0.28	0.28
Q5_Central	-0.05	0.14	-0.14
Burned_Area	1	-0.13	0.13
Low_Burned_Ar	-0.13	1	-1
Large_Burned_	0.13	-1	1
Fireman	0.98	-0.23	0.23
TEMP	0.08	-0.13	0.13
HEAT_DEG_DAY	-0.04	0.1	-0.1
COOL_DEG_DAY	0.12	-0.18	0.18
RAIN	-0.01	0.01	-0.01
SPEED	-0.03	-0.13	0.13
ESE	-0.01	0.06	-0.06
SE	-0.01	-0.07	0.07
NE	0.1	0.08	-0.08
NNE	-0.01	-0.04	0.04
SSE	-0.01	0	0
NNW	-0.01	0.06	-0.06
W	-0.01	0.04	-0.04
ENE	-0.01	0.04	-0.04
S	-0.01	0.04	-0.04
N	-0.01	-0.03	0.03
SSW	-0.02	-0.13	0.13
SW	-0.01	0.03	-0.03
WNW	-0.01	0	0
NW	-0.01	0.02	-0.02
WSW	-0.01	-0.03	0.03

Εικόνα 4. 12 Πίνακας συσχετίσεων καμένης γης, Low_Burned_Area και Large_Burned_Area

Πιο αναλυτικά ισχυρή συσχέτιση υπάρχει όπως ήταν αναμενόμενο, με τις πυροσβεστικές δυνάμεις ενώ μικρή είναι και η συσχέτιση με την θερμοκρασία, την βλάστηση Q2 καθώς και με τον NE άνεμο. Μεγάλη μας έκπληξη μας προκαλεί το γεγονός ότι η ταχύτητα του ανέμου συσχετίζεται αρνητικά με την έκταση της καμένης γης.

Στην συνέχεια μέσω του αλγόριθμου Linear_Regression θα προσπαθήσουμε να προβλέψουμε την καμένη έκταση (Burned_Area).

No.	Name
1	Mon
2	Tue
3	wed
4	Thu
5	Friday
6	Sat
7	Sun
8	Winter
9	spring
10	summer
11	Autumn
12	Early
13	Morning
14	Noon
15	Afternoon
16	MidNight
17	Q1_skuros
18	Q2_Vraioi_Mantoudi
19	Q3_Istiaia
20	Q4_karustos
21	Q5_Central
22	Burned_Area
23	TEMP
24	RAIN
25	SPEED
26	ESE
27	SE
28	NE
29	NNE
30	SSE
31	NNW
32	W
33	E
34	ENE
35	S
36	N
37	SSW
38	SW
39	WNW
40	NW
41	WSW

Εικόνα 4. 13 Οι ανεξάρτητες και η εξαρτημένη μεταβλητή (μπλε χρώμα) της παλινδρόμησης καμένης γης

Επίσης προχωρήσαμε μέσω του αντίστοιχου φίλτρου στο weka σε normalization οπού η ελάχιστη τιμή έγινε μηδέν και η μέγιστη 1 καθώς τα δεδομένα μας έχουν πολύ μεγάλο εύρος. Ο μαθηματικός τύπος είναι ο ακόλουθος:

x_normalized = (x - x_min) / (x_max - x_min)

όπου x είναι η αρχική τιμή του χαρακτηριστικού, x_min είναι η ελάχιστη τιμή του χαρακτηριστικού και x_max είναι η μέγιστη τιμή του χαρακτηριστικού. Αυτό έχει αποτελέσματα τα δεδομένα μας είναι στο ίδιο επίπεδο και μπορούν να συγκριθούν.

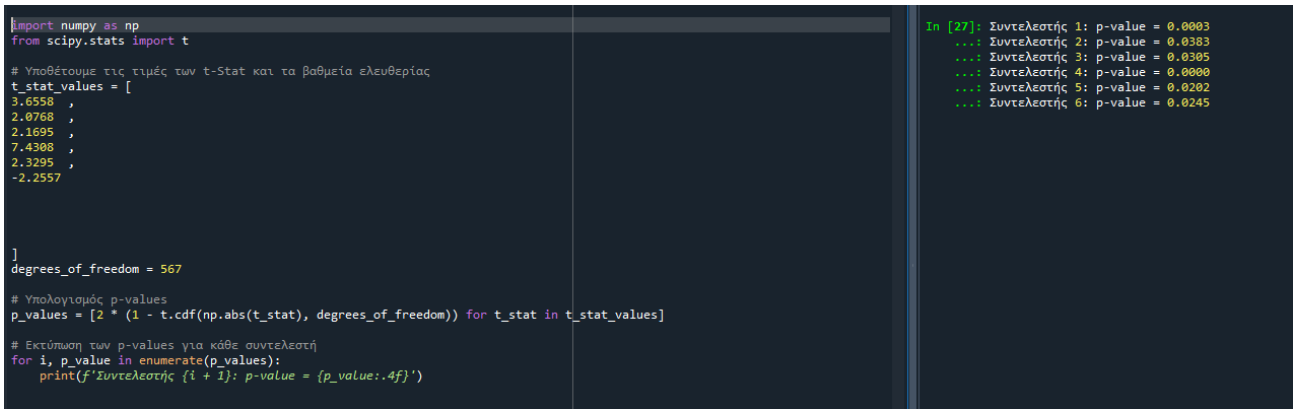
Για αρχή ως ανεξάρτητες μεταβλητές έχουμε την κατεύθυνση (πχ NE) και την ταχύτητα του ανέμου (Speed), την βροχή (Rain), την θερμοκρασία (Temp), την βλάστηση που αναλύσαμε στην προηγουμένη ενότητα, την ώρα και την ημέρα που ξεκίνησε η πυρκαγιά και τέλος την εποχή. Δυστυχώς, τα αποτελέσματα δεν ήταν θετικά, και ως εκ τούτου προχωρήσαμε στην εφαρμογή γραμμικής παλινδρόμησης ξεχωριστά σε δύο νέες μεταβλητές, τις Low_Burned_Area και Large_Burned_Area. Η Low_Burned_Area περιλαμβάνει τις πυρκαγιές με έκταση από 9 στρέμματα και κάτω και η

Large_Burned_Area έχει τις υπόλοιπες. Το συγκεκριμένο κατώφλι προέκυψε μέσω αρκετών δοκιμών που πραγματοποιήσαμε πάνω στον αλγόριθμο linear regression και δεν βασίζεται σε στατιστικούς υπολογισμούς, αλλά σε αποτελέσματα που βρήκαμε. Πρώτα παρουσιάζεται το μοντέλο παλινδρόμησης για το Large_Burned_Area (εξαρτημένη μεταβλητή). Από τις παραπάνω ανεξάρτητες μεταβλητές που περιγράψαμε μέσω της διαδικασίας backward elimination αφαιρούσαμε κάθε φορά τις μεταβλητές που είχαν p value πάνω από 0,05% και καταλήξαμε στην παρακάτω εξίσωση που φαίνεται στην εικόνα (4. 14). Επιπρόσθετα, σε όλα τα μοντέλα παλινδρόμησης που δημιουργήσαμε δεν χρησιμοποιήσαμε όλα τα δεδομένα ως training set, αλλά εφαρμόσαμε στα test_options του weka το cross-validation με folds 10 η οποία είναι μια τεχνική για την αξιολόγηση ενός μοντέλου και τον έλεγχο της απόδοσής του. Στην επιλογή αυτή το weka «σπάει» σε δέκα segment τα δεδομένα και μετά χρησιμοποιεί τα εννιά για να κτίσει το μοντέλο και το ένα για πρόβλεψη. Η συγκεκριμένη διαδικασία πραγματοποιείται δέκα φορές και κάθε φορά δημιουργεί διαφορετικούς συνδυασμούς. Επίσης διαδεδομένη τεχνική αξιολόγησης του μοντέλου είναι το percentage-split, η οποία διαχωρίζει τα δεδομένα ανάμεσα σε δεδομένα εκπαίδευσης και πρόβλεψης. Πχ 80-20 είναι 80% δεδομένα εκπαίδευσης και 20% δεδομένα πρόβλεψης. Στην συγκεκριμένη ερευνά επιλέξαμε το cross-validations με folds 10 το οποίο είχε ως προ επιλογή, αλλά και στα υπόλοιπα τεστ δεν παρατηρήθηκαν σημαντικές διαφορές μεταξύ των ποτελεσμάτων.

```
=== Classifier model (full training set) ===


Linear Regression Model

Burned_Area =

      0.3329 * Q2_Vraioi_Mantoudi +
     -0.3333 * WNW +
      0.0004

Regression Analysis:

Variable               Coefficient     SE of Coef        t-Stat
Q2_Vraioi_Mantoudi          0.3329         0.0647         5.1462
WNW                        -0.3333         0.126         -2.6458
const                       0.0004         0.0147         0.0298

Degrees of freedom = 56
R^2 value = 0.3213
Adjusted R^2 = 0.29705
F-statistic = 13.2545

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                -0.0303
Mean absolute error                     0.0429
Root mean squared error                 0.1656
Relative absolute error               126.6944 %
Root relative squared error           126.0703 %
Total Number of Instances              59
```

Εικόνα 4. 15 Αποτελέσματα Linear Regression Model για Large_Burned_Area

Στη συγκεκριμένη εξίσωση (εικόνα 4.14) βλέπουμε ότι οι μονές ανεξάρτητες μεταβλητές που βοηθούν στην πρόβλεψη της Large_Burned_Area είναι η βλάστηση Q2 και η κατεύθυνση του ανέμου WNW. Οι αριθμοί που βλέπουμε (το 0,33 και το -0,33) είναι οι συντελεστές που πολλαπλασιάζονται οι αντίστοιχες ανεξάρτητες μεταβλητές (beta) για τον υπολογισμό της εξαρτημένης μεταβλητής.



Εικόνα 4. 16 Υπολογισμός p-value python

Όλες οι παράμετροι είναι στατιστικά σημαντικές, δηλαδή το p_value που υπολογίσαμε από τον παρακάτω κώδικα είναι κάτω από 0,05% (δεν μας το δίνει έτοιμο το weka και για αυτό το υπολογίσαμε).

Την ίδια διαδικασία ακολουθήσαμε και για την εξαρτημένη μεταβλητή Low_Burned_Area:

```
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


Linear Regression Model

Burned_Area =

     -0.08   * spring +
     -0.0353 * Afternoon +
     -0.0722 * Q2_Vraioi_Mantoudi +
      0.1305 * Q4_karustos +
     -0.13   * TEMP +
      0.2356

Regression Analysis:

Variable               Coefficient     SE of Coef        t-Stat
spring                    -0.08           0.0204        -3.9173
Afternoon                 -0.0353         0.0171        -2.0668
Q2_Vraioi_Mantoudi        -0.0722         0.0268        -2.6953
Q4_karustos                0.1305         0.0259         5.04
TEMP                      -0.13           0.0394        -3.2998
const                      0.2356         0.0246         9.5602

Degrees of freedom = 511
R^2 value = 0.111
Adjusted R^2 = 0.1023
F-statistic = 12.7604

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.2786
Mean absolute error                      0.1398
Root mean squared error                  0.1894
Relative absolute error                 94.1381 %
Root relative squared error             96.0188 %
Total Number of Instances              517
```

Εικόνα 4. 17 Αποτελέσματα Linear Regression Model για Low_Burned_Area

Σε αντίθεση με το προηγούμενο μοντέλο, εδώ έχουμε περισσότερες ανεξάρτητες μεταβλητές που βοηθούν στην ερμηνεία του Low_Burned_Area. Στον πίνακα 4.2 παρουσιάζονται τα p_value κάθε ανεξάρτητες μεταβλητής.

Συντελεστής	p-value
spring	0.0001
Afternoon	0.0393

Q2_Vraioi_Mantoudi	0.0073
Q4_karustos	0.0000
TEMP	0.0010

Πίνακας 4. 2 Αποτελέσματα P_Value για τις ανεξάρτητες μεταβλητές Low_Burned_Area

Συγκεντρωτικά τα αποτελέσματα των δυο μοντέλων.

Στήλη1	**Large_Burned_Area**	**Low_Burned_Area**
Degrees of freedom	56	568
R^2 value	0.32	0.11
Adjusted R^2	0.30	0.10
F-statistic	13.25	12.7604
Correlation coefficient	-0.03	0.2786
Mean absolute error	0.04	0.1398
Root mean squared error	0.16	0.1894
Relative absolute error	126.69%	94.1381 %
Root relative squared error	126.07%	96.8375 %
Total Number of Instances	59	517

Πίνακας 4. 3 Σύγκριση Linear Regression

Το R^2 Value υπολογίζει το ποσοστό διακύμανσης στην εξαρτημένη μεταβλητή και όσο υψηλότερο είναι τόσο καλύτερη η επεξηγηματική ικανότητα του μοντέλου. Εδώ το Large_Burned_Area έχει καλύτερο R^2 από το Low_Burned_Area. Το Adjusted R^2 δίνει περισσότερη βαρύτητα σε σημαντικές συσχετίσεις και θεωρείται πιο αξιόπιστο από το R^2 Value. Το Large_Burned_Area είναι και εδώ υψηλότερο. Επίσης το F-Statistic παρουσιάζει την σημαντικότητα του μοντέλου και όσο μεγαλύτερες είναι οι τιμές, τόσο πιο σημαντικό είναι το μοντέλο. Το Large_Burned_Area έχει μια μεγαλύτερη τιμή, υποδηλώνοντας μεγαλύτερη σημαντικότητα. Το Correlation Coefficient (συντελεστής συσχέτισης) αποκαλύπτει τον βαθμό συσχέτισης μεταξύ των μεταβλητών. Μεγαλύτερη συσχέτιση έχει το Low_Burned_Area. Τέλος τα σφάλματα (Mean Absolute Error και το Root Mean Squared Error μας δείχνουν την απόσταση των πραγματικών δεδομένων με τις προβλέψεις των μοντέλων. Στο Large_Burned_Area, αυτές οι τιμές είναι σημαντικά χαμηλότερες, που σημαίνει καλύτερη ακρίβεια.

Συνεπώς το μοντέλο Large_Burned_Area έχει καλύτερη επίδοση από το μοντέλο Low_Burned_Area αν και τα δυο έχουν πολύ μικρό R^2 Value

b) Random Forest

Στα παραπάνω δεδομένα και με τις ιδίες ανεξάρτητες και εξαρτημένες μεταβλητές εφαρμόσαμε έναν διαφορετικό αλγόριθμο, τον Random Forest, με σκοπό να ελέγξουμε εάν προκύψουν καλυτέρα αποτελέσματα.

Στήλη1	**RandomForest_Large_Burned_Area**	**RandomForest_Low_Burned_Area**
Correlation coefficient	-0.0357	0.26
Mean absolute error	0.029	1.29
Root mean squared error	0.14	1.74
Relative absolute error	85.6167%	96.29%
Root relative squared error	104.90%	98.37%
Total Number of Instances	59	517

Πίνακας 4. 4 Σύγκριση Random Forest

Εδώ φαίνεται ότι ο αλγόριθμος έχει καλύτερη προσαρμογή στο Large_Burned_Area καθώς έχει μικρότερο MAE και RMSE, αλλά έχει εκπαιδευτεί με λιγότερα δεδομένα.
Και συγκεντρωτικά οι αλγόριθμοι:

Στήλη1	**RandomForest_Large_Area**	**RandomForest_Low_Area**	**Linear_Low_Area**	**Linear**_Large_Area
Correlation coefficient	-0.0357	0.26	0.2786	-0.03
Mean absolute error	0.029	1.29	0.1398	0.04
Root mean squared error	0.14	1.74	0.1894	0.16
Relative absolute error	85.6167%	96.29%	94.1381 %	126.69%
Root relative squared error	104.90%	98.37%	96.8375 %	126.07%
Total Number of Instances	59	517	517	59

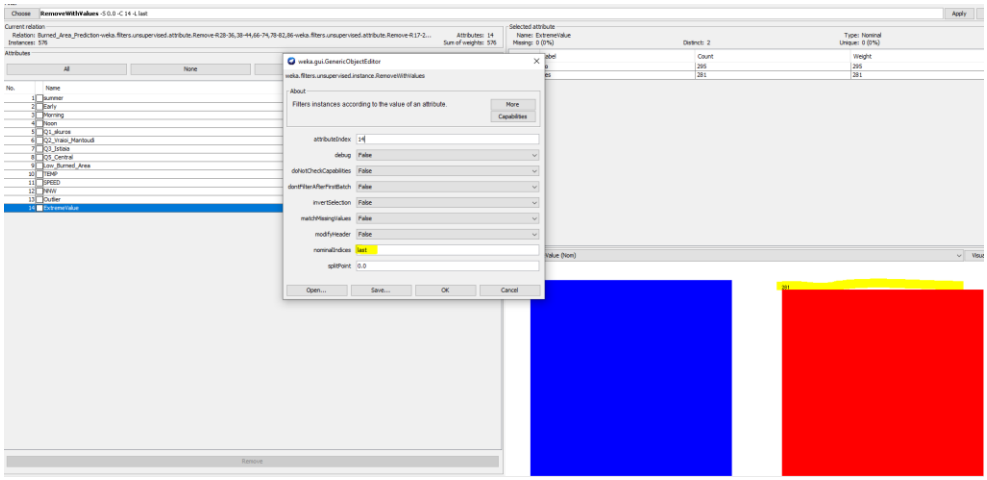
Πίνακας 4. 5 Σύγκριση περιγραφικών Αλγορίθμων

Το Linear Regression έχει καλύτερη απόδοση από το Random Forest όσον αφορά το Low_Burned_Area ενώ για το Large_ Burned _Area ο αλγόριθμος random forest είναι καλύτερος.

c) Απαλοιφή Extreme_Value και Outlier

Το weka έχει μια επιλογή στο filter που βρίσκει τα outliers με την μέθοδο InterquartileRange και στο πεδίο attribute επιλέγεις σε ποια στήλη θέλεις να γίνει ο έλεγχος. Επίσης η επιλογή first-last σημαίνει όλες τις στήλες και εμφανίζει δυο στήλες στο τέλος που περιέχουν τα outliers και τα extreme value.

Στη συνέχεια, πάλι από το filter-instance-removeWithvalue στην επιλογή attribute επιλέγουμε την στήλη 14 και στην επιλογή nominal indices βάζουμε την επιλογή last καθώς είναι το τελευταίο πεδίο όπως δείχνει και η εικόνα και αφαιρούμε 48 τιμές. Δυστυχώς όμως το μοντέλο μας δεν βελτιώθηκε.



Εικόνα 4. 18 Απαλοιφή Extreme_Value και Outlier μέσω weka

```
Linear Regression Model

Large_Burned_Area =

     -0.2155 * spring +
      0.0573 * Noon +
     -0.1388 * Q2_Vraioi_Mantoudi +
      0.2935 * Q4_karustos +
     -0.0101 * TEMP +
     -0.2399 * E +
     -0.1353 * NW +
      0.487

Regression Analysis:

Variable               Coefficient     SE of Coef        t-Stat
spring                    -0.2155         0.046         -4.687
Noon                       0.0573         0.0385         1.49
Q2_Vraioi_Mantoudi        -0.1388         0.0613        -2.2626
Q4_karustos                0.2935         0.0579         5.0709
TEMP                      -0.0101         0.003         -3.3647
E                         -0.2399         0.1436        -1.6704
NW                        -0.1353         0.0769        -1.7584
const                      0.487          0.0669         7.2815

Degrees of freedom = 520
R^2 value = 0.1135
Adjusted R^2 = 0.10154
F-statistic = 9.5084

Time taken to build model: 0.01 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.2167
Mean absolute error                      0.3737
Root mean squared error                  0.4413
Relative absolute error                 92.8215 %
Root relative squared error             98.2508 %
Total Number of Instances              528
```

Εικόνα 4. 19 Αποτελέσματα Linear Regression χωρίς outlier και extreme value

4.3 Algorithm J48

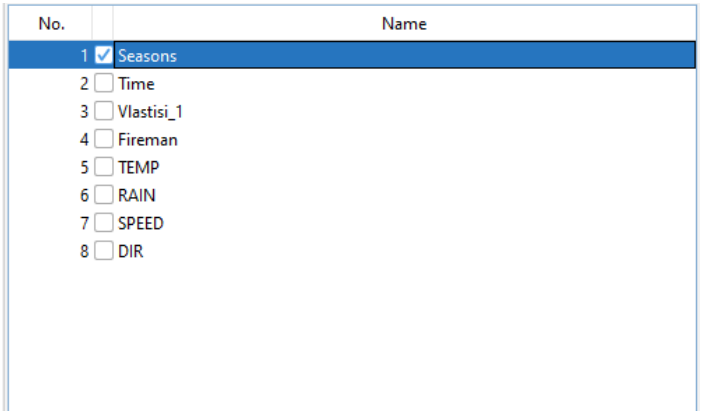
Στην ενότητα αυτή θα εφαρμόσουμε τον αλγόριθμο J48 στα δεδομένα μας με στόχο την δημιουργία ενός decision tree για τον καλύτερο καταμερισμό των δυνάμεων της πυροσβεστικής. Για τον σκοπό αυτό δημιουργήσαμε 2 διαφορετικά σενάρια. Στο πρώτο σενάριο το μοντέλο μας δεν θα γνωρίζει αρχικά ποσά στρέμματα έχουν καεί στην περιοχή παλαιοτέρα, ούτε την διάρκεια της πυρκαγιάς ενώ στο δεύτερο σενάριο θα περιλαμβάνονται όλες οι διαθέσιμες πληροφορίες. Ο διαχωρισμός αυτός γίνεται ώστε να συγκρίνουμε την βελτίωση του μοντέλου στην προσθήκη όλων των παραμέτρων. Οι μεταβλητές που συμμετέχουν στο δέντρο απόφασης είναι οι παρακάτω (σχήμα 4.19) :

- Fireman (εξαρτημένη μεταβλητή): Είναι η μεταβλητή που θέλουμε να προβλέψουμε και αρχικά την χωρίσαμε σε:
    - Low_Force 2-9 πυροσβέστες
    - Normal_Force 10-27 πυροσβέστες
    - Full_Force 28-2661 πυροσβέστες

Το συγκεκριμένο κατώφλι τιμών προέκυψε μέσω δοκίμων και αξιολόγησης των αποτελεσμάτων για την παρούσα έρευνα καθώς στην βιβλιογραφία δεν υπάρχει συγκεκριμένος αριθμός που να χαρακτηρίζει μεγάλη η μικρή μια πυροσβεστική δύναμη. Συνεπώς δεν βασίζεται σε στατιστικούς υπολογισμούς αλλά είναι μια πιο πειραματική προσέγγιση όπου δοκιμάσαμε διάφορες τιμές ως κατώτατα όρια και επιλέξαμε την βέλτιστη. Τέλος, επιλέξαμε να μην προσθέσουμε τα μηχανοκίνητα μέσα καθώς δεν ήταν συμπληρωμένα για όλες τις πυρκαγιές και αυτό θα επηρέαζε τα αποτελέσματα.

- Seasons (ανεξάρτητη μεταβλητή): Είναι οι εποχές του χρόνου κωδικοποιημένες ανάλογα με τον μηνά που ξεκίνησε μια πυρκαγιά. Η κωδικοποίηση είναι Winter (Δεκέμβριος, Ιανουάριος, Φεβρουάριος), Spring (Μάρτιος, Απρίλιος, Μάιος), Summer (Ιούνιος, Ιούλιος, Αύγουστος), Autumn (Σεπτέμβριος, Οκτώβριος, Νοέμβριος). Προχωρήσαμε στην παραπάνω κωδικοποίηση καθώς, τα αποτελέσματα στο Weka ήταν πολύ καλυτέρα σε σύγκριση με την εξέταση του κάθε μηνά ξεχωριστά.
- Time (ανεξάρτητη μεταβλητή): Στην μεταβλητή αυτή χωρίσαμε σε ίσα διαστήματα τον χρόνο που ξεκίνησε η πυρκαγιά ώστε να μελετηθεί καλυτέρα η συγκεκριμένη μεταβλητή. Η τιμές που ορίσαμε είναι: Afternoon (16:00-19:59), Morning (08:00-11:59), Noon (12:00-15:59), Early (04:00-07:59), MidNight (00:00-03:59) και η συγκεκριμένη ομαδοποίηση έγινε πάλι μέσω διάφορων δοκίμων και καταλήξαμε σε αυτό το χρονικό διαχωρισμό.
- Vlastisi_1 (ανεξάρτητη μεταβλητή): Είναι η κωδικοποίηση της βλάστησής που αναλύσαμε σε προηγουμένη ενότητα και παίρνει τις παρακάτω τιμές Q3_Istiaia, Q4_karustos, Q5_Central, Q1_skuros, Q2_Vraioi_Mantoudi
- Temp (ανεξάρτητη μεταβλητή): Είναι η θερμοκρασία σε βαθμούς κελσίου που αποκτήσαμε από τους μετεωρολογικός σταθμούς της Ευβοίας.
- Rain (ανεξάρτητη μεταβλητή): Η βροχή υπολογισμένη σε χιλιοστά (mm)
- SPEED (ανεξάρτητη μεταβλητή): Η ταχύτητα του ανέμου υπολογισμένη σε χιλιόμετρα ανά ώρα (km/h)
- Dir (ανεξάρτητη μεταβλητή): Η κατεύθυνση του ανέμου υπολογισμένη σε μοίρες και πιο συγκεκριμένα:

- 0° — βόρειος άνεμος (N)
- 22.5° — βόρειος-βορειοανατολικός άνεμος (NNE)
- 45° — βορειοανατολικός άνεμος (NE)
- 67.5° — ανατολικός-βορειοανατολικός άνεμος (ENE)
- 90°— ανατολικός άνεμος (E)
- 112.5° — ανατολικός-νοτιοανατολικός άνεμος (ESE)
- 135° — νοτιοανατολικός άνεμος (SE)
- 157.5° — νότιος-νοτιοανατολικός άνεμος (SSE)
- 180° — νότιος άνεμος (S)
- 202.5° — νότιος-νοτιοδυτικός άνεμος (SSW)
- 225° — νοτιοδυτικός άνεμος (SW)
- 247.5° — δυτικός-νοτιοδυτικός άνεμος (WSW)
- 270° — δυτικός άνεμος (W)
- 292.5° — δυτικός-βορειοδυτικός άνεμος (WNW)
- 315° —βορειοδυτικός άνεμος (NW)
- 337.5° — βόρειος-βορειοδυτικός άνεμος (NNW)
- 360° — βόρειος άνεμος (N)



Εικόνα 4. 20 Μεταβλητές για την παραγωγή του δέντρου απόφασης

Στο πρώτο σενάριο, υποθέτουμε ότι έχουμε στην διάθεσή μας μόνο τα περιβαλλοντικά στοιχεία, την εποχή, τον χρόνο και βλάστηση. Το arff αρχείο έχει την παρακάτω μορφή:

@attribute Seasons {Winter, spring, summer, Autumn}

@attribute Time {Afternoon, Morning, Noon, Early, MidNight, Night}

@attributeVlastisi_1{Q3_Istiaia,Q4_karustos,Q5_Central,Q1_skuros,Q2_Vraioi_Mantoudi}

@attribute Fireman {Low_Force, Normal_Force, Full_Force}

@attribute TEMP numeric

@attribute Rain numeric

@attribute SPEED numeric

@attribute DIR{ESE, SE,NE, NNENNE, SSE,NNW, WW, E,ENE,S,N,SSW,SW,WNW,NW,WSW}

Τα αποτελέσματα του μοντέλου παρουσιάζονται στον παρακάτω πίνακα:

```
Number of Leaves  :     52

Size of the tree :      70


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         387               67.6573 %
Incorrectly Classified Instances       185               32.3427 %
Kappa statistic                          0.1846
Mean absolute error                      0.2655
Root mean squared error                  0.3877
Relative absolute error                 85.3568 %
Root relative squared error             98.4332 %
Total Number of Instances              572

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,894    0,707    0,727      0,894    0,802      0,236    0,674     0,764     Low_Force
                 0,209    0,107    0,416      0,209    0,278      0,132    0,610     0,391     Normal_Force
                 0,258    0,018    0,444      0,258    0,327      0,311    0,685     0,309     Full_Force
Weighted Avg.    0,677    0,509    0,629      0,677    0,636      0,212    0,658     0,640

=== Confusion Matrix ===

   a   b   c   <-- classified as
 347  40   1 |   a = Low_Force
 112  32   9 |   b = Normal_Force
  18   5   8 |   c = Full_Force
```

Εικόνα 4. 21 Αποτελέσματα πρώτου δέντρου απόφασης με την χρήση J48

Το μοντέλο μας έχει ποσοστό σωστών προβλέψεων 67,65 το οποίο δεν είναι υψηλό με δεδομένο ότι το precision είναι στο 62,9%.

Στην συνέχεια προσπαθήσαμε να βελτιώσουμε τα στατιστικά μεγέθη του μοντέλου μας, ομαδοποιώντας τα δεδομένα μας και πιο συγκεκριμένα ενοποιήσαμε τις μεταβλητές Normal_Force και το Full_Force και η καινούργια μεταβλητή ονομάστηκε Full_Force η οποία έχει τιμή από 10 πυροσβέστες και πάνω. Επιπρόσθετα, προχωρήσαμε και σε μια ομαδοποίηση των τιμών της κατεύθυνσης του αέρα που περιγράφεται στον πίνακα 4.6

και πιο συγκεκριμένα την ονομάσαμε Dir και οι τιμές της είναι οι ακόλουθες { new_S, new_E, new_N, new_W, new_SE, new_NE, new_SW, new_NW }. Στην συνέχεια τρέξαμε τον αλγόριθμο με τις νέες μεταβλητές Παρακάτω παρατίθενται όλες οι μετατροπές που πραγματοποιήσαμε.

old_Dir	New_DIR	Διάστημα-ωρών	Κωδικοποίηση_Weka	WIND_SPEED_Weka	Τιμές	Temp_Weka	Τιμές	Fireman_weka	Πυροσβέστες
ESE	new_SE	00:00-03:59	Midnight	No wind	0.0-1.0 BF	Cold	<=13	Low_Force	2-9
NNE	new_NE	04:00-07:59	Early	Medium	1.1-4.0 BF	Cool	<=25	Normal_Force	10-27
SSE	new_SE	08:00-11:59	Morning	Strong	4.1-7.0 BF	Hot	>25	Full_Force	28-2661
NNW	new_NW	12:00-15:59	Noon	Powerful	7.1-9.0 BF				
ENE	new_SE	16:00-19:59	Afternoon	Stormy	>9.0 BF				
SSW	new_SW	20:00-23:59	Night						
WNW	new_NW								
WSW	new_SW								

Πίνακας 4. 6 Κωδικοποίηση Weka Μεταβλητών για δέντρο απόφασης και Apriori Rules

Τα αποτελέσματα παρουσιάζονται στην παρακάτω εικόνα

```
Number of Leaves  :     45

Size of the tree :      63


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         419               73.2517 %
Incorrectly Classified Instances       153               26.7483 %
Kappa statistic                          0.345
Mean absolute error                      0.3496
Root mean squared error                  0.4494
Relative absolute error                 80.0647 %
Root relative squared error             96.2044 %
Total Number of Instances              572

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,861    0,538    0,771      0,861    0,814      0,352    0,657     0,735     Low_F
                 0,462    0,139    0,612      0,462    0,526      0,352    0,657     0,533     Full_F
Weighted Avg.    0,733    0,410    0,720      0,733    0,721      0,352    0,657     0,670

=== Confusion Matrix ===

   a   b   <-- classified as
 334  54 |   a = Low_F
  99  85 |   b = Full_F
```

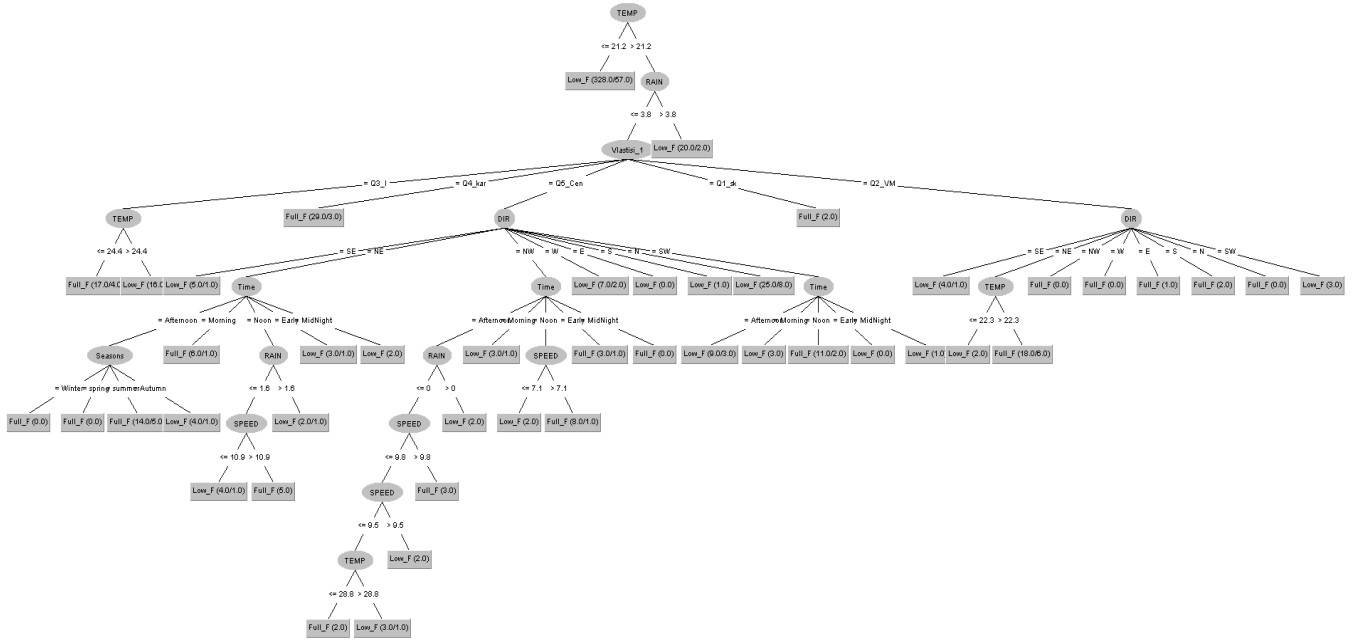
Εικόνα 4. 22 Αποτελέσματα δεύτερου δέντρου απόφασης με την χρήση J48

Το ποσοστό σωστών επιλογών βελτιώθηκε στο 73,25% και το Precision στο 72,00%.

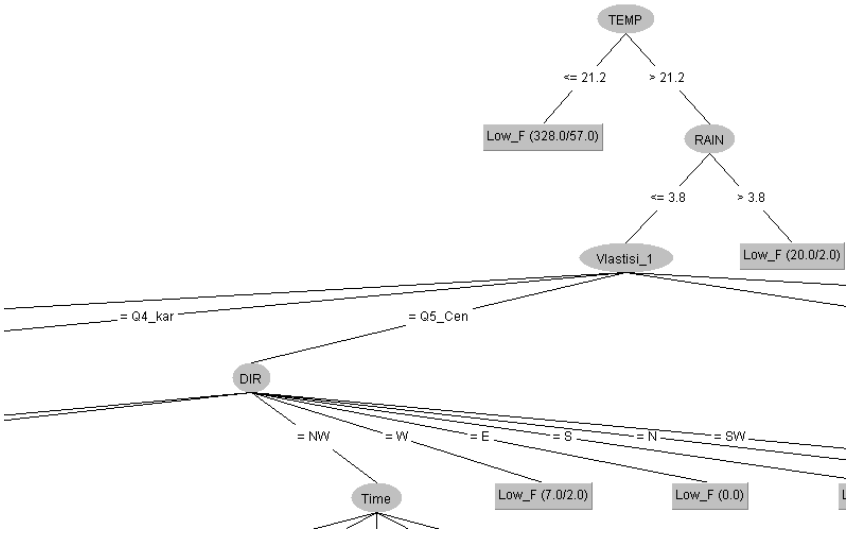
Τέλος, ελέγξαμε και την περίπτωση το Dir να μην υποστεί κάποια επεξεργασία και ενοποιήσαμε το Normal_Force και το Full_Force όπως αναλύθηκε πιο πάνω. Τα αποτελέσματα είναι ελαφρώς πιο χαμηλά σε σχέση με το δεύτερο μοντέλο.

```
Number of Leaves  :     52

Size of the tree :      62


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         417               72.9021 %
Incorrectly Classified Instances       155               27.0979 %
Kappa statistic                          0.3384
Mean absolute error                      0.3524
Root mean squared error                  0.4493
Relative absolute error                 80.694  %
Root relative squared error             96.1923 %
Total Number of Instances              572

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,856    0,538    0,770      0,856    0,811      0,344    0,696     0,785     Low_Force
                 0,462    0,144    0,603      0,462    0,523      0,344    0,696     0,512     Full_Force
Weighted Avg.    0,729    0,411    0,716      0,729    0,718      0,344    0,696     0,697

=== Confusion Matrix ===

   a   b   <-- classified as
 332  56 |   a = Low_Force
  99  85 |   b = Full_Force
```

Εικόνα 4. 22 Αποτελέσματα τρίτου δέντρου απόφασης με την χρήση J48



Εικόνα 4. 23 Οπτικοποίηση δέντρου απόφασης για το δεύτερο μοντέλο.



Εικόνα 4. 24 Δείγμα Δέντρου

Η εικόνα 4.24 απεικονίζει ένα κομμάτι του δέντρου. Στην κορυφή (root) είναι η σημαντικότερη μεταβλητή, η θερμοκρασία. Εάν αυτή είναι κάτω από 21,2 τότε θα χρησιμοποιήσουμε μικρή δύναμη πυρόσβεσης. Οι αριθμοί στο φύλλο με ετικέτα Low_F σημαίνει ότι ταξινομήθηκαν 328 παρατηρήσεις από τις οποίες οι 57 ταξινομήθηκαν εσφαλμένα. Στην συνέχεια για θερμοκρασίες πάνω από 21,2 °C, το δέντρο χρειάζεται να ερμηνεύσει την μεταβλητή rain για να εξηγήσει τι συμβαίνει στην περίπτωση που η

θερμοκρασία ξεπεράσει τους 21,2 °C. Για ποσό μεγαλύτερο του 3.8 βροχής, χρειάζεται μικρή δύναμη για να σβήσει η φωτιά ενώ σε αντίθετη περίπτωση πρέπει να ελέγξει την μεταβλητή βλάστηση. Η παραπάνω διαδικασία συνεχίζεται μέχρι να εξηγηθούν όλες οι μεταβλητές.

Παρακάτω παρουσιάζονται τα αποτελέσματα συγκεντρωτικά για το πρώτο σενάριο:

=== Summary ===

Στήλη1	**Modelo_1**	**Modelo_2**	**Modelo_3**
Correctly Classified Instances	387 || 67,66%	419 || 73,25%	417 || 72,90 %
Incorrectly Classified Instances	185 || 32,34%	153 || 26,74%	155 || 27,10 %
Kappa statistic	0,18	0,35	0,34
Mean absolute error	0,27	0,35	0,35
Root mean squared error	0,39	0,45	0,45
Relative absolute error	85,36%	80,06%	80,70%
Root relative squared error	98,43%	96,20%	96,20%
Total Number of Instances	572	572	572

Πίνακας 4. 7 Σύγκριση «Summary» των 3 δέντρων απόφασης

=== Detailed Accuracy By Class ===

Μοντέλο	**TP_Rate**	**FP_Rate**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC_Area**	**PRC_Area**	**Class**
2	0,861	0,538	0,771	0,861	0,814	0,352	0,657	0,735	Low_Force
2	0,462	0,139	0,612	0,462	0,526	0,352	0,657	0,533	Full_Force
2	0,733	0,410	0,720	0,733	0,721	0,352	0,657	0,670	Weighted_Avg.
3	0,856	0,538	0,770	0,856	0,811	0,344	0,696	0,785	Low_Force
3	0,462	0,144	0,603	0,462	0,523	0,344	0,696	0,512	Full_Force
3	0,729	0,411	0,716	0,729	0,718	0,344	0,696	0,697	Weighted_Avg.
1	0,894	0,707	0,727	0,894	0,802	0,236	0,674	0,764	Low_Force
1	0,209	0,107	0,416	0,209	0,278	0,132	0,610	0,391	Normal_Force
1	0,258	0,018	0,444	0,258	0,327	0,311	0,685	0,309	Full_Force
1	0,677	0,509	0,629	0,677	0,636	0,212	0,658	0,640	Weighted_Avg.

Πίνακας 4. 8 Σύγκριση «Detailed Accuracy By Class» των 3 δέντρων απόφασης

=== Confusion Matrix ===

=== Confusion Matrix ===		
2_μοντελο	**3_μοντελο**	**1_μοντελο**

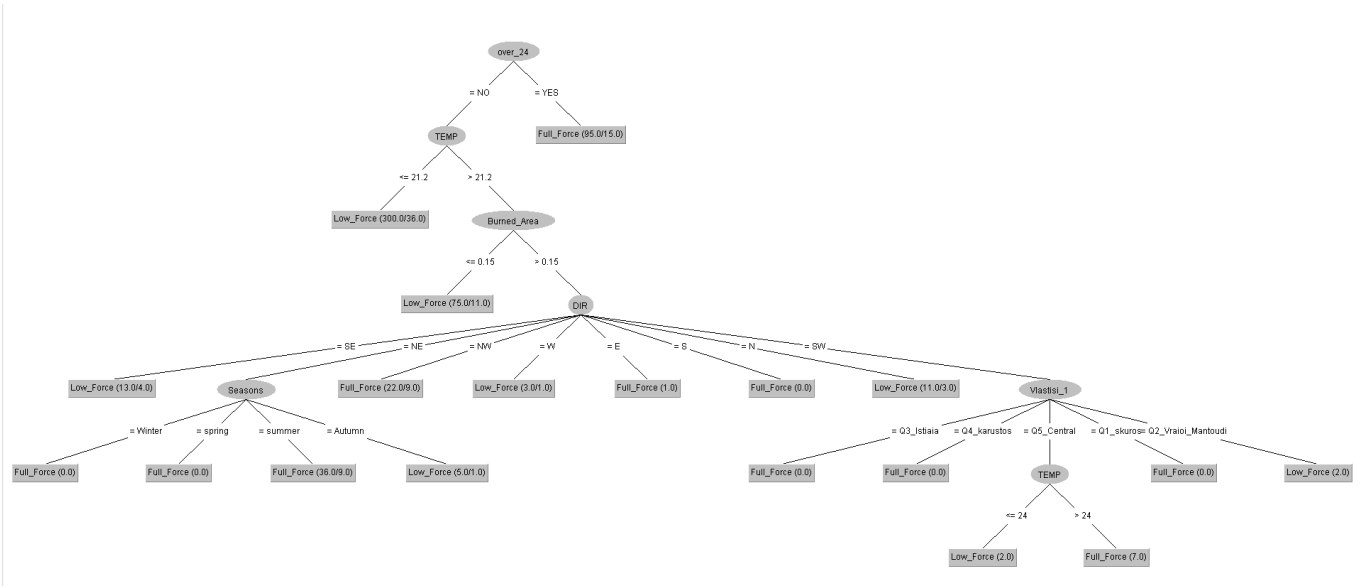
a b <-- classified as	a b <-- classified as	a b c <-- classified as
334 54 | a = Low_F	332 56 | a = Low_Force	347 40 1 | a = Low_Force
99 85 | b = Full_F	99 85 | b = Full_Force	112 32 9 | b = Normal_Force
		18 5 8 | c = Full_Force

Πίνακας 4. 9 Σύγκριση «Confusion Matrix» των 3 δέντρων απόφασης

Για το δεύτερο σενάριο στο οποίο είναι γνωστό ο αριθμός των στρεμμάτων γης που κάηκαν στην περιοχή κατά το παρελθόν και η διάρκεια (duration) της πυρκαγιάς , το μοντέλο μας βελτιώνεται.

```
Number of Leaves  :     19

Size of the tree :      26


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         458               80.0699 %
Incorrectly Classified Instances       114               19.9301 %
Kappa statistic                          0.5271
Mean absolute error                      0.2865
Root mean squared error                  0.4065
Relative absolute error                 65.6128 %
Root relative squared error             87.03   %
Total Number of Instances              572

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,884    0,375    0,833      0,884    0,857      0,530    0,747     0,798     Low_Force
                 0,625    0,116    0,719      0,625    0,669      0,530    0,747     0,593     Full_Force
Weighted Avg.    0,801    0,292    0,796      0,801    0,797      0,530    0,747     0,732

=== Confusion Matrix ===

   a   b   <-- classified as
 343  45 |   a = Low_Force
  69 115 |   b = Full_Force
```

Εικόνα 4. 23 Αποτελέσματα αλγορίθμου j48 με την προσθήκη του Burned Area και duration



Εικόνα 4. 24 Δέντρο απόφασης με την προσθήκη του Burned Area και duration

Συμπερασματικά, τα δυο σενάρια μας δείχνουν ότι με την προσθήκη της καμένης έκτασης και της διάρκειας μιας πυρκαγιάς το δέντρο απόφασης έχει καλύτερη συμπεριφορά και βοηθάει περισσότερο στον καταμερισμό των πυροσβεστικών δυνάμεων από τους αρμοδίους.

4.4 Apriori rules

Με τα δεδομένα που δημιουργήσαμε το δέντρο απόφασης στην προηγουμένη ενότητα, μέσω του αλγορίθμου Apriori θα δημιουργήσουμε κανόνες οι οποίοι θα συνδέουν την καμένη γη, την εποχή, την ώρα, την διάρκεια μιας πυρκαγιάς, την βλάστηση, την πυροσβεστική δύναμη την θερμοκρασία, την ένταση του ανέμου και την κατεύθυνση. Στο συγκεκριμένο κεφάλαιο αφαιρέσαμε την μεταβλητή Βροχή καθώς δεν βρέθηκε κάποιος ενδιαφέρων κανόνας που να την συμπεριλαμβάνει. Η ιδιαιτερότητα του αλγορίθμου Apriori είναι ότι έπρεπε να μετατρέψουμε τις τιμές από αριθμητικές (numeric) σε ονομαστικές (nomimal). Πιο συγκεκριμένα οι μεταβλητές Seasons, Time, Vlastisi, και Dir παρέμειναν ιδίες με την προηγουμένη ενότητα. Οι υπόλοιπες μεταβλητές που χρησιμοποιήσαμε είναι:

- Over_24: Περιέχει τιμές Yes οπού σημαίνει πάνω από 24 ώρες η διάρκεια μιας πυρκαγιάς και No μικρότερη.

- Burn Area (καμένη Γη): Οι πυρκαγιές που έχουν burn area κάτω από 9 στρέμματα χαρακτηριστήκαν ως Small Burn Area ενώ οι υπόλοιπες ως Large Burn Area. Το συγκεκριμένο κατώφλι βρέθηκε και αυτό μέσω δοκίμων με σκοπό την βελτίωση των αποτελεσμάτων.
- Fireman: Όπως και το Burn Area ετσι και η μεταβλητή Fireman βρέθηκε μέσω δοκίμων και περιέχει δυο τιμές το Low_Force το οποίο χαρακτηριστήκαν οι δυνάμεις κάτω από 10 άτομα και το Full_Force πάνω από 10 άτομα.
- Temperature: Επίσης μέσω δοκίμων καταλήξαμε στην παρακάτω κωδικοποίηση της θερμοκρασίας.
    - Cold <=13°C
    - Cool <=25°C
    - Hot >25°C
- Wind_Entasi: Για τον χαρακτηρισμό της συγκεκριμένης μεταβλητής μετατρέψαμε την ταχύτητα του ανέμου της προηγουμένης ενότητας η οποία ήταν υπολογισμένη σε χιλιόμετρα ανά ώρα (km/h) σε κλίμακα Bofor όπως αναφέρεται παρακάτω :
    - No wind: 0.0-1.0 BF (<1 km/h)
    - Medium: 1.1-4.0 BF (1-28 km/h)
    - Strong : 4.1-7.0 BF (29-61 km/h)
    - Powerful : 7.1-9.0 BF (62-88 km/h)
    - Stormy : >9.0 BF (>89 km/h

Στην συνέχεια μετά την κωδικοποίηση των παραπάνω μεταβλητών με σκοπό την παραγωγή καλυτέρων κανόνων, ξεχωρίσαμε τους κάτωθι:

DIR=NE 84 ==> Burned_Area=Small_Burn_Area 64 <conf:(0.76)> lift:(1.2) lev:(0.02) [10] conv:(1.46)

Ο κανόνας αυτός μας δείχνει ότι από τις συνολικά 84 παρατηρήσεις που εμφανίστηκε Βόρειο Ανατολικός Άνεμος, οι 64 είχαν και την τιμή Small_Burn_Area (64/80=0.76). Το conf σημαίνει ότι ο κανόνας επαληθεύεται στο 76% των περιπτώσεων, το 1,2 του lift (δείχνει τον αριθμό της αύξησης της πιθανότητας σε σχέσης με την τυχαιότητα) σημαίνει ότι είναι 1,2 φορές πιθανότερη να συμβεί αυτή η σχέση από την τυχαιότητα και το θέλουμε πάνω από 1. Το Lev (0.02) σημαίνει ότι υπάρχει μια μικρή

απόκλιση της σχέσης από την τυχαιότητα και πρέπει να έχει τιμή πάνω από 0. Το [10] σημαίνει ότι χρησιμοποιήθηκαν δέκα παρατηρήσεις για να υπολογιστούν τα συγκεκριμένα στατιστικά μεγέθη του κανόνα. Τέλος, το Conviction που είναι το γινόμενο της πιθανότητας L να συμβεί και της πιθανότητας R να μην συμβεί, προς την πιθανότητα να συμβούν ταυτόχρονα και το θέλουμε όσο πιο κοντά στο 1, ώστε η σχέση να είναι αξιόπιστη.

Vlastisi_1=Q3_Istiaia 85 ==> Burned_Area=Small_Burn_Area 59 <conf:(0.69)> lift:(1.09) lev:(0.01) [5] conv:(1.15)

Εδώ ο κανόνας επαληθεύτηκε στο 69% τω περιπτώσεων και το lift και το lev είναι οριακά πάνω από τις τιμές που θέλουμε και το conv κοντά στο 1. Εδώ μας λέει ότι η ομάδα βλάστησης Q3, μας δίνει μικρή σε έκταση πυρκαγιά

Vlastisi_1=Q4_karustos 86 ==> Burned_Area=Large_Burn_Area 58 <conf:(0.67)> lift:(1.85) lev:(0.05) [26] conv:(1.88)

Ο κανόνας αυτός μας δείχνει ότι ενώ το conf επίπεδο δεν είναι και τόσο μεγάλο, εάν συγκρίνουμε και τα αλλά στατιστικά μεγέθη δείχνει ότι είναι σημαντικός και η βλάστηση Q4 στην περιοχή της Καρύστου παράγει μεγάλες πυρκαγιές.

Seasons=Winter 100 ==> over_24=NO Fireman=Low_Force 88<conf:(0.88)> lift:(1.23) lev:(0.03) [16] conv:(2.19)

Ο κανόνας αυτός μας δείχνει ότι ο χειμώνας δεν ευνοεί τις μεγάλης διάρκειας φωτιές και ότι χρειάζεται μικρή δύναμη για να τις καταπολεμήσουν.

Seasons=Autumn Vlastisi_1=Q5_Central 78 ==> over_24=NO 71 <conf(0.91)> lift(1.09) lev(0.01) [5] conv(1.61)

Με conf 91% και conv κοντά στο 1 μας λέει οτι οι πυρκαγιές που λαμβάνουν χώρα το φθινόπωρο με βλάστηση Q5_central δεν έχουν μεγάλη διάρκεια και σβήνουν γρηγορά.

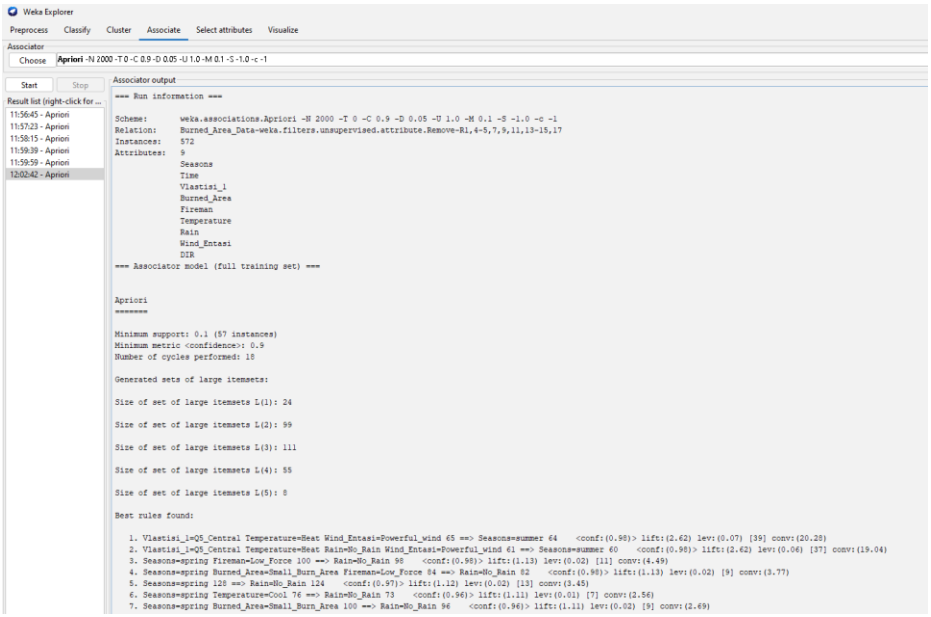
over_24=YES 95 ==>Seasons=summer 62 <conf:(0.65)> lift:(1.75) lev:(0.05) [26] conv:(1.75)

Αυτός ο κανόνας μας λέει ότι όσες πυρκαγιές έλαβαν χώρα στην Εύβοια και είχαν διάρκεια πάνω από μια ημέρα, έγιναν το καλοκαίρι.

Time=Noon Burned_Area=Large_Burn_Area 93 ==> Vlastisi_1=Q5_Central 58 <conf:(0.62)> lift:(1.09) lev:(0.01) [4] conv:(1.1)

Τέλος ο κανόνας αυτός μας αναφέρει με conf 62% (τα υπόλοιπα μετρά lift, lev και conv πληρούν τα κριτήριά μας) ότι τις απογευματινές ώρες όπου κάηκε μεγάλη περιοχή η βλάστηση ήταν Q5_Central.

Παρακάτω δείχνουμε και στο weka πως παράγονται οι κανόνες.



Εικόνα 4. 25 Παράδειγμα χρήσης Apriori rules στο weka

4.4 SequentialPatterns

Σε αντίθεση με τον αλγόριθμο Apriori ο οποίος εξάγει συμπεράσματα αναλογα με την συχνότητα που εμφανίζονται οι συνδυασμοί μεταξύ των μεταβλητών στα υποσύνολα που ελέγχει, ο αλγόριθμος GSP είναι κατάλληλος για ακολουθιακά δεδομένα όπως πχ

χρονοσειρές οι ακολουθίες ενεργειών και αναζητά μοτίβα που εμφανίζονται με συγκεκριμένη σειρά.

Τρέξαμε τον αλγόριθμο στα ιδιά δεδομένα με την προηγουμένη ενότητα και τα αποτελέσματα παρουσιάζονται παρακάτω:

Εικόνα 4. 26 Χρήση του αλγορίθμου GSP στο weka

Ενδιαφέρουσα ακολουθία είναι το {Q3_Istiaia,Small_Burn_Area,Low_Force,ESE} η οποία εμφανίζεται 4 φορές στα δεδομένα μας καθώς μας αποκαλύπτει ότι ίσως και να υπάρχει κάποια σχέση εδώ αναμεσά στην βλάστηση της ευρύτερης περιοχή της Ιστιαίας με τον ESE (Ανατολικός-Νοτιοανατολικό άνεμο) που παράγει Small_Burn_Area και χρειάζεται μικρή δύναμη κατάσβεσης. Αυτή η ακολουθία έχει χρησιμότητα στον καταμερισμό των πυροσβεστικών δυνάμεων σε περίπτωση πυρκαγιάς στην συγκεκριμένη περιοχή.

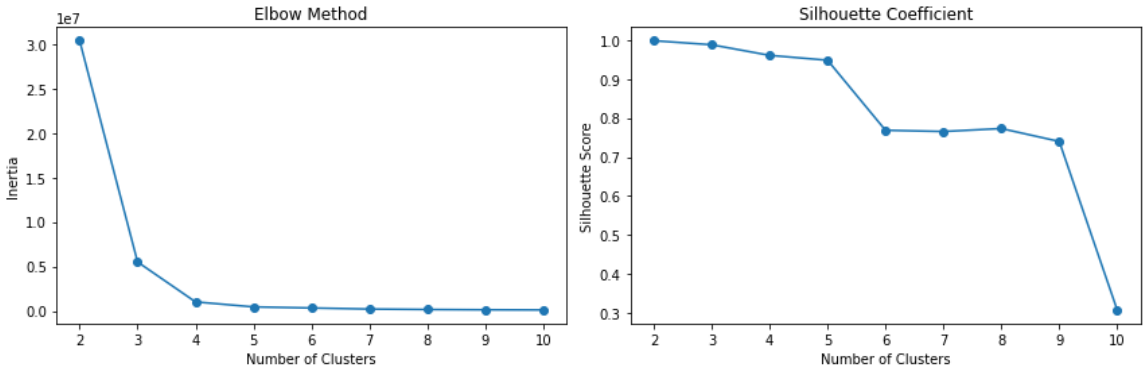
4.5 K-Mean Algorithm Εύβοια

Στην ενότητα αυτή πραγματοποιήσαμε χρήση του αλγορίθμου k-mean με σκοπό να ομαδοποιήσουμε τα δεδομένα μας και να εντοπίσουμε μοτίβα. Στην ενότητα αυτή χρησιμοποιήσαμε τις παρακάτω μεταβλητές:

- Date: Η ημέρα της εβδομάδας που ξεκίνησε η πυρκαγιά. (nomimal)

- Seasons: Η εποχή που ξεκίνησε η πυρκαγιά (ιδιά κωδικοποίηση με τις προηγούμενες ενότητες) (nomimal)
- Time: Η ώρα έναρξης της πυρκαγιάς (ιδιά κωδικοποίηση με τις προηγούμενες ενότητες) (nomimal)
- Vlastisi: Η βλάστηση στην περιοχή της φωτιάς (ιδιά κωδικοποίηση με τις προηγούμενες ενότητες) (nomimal)
- Burn_Area: Η έκταση της καμένης γης (numeric)
- Fireman: Οι πυροσβεστικές δυνάμεις που έλαβαν μέρος στην καταστολή της πυρκαγιάς. (numeric)
- Temp: Η θερμοκρασία σε βαθμούς κελσίου που αποκτήσαμε από τους μετεωρολογικός σταθμούς της Ευβοίας. (numeric)
- SPEED: Η ταχύτητα του ανέμου υπολογισμένη σε χιλιόμετρα ανά ώρα (km/h) (numeric)
- Dir: Η κατεύθυνση του ανέμου υπολογισμένη σε μοίρες (nomimal)

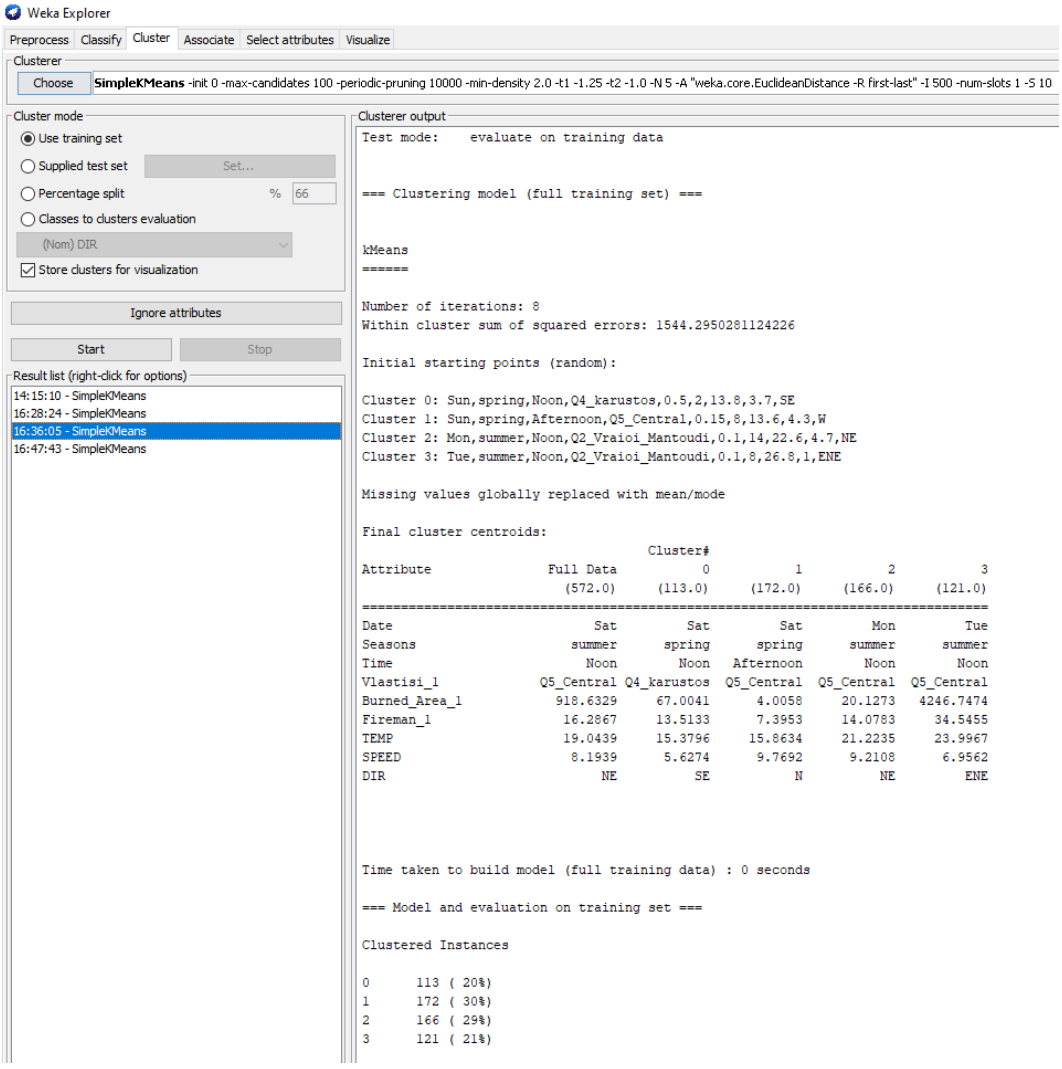
Για να βρούμε τον αριθμό των cluster που θα εισάγουμε στο weka πραγματοποιήσαμε σε python, την μέθοδο Elbow και με το Silhouette Coefficient και ελέγξαμε την σημαντικότητα του. Όσο πιο κοντά στο 1 τόσο το καλύτερο. Επίσης για τις ανάγκες του συγκεκριμένου ελέγχου αλλάξαμε το αρχείο που κάναμε εισαγωγή στο weka ώστε οι κατηγορικές μεταβλητές (πχ η Δευτέρα) είχε την μορφή 1 0 0 0 0 0 0.



Εικόνα 4. 27 Elbow Method και η αξιολόγηση τους

Από το σχήμα 4.27 βλέπουμε ότι το κ=3 και κ=4 είναι οι αριθμοί με το καλύτερο silhouette Coefficient (κοντά στο 1). Στην συνέχεια επιλέξαμε να δημιουργήσουμε 4

cluster αντί 3 καθώς τα 4 cluster έχουν μικρότερο σφάλμα στο weka (Within cluster sum of squared errors)



Εικόνα 4. 28 K-means algorithm για την Εύβοια

Τα αποτελέσματα συνοπτικά σε έναν πίνακα:

Attribute	Full_Data (572.0)	0 (113.0)	1 (172.0)	2 (166.0)	3 (121.0)
Date	Sat	Sat	Sat	Mon	Tue
Seasons	summer	spring	spring	summer	summer
Time	Noon	Noon	Afternoon	Noon	Noon
Vlastisi_1	Q5_Central	Q4_karustos	Q5_Central	Q5_Central	Q5_Central
Burned_Area_1	918.6329	67.0041	4.0058	20.1273	4246.7474
Fireman_1	16.2867	13.5133	7.3953	14.0783	34.5455
TEMP	19.0439	15.3796	15.8634	21.2235	23.9967
SPEED	8.1939	5.6274	9.7692	9.2108	6.9562

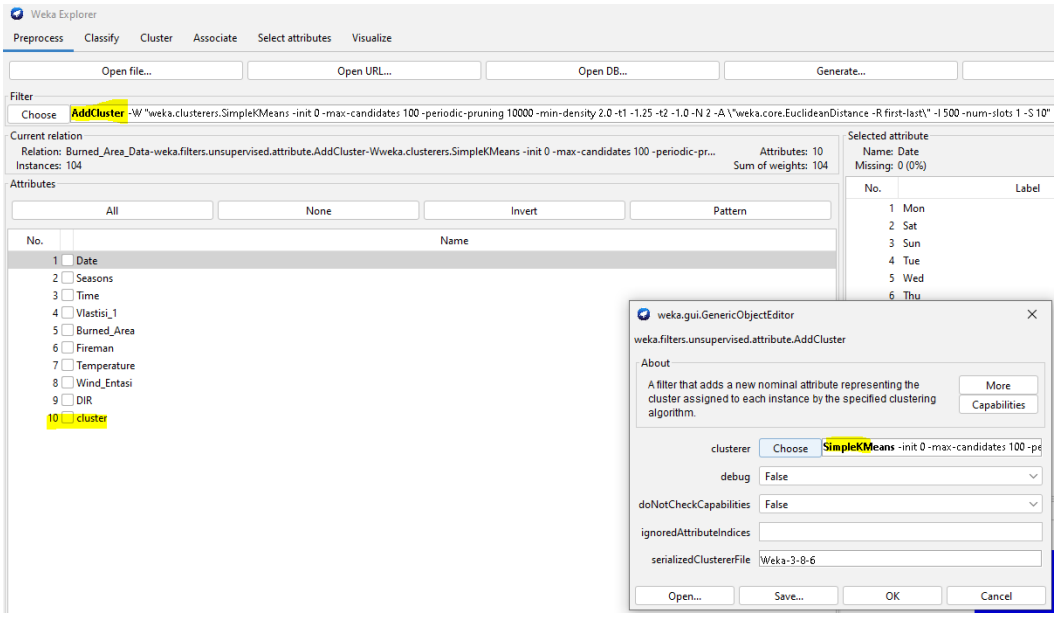
DIR	NE	SE	N	NE	ENE

Πίνακας 4. 10 Αποτελέσματα K-mean Algorithm στην Εύβοια

Ο αλγόριθμος τέλειωσε μετά από 8 επαναλήψεις και το Within cluster sum of squared errors που μας δείχνει πόσο καλά ομαδοποιήθηκαν τα δεδομένα μας είναι 1544,29. Το cluster 3 είναι το πιο επικίνδυνο από τα υπόλοιπα καθώς έχει το μεγαλύτερο Burning_Area και μια από τις μικρότερες ταχύτητες ανέμου. Επίσης η κατεύθυνση του ανέμου είναι ENE και η θερμοκρασία είναι στους 23,99°C. Πολύ ενδιαφέρον είναι και το cluster 0 το οποίο λαμβάνει χωρά την άνοιξη και έχει την μικρότερη θερμοκρασία δείχνοντας ότι οι πυρκαγιές δεν είναι μόνο καλοκαιρινό φαινόμενο αλλά λαμβάνουνε χωρά και τους υπολοίπους μήνες του χρόνου. Επίσης μας δείχνει ότι και με χαμηλή θερμοκρασία και με έναν μέτριο άνεμο η Βλάστηση Q4_karustos μας δίνει μεγάλή έκταση καμένης γης. Αυτή η ανάλυση είναι πολύ σημαντική καθώς δίνει την ευκαιρία σε κάθε δήμο να δημιουργήσει μια τοπική στρατηγική πρόληψης, γνωρίζοντας σε ποια συστάδα ανήκει και εκμεταλλευόμενος την ιδιότητα ότι κάθε συστάδα έχει πολύ διαφορετικά στοιχεία από τις υπόλοιπες. Για παράδειγμα, εάν ένας δήμος ανήκει στην πρώτη συστάδα θα πρέπει μέχρι την άνοιξη να έχει τελειώσει με τον καθαρισμό των δασών του καθώς η συγκεκριμένη εποχή είναι χαρακτηριστικό της συστάδας. Στην συνέχεια θα πραγματοποιήσουμε συνδυασμό του k-mean με τον αλγόριθμο apriori ώστε να ανακαλύψουμε συσχετίσεις εντός της κάθε ομάδας.

4.6 Apriori rule εντός κάθε διαφορετικού cluster

Στην ενότητα αυτή θα πραγματοποιήσουμε εφαρμογή του αλγορίθμου apriori σε κάθε cluster της προηγουμένης ενότητας ξεχωριστά, με σκοπό να εντοπίσουμε pattern αλλά και κανόνες μέσα σε αυτά. Μέσω της δυνατότητας που μας δίνεται από το weka με την εντολή add_cluster, επιλέγουμε τον αλγόριθμο k-mean algorithm και την ευκλείδεια απόσταση (δοκιμάσαμε και την απόσταση Manhattan και τα αποτελέσματα ήταν παρόμοια) και πλέον γνωρίζουμε σε ποιο cluster ανήκει κάθε πυρκαγιά. Οι κανόνες που ξεχωρίσαμε είναι οι ακόλουθοι:



Εικόνα 4. 29 Εισαγωγή της κλάσης cluster στα δεδομένα μέσω weka

rule	**Cluster**
Temperature=Cool DIR=NE 13 ==> Burned_Area=Small_Burn_Area 10 <conf:(0.77)> lift:(1.21) lev:(0.02) [1] conv:(1.19)	cluster 1
Wind_Entasi=Medium_wind DIR=NE 13 ==> Burned_Area=Small_Burn_Area 10 <conf:(0.77)> lift:(1.21) lev:(0.02) [1] conv:(1.19)	cluster 1
DIR=NE 23 ==> Burned_Area=Small_Burn_Area 22 <conf:(0.96)> lift:(1.37) lev:(0.03) [5] conv:(3.46)	cluster 2
Vlastisi_1=Q5_Central DIR=NW 13 ==> Burned_Area=Small_Burn_Area 11 <conf:(0.85)> lift:(1.63) lev:(0.04) [4] conv:(2.08)	cluster 3
DIR=NW 14 ==> Burned_Area=Small_Burn_Area 11 <conf:(0.79)> lift:(1.51) lev:(0.04) [3] conv:(1.68)	cluster 3
DIR=WSW 16 ==> Burned_Area=Small_Burn_Area 10 <conf:(0.63)> lift:(1.2) lev:(0.02) [1] conv:(1.1)	cluster 3

Πίνακας 4. 11 Αποτελέσματα Apriori rule σε ξεχωριστά cluster

Sequent	**Cluster**
Afternoon,NO,Q3_Istiaia,Small_Burn_Area,Low_Force	cluster 3
Afternoon,NO,Q3_Istiaia,Small_Burn_Area,Cool	cluster 3
NO,Q3_Istiaia,Small_Burn_Area,Low_Force	cluster 4

Πίνακας 4. 12 Pattern μεσα στα διαφορετικα cluster.

5 Συμπεράσματα

Η διαχείριση των πυρκαγιών και η βελτίωση των συστημάτων διαχείρισης πυρκαγιών τις τελευταίες δυο δεκαετίες, είναι κρίσιμης σημασίας καθώς οι επιπτώσεις τόσο στο περιβάλλον όσο και στις ανθρώπινες ζωές είναι μεγάλες. Οι κυριότερες τάσεις που υπάρχουν προς σε αυτή την κατεύθυνση σήμερα είναι η χρήση δορυφορικών δεδομένων, η χρήση υπέρυθρων/καπνού και τέλος η χρήση τοπικών αισθητήρων που συλλέγουν δεδομένα (π.χ. μετεωρολογικά). (Cortez & Morais, 2007)

Στο πλαίσιο της παρούσας εργασίας, εφαρμόσαμε τεχνικές εξόρυξης δεδομένων (ΕΔ) χρησιμοποιώντας μετεωρολογικά δεδομένα, εμπλουτίζοντας αυτά των πυρκαγιών που αποκτήσαμε από την πυροσβεστική υπηρεσία για την περιοχή της Ευβοίας. Το πλεονέκτημα της προσέγγισης αυτής είναι ότι σε πραγματικό χρόνο και με χαμηλό κόστος (σε σύγκριση με αυτή του δορυφόρου), επιτρέπει στους τοπικούς αρμοδίους φορείς να εφαρμόζουν αποτελεσματικά τοπικές στρατηγικές πρόληψης και καταστολής των πυρκαγιών.

Πιο συγκεκριμένα, με την ομαδοποίηση των πυρκαγιών σε 4 συστάδες (ο αριθμός προέκυψε από την μέθοδο Elbow και silhouette), δίνεται η δυνατότητα ανάλογα σε ποια συστάδα βρίσκεται ο κάθε δήμος να λάβει τοπικά μέτρα πρόληψης και καταστολής. Επίσης, εφαρμόζοντας τον αλγόριθμο apriori ανακαλύψαμε ορισμένους κανόνες από τα δεδομένα μας όπως ότι τον χειμώνα οι φωτιές έχουν μικρότερη ένταση και καταστέλλονται πιο ευκολά, ότι η βλάστηση Q4_karustos παράγει μεγάλες πυρκαγιές ενώ αντίθετα η Q3_istiaia όχι. Δυστυχώς, τον μόνο κανόνα που δημιουργήσαμε που να συνδέει την καμένη περιοχή με την κατεύθυνση του ανέμου είναι ότι με κατεύθυνση NE καίγεται μια μικρή περιοχή.

Τα αποτελέσματα από τα μοντέλα που προέκυψαν τον αλγόριθμο Random Forest και τον Linear Regression που είχαν ως στόχο την εκτίμηση της καμένης έκτασης δασικών πυρκαγιών, δεν είναι υψηλά και για αυτό τον λόγο δεν μπορούν να αποτελέσουν επιχειρησιακά εργαλεία.

Τα αίτια για τα χαμηλά ποσοστά τόσο των δυο μοντέλων όσο και της αδυναμίας εύρεσης περισσοτέρων χρήσιμων κανόνων συνοψίζονται παρακάτω:

- Τα δεδομένα που αποκτήσαμε από το site της πυροσβεστικής υπηρεσίας, είχαν αρκετές ανακρίβειες τόσο ως προς τον αριθμό των καμένων στρεμμάτων όσο και

ως προς το σημείο εκκίνησης της πυρκαγιάς. Το τελευταίο έχει ως αποτέλεσμα να μην γίνει σωστή καταγραφή της βλάστησης που θα μπορούσε να αποτελέσει σημαντικό στοιχείο για την πρόβλεψη της εξέλιξης μιας πυρκαγιάς.

- Περιορίσαμε το δείγμα μας μόνο στην περίοδο 2020-2022 καθώς για τα προηγούμενα χρόνια τα μετεωρολογικά δεδομένα είτε δεν υπήρχαν είτε ήταν ελλιπέστατα και κατέστη αδύνατον να τα αξιοποιήσουμε. Πιστεύουμε ότι σε μεγαλύτερο δείγμα, θα αποκαλύπτονταν πιο ενδιαφέροντες συσχετίσεις και παραπάνω κανόνες.
- Σε αρκετές περιοχές τα δεδομένα των μετεωρολογικών σταθμών είχαν χαθεί και τα αντικαταστήσαμε με τις τιμές από γειτονικούς σταθμούς. Αυτό έχει ως αποτέλεσμα η ποιότητα των δεδομένων να μην είναι η καλύτερη δυνατή και να μην αντικατοπτρίζει τις πραγματικές συνθήκες έναρξης πυρκαγιάς.
- Το κυριότερο πρόβλημα όμως ήταν ότι οι μετεωρολογικοί σταθμοί στην Εύβοια, δεν καταγράφουν την υγρασία η οποία αποτελεί σημαντική μεταβλητή για την παλινδρόμηση σύμφωνα με την μελέτη του (Guan, 2023), αλλά τους δείκτες Cooling and Heating Degree Days οι οποίοι είναι δείκτες ενέργειας που μας φανερώνουν, την ενέργεια που απαιτείται για την ψύξη και την θέρμανση ενός κτηρίου.
- Ενδιαφέρον θα είχε και η προσθήκη των παραμέτρων της κλίσης του εδάφους καθώς και ο αριθμός παράλληλων πυρκαγιών που ήταν σε εξέλιξη σε κοντινές περιοχές. Για παράδειγμα, παράλληλα με τις φωτιές στην Εύβοια υπήρχε και ένα πολύ μεγάλο μέτωπο στην περιοχή της Βαρυμπόμπης λίγο έξω από την Αθηνά και στην αρχαιά Ολυμπία που επηρέασε τον διαθέσιμο αριθμό πυροσβεστικών οχημάτων και εναέριων μέσων.
- Τέλος μια παράμετρος που δεν υπήρχε τρόπος να την υπολογίσουμε με τα δεδομένα που είχαμε στην διάθεση μας και επηρέασε τα αποτελέσματα της έρευνάς μας, είναι ότι οι πυρκαγιές δημιουργούν το δικό τους κλίμα γύρω από την φωτιά. Αυτό έχει ως αποτέλεσμα να διαφέρει σημαντικά η ταχύτητα του άνεμου που κατέγραψαν οι μετεωρολογικοί σταθμοί με αυτή που υπήρχε πολύ κοντά στην φωτιά.

Σε αντίθεση με τα μοντέλα πρόβλεψης, το δέντρο απόφασης που κατασκευάσαμε με την βοήθεια του αλγορίθμου j48 είχε καλυτέρα αποτελέσματα και στην περίπτωση που τροφοδοτηθεί με καλύτερης ποιότητας δεδομένα, μπορεί να αποτελέσει επιχειρησιακό εργαλείο. Πιο συγκεκριμένα ,προχωρήσαμε στην δημιουργία δυο διαφορετικών σεναρίων, στο πρώτο δεν ήταν διαθέσιμα τα δεδομένα της έκτασης καμένης γης και της διάρκειας της πυρκαγιάς πάρα μόνο τα μετεωρολογικά δεδομένα, η βλάστηση και η εποχή. Για την βελτίωση των αποτελεσμάτων προχωρήσαμε σε μια ομαδοποίηση της κατεύθυνσης του ανέμου και το ποσοστό των σωστών επιλογών έφτασε 73,25%. Με την προσθήκη της καμένης έκτασης και της διάρκειας το ποσοστό έφτασε στο 80%.
Οι παραπάνω Τεχνικές εξόρυξης, μέσω του εργαλείου Weka και με τον συνδυασμό καλύτερης ποιότητας δεδομένων, μπορούν να βελτιώσουν σημαντικά τα αποτελέσματα. Για αυτό και η καλύτερη συλλογή και οργάνωση των δεδομένων στην Ελλάδα είναι κάτι παραπάνω από ζωτικής σημασίας και θα αποτελούσε ένα πολύ χρήσιμο εργαλείο σε εθνικό επίπεδο για την πρόληψη και την καταπολέμηση των πυρκαγιών.

Bibliography

Ahmad, P. H., & Dang, S. (2015). Performance Evaluation of Clustering Algorithm Using Different Datasets.

Alexandridis, A., Russo, L., Vakalis, D., & Siettos, C. I. (2011). Simulation of Wildland Fires in Large-Scale Eterogeneous. *10th International Conference on Chemical and Process Engineering.*

Alexandridis, A., Vakalis, D., Siettos, C., & Bafa, G. (2008). A cellular automata model for forest fire spread prediction: The case.

Bowman, D., Kolden, C., Abatzoglou, J., Johnston, F., Werf, G., Flannigan, M., & Bowman, D. (2020). Vegetation fires in the Anthropocene. *Nature Reviews Earth & Environment*.

Chandel, A., Sarwat, W., Najah, A., Dhanagare, S., & Agarwala, M. (2022). Evaluating methods to map burned area at 30-meter resolution in forests and agricultural areas of Central India.

Corrales-Suastegui, A., Ruiz-Alvarez, O., Torres-Alavez, J. A., & Pavia, E. G. (2021). Analysis of Cooling and Heating Degree Days over Mexico in Present and Future Climate.

Cortez, P., & Morais, A. (2007). A Data Mining Approach to Predict Forest Fires.

Dafallah, F., Elhassan, M., & Ahamed, Y. M. (2020). Compare Clustering Algorithms of Weka Tool.

Daniau, A., d'Errico, F., & Sánchez Goñi, M. (2010). Testing the hypothesis of fire use for ecosystem management by neanderthal and upper palaeolithic modern human populations. *PLoS One*.

Dimitrakopoulos, A., Gogi, C., Stamatelos, G., & Mitsopoulos, I. (2011). Statistical analysis of the fire environment of large forest fires (1000 ha) in Greece.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Data Mining to Knowledge Discovery in Databases.

Gayathri, N. S. (2018). Performance and Classification Evaluation of J48 Algorithm and Kendall's Based J48 Algorithm (KNJ48).

Guan, R. (2023). Predicting Forest Fire with Linear Regression and Random Forest.

Hidayati, I. C., Nalaratih, N., Shabrina, A., Wahyuni, I. N., & Latifah, A. L. (2020). Correlation of Climate Variability and Burned Area in Borneo using Clustering Methods.

Kanageswari, V., & Pethalakshmi, A. (2017). A Novel Approach of Clustering Using COBWEB.

Khairani, N. A., & Sutoyo, E. (2020). Application of K-Means Clustering Algorithm for Determination of Fire-Prone Areas Utilizing Hotspots in West Kalimantan Province.

Kumar, S. (n.d.). *towardsdatascience.com*. Retrieved from https://towardsdatascience.com/hierarchical-clustering-agglomerative-and-divisive-explained-342e6b20d710.

Kusak, L., Unel, F. B., Alptekin, A., Celik, M. O., & Yakar, M. (2021). Apriori association rule and K-means clustering algorithms for interpretation of pre-event landslide areas and landslide inventory mapping.

Liu, B. (2011). Web data mining. Exploring hyperlinks contents and usage data .

Mann, P. S. (2011). Introductory Statistics- In Box-and-Whisker Plot. pp. 123-125.

McCallum, A., Nigam, K., & Ungar, L. H. (2000). Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching.

Mitsopoulos, I., Mallinis, G., Karali, A., Giannakopoulos, C., & Arianoutsou, M. (2015). Mapping fire behaviour in a Mediterranean landscape under different future climate.

Panda, M., & Patra, M. R. (2009). Hudrid Clustering Approach for Network Instrision detection using Cobweb and FFT.

Penney, G., & Richardson, S. (2019). Modelling of the Radiant Heat Flux and Rate of Spread of Wildfire within the Urban Environment.

San-Miguel-Ayanz, J., Moreno, J. M., & Camia, A. (2013). Analysis of large fires in European Mediterranean landscapes: Lessons learned and perspectives. *Forest Ecology and Management*.

Scott, J. H. (2012). *Introduction to Wildfire Behavior Modeling. National Interagency Fuels, Fire, & Vegetation Technology Transfer.*

Slimani, T., & Lazzez, A. (2013). SEQUENTIAL MINING: PATTERNS AND ALGORITHMS ANALYSIS.

Trucchia, A., D'Andrea, M., Baghino, F., & Fiorucci, P. (2020). PROPAGATOR: An Operational Cellular-Automata Based.

Tutmez, B., Ozdogan, M. G., & Boran, A. (2016). Mapping forest fires by nonparametric clustering analysis.

Γκουρμπάτσης, Α. (2014). Εγκληματικότητα των Εμπρησμών και Προφίλ των Εμπρηστών (2000 – 2010).

Γκουρμπάτσης, Α. (2015). *Το κόστος της δασοπυρόσβεσης στην Ελλάδα.*

Καρδαρά, Α. (2020). *Απεικονίζοντας το εγκληματικό προφίλ των εμπρηστών*. Retrieved from postmodern.gr: https://www.postmodern.gr/2020/07/19/apeikonizontas-to-egklimatiko-profi/

Ξανθόπουλος, Γ. (2016). Οι δασικές πυρκαγιές, η διαχείρισή τους στην Ελλάδα και το αποτύπωμά της στην Αττική.

Παπαγεωργίου, Α., Καρέτσος, Γ., & Κατσαδωράκης, Γ. (2012). Το δάσος: Μια ολοκληρωμένη προσέγγιση.

Ταμπάκη, Σ., & Καρανικόλα, Π. (2015). Δασικές Πυρκαγιές και Κοινωνία.

Τσαγκάρη, Κ., Καρέτσος, Γ., & Προύτσος, Ν. (2011). Δασικές Πυρκαγιές Ελλάδας 1983-2008. *WWF Ελλάς και ΕΘΙΑΓΕ-ΙΜΔΟ*.

Φωτιές – Meteo: Κάηκε το ένα τρίτο των δασών στην Εύβοια. (n.d.). Retrieved from https://www.kathimerini.gr/society/561478774/foties-meteo-kaike-to-ena-trito-ton-dason-stin-eyvoia/

Ηλεκτρονικές Διευθύνσεις

- https://dasarxeio.com/2017/11/10/51037/ (τελευταία πρόσβαση στις 9/07/2023)
- http://wwww.minagric.gr/greek/agro_pol/DASIKA/Forests/Forests1.htm (τελευταία πρόσβαση στις 9/07/2023)
- https://wwfeu.awsassets.panda.org/downloads/acf_wwf_miir_policy_report_may_2022_high.pdf (τελευταία πρόσβαση στις 9/07/2023)
- https://www.klimaka.org.gr/milontas-gia-tin-puromania (τελευταία πρόσβαση στις 9/07/2023)
- https://www.kathimerini.gr/society/561478774/foties-meteo-kaike-to-ena-trito-ton-dason-stin-eyvoia (τελευταία πρόσβαση στις 9/07/2023)
- https://www.firesecurity.gr/paragontesdas.html (τελευταία πρόσβαση στις 9/07/2023)
- https://www.huffingtonpost.gr/entry/poia-einai-oi-treis-tepoi-perosvestikon-aeroplanon-poe-chresimopoiei-e-ellada_gr_5d53e43ce4b0c63bcbf043f0 (τελευταία πρόσβαση στις 9/07/2023)

- https://dasarxeio.com/2022/07/17/114579/ (τελευταία πρόσβαση στις 9/07/2023)
- https://www.efsyn.gr/ellada/koinonia/340460_antipyriki-2022-me-kostos-sta-ypsi-kai-tragikes-elleipseis (τελευταία πρόσβαση στις 9/07/2023)
- https://inbox.gr/posa-einai-ta-enaeria-kai-epigeia-mesa-tis-pyrosvestikis/ (τελευταία πρόσβαση στις 9/07/2023)
- https://www.autotypos.gr/auto-news/633730-pyrosbestiki-aytos-tha-einai-o-stolos-se-enaeria-kai-epigeia-mesa-gia-to-kalokairi/ (τελευταία πρόσβαση στις 9/07/2023)
- https://www.orinimelissa.com/2020/08/blog-post_50.html (τελευταία πρόσβαση στις 9/07/2023)
- https://www.epiruspost.gr/giati-den-timoroyntai-oi-empristes-ton-dason/ (τελευταία πρόσβαση στις 9/07/2023)
- https://dasarxeio.com/2021/12/20/106118/ (τελευταία πρόσβαση στις 9/07/2023)
- https://greenagenda.gr/%CE%B7-%CE%B4%CE%B9%CE%B1%CF%87%CE%B5%CE%AF%CF%81%CE%B9%CF%83%CE%B7-%CF%84%CE%BF%CF%85-%CE%BE%CF%8D%CE%BB%CE%BF%CF%85-%CF%83%CF%84%CE%B9%CF%82-%CE%BA%CE%B1%CE%BC%CE%AD%CE%BD%CE%B5%CF%82-%CE%B4%CE%B1/ (τελευταία πρόσβαση στις 9/07/2023)
- https://www.evima.gr/oikonomia/evvoia-800-evro-to-mina-gia-tous-ritinoparagogous-pos-tha-lavoun/ (τελευταία πρόσβαση στις 9/07/2023)
- https://lab.imedd.org/oi-dasikes-pyrkagies-tou-2019/
- https://www.fireservice.gr/el/synola-dedomenon
- https://kr-uttam.medium.com/hierarchical-clustering-a-practical-introduction-of-agglomerative-and-divisive-methods-f7c173158d5b (τελευταία πρόσβαση στις 21/08/2023)
- https://lab.imedd.org/oi-dasikes-pyrkagies-tou-2019/ (τελευταία πρόσβαση στις 21/08/2023)
- https://helppost.gr/kairos/anemologio-dieythinsi-onomata-anemon/ (τελευταία πρόσβαση στις 21/08/2023)
- https://www.mykosmos.gr/loc_mk/metatropeas-anemou.asp (τελευταία πρόσβαση στις 21/08/2023)
- http://learnline.cdu.edu.au/units/env207/fundamentals/weather.html (τελευταία πρόσβαση στις 02/09/2023)
- https://www.evima.gr/oikonomia/evvoia-800-evro-to-mina-gia-tous-ritinoparagogous-pos-tha-lavoun (τελευταία πρόσβαση στις 21/08/2023)
- *Φωτιές – Meteo: Κάηκε το ένα τρίτο των δασών στην Εύβοια*. (χ.χ.). Ανάκτηση από https://www.kathimerini.gr/society/561478774/foties-meteo-kaike-to-ena-trito-ton-dason-stin-evvoia/ (τελευταία πρόσβαση στις 21/08/2023)