ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«ΣΥΓΚΡΙΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ: Η ΠΕΡΙΠΤΩΣΗ ΤΗΣ ΠΡΟΒΛΕΨΗΣ ΔΑΣΙΚΩΝ ΠΥΡΚΑΓΙΩΝ»

Βασίλειος Παναγιώτης Τιμούδας
ΑΜ: 171066

Επιβλέπων: Χρήστος Τρούσσας

UNIVERSITY OF WEST ATTICA

FACULTY OF ENGINEERING

DEPARTMENT OF INFORMATICS AND COMPUTER ENGINEERING

DIPLOMA THESIS

« COMPARATIVE EVALUATION OF MACHINE LEARNING ALGORITHMS: THE CASE OF FOREST FIRES PREDICTION»

Vasileios Panagiotis Timoudas
RN: 171066

Supervisor: Christos Troussas

Η Διπλωματική Εργασία έγινε αποδεκτή και βαθμολογήθηκε από την εξής τριμελή επιτροπή:

Χρήστος Τρούσσας Επ. Καθηγητής	Ακριβή Κρούσκα Μεταδιδακτορική Ερευνήτρια	Παναγιώτα Τσελέντη ΕΔΙΠ
Christos Troussas Digitally signed by Christos Troussas Date: 2022.07.12 17:25:45 +03'00'	Akrivi Krouska Digitally signed by Akrivi Krouska Date: 2022.07.14 12:36:19 +03'00'	Panagiota Tselenti Digitally signed by Panagiota Tselenti Date: 2022.07.13 13:20:21 +03'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο κάτωθι υπογεγραμμένος Βασίλειος Παναγιώτης Τιμούδας του Σπυρίδωνος, με αριθμό μητρώου 171066 φοιτητής του Πανεπιστημίου Δυτικής Αττικής της Σχολής ΜΗΧΑΝΙΚΩΝ του Τμήματος ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ,

δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.
Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του διπλώματός μου.

Ο Δηλών
Βασίλειος Παναγιώτης Τιμούδας

(Υπογραφή φοιτητή)

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω ιδιαίτερα την μητέρα μου Αναστασία καθώς και τον πατέρα μου Σπύρο για την πολύτιμη βοήθεια τους στην ζωή μου. Επίσης θα ήθελα να ευχαριστήσω την οικογένεια μου για την συνεχή υποστήριξη τους καθώς επίσης και τους φίλους μου από το Erasmus για όλες τις εμπειρίες που περάσαμε μαζί. Τέλος θα ήθελα να ευχαριστήσω τον καθηγητή μου για την ανάθεση του θέματος της παρούσας διπλωματικής εργασίας.

ΠΕΡΙΛΗΨΗ

Οι δασικές πυρκαγιές είναι ένα από τα σημαντικότερα προβλήματα στον πλανήτη μας. Εξαιτίας των δασικών πυρκαγιών καταστρέφονται μεγάλες εκτάσεις δασών. Αυτό έχει ως συνέπεια την καταστροφή τους, την μόλυνση του περιβάλλοντος, την αύξηση της κλιματικής αλλαγής προκαλώντας οικονομικά προβλήματα και απειλώντας την ανθρώπινη ζωή. Με την άνοδο της τεχνολογίας, η μηχανική μάθηση μπορεί να δώσει λύσεις σε ολοένα και περισσότερα προβλήματα όπου ένα από αυτά τα προβλήματα είναι οι δασικές πυρκαγιές. Η μηχανική μάθηση είναι μέρος της τεχνητής νοημοσύνης και έχει την δυνατότητα να βελτιώνεται αυτόματα μέσω της εμπειρίας και της χρήσης των δεδομένων.

Στην παρούσα διπλωματική εργασία, θα γίνει συγκριτική αξιολόγηση αλγόριθμων μηχανικής μάθησης για την πρόβλεψη δασικών πυρκαγιών. Η πρόβλεψη και η συγκριτική αξιολόγηση των αλγόριθμων μηχανικής μάθησης θα γίνει με χρήση της γλώσσας προγραμματισμού Python και μέσω της βιβλιοθήκης Scikit-learn. Αρχικά θα επιλεχθούν ορισμένοι αλγόριθμοι μηχανικής μάθησης για κατηγοριοποίηση. Αυτοί οι αλγόριθμοι μηχανικής μάθησης, ή αλλιώς κατηγοριοποιητές, που θα χρησιμοποιηθούν είναι οι Κ-κοντινότεροι-γείτονες, τα Δέντρα απόφασης, τα Τυχαία δάση, ο AdaBoost, ο Gradient tree bosting, η Λογιστική παλινδρόμηση, το Νευρωνικό δίκτυο πολλών επιπέδων και ο Απλοϊκός Bayes εφαρμόζοντας την κατανομή Bernoulli. Στην συνέχεια αυτοί οι αλγόριθμοι μηχανικής μάθησης για κατηγοριοποίηση θα χρησιμοποιηθούν πάνω σε δεδομένα δύο πόλεων της Αλγερίας. Η έξοδος του συνόλου δεδομένων είναι “not fire” ή “fire”. Αφού χρησιμοποιηθούν στο συγκεκριμένο σύνολο δεδομένων θα γίνει συγκριτική αξιολόγηση μεταξύ τους. Πιο συγκεκριμένα, θα υπολογιστεί η ορθότητα, η ακρίβεια, η ανάκληση και ο αρμονικός μέσος του κάθε αλγόριθμου και στο τέλος θα γίνει συγκριτική αξιολόγηση μεταξύ τους. Ο αλγόριθμος ο οποίος έχει την υψηλότερη απόδοση είναι ο πιο βέλτιστος και είναι ο πιο κατάλληλος για την πρόβλεψη των δασικών πυρκαγιών στο συγκεκριμένο σύνολο δεδομένων. Ο κατηγοριοποιητής RandomForestClassifier είναι ο πιο βέλτιστος έχοντας την υψηλότερη απόδοση συγκριτικά και είναι ο πιο κατάλληλος για την πρόβλεψη των δασικών πυρκαγιών στο σύνολο των δεδομένων της Αλγερίας.

Λέξεις Κλειδιά: δασικές πυρκαγιές, τεχνητή νοημοσύνη, μηχανική μάθηση, αλγόριθμοι μηχανικής μάθησης, συγκριτική αξιολόγηση, κατηγοριοποίηση, κατηγοριοποιητές, Python, Scikit-learn.

ABSTRACT

Forest fires are one of the most important problems in our planet. Due to the forest fires, large areas of forests are destroyed. This has effect on polluting the environment, increasing climate change as well as causing economic problems and threatening human life. With the rise of the technology, machine learning can provide solutions to more and more problems; one of these problems is forest fires. Machine learning is part of artificial intelligence and has the potential to be automatically enhanced through experience and the use of big data.

In the present thesis a comparative evaluation of machine learning algorithms for the prediction of forest fires will be presented. The prediction and comparative evaluation of the machine learning algorithms will be conducted with the use of the Python programming language and through the Scikit-learn library. Initially some machine learning algorithms will be selected for classification. These machine learning algorithms, or classifiers, that will be used are K-nearest-neighbors, Decision Trees, Random Forests, AdaBoost, Gradient tree bosting, Logistic Regression, Multilevel Neural Networks and Naïve Bayes applying the Bernoulli distribution. Then these machine learning algorithms for classification will be used on data from two Algerian cities. The data set output is “not fire” or “fire”. After being applied in this data set, a comparative evaluation will be made between them. More specifically, the accuracy, precision, recall and f1 score of each algorithm will be calculated and at the end a comparative analysis will be performed. The algorithm that has the highest performance is the most optimal and is the most suitable for predicting forest fires in this data set. The classifier RandomForestClassifier is the most optimal having the highest performance and is the most suitable for the prediction of the forest fires in the data set of Algeria.

Key Words: forest fires, artificial intelligence, machine learning, machine learning algorithms, comparative evaluation, classification, classifiers, Python, Scikit-learn.

ΠΕΡΙΕΧΟΜΕΝΑ

Πίνακας Πινάκων

Πίνακας Εικόνων

Πίνακας Εξισώσεων

Πίνακας Γραφικών Παραστάσεων

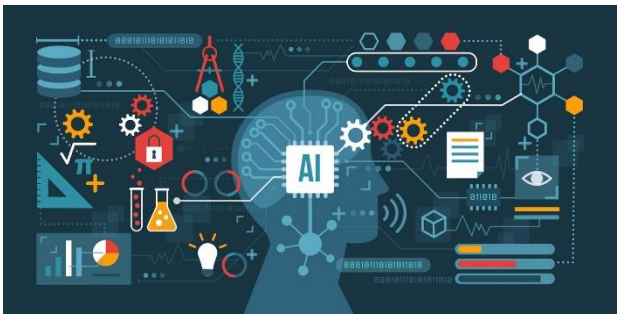
Αλφαβητικό Ευρετήριο

Artificial Intelligence	AI
Artificial Neural Networks	ANN
False Negative	FN
False Positive	FP
K-Nearest Neighbors	kNN
Linear Discriminant Analysis	LDA
Machine Learning	ML
Non-Negative Matrix Factorization	NMF
Principal Component Analysis	PCA
Singular Value Decomposition	SVD
True Negative	TN
True Positive	TP

Κεφάλαιο 1: Εισαγωγή

Σε αυτό το κεφάλαιο θα γίνει εισαγωγή στην διπλωματική εργασία που έχει σκοπό την συγκριτική αξιολόγηση αλγόριθμων μηχανικής μάθησης για την πρόβλεψη των δασικών πυρκαγιών. Αρχικά θα παρουσιαστεί τι είναι η τεχνητή νοημοσύνη και πως μπορεί να συμβάλει σε ορισμένους τομείς όπως η εκπαίδευση και η υγεία με ορισμένα παραδείγματα από έρευνες. Συνεχίζοντας θα παρουσιαστεί το πρόβλημα των δασικών πυρκαγιών εξιγόντας τις συνέπειες συνέπειες που μπορούν να προκληθούν από αυτές. Επίσης θα αναφερθούν οι δασικές πυρκαγιές στην Αλγερία και θα εξηγηθεί ο λόγος ο οποίος επιλέχθηκαν δεδομένα από περιοχές της Αλγερίας για την παρούσα διπλωματική εργασία. Στην συνέχεια θα θα αναφερθεί η αναγκαιότητα του προβλήματος των δασικών πυρκαγιών και θα αναλυθούν τρόπου για το πως μπορεί να αντιμετωπιστεί. Τέλος θα παρουσιαστεί η δομή της διπλωματικής εργασίας.

Με την άνοδο των χρόνων και της τεχνολογίας η τεχνητή νοημοσύνη (artificial intelligence – AI) έχει αρχίσει να εξελίσσεται και να χρησιμοποιείται όλο και περισσότερο στην καθημερινότητα. Η τεχνητή νοημοσύνη είναι η ικανότητα ενός υπολογιστή να εκτελεί εργασίες χωρίς κάποια ανθρώπινη διεπαφή χρησιμοποιώντας νοημοσύνη και ουσιαστικά προσπαθεί να αντιγράψει την ανθρώπινη συμπεριφορά.



Εικόνα 1: *Τεχνητή νοημοσύνη*

Στην τεχνητή νοημοσύνη υπάρχουν πολλές διαφορετικές μορφές μάθησης. Μια μορφή μάθησης είναι η δομική και λάθος (trial and error) όπου ένα απλό πρόγραμμα μπορεί να δοκιμάσει τυχαίες κινήσεις με σκοπό να μάθει από τα λάθη του και να μην τα ξανακάνει. Επίσης το πρόγραμμα να αποθηκεύει κάθε λύση έτσι ώστε αν ξανασυναντήσει την ίδια κατάσταση να γνωρίζει ποιες κινήσεις πρέπει να κάνει για την λύση.

Οι έρευνες πάνω στην τεχνητή νοημοσύνη επικεντρώνονται κυρίως στην μάθηση (learning), στην επίλυση προβλημάτων (problem solving), στον συλλογισμό (reasoning) και στην αντίληψη και χρήση της γλώσσας (perception and use of language). Ένα παράδειγμα τεχνητής νοημοσύνης είναι η αναγνώριση προσώπου και είναι μία από τις κορυφαίες εφαρμογές της. Η έρευνα [1] χρησιμοποίησε ανίχνευση προσώπου έτσι ώστε να εντοπίσει ανθρώπινα πρόσωπα και να ανιχνεύσει εάν φοράνε μάσκα προσώπου σε πραγματικό χρόνο εν μέσω της πανδημίας COVID-19.

Η τεχνητή νοημοσύνη μπορεί να συμβάλει και συνεισφέρει σε πολλούς τομείς. Ένας από αυτούς τους τομείς είναι η εκπαίδευση. Η έρευνα [2] παρουσιάζει ένα παιχνίδι quiz για κινητό όπου είναι μια εφαρμογή εκμάθησης μαθητών στην τριτοβάθμια εκπαίδευση πάνω στον προγραμματισμό χρησιμοποιώντας τεχνικές τεχνητής νοημοσύνης. Μια άλλη έρευνα [3] παρουσιάζει πως μπορεί να γίνει εξόρυξη δεδομένων (data mining) για την βελτίωση της τριτοβάθμιας εκπαίδευσης εν μέσω της πανδημίας COVID-19. Ακόμη μια άλλη έρευνα [4] που χρησιμοποιείται για την εκπαίδευση ενισχύει την αποτελεσματικότητα ευφυών συστημάτων διδασκαλίας (Intelligent Tutoring Systems) με χρήση προσαρμογής και μοντελοποίησης γνωσιακής διάγνωσης.

Εκτός από την εκπαίδευση ένας άλλος πολύ σημαντικός τομής είναι η υγεία. Η έρευνα [5] συμβάλει και στην υγεία αλλά και στην εκπαίδευση. Παρουσιάζει ένα framework επαυξημένης πραγματικότητας με σκοπό την βελτιώση της αναγνωστικής κατανόησης στην ειδική εκπαίδευση (special education).

Ακόμη μερικές ακόμη έρευνες [6], [7] είναι παραδείγματα τεχνητής νοημοσύνης και πιο συγκεκριμένα της μηχανικής μάθησης της οποίας θα γίνει βιβλιογραφική ανασκόπηση στο επόμενο κεφάλαιο. Αυτές οι έρευνες χρησιμοποιούν αλγόριθμους κατηγοριοποίησης μηχανικής μάθησης (classification machine learning algorithms) με σκοπό την συγκριτική αξιολόγηση ανάλυσης. Αρκετές από τους μεθόδους και αλγορίθμους που χρησιμοποιούν θα χρησιμοποιηθούν στην παρούσα διπλωματική εργασία και θα παρουσιαστούν αναλυτικά στα επόμενα κεφάλαια.

Συνεχίζοντας ένας άλλος πολύ σημαντικός τομής που μπορεί να συνεισφέρει η τεχνητή νοημοσύνη το περιβάλλον. Το παρόν θέμα της διπλωματικής εργασίας θα ασχοληθεί με την αντιμετώπιση των δασικών πυρκαγιών οι οποίες είναι ένα πολύ σημαντικό ζήτημα για την προστασία του περιβάλλοντος.

Με την πάροδο των χρόνων η αύξηση των δασικών πυρκαγιών είναι δραματική. Κάθε χρόνο στον πλανήτη μας προκαλούνται χιλιάδες δασικές πυρκαγιές καταστρέφοντας μεγάλες εκτάσεις δασών. Αυτό έχει ως συνέπεια την καταστροφή των δασών, την μόλυνση του περιβάλλοντος και την αύξηση της κλιματικής αλλαγής. Επίσης οι δασικές πυρκαγιές εκτός από οικολογική ζημιά μπορούν να προκαλέσουν οι οικονομικές ζημιές και απειλή προς την ανθρώπινη ζωή [8].

Εικόνα 2: *Δασικές πυρκαγιές*

Μία από τις μεσογειακές χώρες που πλήττεται περισσότερο από τις δασικές πυρκαγιές είναι η Αλγερία. Η Αλγερία επιλέχθηκε καθώς υπάρχουν αρκετά δεδομένα για έρευνα σε πολλές περιοχές της και είναι μια από τις χώρες όπου έχουν χαθεί τεράστιες εκτάσεις δασών εξετίας των δασικών πυρκαγιών. Τα τελευταία χρόνια η Αλγερία είναι μία από τις μεσογειακές χώρες που πλήττεται περισσότερο από καύσωνες κάνοντας τις δασικές πυρκαγιές ακόμα πιο επικίνδυνες. Εξαιτίας των καυσώνων, οι δασικές πυρκαγιές εξαπλώνονται ταχύτερα και σε σημεία όπου έχουν δύσκολη προσβασιμότητα. Με ελάχιστη περιορισμένη πρόσβαση ο εντοπισμός και η αποτελεσματική επέμβαση των πυροσβεστών γίνεται πάρα πολύ δύσκολη. Σύμφωνα με έρευνα [9], στο δάσος Tlemcen της Αλγερίας μετρούνται 1600 πυρκαγιές στην περίοδο 1980 έως 2015. Επιπρόσθετα η Αλγερία είναι μία από τις χώρες που κινδυνεύουν να χάσουν όλες τις δασικές εκτάσεις τους εξαιτίας όλων αυτών των φαινομένων. Δύο μεγάλα ερευνητικά ερωτήματα είναι τι θα γινόταν αν θα μπορούσαμε να προβλέψουμε τις δασικές πυρκαγιές και πως θα γινόταν; Την απάντηση σε αυτά τα δύο ερευνητικά ερωτήματα θα δώσει η παρούσα διπλωματική εργασία.

Οι δασικές πυρκαγιές είναι ένα από τα σημαντικότερα προβλήματα που ταλαιπωρεί τον πλανήτη μας εδώ και πολλά χρόνια. Πιο πριν αναφέρθηκαν τα προβλήματα των δασικών πυρκαγιών. Αν δεν λάβουμε κάποια μέτρα η κατάσταση στον πλανήτη μας θα χειροτερέψει και μπορεί να γίνει μοιραία. Για αυτόν τον λόγω είναι αναγκαίο η αντιμετώπιση των δασικών πυρκαγιών. Η επόμενη παράγραφος θα μιλήσει για την αντιμετώπιση των δασικών πυρκαγιών.

Στην προ προηγούμενη παράγραφο αναφέρθηκαν τα ερευνητικά ερωτήματα τι θα γινόταν αν θα μπορούσαμε να προβλέψουμε τις δασικές πυρκαγιές και πως θα γινόταν. Με την άνοδο της τεχνολογίας η μηχανική μάθηση μπορεί να δώσει λύσεις σε ολοένα και περισσότερα προβλήματα. Έτσι λοιπόν μέσω της μηχανικής μάθησης μπορεί να απαντηθεί το ερευνητικό ερώτημα για το πως μπορούν να προβλεφθούν οι δασικές πυρκαγιές. Στόχος είναι η πρόβλεψη των δασικών πυρκαγιών έτσι ώστε αφού μπορούν να προβληθούν οι δασικές πυρκαγιές να μπορέσουν να ελαχιστοποιηθούν αλλά και γιατί όχι να αποτραπούν. Έτσι θα εξασφαλίζαμε την προστασία των δασικών περιοχών στον πλανήτη μας μειώνοντας την κλιματική αλλαγή καθώς επίσης θα σωνόντουσαν ανθρώπινες ζωές και θα μειωνόντουσαν οι οικονομικές ζημιές που προκαλούν οι δασικές πυρκαγιές. Αυτή είναι η απάντηση στο ερευνητικό ερώτημα τι θα γινόταν αν θα μπορούσαμε να προβλέψουμε τις δασικές πυρκαγιές αλλά για να γίνει αυτό θα πρέπει να δοθεί έμφαση στο πρώτο ερευνητικό ερώτημα για το πως μπορούμε να προβλέψουμε τις δασικές πυρκαγιές. Παρακάτω στην παρούσα διπλωματική εργασία θα ερευνηθεί και θα δοθεί η απάντηση σε αυτήν την πολύ σημαντική ερώτηση.

Στην παρούσα διπλωματική εργασία θα προβλεφθούν δασικές πυρκαγιές με την χρήση των αλγόριθμων μηχανικής μάθησης. Στην συνέχεια τα αποτελέσματα των αλγορίθμων θα αξιολογηθούν βρίσκοντας τον πιο βέλτιστο αλγόριθμο που θα μπορέσει να δώσει λύσει στο ερευνητικό ερώτημα πως μπορεί να προβλεφθούν οι δασικές πυρκαγιές. Αποφάσισα να χωρίσω την διπλωματική εργασία σε 5 κεφάλαια ξεκινώντας από την εισαγωγή στο πρόβλημα των δασικών πυρκαγιών και φτάνοντας στο συμπέρασμα της πρόβλεψης των δασικών πυρκαγιών. Παρακάτω αναφέρεται η δομή της διπλωματικής εργασίας.

- Στο κεφάλαιο 2 θα αναφερθεί όλη η βιβλιογραφική ανασκόπηση της διπλωματικής εργασίας. Θα παρουσιαστεί η μηχανική μάθηση και η θεωρία της. Αφού καταλάβουμε τι είναι μηχανική μάθηση και για ποιον λόγω χρησιμοποιείται θα παρουσιαστούν οι τύποι της. Στην συνέχεια θα αναφερθεί η προεπεξεργασία των δεδομένων όπου είναι απαραίτητη πριν από την χρήση των αλγόριθμων μηχανικής μάθησης και μετά θα παρουσιαστούν οι αλγόριθμοι της μηχανικής μάθησης όπου οι περισσότεροι από αυτούς είναι αλγόριθμοι κατηγοριοποίησης. Τέλος θα αναφερθούν τρόποι για το πως μπορούν να αξιολογηθούν οι αλγόριθμοι μηχανικής μάθησης για κατηγοριοποίηση όπου είναι ο σκοπός της διπλωματικής εργασίας.

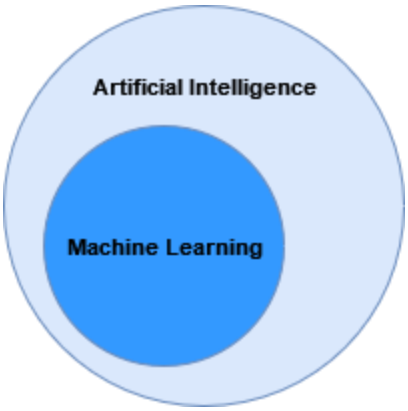
- Στο κεφάλαιο 3 θα αναφερθεί η μεθοδολογία της διπλωματικής εργασίας. Πιο αναλυτικά θα αναφερθεί η μεθοδολογία που θα χρησιμοποιηθεί για να δώσει λύση στο πρόβλημα των δασικών πυρκαγιών. Αρχικά παρουσιάζεται πως μπορεί να γίνει χρήση μηχανικής μάθησης με την γλώσσα προγραμματισμού Python. Εκεί θα παρουσιαστεί η επιλογή περιβάλλοντος, η γλώσσα προγραμματισμού Python και οι βιβλιοθήκες που θα χρησιμοποιηθούν. Στην συνέχεια θα παρουσιαστεί το σύνολο δεδομένων της Αλγερίας και θα γίνει πρόβλεψη των δασικών πυρκαγιών. Συνεχίζοντας θα παρουσιαστούν οι αλγόριθμοι μηχανικής μάθησης ή αλλιώς κατηγοριοποιητές που θα χρησιμοποιηθούν και γιατί καθώς επίσης και οι παράμετροί τους. Τέλος θα παρουσιαστούν τα βήματα υλοποίησης της διπλωματικής εργασίας.
- Στο κεφάλαιο 4 θα παρουσιαστούν τα αποτελέσματα της διπλωματικής εργασίας. Αρχικά θα παρουσιαστούν οι προβλέψεις των αλγόριθμων μηχανικής μάθησης ή αλλιώς κατηγοριοποιητών και στην συνέχεια θα γίνει συγκριτική αξιολόγηση μεταξύ τους. Τέλος θα επιλεχθεί ο πιο βέλτιστος αλγόριθμος μηχανικής μάθησης ή αλλιώς κατηγοριοποιητής που μπορεί να δώσει λύση στο πρόβλημα της πρόβλεψης των δασικών πυρκαγιών για το συγκεκριμένο σύνολο δεδομένων της Αλγερίας.
- Στο κεφάλαιο 5 θα παρουσιαστούν τα συμπεράσματα και προτάσεις για μελλοντικές κατευθύνσεις της διπλωματικής εργασίας. Επίσης θα παρουσιαστούν μερικές προτάσεις για το πως μπορεί να εξελιχθεί η παρούσα διπλωματική εργασία
- Τέλος παρουσιάζεται όλη η βιβλιογραφία της παρούσας διπλωματικής εργασίας όπου χωρίς αυτές θα ήταν αδύνατη η συγγραφή της διπλωματικής εργασίας.

Κεφάλαιο 2: Μηχανική μάθηση: Θεωρητική θεμελίωση και Ανασκόπηση της Βιβλιογραφίας

Στο παρόν κεφάλαιο θα γίνει βιβλιογραφική ανασκόπηση της διπλωματικής εργασίας. Ξεκινώντας θα αναλυθεί τι είναι η μηχανική μάθηση και γιατί χρησιμοποιούμε μηχανική μάθηση με ορισμένα παραδείγματα. Συνεχίζοντας θα παρουσιαστούν οι τύποι της μηχανικής μάθησης με τις υποκατηγορίες τους και θα αναλυθούν αναλυτικά παρουσιάζοντας εφαρμογές και μερικές εταιρίες τις οποίες χρησιμοποιούν σήμερα. Επίσης ένα πολύ σημαντικό κομμάτι του συγκεκριμένου κεφαλαίου που θα αναλυθεί είναι οι αλγόριθμοι μηχανικής μάθησης για κατηγοριοποίηση οι οποίοι θα χρησιμοποιηθούν στην παρούσα διπλωματική εργασία. Τέλος θα αναλυθούν οι τύποι αξιολόγησης των αλγορίθμων μηχανικής μάθησης οι οποίοι αποτελούν το κύριο υλικό της διπλωματικής εργασίας.

2.1 Τι είναι μηχανική μάθηση

Μηχανική μάθηση (Machine Learning - ML) είναι η επιστήμη των υπολογιστών να μπορούν να μαθαίνουν από τα δεδομένα [11]. Αποτελούν ένα κομμάτι της τεχνητής νοημοσύνης (Artificial Intelligence - AI) όπως βλέπουμε στην παρακάτω εικόνα 1 και εστιάζει στην χρήση δεδομένων και αλγορίθμων προσπαθώντας να μάθουν τον τρόπο με τον οποίο μαθαίνουν οι άνθρωποι, βελτιώνοντας σταδιακά την ακρίβειά τους. Το 1959 ο Arthur Samuel όρισε την μηχανική μάθηση ως "πεδίο σπουδών που δίνει στους υπολογιστές τη δυνατότητα μάθησης χωρίς να έχει προγραμματιστεί ρητά" [11]. Ένας άλλος ορισμός που έδωσε ο Tom Mitchell το 1997 "ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από την εμπειρία E σε σχέση με κάποια εργασία T και κάποιο μέτρο απόδοσης P, εάν η απόδοσή του στο T, όπως μετράται με το P, βελτιώνεται με εμπειρία E" [12].



***Εικόνα 3**: Μηχανική μάθηση μέρος της τεχνητής νοημοσύνης*

Οι αλγόριθμοι της μηχανικής μάθησης (machine learning algorithms) δημιουργούν ένα μοντέλο χρησιμοποιώντας δεδομένα τα οποία είναι γνωστά ως δεδομένα εκπαίδευσης με σκοπό να κάνουν προβλέψεις και να λαμβάνουν αποφάσεις χωρίς να είναι προγραμματισμένοι να το κάνουν αυτό. Βλέπουμε την παρακάτω εικόνα η οποία το παρουσιάζει αυτό. Πολλές εταιρίες χρησιμοποιούν αλγόριθμους μηχανικής μάθησης. Μερικές από αυτές είναι η YouTube η οποία είναι μια πλατφόρμα με βίντεο, κατηγοριοποιώντας τα βίντεο τα οποία παρακολουθεί κάθε χρήστης και εμφανίζει προτεινόμενα βίντεο μέσω των αλγόριθμων μηχανικής μάθησης [13]. Άλλη εταιρία είναι η Netflix η οποία με παρόμοιο τρόπο με την YouTube κατηγορποιεί τις ταινίες που παρακολουθούν οι χρήστες και μέσω των αλγόριθμων μηχανικής μάθησης εμφανίζει προτεινόμενες ταινίες [14]. Τέλος οι αλγόριθμοι της μηχανικής μάθησης μπορούν να χρησιμοποιηθούν και σε άλλους τομείς όπως είναι η υγεία, η εκπαίδευση, η αναγνώριση προσώπων και ομιλίας καθώς και πολλά άλλα.

2.2 Γιατί χρησιμοποιούμε μηχανική μάθηση

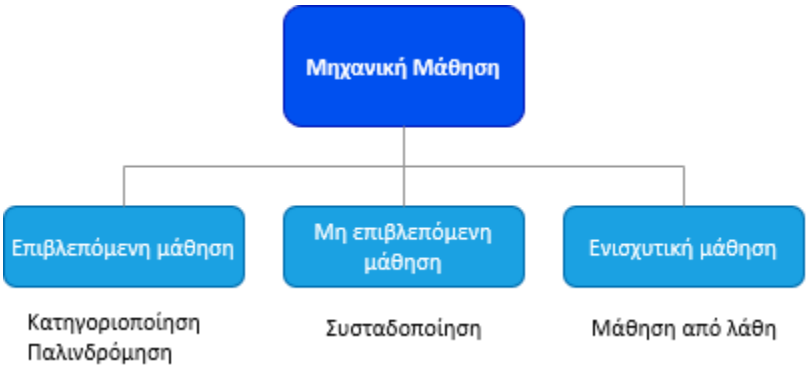
Όπως αναφέρθηκε προηγουμένως η μηχανική μάθηση μπορεί να μαθαίνει από τα δεδομένα και μπορεί να προβλέπει και να βγάζει αποφάσεις. Συγκρίνοντας με το παραδοσιακό μοντέλο όπου τα προγράμματα των υπολογιστών προγραμματίζονται και ξαναπρογραμματίζονται από την αρχή η μηχανική μάθηση δίνει την δυνατότητα στους υπολογιστές να μαθαίνουν αυτόματα από τα δεδομένα χωρίς να χρειαστεί να ξαναπρογραμματιστούν. Αυτός είναι ένας από τους βασικούς λόγους όπου η μηχανική μάθηση χρησιμοποιείται σχεδόν παντού στις μέρες μας. Ορισμένα πλεονεκτήματα της μηχανικής μάθησης είναι [14] [15]:

- μπορεί με ευκολία να αναγνωρίζει μοτίβα

- δεν χρειάζεται ανθρώπινη παρέμβαση
- μπορεί να βελτιώνεται διαρκώς
- δίνει λύσεις σε προβλήματα τα οποία είναι περίπλοκα
- μπορεί να λειτουργήσει σε περίπλοκα προβλήματα τα οποία χρησιμοποιούν τεράστιο αριθμό δεδομένων

2.3 Τύποι μηχανικής μάθησης

Στην μηχανική μάθηση υπάρχουν τρεις τύποι μηχανικής μάθησης όπου η κάθε μία από αυτές είναι διαφορετική. Διακρίνονται στην επιβλεπόμενη μάθηση (supervised learning), στην μη επιβλεπόμενη μάθηση (unsupervised learning) και στην ενισχυτική μάθηση (reinforcement learning). Μπορεί να παρατηρηθεί και η παρακάτω εικόνα η οποία δείχνει όλους αυτούς τους τύπους της μηχανικής μάθησης. Κάθε τύπος είναι διαφορετικός και χρησιμοποιείται διαφορετικά σε κάθε πρόβλημα και περιέχει κάποιες υποκατηγορίες. Τέλος εκτός από τους τρεις τύπους μάθησης θα παρουσιαστούν και ορισμένοι άλλοι τύποι μάθησης οι οποίοι είτε χρησιμοποιούν δύο διαφορετικούς τύπους μάθησης είτε ανήκουν σε κάποια κατηγορία από τους βασικούς τύπους.



Εικόνα 4: *Τύποι μηχανικής μάθησης*

2.3.1 Επιβλεπόμενη μάθηση

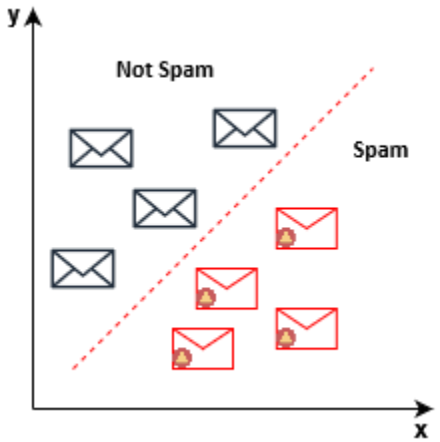
Η επιβλεπόμενη μάθηση (supervised learning) χρησιμοποιεί δεδομένες εισόδους (όπου ονομάζονται σύνολο εκπαίδευσης) σε γνωστές επιθυμητές εξόδους [15]. Στόχος της είναι η εύρεση λύσης σε εισόδους που έχουν άγνωστη έξοδο. Περιέχει δύο υποκατηγορίες, την κατηγοριοποίηση (classification) και την παλινδρόμηση (regression) [18].

2.3.1.1 Κατηγοριοποίηση

Η κατηγοριοποίηση (classification) στην μηχανική μάθηση είναι μια επιβλεπόμενη μάθηση. Βασικός της στόχος είναι να κατηγοριοποιεί τα νέα δεδομένα που προκύπτουν. Ορισμένοι από τους πιο γνωστούς αλγόριθμους της κατηγοριοποίησης είναι οι:

- Κ-κοντινότεροι-γείτονες
- Δέντρο απόφασης
- Λογιστική παλινδρόμηση
- Απλοϊκός Bayes
- Νευρωνικά δίκτυα
- Τυχαία δάση

Για παράδειγμα ένα πρόβλημα το οποίο μπορεί να λυθεί με κατηγοριοποίηση είναι τα spam emails. Στο email τα emails τα οποία λαμβάνονται κατηγοριοποιούνται σε μη spam email και σε spam email. Κάθε ένα νέο email το οποίο λαμβάνεται κατηγοριοποιείται σε μη spam email ή σε spam email [17]. Το συγκεκριμένο πρόβλημα ανήκει στην δυαδική κατηγοριοποίηση (binary classification) δηλαδή ή αληθές (true) ή ψευδές (false). Όμως υπάρχουν και διακρίσεις όπως η πολλαπλή κατηγοριοποίηση (multiclass classification) όπου κατηγοριοποιεί τα δεδομένα σε περισσότερες από δύο κλάσεις (classes).



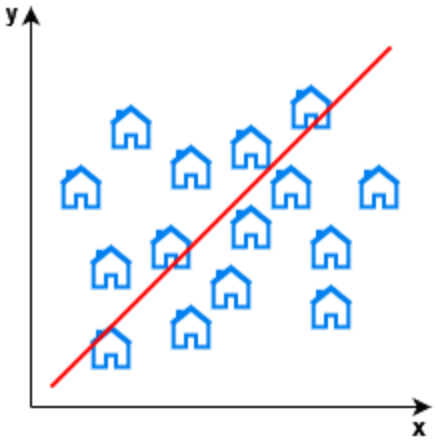
Εικόνα 5: *Παράδειγμα κατηγοριοποίησης για spam emails*

2.3.1.2 Παλινδρόμηση

Η παλινδρόμηση (regression) στην μηχανική μάθηση είναι μια επιβλεπόμενη μάθηση. Στόχος της είναι να μπορεί να προβλέψει μια συνεχή μεταβλητή δεδομένου κάποιας εισόδου. Οι πιο γνωστοί αλγόριθμοι παλινδρόμησης είναι οι:

- Γραμμική παλινδρόμηση
- Πολυωνυμική γραμμική παλινδρόμηση
- Κ-κοντινότεροι-γείτονες
- Δέντρα Απόφασης
- Τυχαία δάση

Ένα παράδειγμα προβλήματος που μπορεί η λύση η παλινδρόμηση είναι η πρόβλεψη της τιμής ενός σπιτιού [11]. Χρησιμοποιώντας δεδομένα όπως το γεωγραφικό μήκος, γεωγραφικό πλάτος, την ηλικία του σπιτιού, τον αριθμό των δωματίων και τον αριθμό των κρεβατιών καθώς και πολλά άλλα έχει την δυνατότητα να προβλέψει την τιμή του.



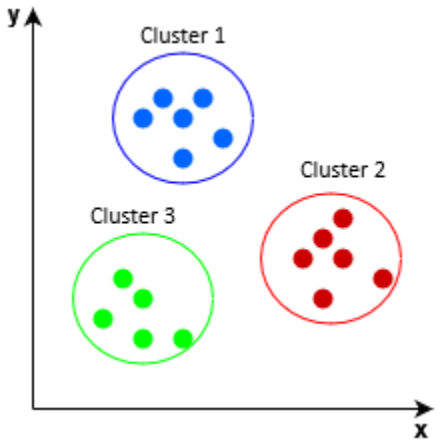
Εικόνα 6: *Παράδειγμα παλινδρόμησης για πρόβλεψη τιμής σπιτιού*

2.3.2 Μη Επιβλεπόμενη μάθηση

Η μη επιβλεπόμενη μάθηση (unsupervised learning) σε σύγκριση με την επιβλεπόμενη μάθηση χρησιμοποιεί δεδομένες εισόδους (όπου ονομάζονται σύνολο εκπαίδευσης) σε άγνωστες εξόδους [16]. Το σύνολο εκπαίδευσης δεν ανήκει σε κάποια κλάση και το σύστημα ουσιαστικά προσπαθεί να μάθει χωρίς καθηγητή [11]. Επίσης η μη επιβλεπόμενη μάθηση όπως και η επιβλεπόμενη μάθηση περιέχει δύο υποκατηγορίες. Οι υποκατηγορίες αυτές είναι η συσταδοποίηση (clustering) και η μείωση διαστάσεων (dimensionality reduction).

2.3.2.1 Συσταδοποίηση

Η συσταδοποίηση (clustering) στην μηχανική μάθηση ανήκει στην μη επιβλεπόμενη μάθηση. Στόχος της είναι το σύνολο των δεδομένων εκπαίδευσης τα οποία περιέχονται σε αυτή να διαχωρίζονται σε ένα σύνολο από υποομάδες (clusters) χωρίς να υπάρχει κάποια προηγούμενη γνώση. Η κάθε υποομάδα (cluster) η οποία προκύπτει κατά την διάρκεια της ανάλυσης [18] είναι μια ομάδα αντικειμένων τα οποία έχουν κάποια όμοια χαρακτηριστικά. Ορισμένες φορές αυτά τα αντικείμενα των υποομάδων (clusters) έχουν ανόμοια αντικείμενα σε σύγκριση με τις άλλες υποομάδες (clusters).



Εικόνα 7: *Παράδειγμα συστάδοποίησης*

Ορισμένοι γνωστοί αλγόριθμοι συσταδοποίησης (clustering) είναι οι παρακάτω:

- K-Means
- Πολυωνυμική Hierarchical Clustering
- DBSCAN
- BIRCH

2.3.2.2 Μείωση διαστάσεων

Η μείωση διαστάσεων (dimensionality reduction) στην μηχανική μάθηση ανήκει στην μη επιβλεπόμενη μάθηση. Είναι ένας μετασχηματισμός (transformation) όπου μετατρέπει τις υψηλές διαστάσεις (high dimensional) σε χαμηλές διαστάσεις (low dimensional). Η μείωση διαστάσεων (dimensionality reduction) χρησιμοποιείται με σκοπό η αναπαράσταση να είναι ευκολότερη στην κατανόηση των ανθρώπων αλλά και των αλγόριθμων της μηχανικής μάθησης [11]. Η εργασία των αλγόριθμων της μηχανικής μάθησης με δεδομένα που ανήκουν σε πίνακες υψηλών

διαστάσεων είναι ανεπιθύμητη για πολλούς λόγους. Ένας από τους λόγους είναι ότι θα χρειαστεί περισσότερος χώρος αποθήκευσης και θα υπάρξει μείωση της υπολογιστικής απόδοσης των αλγόριθμων της μηχανικής μάθησης. Μερικοί από τους αλγόριθμους της μείωσης διαστάσεων είναι οι:

- Principal component analysis (PCA)
- Linear discriminant analysis (LDA)
- Non-negative matrix factorization (NMF)
- Singular value decomposition (SVD)

2.3.3 Ενισχυτική μάθηση

Η ενισχυτική μάθηση (reinforcement learning) σε σύγκριση με την επιβλεπόμενη μάθηση και την μη επιβλεπόμενη μάθηση είναι ένας διαφορετικός τύπος μηχανικής μάθησης. Το σύστημα μαθαίνει μια στρατηγική ενεργειών αλληλοεπιδρώντας με το περιβάλλον [16]. Στην ενισχυτική μάθηση το σύστημα περιέχει τον πράκτορα (agent), το περιβάλλον (environment) και την ενέργεια (action) [18]. Το σύστημα εκμάθησης χρησιμοποιεί τον πράκτορα στο περιβάλλον εκτελώντας ορισμένες ενέργειες. Από κάθε ενέργεια ο πράκτορας παίρνει κάποια ανταμοιβή (reward) ή ποινή (penalty) με βάση τον σκοπό του συστήματος. Ο πράκτορας πρέπει να μάθει από τον αυτό του ποια είναι η καλύτερη στρατηγική (strategy) η οποία ονομάζεται πολιτική (policy) [11] με βάση τις ανταμοιβές ή τις τιμές από τις οποίες λαμβάνει. Έτσι ο πράκτορας με βάση την πολιτική καταλαβαίνει ποια ενέργεια πρέπει να εκτελέσει σε κάθε διαφορετική κατάσταση.

Εικόνα 8: *Διαδικασία ενισχυτικής μάθησης*

2.3.4 Άλλοι τύποι μάθησης

Έκτος από την επιβλεπόμενη μάθηση, την μη επιβλεπόμενη μάθηση και την ενισχυτική μάθηση δηλαδή τους τρεις τύπους μάθησης υπάρχουν και ορισμένοι άλλοι τύποι μάθησης. Αυτοί οι τύποι μάθησης είτε ανήκουν σε κάποια υποκατηγορία από τους τρεις τύπους μάθησης είτε χρησιμοποιούν δύο διαφορετικούς τύπους μάθησης. Ένας από τους τύπους μάθησης που χρησιμοποιούν δύο διαφορετικούς τύπους μάθησης είναι η ημι-επιβλεπόμενη μάθηση. Άλλοι τύποι μάθησης είναι η τεμπέλικη μάθηση (lazy learning) και η πρόθυμη μάθηση (eager learning).

2.3.4.1 Ημι-επιβλεπόμενη μάθηση

Η ημι-επιβλεπόμενη μάθηση (semi-supervised learning) είναι ένας τύπος μάθησης που είναι ανάμεσα στην επιβλεπόμενη μάθηση και στην μη επιβλεπόμενη μάθηση. Ουσιαστικά στην ημι-επιβλεπόμενη μάθηση στο σύνολο εκπαίδευσης ορισμένοι έξοδοι ανήκουν σε κάποια κλάση και κάποιοι έξοδοι έχουν άγνωστη έξοδο δηλαδή δεν ανήκουν σε κάποια κλάση.

2.3.4.2 Lazy learning

Η μάθηση lazy learning είναι μια μέθοδος όπου γενικεύει τα δεδομένα εκπαίδευσης και καθυστερεί μέχρι να γίνει κάποιο ερώτημα στο σύστημα [19]. Ονομάζονται lazy επειδή περιμένουν αρκετά όσο μπορούν μέχρι να δημιουργήσουν κάποιο μοντέλο [20]. Ένα από τα θετικά της μάθησης lazy learning είναι ότι μαθαίνει γρήγορα. Απαιτεί μεγάλο χώρο για να αποθηκεύσουν όλα τα δεδομένα εκπαίδευσης ταξινομώντας τα με αργό ρυθμό. Ένας αλγόριθμος μηχανικής μάθησης ο οποίος χρησιμοποιεί τεμπέλικη μάθηση είναι ο Κ-κοντινότεροι-γείτονες (K-Nearest Neighbors - kNN).

2.3.4.3 Eager learning

Σε αντίθεση με την μάθηση lazy learning, στην μάθηση eager learning το σύστημα κατά την διάρκεια της εκπαίδευσης του συστήματος προσπαθεί να κατασκευάσει έναν ανεξάρτητο στόχο εισόδου [20]. Ένα κύριο πλεονέκτημα της μάθησης eager learning είναι ότι απαιτεί λιγότερο χώρο για την αποθήκευση των δεδομένων εκπαίδευσης [19] σε σύγκριση με την τεμπέλικη μάθηση η οποία χρειάζεται μεγάλο χώρο για την αποθήκευση των δεδομένων εκπαίδευσης. Όμως παρόλα

αυτά η δημιουργία του μοντέλου είναι αργή. Ορισμένοι αλγόριθμοι μηχανικής μάθησης οι οποίοι χρησιμοποιούν πρόθυμη μάθηση είναι ο Support Vector Machines και τα Neural Networks.

2.4 Προεπεξεργασία δεδομένων

Η προεπεξεργασία δεδομένων (data preprocessing) είναι μια διαδικασία προετοιμασίας των ακατέργαστων δεδομένων (raw data) κάνοντάς τα κατάλληλα για τα μοντέλα των αλγορίθμων μηχανικής μάθησης. Κατά την δημιουργία ενός έργου μηχανικής μάθησης τα δεδομένα που θα χρησιμοποιηθούν τις περισσότερες φορές δεν είναι κατάλληλα μορφοποιημένα και μπορούν να φέρουν ανακριβή αποτελέσματα στους αλγόριθμους μηχανικής μάθησης αν δεν υπάρξει κάποια προεπεξεργασία. Για την προεπεξεργασία των δεδομένων υπάρχουν δύο βασικά βήματα, ο καθαρισμός των δεδομένων (data cleaning) και η κλιμάκωση των χαρακτηριστικών (feature scaling).

2.4.1 Καθαρισμός δεδομένων

Ο καθαρισμός δεδομένων (data cleaning) είναι η διαδικασία της προετοιμασίας των δεδομένων για να χρησιμοποιηθούν από τους αλγόριθμους μηχανικής μάθησης. Αρκετές φορές τα δεδομένα είναι λανθασμένα, ελλιπή, άσχετα ή διπλότυπα. Αυτά τα δεδομένα δεν είναι χρήσιμα καθώς αν χρησιμοποιηθούν θα φέρουν ανακριβή αποτελέσματα στους αλγόριθμους μηχανικής μάθησης. Για αυτόν τον λόγω υπάρχουν αρκετές επιλογές για τον καθαρισμό των δεδομένων (data cleaning). Μερικές από αυτές είναι οι παρακάτω:

- Διαγραφή διπλότυπων δεδομένων
- Διαγραφή άσχετων στηλών
- Χειρισμός ελλιπών τιμών

2.4.2 Κλιμάκωση χαρακτηριστικών

Ένας από τους πιο σημαντικούς μετασχηματισμούς που πρέπει να εφαρμοστούν στο σύνολο δεδομένων είναι η κλιμάκωση χαρακτηριστικών (feature scaling) δηλαδή τα δεδομένα να έχουν την ίδια κλίμακα. Αρκετές φορές οι αλγόριθμοι μηχανικής μάθησης δεν έχουν καλή απόδοση επειδή τα χαρακτηριστικά εισόδου έχουν πολύ διαφορετικές κλίμακες (scales) [11]. Δύο

συνηθισμένοι τρόποι για κλιμάκωση χαρακτηριστικών είναι η κανονικοποίηση (normalization) και η τυποποίηση (standardization).

2.4.2.1 Κανονικοποίηση

Η κανονικοποίηση (normalization) ή αλλιώς ελάχιστη-μέγιστη κλιμάκωση (min-max scaling) είναι μια τεχνική κλιμάκωσης χαρακτηριστικών (feature scaling). Οι τιμές των χαρακτηριστικών μετατοπίζονται και επανακλιμακώνονται (rescaled) έτσι ώστε να κυμαίνονται στο διάστημα μεταξύ 0 και 1. Για τον υπολογισμό της κανονικοποίησης υπολογίζεται η ελάχιστη (minimum) και η μέγιστη (maximum) του χαρακτηριστικού της συγκεκριμένης στήλης και υπολογίζεται από τον παρακάτω τύπο:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Εξίσωση 1: *Τύπος κανονικοποίησης*

2.4.2.2 Τυποποίηση

Η τυποποίηση (standardization) είναι μια άλλη τεχνική κλιμάκωσης χαρακτηριστικών αρκετά διαφορετική από την κανονικοποίηση (normalization). Εδώ οι τιμές επικεντρώνονται γύρω από τον μέσο όρο (mean) με μια τυπική απόκλιση (standard deviation) μονάδας. Αρχικά αφαιρείται η μέση τιμή (mean) όπου οι τιμές της πάντα έχουν μηδενικό μέσο όρο και διαιρείται με την τυπική απόκλιση (standard deviation) έτσι ώστε το αποτέλεσμα της προκύπτουσας κατανομής να έχει μοναδιαία διακύμανση [11]. Για τον υπολογισμό της τυποποίησης χρησιμοποιείται ο παρακάτω τύπος:

$$X_{stand} = \frac{X - \mu_X}{\sigma_X}$$

Εξίσωση 2: *Τύπος τυποποίησης*

Για τον υπολογισμό του μέσου όρου (mean) χρησιμοποιείται ο παρακάτω τύπος:

$$\mu = \frac{\sum_i^n x_i}{n}$$

Εξίσωση 3: *Τύπος μέσου όρου*

Ο υπολογισμός της τυπικής απόκλισης (standard deviation) υπολογίζεται από τον παρακάτω τύπο:

$$\sigma = \sqrt{\frac{\sum_1^n (x_i - \mu)^2}{n}}$$

Εξίσωση 4: *Τύπος τυπικής απόκλισης*

2.5 Αλγόριθμοι μηχανικής μάθησης

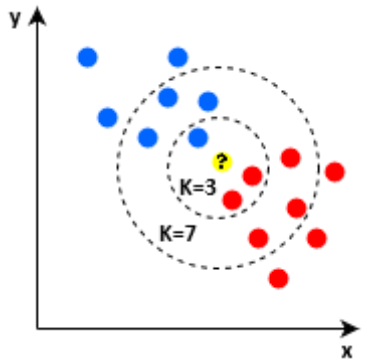
Η επιβλεπόμενη μάθηση (supervised learning) και η μη επιβλεπόμενη μάθηση (unsupervised learning) χρησιμοποιούν διαφορετικούς αλγόριθμους μηχανικής μάθησης. Επειδή η κλάση (class) του συνόλου δεδομένων ανήκει στην δυαδική κατηγοριοποίηση (binary classification) παρακάτω θα περιγραφούν αλγόριθμοι κατηγοριοποίησης (classification). Θα περιγραφούν οι αλγόριθμοι K-κοντινότεροι γείτονες (K-nearest-neighbors - kNN), η συλλογιστική μάθηση (ensemble learning), η γραμμική παλινδρόμηση (linear regression) η οποία είναι ένας αλγόριθμος παλινδρόμησης αλλά θα περιγραφεί επειδή θα χρησιμοποιηθεί η λογιστική παλινδρόμηση (logistic regression) η οποία είναι όμοια της γραμμικής παλινδρόμηση (linear regression). Τέλος θα περιγραφούν τα νευρωνικά δίκτυα (neural networks), ο απλοϊκός Bayes (Naïve Bayes).

2.5.1 Κ-κοντινότεροι-γείτονες

Ο αλγόριθμος Κ-κοντινότεροι-γείτονες (K-Nearest-Neighbors - kNN) είναι ένας από τους πιο δημοφιλείς αλγόριθμους της επιβλεπόμενης μάθησης. Αναπτύχθηκε το 1951 από την Evelyn Fix και τον Joseph Hodges και αργότερα επεκτάθηκε από τον Thomas Cover. Επίσης ο αλγόρθμος K-κοντινότεροι-γείτονες είναι ένα μη γραμμικό μοντέλο και παρέχεται ένα επισημασμένο (labeled) σύνολο δεδομένων εκπαίδευσης όπου τα δεδομένα κατηγοριοποιούνται σε κατηγορίες με σκοπό να μπορεί να προβλεθεί η κατηγορία των μη επισημασμένων δεδομένων (unlabeled data) [23]. Ακόμη είναι ένας αλγόριθμος τεμπέλικης μάθησης επειδή απομνημονεύει το σύνολο δεδομένων εκπαίδευσης. Χρησιμοποιείται για κατηγοριοποίηση αλλά και για παλινδρόμηση.

Στην κατηγοριοποίηση ο αλγόριθμος Κ-κοντινότεροι-γείτονες χρησιμοποιείται πιο συχνά ως κατηγοριοποιητής (classifier) και κύριος σκοπός του είναι να κατηγοριοποιεί τα μη επισημασμένα δεδομένα (unlabeled data). Κατηγοριοποιεί τα δεδομένα με βάση τα κοντινότερα ή τα γειτονικά δεδομένα εκπαίδευσης [23]. Υπολογίζονται οι Κ κοντινότεροι γείτονες και η πλειοψηφία μεταξύ των γειτονικών δεδομένων καθορίζει την κατηγοριοποίηση των νέων δεδομένων. Η τιμή του Κ είναι σημαντική [19] για την κατηγοριοποίηση των μη επισημασμένων δεδομένων (unlabeled

data). Για τον υπολογισμό της καλύτερης τιμής του K πρέπει να εκτελεστεί ο αλγόριθμος K-κοντινότεροι-γείτονες αρκετές φορές με διαφορετική τιμή στο K. Έτσι θα παρατηρήσουμε ποια τιμή στο K δίνει το αποτελεσματικότερο αποτέλεσμα.



Εικόνα 9: *Κατηγοριοποίηση αλγόριθμου K-κοντινότεροι-γείτονες*

Στην παλινδρόμηση ο αλγόριθμος K-κοντινότεροι-γείτονες προβλέπει συνεχές τιμές [23] με τον μέσο όρο των παρατηρήσεων που είναι στην ίδια γειτονιά. Όπως και στην κατηγοριοποίηση και στην παλινδρόμηση η τιμή του K είναι σημαντική και αλλάζοντας τιμή στον K κάθε φορά που εκτελείται ο αλγόριθμος η έξοδος του μπορεί να είναι διαφορετική.

Τα βασικά βήματα για να χρησιμοποιηθεί ο ο αλγόριθμος K-κοντινότεροι-γείτονες είναι τα παρακάτω:

1. Επιλογή των αριθμών των K γειτόνων
2. Επιλογή μέτρου απόστασης
3. Παραλάβει των K πλησιέστερων γειτόνων με βάση το μέτρο απόστασης
4. Μεταξύ των K γειτόνων μέτρηση του αριθμών των δεδομένων σε κάθε κατηγορία
5. Αντιστοίχιση των νέων δεδομένων στην κατηγορία την οποία ο αριθμός του γείτονα είναι ο μέγιστος

Το μέτρο απόστασης χρησιμοποιείται για την μέτρηση της απόστασης ή της ομοιότητας μεταξύ ορισμένων σημείων D(x, y). Μερικά από τα μέτρα απόστασης είναι τα παρακάτω:

- **Minkowski απόσταση (Minkowski distance):** ανάλογα με τον αριθμό p του τύπου μπορεί να τροποποιηθεί και να δώσει διαφορετικές αποστάσεις. Για p=1 είναι η Manhattan απόσταση και για p=2 είναι η Ευκλείδεια απόσταση. Για p=∞ είναι η Chebyshev απόσταση. Υπολογίζεται από τον τύπο:

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

Εξίσωση 5: *Minkowski απόσταση*

- **Μανχάταν απόσταση (Manhattan distance):** η απόσταση μεταξύ δύο σημείων είναι το άθροισμα των απόλυτων διαφορών των καρτεσιανών συντεταγμένων τους. Υπολογίζεται από τον τύπο:

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

Εξίσωση 6: *Μανχάταν απόσταση*

- **Ευκλείδεια απόσταση (Euclidean distance):** είναι ένα μέτρο της πραγματικής ευθείας της απόστασης μεταξύ δύο σημείων στον Ευκλείδειο χώρο. Υπολογίζεται από τον τύπο:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Εξίσωση 7: *Ευκλείδεια απόσταση*

- **Chebyshev απόσταση (Chebyshev Distance):** ονομάζεται ή αλλιώς μέγιστη τιμή απόστασης και εξετάζει το απόλυτο μέγεθος των διαφορών μεταξύ των συντεταγμένων ενός ζεύγους αντικειμένων. Υπολογίζεται από τον τύπο:

$$d(x,y) = \max_{i=1,\dots,n} |x_i - y_i|$$

Εξίσωση 8: *Chebyshev απόσταση*

- **Απόσταση συνημιτόνου (Cosine distance):** μετράται από το συνημίτονο της γωνίας μεταξύ δύο διανυσμάτων και καθορίζει εάν δύο διανύσματα δείχνουν προς την ίδια κατεύθυνση. Υπολογίζεται από τον τύπο:

$$d(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Εξίσωση 9: *Απόσταση συνημιτόνου*

- **Hamming απόσταση (Hamming distance):** εξετάζει όλα τα δεδομένα και βρίσκει πότε τα σημεία δεδομένων είναι παρόμοια και πότε είναι ανόμοια ένα προς ένα. Υπολογίζεται από τον τύπο:

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

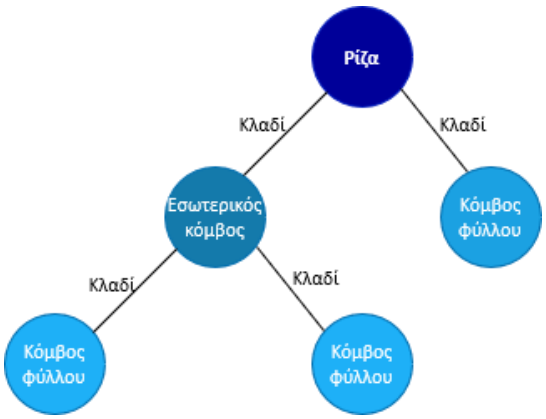
Εξίσωση 10: *Hamming απόσταση*

2.5.2 Δέντρα απόφασης

Ο αλγόριθμος δέντρα απόφασης (decision trees) όπως και ο αλγόριθμος K-κοντινότεροι-γείτονες είναι ένας από τους πιο γνωστούς αλγόριθμους της επιβλεπόμενης μάθησης. Δημιουργήθηκε το 1975 από τον J. Ross Quinlan. Σε σύγκριση με άλλους αλγόριθμους επιβλεπόμενης μάθησης τα δέντρα απόφασης μπορούν να χρησιμοποιηθούν για την επίλυση προβλημάτων κατηγοριοποίησης αλλά και παλινδρόμησης. Επίσης είναι ο πιο εύκολος τρόπος κατηγοριοποίησης και πρόβλεψης καθώς είναι πολύ εύκολος στην κατανόησή του [18] και έχει την δυνανότητα κανοί να προσαρμοστεί σε πολύπλοκα σύνολα δεδομένων [11].

Κάθε δέντρο απόφασης (decision tree) αποτελείται από:

- **Ρίζα κόμβου (Root node):** όπου ουσιαστικά είναι η αρχική ρίζα ολόκληρου του δένδρου και είναι στην κορυφή. Είναι ένα από τα χαρακτηριστικά που ο αλγόριθμος επιλέγει πρώτα [16].
- **Εσωτερικός κόμβος (Internal node):** είναι οι ενδιάμεσος κόμβος και ουσιαστικά είναι ένα από τα χαρακτηριστικά που δεν έχει επιλεχθεί ως ρίζα κόμβου (root node).
- **Κλαδί (Branch):** είναι το αποτέλεσμα του κάθε κόμβου. Κάθε κόμβος μπορεί να έχει διαφορετικό αποτέλεσμα από τον προηγούμενο.
- **Κόμβος φύλλου (Leaf node):** είναι τα τελικά αποτελέσματα δηλαδή η έξοδος (output) του δένδρου απόφασης



Εικόνα 10: *Δέντρο απόφασης*

Το δέντρο απόφασης αρχικά δέχεται ένα σύνολο δεδομένων. Αυτό το σύνολο δεδομένων περιέχει ορισμένα χαρακτηριστικά. Ο αλγόριθμος επιλέγει ένα από αυτά τα χαρακτηριστικά ως ρίζα κόμβου (root node). Στην συνέχεια τα κλαδιά (branches) περιέχουν κάποια αποτελέσματα του ρίζα κόμβου (root node) όπου μπορεί να είναι διακριτά ή συνεχή [18] και συνδέονται με τους εσωτερικούς κόμβους (internal nodes). Οι εσωτερικοί κόμβοι (internal nodes) είναι τα υπόλοιπα χαρακτηριστικά που δεν επιλέχθηκαν ως ρίζα κόμβος (root node). Και εκεί όπως και προηγουμένος τα κλαδιά (branches) περιέχουν το αποτέλεσμα των εσωτερικών κόμβων και καταλήγουν στον κόμβο φύλλου (leaf node). Τέλος οι κόμβοι φύλλου (leaf nodes) είναι το αποτέλεσμα δηλαδή η έξοδος του δένδρου απόφασης (decision tree). Αν η έξοδος τους είναι διακριτική τότε είναι κατηγοριοποίηση δηλαδή για παράδειγμα αν η έξοδος είναι 0 ή 1 τότε είναι δυαδική κατηγοριοποίηση αλλιώς αν η έξοδος είναι 0 ή 1 ή 2 έως n τότε είναι πολλαπλή κατηγοριοποίηση. Αλλιώς αν η έξοδος τους είναι συνεχής συνάρτηση δηλαδή για παράδειγμα 0.1, 0.2 έως n τότε είναι παλινδρόμηση.

Μερικοί από τους πιο γνωστούς αλγόριθμους δένδρων απόφασης (decision trees) είναι ο ID3, ο C4.5, ο CHAID, ο CART και ο MARS.

2.5.2.1 ID3

Στον αλγόριθμο ID3 το σύνολο δεδομένων αποτελείται από N χαρακτηριστικά και αποφασίζει πιο χαρακτηριστικό θα τοποθετηθεί στην κορυφή ως ρίζα κόμβου (root node). Για να αποφασίσει πιο χαρακτηριστικό θα είναι στην κορυφή ως ρίζα κόμβου εφαρμόζει αναδρομικά μια άπληστη αναζήτηση [10] δηλαδή επιλέγει το καλύτερο χαρακτηριστικό. Για την επιλογή του καλύτερου χαρακτηριστικού ο αλγόριθμος ID3 χρησιμοποιεί τις παρακάτω βασικές έννοιες:

- **Εντροπία (Entropy):** είναι ένα μέτρο τυχαιότητας και χαρακτηρίζει το βαθμό αβεβαιότητας ενός συνόλου δεδομένων S [10] και. κάθε p_i είναι η πιθανότητα του ενδεχόμενου που περιλαμβάνει αυτό το σύνολο δεδομένων S. Υπολογίζεται από τον τύπο:

$$E(S) = -\sum_{i=1}^{n} p_i log_2 p_i$$

Εξίσωση 11: *Εντροπία*

Σε περίπτωση που υπάρχει δυαδική κατηγοριοποίηση τότε η εντροπία υπολογίζεται από τον παρακάτω τύπο:

$$E(p) = -plog_2 p - (1-p)log_2(1-p)$$

Εξίσωση 12: *Εντροπία δυαδικής κατηγοριοποίησης*

- **Κέρδος πληροφορίας (Information gain):** είναι μια στατιστική ιδιότητα που μετρά πόσο καλά ένα δεδομένο χαρακτηριστικό διαχωρίζει το σύνολο δεδομένων εκπαίδευσης σύμφωνα με τον στόχο ταξινόμησης. Αποτελείται από το σύνολο δεδομένων S όπου κάθε περιέχει το χαρακτηριστικό A. Υπολογίζεται από τον τύπο:

$$IG(S) = E(S) - E(S, A)$$

Εξίσωση 13: *Κέρδος πληροφορίας*

- **Δείκτης Gini (Gini Index):** είναι μια συνάρτηση κόστους που χρησιμοποιείται για την αξιολόγηση των διαχωρισμών στο σύνολο δεδομένων S. Υπολογίζεται από τον τύπο:

$$Gini(S) = 1 - \sum_{i=1}^{n} p_i^2$$

Εξίσωση 14: *Δείκτης Gini*

Τα βήματα υλοποίησης του αλγόριθμου ID3 είναι τα παρακάτω:

1. Ξεκινά με το αρχικό σύνολο δεδομένων S ως ρίζα κόμβο (root node)
2. Σε κάθε επανάληψη του αλγορίθμου υπολογίζει στο σύνολο δεδομένων S την εντροπία (entropy) και το κέρδος πληροφορίας (information gain) σε κάθε χαρακτηριστικό
3. Επιλέγει το χαρακτηριστικό με την μικρότερη εντροπία (entropy) ή το μεγαλύτερο κέρδος πληροφορίας (information gain)
4. Το σύνολο δεδομένων S διαιρεί το επιλεγμένο χαρακτηριστικό και το προσθέτει σε ένα υποσύνολο δεδομένων

5. Ο αλγόριθμος συνεχίζει να επαναλαμβάνεται σε κάθε υποσύνολο λαμβάνοντας υπόψη μόνο χαρακτηριστικά που δεν έχουν επιλεγεί ποτέ πριν

2.5.2.2 C4.5

Ο αλγόριθμος C4.5 βασίζεται στον ID3 και ουσιαστικά είναι μια επέκτασή του. Χρησιμοποιεί και αυτός το κέντρο πληροφορίας (information gain) και τον δείκτη gini (gini index). Η κύρια διαφορά του με τον αλγόριθμο ID3 είναι ότι λειτουργεί τόσο με διακριτά όσο και με συνεχή δεδομένα ενώ ο ID3 λειτουργεί μόνο με διακριτά ή ονομαστικά δεδομένα [24]. Τέλος χρησιμοποιεί την διαδικασία του κλαδέματος με σκοπό να μην υπάρξει υπερπροσαρμογή (overfitting).

2.5.2.3 CHAID

Ο αλγόριθμος CHAID είναι ένας αλγόριθμος δέντρων αποφάσεων παλαιού τύπου. Χρησιμοποιεί την μέτρηση Chi-Τετράγωνο (Chi-Square) με σκοπό να ανακαλύψει το πιο σημαντικό χαρακτηριστικό στο σύνολο δεδομένων. Στην συνέχεια εφαρμόζει αυτήν την μέτρηση αναδρομικά μέχρι τα υποσύνολα δεδομένων να έχουν μια ενιαία απόφαση. Το Υπολογίζεται Chi-Τετράγωνο (Chi-Square) από τον παρακάτω τύπο:

$$\chi^2 = \sum_{i}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Εξίσωση 15: *Chi-Τετράγωνο*

Όπου Ο είναι η παρατηρηθείσα βαθμολογία (observed score) και Ε είναι η προβλεπόμενη βαθμολογία (expected score).

2.5.3 Συλλογιστική μάθηση

Η συλλογιστική μάθηση (ensemble learning) είναι μια μέθοδος που χρησιμοποιεί πολλαπλούς αλγόριθμους μηχανικής μάθησης έτσι ώστε να πετύχουν την καλύτερη προγνωστική απόδοση από ότι θα μπορούσε να επιτευχθεί. Στόχος του είναι να συνδυαστούν οι προβλέψεις πολλών εκτιμητών βάσης που έχουν κατασκευαστεί με έναν δεδομένο αλγόριθμο εκμάθησης προκειμένου να βελτιωθεί η γενίκευση από έναν μόνο εκτιμητή.

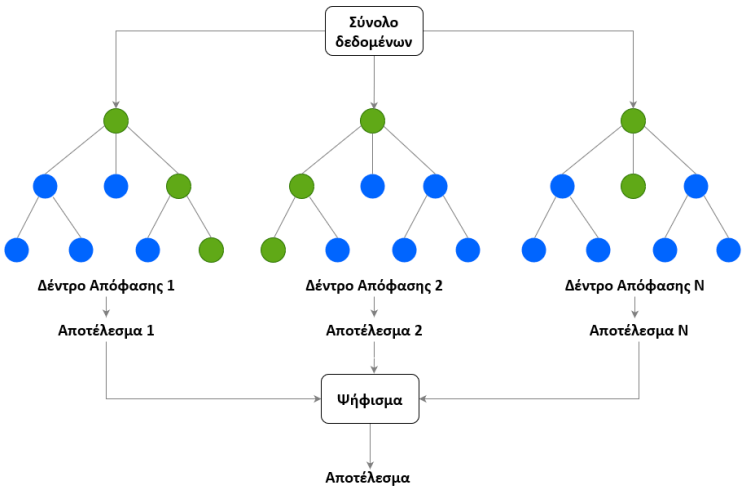
Συνήθως η συλλογιστική μάθηση διακρίνεται σε δύο κατηγορίες. Τις μεθόδους υπολογισμού του μέσου όρου (averaging methods) και τις μεθόδους ενίσχυσης (boosting methods). Στις μεθόδους υπολογισμού του μέσου όρου δημιουργούνται πολλοί εκτιμητές ανεξάρτητα και στην συνέχεια υπολογίζεται ο μέσος όρος των προβλέψεων τους. Από την άλλη μεριά αντίθετα στις μεθόδους ενίσχυσης οι βασικοί εκτιμητές κατασκευάζονται διαδοχικά προσπαθώντας να μειωθεί η προκατάληψη του συνδυασμένου εκτιμητή. Ουσιαστικά πολλά αδύναμα μοντέλα συνδυάζονται με σκοπό να δημιουργήσουν ένα δυνατό μοντέλο.

Ένα παράδειγμα αλγόριθμου των μεθόδων υπολογισμού του μέσου όρου είναι τα τυχαία δάση (random forests) ενώ μερικά παραδείγματα αλγορίθμων των μεθόδων ενίσχυσης είναι ο AdaBoost και ο gradient tree boosting.

2.5.3.1 Τυχαία δάση

Τα τυχαία δάση (random forests) δημιουργήθηκαν από τον Leo Breiman το 2001 και είναι ένας αλγόριθμος συλλογιστικής μάθησης για κατηγοριοποίηση και για παλινδρόμηση. Κατασκευάζει ένα πλήθος δέντρων αποφάσεων κατά την διάρκεια της εκπαίδευσης. Για κατηγοριοποίηση η έξοδος του τυχαίου δάσους (random forest) είναι η κλάση που έχει επιλεχθεί από τα περισσότερα δέντρα. Για παλινδρόμηση επιστρέφεται η μέση πρόβλεψη των μεμονωμένων δέντρων.

Ουσιαστικά τα τυχαία δάση βασίζονται σε δέντρα που αξιοποιούν τη δύναμη πολλαπλών δέντρων αποφάσεων για τη λήψη αποφάσεων. Για την έξοδο τους συνδυάζονται πολλά τυχαία δέντρα απόφασης για το τελικό αποτέλεσμα.



Εικόνα 11: *Τυχαία δάση*

2.5.3.2 AdaBoost

Ο AdaBoost ή αλλιώς Adaptive Boosting είναι ένας αλγόριθμος συλλογιστικής μάθησης και χρησιμοποιεί την μέθοδο της ενδυνάμωσης από την οποία και προκύπτει το όνομά του. Είναι μια τεχνική ενίσχυσης (boosting) συνδυάζοντας ένα σύνολο αδύναμων κατηγοριοποιητών σε έναν ισχυρό κατηγοριοποιητή [26]. Λόγω της χαμηλής πολυπλοκότητας υλοποίησης του και της γρήγορης απόδοσης του η ενίσχυση έχει γίνει ένα από τα πιο σημαντικά εργαλεία κατηγοριοποίησης [26].

Ουσιαστικά ο AdaBoost λειτουργεί με την διαδοχική ανάπτυξη των κατηγοριοποιητών. Εκτός από τον πρώτο κατηγοριοποιητή ο επόμενος κατηγοριοποιητής αναπτύσσεται από τους προηγούμενους αναπτυγμένους κατηγοριοποιητές. Οπότε οι αδύναμοι κατηγοριοποιητές μετατρέπονται σε ισχυρούς κατηγοριοποιητές.

2.5.3.3 Gradient tree boosting

Ο Gradient tree boosting είναι ένας αλγόριθμος συλλογιστικής μάθησης (ensemble learning) και χρησιμοποιείται για κατηγοριοποίηση αλλά και για παλινδρόμηση. Δίνει ένα μοντέλο πρόβλεψης με τη μορφή ενός συνόλου αδύναμων μοντέλων πρόβλεψης τα οποία είναι συνήθως δέντρα απόφασης. Συγκρίνοντας με τα τυχαία δάση ο gradient tree boosting προσπαθεί να διορθώσει τα λάθη του και το βάθος των δέντρων που προκύπτει είναι πολύ μικρότερο [18].

2.5.4 Γραμμική παλινδρόμηση

Η γραμμική παλινδρόμηση (linear regression) είναι ένας αλγόριθμος επιβλεπόμενης μάθησης και είναι ένα γραμμικό μοντέλο. Είναι ένα απίστευτα ισχυρό εργαλείο για την ανάλυση δεδομένων και περιλαμβάνει την απλή γραμμική παλινδρόμηση (simple linear regression) στην οποία υπάρχουν 2 μεταβλητές και την πολλαπλή γραμμική παλινδρόμηση (multiple linear regression) όπου εκεί υπάρχουν περισσότερες μεταβλητές. Χρησιμοποιείται για να κάνει προβλέψεις υπολογίζοντας ένα ένα σταθμισμένο άθροισμα χαρακτηριστικών εισόδου [11]. Είναι κατάλληλη για την πρόβλεψη μιας συνεχούς μεταβλητής.

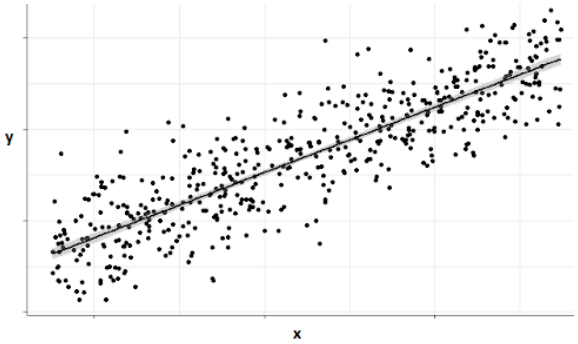
2.5.4.1 Απλή γραμμική παλινδρόμηση

Η απλή γραμμική παλινδρόμηση (simple linear regression) περιλαμβάνει την τιμή x όπου είναι μια ανεξάρτητη μεταβλητή ή αλλιώς μεταβλητή πρόβλεψης και την τιμή y η οποία είναι μια εξαρτημένη μεταβλητή δηλαδή η τιμή εξόδου (output). Επίσης περιλαμβάνει τιμή α όπου είναι είναι η τομή της γραμμής και την τιμή β όπου είναι η είναι η κλίση της γραμμής. Παρακάτω είναι ο τύπος της απλής γραμμικής παλινδρόμησης:

$$y = \alpha + \beta x$$

Εξίσωση 16: *Απλή γραμμική παλινδρόμηση*

Η παρακάτω εικόνα δείχνει πως εφαρμόζεται ο τύπος της απλής γραμμικής παλινδρόμησης.



Εικόνα 12: *Απλή γραμμική παλινδρόμηση*

2.5.4.2 Πολλαπλή γραμμική παλινδρόμηση

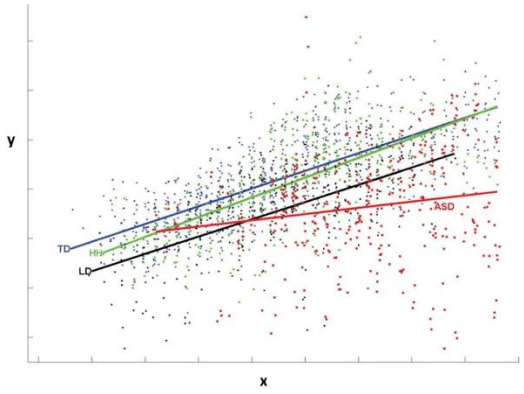
Η πολλαπλή γραμμική παλινδρόμηση (multiple linear regression) περιλαμβάνει τις τιμές y και α όπου είναι ίδιες τιμές με την απλή γραμμική παλινδρόμηση (simple linear regression). Αντί της τιμής x περιλαμβάνει ένα διάνυσμα τιμών x_1, ... , x_n όπου κάθε διαφορετικό n είναι ένας

διαφορετικός προβλεπτής δηλαδή η τιμή ενός χαρακτηριστικού [11]. Ακόμη αντί της τιμής β περιλαμβάνει ένα διάνυσμα τιμών $\beta_1, \ldots, \beta_n$ όπου n είναι μια διαφορετική τιμή κλίσης της γραμμής. Παρακάτω είναι ο τύπος της πολλαπλής γραμμικής παλινδρόμησης:

$$y = \alpha + \beta_1 x_1 + \cdots + \beta_n x_n$$

Εξίσωση 17: *Πολλαπλή γραμμική παλινδρόμηση*

Η παρακάτω εικόνα δείχνει πως εφαρμόζεται ο τύπος της πολλαπλής γραμμικής παλινδρόμησης.



Εικόνα 13: *Πολλαπλή γραμμική παλινδρόμηση*

2.5.5 Λογιστική παλινδρόμηση

Η λογιστική παλινδρόμηση (logistic regression) είναι ένας αλγόριθμος επιβλεπόμενης μάθησης και είναι ένα γραμμικό μοντέλο κατηγοριοποίησης. Χρησιμοποιείται συνήθως για την εκτίμηση της πιθανότητας ότι ένα στιγμιότυπο ανήκει σε μια συγκεκριμένη κλάση [11]. Η διαφορά της με την γραμμική παλινδρόμηση (linear regression) είναι ότι η λογιστική παλινδρόμηση (logistic regression) χρησιμοποιείται για την πρόβλεψη μιας τιμής κατηγοριοποίησης ενώ η γραμμική παλινδρόμηση (linear regression) για την πρόβλεψη μιας συνεχούς μεταβλητής. Ο αλγόριθμος της λογιστικής παλινδρόμησης (logistic regression) βασίζεται στο μοντέλο της γραμμικής παλινδρόμησης [18]. Μερικοί τύποι λογιστικής παλινδρόμησης είναι η δυαδική λογιστική παλινδρόμηση (binary logistic regression) και η πολυωνυμική λογιστική παλινδρόμηση (multinomial logistic regression).

2.5.5.1 Δυαδική λογιστική παλινδρόμηση

Η δυαδική λογιστική παλινδρόμηση (binary logistic regression) χρησιμοποιείται όταν η κλάση του συνόλου δεδομένων δηλαδή η έξοδος έχει 2 πιθανά αποτελέσματα δηλαδή είναι μια δυαδική

κατηγοριοποίηση. Για παράδειγμα στο σύνολο της παρούσας διπλωματικής το οποίο θα περιγραφεί στο επόμενο κεφάλαιο προβλέπτει δασικές πυρκαγιές και η έξοδος είναι φωτιά (fire) ή όχι φωτιά (not fire). Βασίζεται στην απλή γραμμική παλινδρόμηση και για την ανάλυση της υπολογίζεται ο τύπος odds είναι ο λόγος των πιθανοτήτων. Ο τύπος αυτός περιλαμβάνει την τιμή p όπου είναι η επιτυχία της πιθανότητας και την τιμή 1-p όπου είναι η αποτυχία της πιθανότητας. Ο τύπος odds εμφανίζεται παρακάτω:

$$odds = \frac{p}{1-p}$$

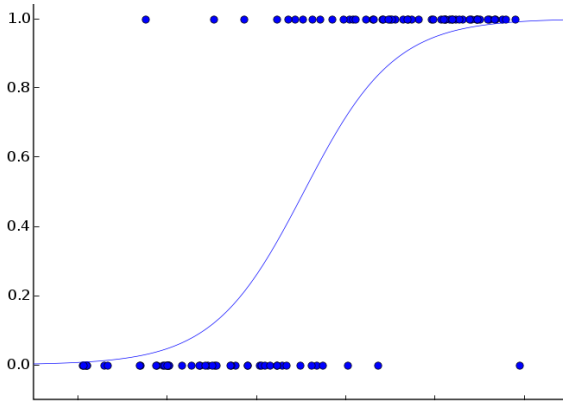
Εξίσωση 18: *Τύπος odds*

Ο τύπος logit είναι ένας φυσικός λογάριθμος της πιθανότητας και χρησιμοποιεί τον τύπο odds μέσα σε λογάριθμο. Παρακάτω εμφανίζεται ο τύπος logit:

$$logit(p) = \ln(\frac{p}{1-p}) = \alpha + \beta x$$

Εξίσωση 19: *Τύπος logit σε δυαδική λογιστική παλινδρόμηση*

Η παρακάτω εικόνα δείχνει πως εφαρμόζεται ο τύπος της δυαδικής λογιστικής παλινδρόμησης.



Εικόνα 14*: Δυαδική λογιστική παλινδρόμηση*

2.5.5.2 Πολυωνυμική λογιστική παλινδρόμηση

Η πολυωνυμική λογιστική παλινδρόμηση (multinomial logistic regression) χρησιμοποιείται όταν η κλάση του συνόλου δεδομένων δηλαδή η έξοδος έχει περισσότερα από 2 πιθανά αποτελέσματα. Δηλαδή είναι μία πολλαπλή κατηγοριοποίηση. Επίσης χρησιμοποιεί τον τύπο ods καθώς και τον τύπο logit αλλά σε πολυωνυμική λογιστική παλινδρόμηση και εμφανίζεται παρακάτω:

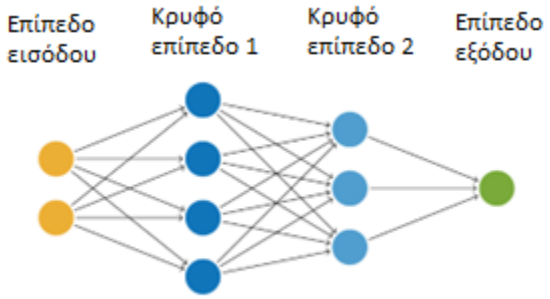
$$logit(p) = \ln(\frac{p}{1-p}) = \alpha + \beta_1 x_1 + \cdots + \beta_n x_n$$

Εξίσωση 20: *Τύπος logit σε πολυωνυμική λογιστική παλινδρόμηση*

2.5.6 Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα (neural networks) [27-30] είναι ένα μοντέλο υπολογισμού για την επίλυση προβλημάτων τεχνητής νοημοσύνης. Η δομή του είναι εμπνευσμένη από την δομή των νευρώνων του ανθρώπινου εγκεφάλου. Τα νευρωνικά δίκτυα εφαρμόστηκαν για πρώτη φορά το 1944 από τους ερευνητές Warren McCullough και Walter Pitts. Τα τεχνητά νευρωτικά δίκτυα (Artificial Neural Networks) χρησιμοποιούνται για την επίλυση προβλημάτων κατηγοριοποίησης ή παλινδρόμησης.

Τα τεχνητά νευρωτικά δίκτυα (Artificial Neural Networks - ANN) οργανώνονται σε 3 επίπεδα (layers), το επίπεδο εισόδου (input layer), τα κρυμμένα επίπεδα (hidden layers) και το επίπεδο εξόδου (output layer). Το επίπεδο εισόδου (input layer) αποτελούν την είσοδο ή τις εισόδους του τεχνητού νευρωνικού δικτύου και επικοινωνούν με τα κρυμμένα επίπεδα (hidden layers) [16]. Τα κρυμμένα επίπεδα (hidden layers) τα οποία μπορούν να είναι από ένα έως πολλά και συνδέονται με το επίπεδο εξόδου (output layer). Τέλος το επίπεδο εξόδου (output layer) είναι η έξοδος του τεχνητού νευρωνικού δικτύου.



Εικόνα 15: *Τεχνητό νευρωνικό δίκτυο*

Οι είσοδοι x_i μπορεί να είναι από έναν έως πολλοί. Κάθε είσοδος x_i περιέχει ένα βάρος (weight) w_i και τα αποτελέσματα τους υπολογίζονται μέσω της συνάρτησης αθροίσματος (summation function) [16]. Η συνάρτηση αθροίσματος υπολογίζεται από τον παρακάτω τύπο:

$$I_j = \sum_{i=1}^{n} x_{ij} w_i$$

Εξίσωση 21: *Συνάρτηση αθροίσματος*

Συνεχίζοντας κάθε ενδιάμεσος νευρώνας περιέχει μια τιμή κατωφλίου (threshold value) θ. Για να δώσει την έξοδο του τεχνητού νευρωτικού δικτύου (Artificial Neural Network) η τιμή τιμή κατωφλίου (threshold value) θ πρέπει να είναι μικρότερη από την συνάρτηση μετάβασης (transfer function) [16] δηλαδή όταν $I_j > \theta$.

Ο τεχνητός νευρώνας Perceptron περιέχει n εισόδους x οι οποίοι δίνονται μέσω της συνάρτησης μετάβασης g [16] και δίνει την έξοδο y. Η συνάρτηση μετάβασης υπολογίζεται από τον παρακάτω τύπο:

$$y_j = g(I_j) = g\left(\sum_{i=1}^{n} x_{ij} w_i\right)$$

Εξίσωση 22: *Συνάρτηση μετάβασης*

Μερικές από τις συναρτήσεις μετάβασης είναι η βηματική συνάρτηση ή συναρτήσεις κατωφλίου (threshold function) όπου η έξοδος είναι 0 ή 1, η συνάρτηση προσήμου (sign function) όπου η έξοδος είναι -1 ή +1, η συνάρτηση αναρρίχησης (ramping function) όπου η έξοδος είναι από 0 έως 1.

Στον παρακάτω πίνακα παρατηρούνται μερικά από τα γνωστότερα τεχνητά νευρωτικά δίκτυα (Artificial Neural Networks).

Πίνακας 1: *Τεχνητά νευρωνικά δίκτυα*

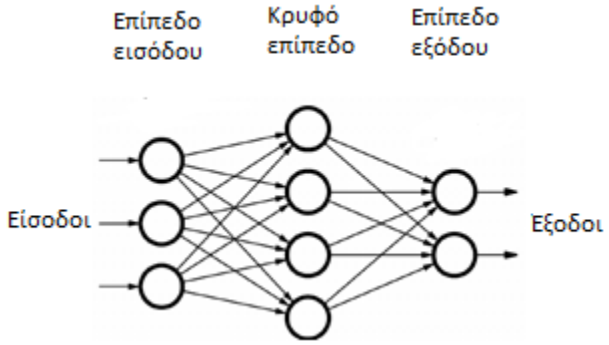
Όνομα Τεχνητού Νευρωνικού δικτύου	**Κατασκευαστής**	**Έτος**	**Τρόπος Εκπαίδευσης**
Perceptron	Rosenblatt (USA)	1957-1962	Με επίβλεψη
Adaline / Madaline	Widrow (USA)	1960-1962	Με επίβλεψη
Back-propagation	Werbow, Rumelhart et al	1974-1986	Με επίβλεψη

Self-organizing map	Kohonen (Finland)	1981	Χωρίς επίβλεψη
Hopfield Net	Hopfield (USA)	1982	Με επίβλεψη
Boltzmann machine	Hinton (Canada), Hopkins, Szu (USA)	1985-1986	Με επίβλεψη

2.5.6.1 Αρχιτεκτονικές νευρωνικών δικτύων

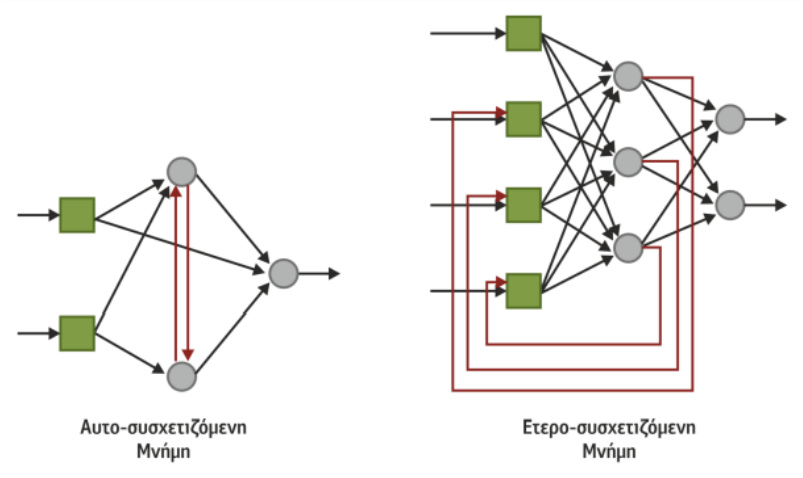
Υπάρχουν δύο βασικές αρχιτεκτονικές νευρωνικών δικτύων, η πρόσθια τροφοδότηση (feed forward) και η οπίσθια τροφοδότηση (feed backward).

Στα νευρωνικά δίκτυα πρόσθιας τροφοδότησης η έξοδος μιας μονάδας αποτελεί την είσοδο της επόμενης μονάδας. Δηλαδή οι μονάδες οργανώνονται σε διαφορετικά επίπεδα με σκοπό έτσι ώστε οι μονάδες του πρώτου επιπέδου να τροφοδοτούν τις μονάδες του επόμενου επιπέδου μέχρι να τροφοδοτηθούν οι μονάδες του τελευταίου επιπέδου [16].



Εικόνα 16: *Νευρωνικό δίκτυο πρόσθιας τροφοδότησης*

Στα νευρωνικά δίκτυα οπίσθιας τροφοδότησης (feed backward) χωρίζονται σε δύο κατηγορίες, στις αυτοσυσχετιζόμενες μνήμες (autoassociated memories) και στις ετεροσυσχετιζόμενες μνήμες (heteroassociated memories). Στις αυτοσυσχετιζόμενες μνήμες η ανατροφοδότηση γίνεται μόνο στους κόμβους του ίδιου επιπέδου. Διαφορετικά τα νευρωνικά δίκτυα οπίσθιας τροφοδότησης (feed backward) είναι ετεροσυσχετιζόμενες μνήμες. Η παρακάτω εικόνα από το βιβλίο [16] παρουσιάζει αυτά τα 2 νευρωνικά δίκτυα οπίσθιας τροφοδότησης.



Εικόνα 17: *Νευρωνικά δίκτυα οπίσθιας τροφοδότης [16]*

2.5.6.2 Νευρωνικά δίκτυα πολλών επιπέδων

Τα νευρωνικά δίκτυα πολλών επιπέδων ανήκουν στην κατηγορία της πρόσθιας τροφοδότησης. Αποτελείται από ένα επίπεδο εισόδου το οποίο μπορεί να έχει από έναν η περισσότερους κόμβους, από ένα ή περισσότερα κρυφά επίπεδα [11] και τέλος από από ένα επίπεδο εξόδου. Τα επίπεδα που βρίσκονται κοντά στο επίπεδο εισόδου ονομάζονται κατώτερα επίπεδα ενώ αυτά που βρίσκονται κοντά στο επίπεδο εξόδου ονομάζονται ανώτερα επίπεδα [11]. Η εικόνα 14 είναι ένα τέτοια παράδειγμα.

Οι κόμβοι του κάθε επιπέδου μπορεί να είναι συνδεδεμένοι με δύο τρόπους δηλαδή πλήρως συνδεδεμένοι (fully connected) ή μερικώς συνδεδεμένοι (partially connected). Στους πλήρως συνδεδεμένους κόμβους κάθε κόμβος του ενός επιπέδου συνδέεται με τους κόμβους του επόμενου επιπέδου [16] ενώ στους μερικώς συνδεδεμένους κόμβους κάθε κόμβος συνδέεται σε μερικούς κόμβους του επόμενου επιπέδου.

2.5.7 Απλοϊκός Bayes

Ο κατηγοριοποιητής απλοϊκός Bayes (Naïve Bayes) είναι ένα αλγόριθμος επιβλεπόμενης μάθησης και βασίζεται στην εφαρμογή του θεωρήματος Bayes. Το θεώρημα Bayes βρίσκει την πιθανότητα να συμβεί ένα γεγονός μιας δεδομένης πιθανότητας ενός άλλου γεγονότος που έχει ήδη συμβεί. Έστω ότι $X = (x_1, x_2, \dots, x_n)$ ένα διάνυσμα με n χαρακτηριστικά και C η κλάση με m κατηγορίες $C_1, C_2, \dots, C_m$. Για την κατηγοριοποίηση ο απλοϊκός Bayes υπολογίζει τις πιθανότητες

$P(C_1|X), P(C_2|X), \dots, P(C_m|X)$ [18] και στην συνέχεια υπολογίζει την βέλτιστη πιθανότητα $P(C|X)$ [25]. Ο τύπος του θεωρήματος του Baynes είναι ο παρακάτω:

$$P(C_i|X) = \frac{P(X|C_i) \cdot P(C_i)}{P(X)}$$

Εξίσωση 23: *Θεώρημα Baynes*

Ο απλοϊκός Bayes είναι ένας αλγόριθμος μηχανικής μάθησης, αλλά πιο συγκεκριμένα χρησιμοποιείται για κατηγοριοποίηση αλλά και για παλινδρόμηση. Το $P(X)$ είναι μια σταθερά και η τιμή της παραμένει ίδια σε όλες τις κατηγορίες οπότε μπορεί να αφαιρεθεί οδηγώντας στην παρακάτω εξίσωση:

$$P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i)$$

Εξίσωση 24: *Απλοϊκός Bayes*

Ακόμη ο απλοϊκός Bayes είναι ένας από τους πιο αποτελεσματικούς αλγόριθμους μηχανικής μάθησης καθώς είναι πολύ εύκολος στην εφαρμογή του και δεν απαιτεί πολλά δεδομένα εκπαίδευσης (training data).

Υπάρχουν τρεις κύριοι τύποι Naive Bayes που χρησιμοποιούνται ο πολυωνυμικός απλοϊκός Bayes (multinomial Naive Bayes), ο Bernoulli απλοϊκός Bayes (Bernoulli Naive Bayes) και ο Gaussian απλοϊκός Bayes (Gaussian Naive Bayes).

2.5.7.1 Gaussian απλοϊκός Bayes

Ο Gaussian απλοϊκός Bayes (Gaussian Naive Bayes) εφαρμόζει την κατανομή του Gauss. Χρησιμοποιείται όταν οι τιμές πρόβλεψης είναι συνεχείς (continues). Για τον υπολογισμό του χρησιμοποιείται ο παρακάτω τύπος και οι τιμές σ_C και μ_C εκτιμώνται χρησιμοποιώντας τη μέγιστη πιθανότητα.

$$P(X_i|C) = \frac{1}{\sqrt{2\pi\sigma_C^2}} \, e^{-\frac{(X_i-\mu_C)^2}{2\sigma_C^2}}$$

Εξίσωση 25: *Gaussian απλοϊκός Bayes*

2.5.7.2 Bernoulli απλοϊκός Bayes

Ο Bernoulli απλοϊκός Bayes (Bernoulli Naive Bayes) εφαρμόζει την κατανομή του Bernoulli. Χρησιμοποιείται όταν η τιμή πρόβλεψης είναι δυαδικού χαρακτήρα δηλαδή αληθής ή ψευδής (αλλιώς 1 ή 0). Ο τύπος που χρησιμοποιεί είναι ο παρακάτω:

$$P(X_i|C) = P(i|C)X_i + (1 - P(i|C))(1 - X_i)$$

Εξίσωση 26: *Bernoulli απλοϊκός Bayes*

2.5.7.3 Πολυωνυμικός απλοϊκός Bayes

Ο πολυωνυμικός απλοϊκός Bayes (multinomial Naive Bayes) εφαρμόζει την πολυωνυμική κατανομή. χρησιμοποιείται συχνά για την επίλυση ζητημάτων που αφορούν την ταξινόμηση εγγράφων ή κειμένου. Για τον υπολογισμό του χρησιμοποιείται ο παρακάτω τύπος όπου $\theta_C = (\theta_{C_1}, \theta_{C_2}, \dots, \theta_{C_n})$ είναι ένα διάνυσμα της κάθε κλάσης C και n ο αριθμός των χαρακτηριστικών. Η τιμή θ_C εκτιμώνται χρησιμοποιώντας τη μέγιστη πιθανότητα και όπου N_{C_i} είναι ο αριθμός που εμφανίζεται το χαρακτηριστικό σε ένα δείγμα τάξης στο σύνολο εκπαίδευσης.

$$\hat{\theta}_{C_i} = \frac{N_{C_i} + \alpha}{N_C + \alpha n}$$

Εξίσωση 27: *Πολυωνυμικός απλοϊκός Bayes*

2.6 Αξιολόγηση αλγορίθμων μηχανικής μάθησης

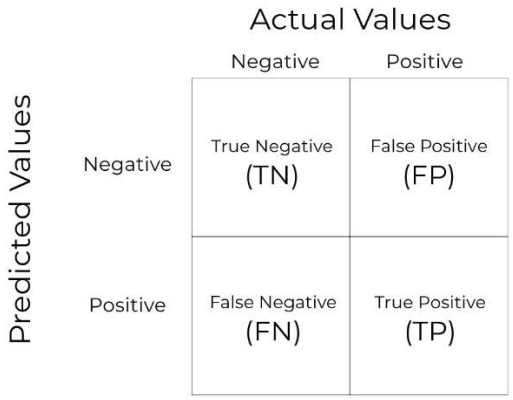
Η αξιολόγηση των επιβλεπόμενων (supervised) και των μη επιβλεπόμενων (unsupervised) αλγόριθμων μηχανικής μάθησης είναι διαφορετική. Επειδή οι αλγόριθμοι που θα χρησιμοποιηθούν ανήκουν στην επιβλεπόμενη μάθηση (supervised learning) και η κλάση του συνόλου δεδομένων ανήκει στην δυαδική κατηγοριοποίηση (binary classification) παρακάτω θα περιγραφούν τα μέτρα της αξιολόγησης των αλγόριθμων επιβλεπόμενης μάθησης. Αυτά είναι ο πίνακας σύγχυσης (confusion matrix) όπου μέσω αυτού θα υπολογιστεί η ορθότητα (accuracy), η ακρίβεια (precision), η ανάκληση (recall) και τέλος ο αρμονικός μέσος (f1 score).

2.6.1 Πίνακας σύγχυσης

Ο πίνακας σύγχυσης (confusion matrix) είναι μια τεχνική για την σύνοψη της απόδοσης ενός αλγόριθμου μηχανικής μάθησης. Ουσιαστικά είναι μια μέτρηση απόδοσης και χρησιμοποιείται σε

προβλήματα κατηγοριοποίησης (classification). Η έξοδος μπορεί να έχει δύο ή περισσότερες κλάσεις ανάλογα με το αν η κατηγοριοποίηση είναι δυαδική (binary) ή πολλαπλή (multiple). Τα δεδομένα του πίνακα σύγχυσης (confusion matrix) χωρίζονται σε TP (True Positive), TN (True Negative), FP (False Positive), και FN (False Negative) και το κάθε ένα συμβολίζει:

- TP (True Positive): είναι ο αριθμός των θετικών και αληθινών προβλέψεων
- TN (True Negative): είναι ο αριθμός των αρνητικών και αληθινών προβλέψεων
- FP (False Positive): είναι ο αριθμός των θετικών και ψευδών προβλέψεων
- FN (False Negative): είναι ο αριθμός των αρνητικών και ψευδών προβλέψεων



Εικόνα 18: *Πίνακας σύγχυσης*

Μέσω του πίνακα σύγχυσης (confusion matrix) το μοντέλο μπορεί να αξιολογηθεί. Έτσι μπορεί να υπολογιστεί η ορθότητα (accuracy), η ακρίβεια (precision), η ανάκληση (recall) και ο αρμονικός μέσος (f1 score).

2.6.2 Ορθότητα

Η ορθότητα (accuracy) είναι το σημαντικότερο κριτήριο αξιολόγησης του μοντέλου μηχανικής μάθησης σε προβλήματα κατηγοριοποίησης (classification). Υπολογίζει την ορθότητα (accuracy) του μοντέλου δηλαδή το ποσοστό των θετικών προβλέψεων σε ένα σύνολο δεδομένων διαιρώντας τον αριθμό τις θετικές προβλέψεις με τις συνολικές προβλέψεις. Υπολογίζεται από τον παρακάτω τύπο:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Εξίσωση 28: *Τύπος ορθότητας*

Όσο πιο υψηλό είναι το ποσοστό της ορθότητας (accuracy) τόσο πιο ορθά είναι τα αποτελέσματα του μοντέλου.

2.6.3 Ακρίβεια

Η ακρίβεια (precision) είναι και αυτή ένα πολύ σημαντικό κριτήριο αξιολόγησης του μοντέλου μηχανικής μάθησης σε προβλήματα κατηγοριοποίησης (classification). Υπολογίζει την ακρίβεια (precision) του μοντέλου δηλαδή το ποσοστό της ορθότητας των θετικών προβλέψεων [11]. Για τον υπολογισμό της διαιρούνται ο αριθμός των θετικών και αληθινών προβλέψεων με τον συνολικό αριθμό όλων των θετικών προβλέψεων. Υπολογίζεται από τον παρακάτω τύπο:

$$Precision = \frac{TP}{TP + FP}$$

Εξίσωση 29: *Τύπος ακρίβειας*

Όσο πιο υψηλό είναι το ποσοστό της ακρίβειας (precision) τόσο πιο ακριβές είναι τα αποτελέσματα του μοντέλου.

2.6.4 Ανάκληση

Η ανάκληση (recall) είναι ένα κριτήριο αξιολόγησης του μοντέλου μηχανικής μάθησης σε προβλήματα κατηγοριοποίησης (classification). Επίσης ονομάζεται αλλιώς και πραγματικό θετικό ποσοστό (TPR) και υπολογίζει το ποσοστό των θετικών προβλέψεων που ανίχνευσε σωστά ο ταξινομητής (classifier) [11]. Για τον υπολογισμό του διαιρείται ο αριθμός των θετικών και αληθινών προβλέψεων με τον αριθμό των αρνητικών και ψευδών προβλέψεων με τον αριθμό των θετικών και αληθινών προβλέψεων. Υπολογίζεται από τον παρακάτω τύπο:

$$Recall = \frac{TP}{TP + FN}$$

Εξίσωση 30: *Τύπος ανάκλησης*

Όσο πιο υψηλό είναι το ποσοστό της ανάκλησης (recall) τόσο πιο σωστά είναι τα αποτελέσματα του μοντέλου.

2.6.5 Αρμονικός μέσος

Είναι αρκετά δύσκολο να συγκριθούν δύο μοντέλα με χαμηλή ακρίβεια (precision) και υψηλή ανάκληση (recall) ή το αντίθετο. Για αυτόν τον λόγω χρησιμοποιείται ο αρμονικός μέσος (f1 score). Χρησιμοποιείται για την μέτρηση της ακρίβειας (precision) και της ανάκλησης (recall) ταυτόχρονα. Για τον υπολογισμό του πολλαπλασιάζεται η ακρίβεια (precision) και η ανάκληση (recall) και διαιρείται με την πρόσθεση της ακρίβειας (precision) και της ανάκλησης (recall). Το αποτέλεσμα πολλαπλασιάζεται με το 2 και υπολογίζεται από τον παρακάτω τύπο:

$$F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Εξίσωση 31: *Τύπος αρμονικού μέσου*

2.7 Αποτελέσματα ερευνών συγκριτικής αξιολόγησης

Παρακάτω θα παρουσιαστούν δύο έρευνες οι οποίες χρησιμοποιούν αλγόριθμους μηχανικής μάθησης για κατηγοριοποίηση και κάνουν συγκριτική αξιολόγηση. Η πρώτη έρευνα κάνει Συγκριτική αξιολόγηση αλγορίθμων για συναισθήματα ανάλυσης υπηρεσιών κοινωνικής δικτύωσης στο Twitter. Η δεύτερη έρευνα προβλέπει δασικές πυρκαγιές και κάνει συγκριτική αξιολόγηση αλγόριθμων μηχανικής μάθησης. Επίσης αυτή η έρευνα χρησιμοποιεί το ίδιο σύνολο δεδομένων με την παρούσα διπλωματική εργασία.

2.7.1 Συγκριτική αξιολόγηση αλγορίθμων για συναισθήματα ανάλυσης υπηρεσιών κοινωνικής δικτύωσης

Ένα παράδειγμα συγκριτικής αξιολόγησης είναι η έρευνα [6]. Η έρευνα αυτή κάνει συγκριτική αξιολόγηση αλγορίθμων για συναισθήματα ανάλυσης υπηρεσιών κοινωνικής δικτύωσης. Πιο συγκεκριμένα χρησιμοποιεί δεδομένα από το Twitter το οποίο είναι ένα από τα μεγαλύτερα κοινωνικά δίκτυα. Σκοπός του είναι να κάνει ανάλυση υπηρεσιών χρησιμοποιώντας αλγόριθμους μηχανικής μάθησης για κατηγοριοποίηση και στο τέλος να κάνει συγκριτική αξιολόγηση μεταξύ τους.

Για την έρευνα [6] χρησιμοποιούνται 3 διαφορετικά σύνολα δεδομένων από το Twitter, το Obama-McCain Debate (OMD), το Health Care Reform (HCR) και το STS Gold Standard (STS-Gold). Κάθε σύνολο δεδομένων περιέχει έναν μεγάλο αριθμό δεδομένων από tweets και η έξοδος δηλαδή

το αποτέλεσμα από κάθε σύνολο δεδομένων είναι positive ή negative. Δηλαδή 0 ή 1 (0 για negative και 1 για positive) δηλαδή είναι ένα πρόβλημα δυαδικής κατηγοριοποίησης.

Το σύνολο δεδομένων Obama-McCain Debate (OMD) περιέχει συνολικά 1904 tweets όπου τα 709 είναι positive και τα υπόλοιπα 1195 είναι negative. Τα δεδομένα αυτά ανιχνεύτηκαν κατά τη διάρκεια της πρώτης προεδρικής τηλεοπτικής συζήτησης των ΗΠΑ τον Σεπτέμβριο του 2008.

Το σύνολο δεδομένων Health Care Reform (HCR) περιέχει συνολικά 1922 tweets όπου τα 541 είναι positive και τα υπόλοιπα 1381 είναι negative. Τα δεδομένα αυτά είναι από tweets που χρησιμοποιούσαν το hashtag #hcr τον Μάρτιο του 2010.

Το σύνολο δεδομένων STS Gold Standard (STS-Gold) περιέχει συνολικά 2034 tweets όπου τα 632 είναι positive και τα υπόλοιπα 1402 είναι negative. Τα δεδομένα αυτά συλλέχθηκαν από tweets από το Stanford Twitter Sentiment Corpus10.

Οι αλγόριθμοι μηχανικής μάθησης για κατηγοριοποίηση είναι ο απλοϊκός bayes, η μηχανή διανυσματικής υποστήριξης, ο K-κοντινότεροι-γείτονες, η λογιστική παλινδρόμηση και ο C4.5.

Για την διαδικασία της αξιολόγησης κάθε σύνολο δεδομένων επεξεργάστηκε και χωρίστε σε κατηγορίες. Στην πρώτη κατηγορία θα χρησιμοποιηθεί η τεχνική split η οποία θα παρουσιαστεί στο επόμενο κεφάλαιο και στην δεύτερη κατηγορία θα χρησιμοποιηθεί η τεχνική cross validation. Κάθε κατηγορία θα χρησιμοποιήσει τους ίδιους αλγόριθμους μηχανικής μάθησης και στο τέλος θα τους συγκρίνει.

Για την αξιολόγηση των αλγόριθμων μηχανικής μάθησης θα χρησιμοποιηθεί η ακρίβεια, η ανάκληση και ο αρμονικός μέσος όπου παρουσιάστηκαν προηγουμένος στο κεφάλαιο 2.6.

Τα αποτελέσματα της αξιολόγησης των αλγόριθμων μηχανικής μάθησης για κατηγοριοποίηση και στα 3 σύνολα δεδομένων παρουσιάζονται παρακάτω [6]:

Πίνακας 2: *Αποτελέσματα έρευνας [6]*

		Split			**Cross validation**		
Σύνολο δεδομένων	**Κατηγοριοποιητές**	Ακρίβεια	Ανάκληση	Αρμονικός μέσος	Ακρίβεια	Ανάκληση	Αρμονικός μέσος
OMD	NB	0.807	0.809	**0.804**	0.810	0.811	**0.806**

	SVM	**0.811**	**0.811**	0.803	**0.822**	**0.812**	0.802
	KNN	0.690	0.625	0.632	0.692	0.636	0.641
	LR	0.768	0.743	0.748	0.753	0.742	0.745
	C4.5	0.726	0.734	0.725	0.750	0.753	0.742
HCR	NB	**0.749**	**0.763**	0.709	0.760	0.767	0.728
	SVM	0.742	0.763	0.719	0.758	**0.770**	**0.737**
	KNN	0.537	0.733	0.620	**0.799**	0.721	0.606
	LR	0.717	0.726	**0.721**	0.713	0.723	0.717
	C4.5	0.690	0.724	0.696	0.720	0.742	0.722
STS-Gold	NB	**0.818**	**0.820**	**0.806**	**0.801**	**0.797**	**0.776**
	SVM	0.770	0.780	0.762	0.790	0.786	0.761
	KNN	0.502	0.708	0.587	0.475	0.689	0.562
	LR	0.797	0.767	0.775	0.738	0.744	0.741
	C4.5	0.722	0.739	0.690	0.731	0.743	0.711

Αυτό που παρατηρείται στο σύνολο δεδομένων OMD με την μέθοδο split η μηχανή διανυσματικής υποστήριξης (SVM) έχει υψηλότερη ακρίβεια και ανάκληση ενώ απλοϊκός bayes (NB) έχει αρμονικός μέσο. Στην μέθοδο cross validation και εδώ η μηχανή διανυσματικής υποστήριξης (SVM) έχει υψηλότερη ακρίβεια και ανάκληση ενώ απλοϊκός bayes (NB) έχει αρμονικός μέσο.

Στο σύνολο δεδομένων HCR με την μέθοδο split ο απλοϊκός bayes (NB) έχει έχει υψηλότερη ακρίβεια και ανάκληση ενώ η γραμμική παλινδρόμηση (LR) έχει υψηλότερο αρμονικό μέσο. Στην μέθοδο cross validation ο K-κοντινότεροι-γείτονες (KNN) έχει υψηλότερη ακρίβεια ενώ η μηχανή διανυσματικής υποστήριξης (SVM) έχει υψηλότερη ανάκληση και αρμονικό μέσο.

Τέλος στο σύνολο δεδομένων STS-Gold με την μέθοδο split και την μέθοδο cross validation ο απλοϊκός bayes (NB) έχει τα υψηλότερα αποτελέσματα.

Συνεπώς αυτό που συμπεραίνει η έρευνα [6] είναι ότι ο απλοϊκός bayes (NB) και η η μηχανή διανυσματικής υποστήριξης (SVM) είναι οι πιο βέλτιστοι.

2.7.2 Πρόβλεψη δασικών πυρκαγιών στην Αλγερία

Η έρευνα [8] προβλέπει δασικές πυρκαγιές στην Αλγερία χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων. Χρησιμοποιεί το ίδιο σύνολο δεδομένων Algerian Forest Fires Dataset με την διπλωματική εργασία το οποίο περιγράφεται στο κεφάλαιο 3.2. Η έξοδος του συνόλου δεδομένων είναι fire ή not fire. Δηλαδή 0 ή 1 (0 για not fire και 1 για fire) δηλαδή είναι ένα πρόβλημα δυαδικής κατηγοριοποίησης.

Χρησιμοποιεί αλγόριθμους μηχανικής μάθησης που χρησιμοποιούν δέντρα απόφασης. Οι αλγόριθμοι αυτοί είναι τα δέντρα απόφασης, τα τυχαία δάση, ο AdaBoost και τέλος ο Bagging. Οι αλγόριθμοι αυτοί παρουσιάστηκαν στο κεφάλαιο 2.5

Για την αξιολόγηση χρησιμοποιεί την ορθότητα, την ακρίβεια και την ανάκληση όπου παρουσιάστηκαν προηγουμένος το κεφάλαιο 2.6.

Τα αποτελέσματα της έρευνας για την τη αξιολόγησης των αλγόριθμων μηχανικής μάθησης για κατηγοριοποίηση στο σύνολο δεδομένων Algerian Forest Fires Dataset παρουσιάζονται στον παρακάτω πίνακα [8].

Πίνακας 3: *Αποτελέσματα έρευνας [8]*

Κατηγοριοποιητές	**Ορθότητα**	**Ακρίβεια**	**Ανάκληση**
Δέντρα απόφασης	82.89	0.92	0.787
Τυχαία δάση	72.36	0.75	0.732
AdaBoost	**84.21**	**0.95**	**0.792**
Bagging	78.94	0.825	0.786

Αυτό που παρατηρείται από τα αποτελέσματα είναι ότι ο AdaBoost δίνει τα καλύτερα αποτελέσματα στην ορθότητα, στην ακρίβεια αλλά και στην ανάκληση. Συνεπώς ο AdaBoost είναι ο πιο βέλτιστος.

Κεφάλαιο 3: Μεθοδολογία

Το συγκεκριμένο κεφάλαιο θα παρουσιάσει την μεθοδολογία της διπλωματικής εργασίας. Αρχικά θα παρουσιαστεί πώς μπορεί να γίνει μηχανική μάθηση μέσω της γλώσσας προγραμματισμού Python. Πιο συγκεκριμένα θα παρουσιαστεί το περιβάλλον, η γλώσσα προγραμματισμού που είναι η Python και οι βιβλιοθήκες της που θα χρησιμοποιηθούν. Συνεχίζοντας θα παρουσιαστεί το σύνολο δεδομένων που θα χρησιμοποιηθεί και θα περιγραφεί και θα παρουσιαστούν τα χαρακτηριστικά του καθώς και ορισμένα στατιστικά του. Επιπλέον θα παρουσιαστούν οι αλγόριθμοι μηχανικής μάθησης που θα χρησιμοποιηθούν καθώς και οι παράμετροί τους. Τέλος θα παρουσιαστούν τα βήματα υλοποίησης.

3.1 Μηχανική Μάθηση με χρήση της Python

Η μηχανική μάθηση μπορεί να χρησιμοποιηθεί μέσω πολλών γλωσσών. Μία από τις πιο γνωστές γλώσσες προγραμματισμού για μηχανική μάθηση είναι η Python και για αυτόν τον σκοπό επιλέχτηκε για την υλοποίηση της παρούσας διπλωματικής εργασίας. Παρακάτω θα αναλυθεί η μεθοδολογία της μηχανικής μάθησης χρησιμοποιώντας την γλώσσα προγραμματισμού Python. Δηλαδή από το περιβάλλον της μέχρι και της βιβλιοθήκες για την μηχανική μάθηση.

3.1.1 Περιβάλλον

Για περιβάλλον της διπλωματικής εργασίας για την υλοποίηση της επιλέχτηκε ο Spyder. Ο Spyder είναι ένα δωρεάν περιβάλλον ανοιχτού κώδικα (open source) για επιστημονικό προγραμματισμό (scientific programming) στην γλώσσα προγραμματισμού Python. Αρχικά δημιουργήθηκε το 2009 από τον Pierre Raybaut και από τότε συντηρείται και βελτιώνεται διαρκώς. Δημιουργήθηκε για επιστήμονες, μηχανικούς και αναλυτές δεδομένων και περιέχει βιβλιοθήκες ανοιχτού κώδικα οι οποίες θα επεξηγηθούν παρακάτω.

3.1.2 Python

Για γλώσσα προγραμματισμού επιλέχτηκε η Python. Η Python είναι μια γλώσσα προγραμματισμού υψηλού επιπέδου και δημιουργήθηκε το 1991 από τον Ολλανδό προγραμματιστή Guido van Rossum. Είναι μια απλή γλώσσα και επιτρέπει στους προγραμματιστές να υλοποιούν κώδικα πιο εύκολα σε σύγκριση με άλλες γλώσσες

προγραμματισμού [21]. Επίσης η γλώσσα προγραμματισμού Python είναι μια από τις πιο δημοφιλής γλώσσες για επιστημονικό προγραμματισμό (scientific programming). Η Python είναι μία από τις πιο γνωστές γλώσσες για μηχανική μάθηση καθώς είναι πολύ ευέλικτη. Περιέχει αρκετές βιβλιοθήκες που χρησιμοποιούνται στην μηχανική μάθηση οι οποίες θα αποτελέσουν το βασικό υλικό για της υλοποίησης της διπλωματικής εργασίας.

3.1.3 NumPy

Η βιβλιοθήκη NumPy είναι μια βιβλιοθήκη λογισμικού για την γλώσσα προγραμματισμού Python και δημιουργήθηκε το 2005 από τον Travis Oliphant. Η κύρια χρήση της είναι η υποστήριξη μεγάλων πολυδιάστατων πινάκων για την αποθήκευση και των χειρισμό δεδομένων [22]. Ένα βασικό πλεονέκτημα της NumPy είναι ταχύτητα δηλαδή είναι πάρα πολύ γρήγορη. Σύγκρινοντας την με τις απλές λίστες (lists) της Python η NumPy είναι πολύ πιο γρήγορη καθώς οι δομές δεδομένων της καταλαμβάνουν λιγότερο χώρο, η εκτέλεσή της είναι πιο γρήγορη και περιλαμβάνει ενσωματωμένες συναρτήσεις όπως οι πράξεις πινάκων γραμμικής άλγεβρας (linear algebra).

3.1.4 Pandas

Η βιβλιοθήκη Pandas είναι μια βιβλιοθήκη λογισμικού για την γλώσσα προγραμματισμού Python. Ξεκίνησε να δημιουργείτε το 2008 από τον προγραμματιστή Wes McKinney. Η βιβλιοθήκη Pandas είναι βασισμένη πάνω στην βιβλιοθήκη NumPy [22]. Είναι μία από τις πιο χρήσιμες βιβλιοθήκες στην μηχανική μάθηση καθώς ο σκοπός της είναι ο χειρισμός και η ανάλυση δεδομένων. Επιτρέπει την εισαγωγή δεδομένων από διάφορες μορφές αρχείων όπως αρχεία Microsoft Excel, JSON και ερωτήματα (queries) ή πίνακες (tables) βάσεων δεδομένων (database). Μερικές από τις δυνατότητες του είναι η εύρεση ελάχιστης, μέσης και μέγιστης τιμής σε μία στήλη. Μπορεί να βρει άμα υπάρχει κάποια συσχέτιση μεταξύ ορισμένων στηλών. Επίσης είναι χρήσιμη και στην προεπεξεργασία των δεδομένων. Μπορεί να διαγράψει στήλες οι οποίες είναι μη σχετικές ή να έχουν λάθος τιμές όπως κενές στήλες. Ορισμένες χρήσιμες συναρτήσεις της βιβλιοθήκης Pandas είναι οι παρακάτω:

- **read_csv():** χρησιμοποιείται για να διαβάσει ένα .csv αρχείο και να το προσθέσει σε ένα πλαίσιο δεδομένων (DataFrame)

- **to_csv():** χρησιμοποιείται για να αποθηκεύσει ένα πλαίσιο δεδομένων (DataFrame) σε ένα .csv αρχείο
- **replace():** χρησιμοποιείται για να αντικαταστήσει τιμές σε κάποια στήλη ή στήλες του πλαισίου δεδομένων.
- **drop():** χρησιμοποιείται για να διαγράψει κάποια ή κάποιες στήλες του πλαισίου δεδομένων.
- **head():** εμφανίζει τις πρώτες 5 στήλες του πλαισίου δεδομένων.
- **tail():** εμφανίζει τις τελευταίες 5 στήλες του πλαισίου δεδομένων.
- **describe():** εμφανίζει ορισμένα στατιστικά του πλαισίου δεδομένων όπως ελάχιστη, μέση, μέγιστη τιμή, τυπική απόκλιση, και κατώτερο, μεσαίο, άνω εκατοστημόριο
- **corr():** εμφανίζει τον πίνακα συσχέτισης

3.1.5 Matplotlib

Η βιβλιοθήκη Matplotlib άρχισε να δημιουργείται το 2002 από τον John Hunter. Χρησιμοποιείται για την οπτικοποίηση δεδομένων σε γραφήματα [22] το οποίο είναι χρήσιμο για την παρουσιάσει των αποτελεσμάτων της διπλωματικής εργασίας. Μερικά από τα γραφήματα τα οποία μπορεί να οπτικοποιήσει τα δεδομένα είναι το ιστόγραμμα (histogram), πολλαπλά ιστογράμματα, το διάγραμμα πίτας (pie chart), το διάγραμμα του κουτιού (boxplot), το διάγραμμα της μπάρας (bar plot) και χρονοσειρές ανά χρόνο. Ακόμη μπορεί να οπτικοποιήσει δεδομένα σε δισδιάστατα (2D) αλλά και σε τρισδιάστατα (3D) γραφήματα. Ορισμένες χρήσιμες συναρτήσεις της βιβλιοθήκης Matplotlib είναι οι παρακάτω:

- **plot():** δημιουργεί ένα γράφημα στους άξονες x και y.
- **show():** εμφανίζει ένα γράφημα που έχει δημιουργηθεί

3.1.6 Seaborn

Η Seaborn είναι μια βιβλιοθήκη οπτικοποίησης δεδομένων Python που βασίζεται στην βιβλιοθήκη Matplotlib. Κύριος στόχος της είναι η παροχή διεπαφής υψηλού επιπέδου με σκοπό την σχεδίαση ελκυστικών και ενημερωτικών στατιστικών γραφημάτων. Θα είναι πολύ χρήσιμη στην παρούσα διπλωματική εργασία μαζί με την Matplotlib για την οπτικοποίηση γραφημάτων για

αποτελέσματα όπως είναι ο πίνακας συσχέτισης των συνόλων δεδομένων ο οποίος θα παρουσιαστεί παρακάτω.

3.1.7 Scikit-learn

Η Scikit-learn είναι μια βιβλιοθήκη μηχανικής μάθησης για την γλώσσα προγραμματισμού Python. Αναπτύχθηκε το 2007 από τον David Cournapeau ως έργο της Google. Είναι μία από τις πιο γνωστές και χρήσιμες βιβλιοθήκες για μηχανική μάθηση σήμερα [22]. Περιέχει διάφορους αλγόριθμους μηχανικής μάθησης για επιβλεπόμενη μάθηση και μη επιβλεπόμενη μάθηση. Χρησιμοποιείται για κατηγοριοποίηση (classification), παλινδρόμηση (regression) συσταδοποίηση (clustering) και μείωση διαστάσεων (dimensionality reduction). Επίσης χρησιμοποιείται και για επιλογή μοντέλου (model selection) και προεπεξεργασία (preprocessing). Ορισμένες χρήσιμες συναρτήσεις της βιβλιοθήκης Scikit-learn είναι οι παρακάτω:

- **train_test_split():** χρησιμοποιείται για να διαχωρίσει τα δεδομένα εκπαίδευσης και δομικής.
- **fit():** είναι ένας εκτιμητής και χρησιμοποιείται από έναν classifier για να εισάγει τα δεδομένα εκπαίδευσης.
- **predict():** χρησιμοποιείται από έναν classifier για να γίνει πρόβλεψη
- **confusion_matrix():** υπολογίζει τον πίνακα συσχέτισης

3.2 Σύνολο δεδομένων Algerian Forest Fires Dataset

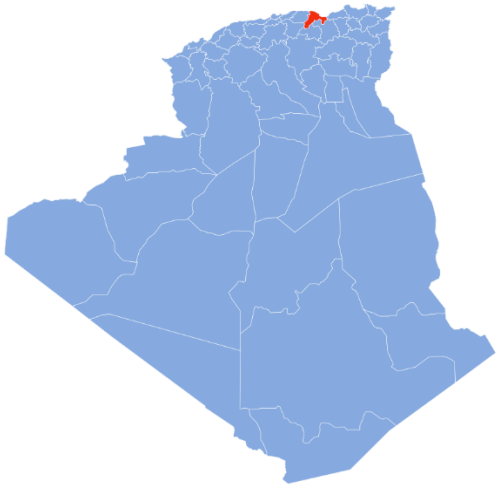
Παρακάτω θα παρουσιαστεί το σύνολο δεδομένων Algerian Forest Fires Dataset όπου θα περιγραφεί και στην συνέχεια θα παρουσιαστούν ορισμένα χρήσιμα χαρακτηριστικά. Αρχικά το σύνολο δεδομένων Algerian Forest Fires Dataset θα περιγραφεί έτσι ώστε να υπάρχει μια γνώση πάνω στα δεδομένα, από ποια ή ποιες περιοχές προέρχονται και πόσες εγγραφές υπάρχουν. Στην συνέχεια θα παρουσιαστούν τα χαρακτηριστικά του ώστε να υπάρχει μια γνώση από το πόσα είναι, από τι τύπο τιμής αποτελούνται, πόσα είναι τα χαρακτηριστικά εισόδου και πόσα εξόδου καθώς και ποια θα χρησιμοποιηθούν. Συνεχίζοντας θα παρουσιαστούν μερικές εγγραφές έτσι ώστε να υπάρχει μια πρώτη ματιά στα δεδομένα. Επίσης θα προβληθούν ορισμένα στατιστικά τα οποία είναι χρήσιμα για τους αλγόριθμους της μηχανικής μάθησης και την συγκριτική αξιολόγησή τους. Τέλος θα παρουσιαστεί το ιστόγραμμα και ο πίνακας συσχέτισης του συνόλου δεδομένων.

3.2.1 Περιγραφή

Οι δασικές πυρκαγιές αποτελούν ένα από τα σημαντικότερα ζητήματα για την μόλυνση του περιβάλλοντος και την αύξηση της κλιματικής αλλαγής και για αυτό είναι απαραίτητο να μπορούν να προβληθούν. Το σύνολο δεδομένων Algerian Forest Fires Dataset επιλέχθηκε για την υλοποίηση της παρούσας διπλωματικής εργασίας. Το σύνολο δεδομένων Algerian Forest Fires Dataset περιλαμβάνει συνολικά 244 εγγραφές δεδομένων όπου είναι δεδομένα δασικών πυρκαγιών από δύο διαφορετικές περιοχής στην Αλγερία.

Το σύνολο δεδομένων Algerian Forest Fires Dataset διατίθεται δωρεάν για έρευνα στον σύνδεσμο https://archive.ics.uci.edu/ml/datasets/Algerian+Forest+Fires+Dataset++#. Ο λόγος που επιλέχθηκε η Αλγερία είναι επειδή είναι μία από τις μεσογειακές χώρες που πλήττεται περισσότερο από τις δασικές πυρκαγιές. Επίσης υπάρχουν αρκετά δεδομένα για έρευνα σε πολλές περιοχές της (παρακάτω αναφέρονται δύο) και είναι μια από τις χώρες όπου έχουν χαθεί τεράστιες εκτάσεις δασών εξετίας των δασικών πυρκαγιών. Ακόμη τα τελευταία χρόνια η Αλγερία είναι μία από τις μεσογειακές χώρες που πλήττεται περισσότερο από καύσωνες κάνοντας τις δασικές πυρκαγιές ακόμα πιο επικίνδυνες. Εξαιτίας των καυσώνων οι δασικές πυρκαγιές εξαπλώνονται ταχύτερα και σε σημεία όπου έχουν δύσκολη προσβασιμότητα. Με ελάχιστη περιορισμένη πρόσβαση ο εντοπισμός και η αποτελεσματική επέμβαση των πυροσβεστών γίνεται πάρα πολύ δύσκολη. Για αυτούς τους λόγους επιλέχθηκε το σύνολο δεδομένων δασικών πυρκαγιών της Αλγερίας.

Η περιοχή Bejaia Region περιλαμβάνει μετεωρολογικά δεδομένα της περιοχής Bejaia η οποία είναι μια βόρεια περιοχή της Αλγερίας (παρατηρείτε παρακάτω η εικόνα της περιοχής). Συνολικά περιλαμβάνει 122 εγγραφές δεδομένων.



Εικόνα 19: *Περιοχή Bejaia Region της Αλγερίας*

Η περιοχή Sidi-Bel Abbes Region περιλαμβάνει μετεωρολογικά δεδομένα της περιοχής Sidi-Bel Abbes η οποία είναι μια βορειοδυτική περιοχή της Αλγερίας Αλγερίας (παρατηρείτε παρακάτω η εικόνα της περιοχής). Συνολικά περιλαμβάνει 122 εγγραφές δεδομένων όπως και η περιοχή Bejaia Region.

Εικόνα 20: *Περιοχή Sidi-Bel Abbes Region της Αλγερίας*

3.2.2 Χαρακτηριστικά

Στο σύνολο δεδομένων Algerian Forest Fires Dataset υπάρχουν 14 διαφορετικά χαρακτηριστικά (όπου 13 είναι τα χαρακτηριστικά εισόδου και 1 είναι η έξοδος). Από τα 13 χαρακτηριστικά που είναι τα χαρακτηριστικά εισόδου 3 από αυτά (ημέρα, μήνας και χρόνος) δεν θα χρησιμοποιηθούν καθώς είναι περιττά και δεν θα ωφελήσουν κάπου για την πρόβλεψη των δασικών πυρκαγιών και

στην συγκριτική αξιολόγηση των αλγορίθμων της μηχανικής μάθησης. Τα χαρακτηριστικά του συνόλου δεδομένων Algerian Forest Fires Dataset περιγράφονται στον παρακάτω πίνακα:

Πίνακας 4: *Περιγραφή χαρακτηριστικών συνόλου δεδομένων Algerian Forest Fires Dataset*

Χαρακτηριστικά	**Περιγραφή**	**Τύπος τιμής**
day	Είναι η ημέρα της εβδομάδας δηλαδή από Δευτέρα μέχρι Κυριακή και η τιμή της είναι ένας ακέραιος αριθμός ο οποίος αντιστοιχεί στην ημέρα του.	Ακέραιος
month	Είναι ο μήνας της εβδομάδας δηλαδή από Ιανουάριος μέχρι Δεκέμβριος και η τιμή του είναι ένας ακέραιος αριθμός ο οποίος αντιστοιχεί στον μήνα του. Λαμβάνει τιμές από το 1 μέχρι το 12.	Ακέραιος
year	Ο χρόνος λαμβάνει ακέραιες τιμές από οποιαδήποτε χρονολογία. Περιέχει μόνο την τιμή 2012.	Ακέραιος
Temperature	Η θερμοκρασία σε βαθμούς κελσίου (Celsius).	Ακέραιος
RH	Η σχετική υγρασία σε ποσοστό.	Ακέραιος
Ws	Η ταχύτητα του ανέμου σε χιλιόμετρα ανά ώρα.	Ακέραιος
Rain	Είναι το ποσοστό της βροχής	Πραγματικός
FFMC	Είναι ο Κωδικός λεπτής υγρασίας καυσίμου.	Πραγματικός
DMC	Είναι ο κωδικός περιεκτικότητας σε υγρασία.	Πραγματικός
DC	Είναι ο κωδικός ξηρασίας.	Πραγματικός
ISI	Είναι η αριθμητική βαθμολογία του αναμενόμενου ρυθμού εξάπλωσης της πυρκαγιάς.	Πραγματικός
BUI	Είναι ηαριθμητική βαθμολογία της συνολικής ποσότητας καυσίμου που είναι διαθέσιμη για καύση.	Πραγματικός
FWI	Είναι ο δείκτης καιρού πυρκαγιάς.	Πραγματικός
Classes	Είναι το αποτέλεσμα της πρόβλεψης και περιλαμβάνει δύο τιμές την not fire (δηλαδή όχι φωτιά) και fire (δηλαδή φωτιά).	Κλάση

Οι τιμές των χαρακτηριστικών FFMC, DMC, DC, ISI, BUI και FWI βαθμολογήθηκαν με βάση το σύστημα FWI (FWI system).

3.2.3 Μερικά δεδομένα

Στον παρακάτω πίνακα θα δούμε τις πρώτες 5 στήλες του συνόλου δεδομένων Algerian Forest Fires Dataset μέσω της συνάρτησης head() της βιβλιοθήκης Pandas. Κάθε σειρά της στήλης είναι μια διαφορετική μέρα στην χώρα της Αλγερίας και οι τιμές είναι διαφορετικές.

Πίνακας 5: *Πρώτες 5 στήλες συνόλου δεδομένων Algerian Forest Fires Data*

day	month	year	Temperature	RH	Ws	Rain	FFMC	DMC	DC	ISI	BUI	FWI	Classes
1	6	2012	29	57	18	0	65.7	3.4	7.6	1.3	3.4	0.5	not fire
2	6	2012	29	61	13	1.3	64.4	4.1	7.6	1	3.9	0.4	not fire
3	6	2012	26	82	22	13.1	47.1	2.5	7.1	0.3	2.7	0.1	not fire
4	6	2012	25	89	13	2.5	28.6	1.3	6.9	0	1.7	0	not fire
5	6	2012	27	77	16	0	64.8	3	14.2	1.2	3.9	0.5	not fire

Στον παρακάτω πίνακα θα δούμε τις τελευταίες 5 στήλες του συνόλου δεδομένων Algerian Forest Fires Dataset μέσω της συνάρτησης tail() της βιβλιοθήκης Pandas. Κάθε σειρά της στήλης είναι μια διαφορετική μέρα στην χώρα της Αλγερίας και οι τιμές είναι διαφορετικές.

Πίνακας 6: *Τελευταίες 5 στήλες συνόλου δεδομένων Algerian Forest Fires Dataset*

day	month	year	Temperature	RH	Ws	Rain	FFMC	DMC	DC	ISI	BUI	FWI	Classes
26	9	2012	30	65	14	0	85.4	16	44.5	4.5	16.9	6.5	fire
27	9	2012	28	87	15	4.4	41.1	6.5	8	0.1	6.2	0	not fire
28	9	2012	27	87	29	0.5	45.9	3.5	7.9	0.4	3.4	0.2	not fire
29	9	2012	24	54	18	0.1	79.7	4.3	15.2	1.7	5.1	0.7	not fire
30	9	2012	24	64	15	0.2	67.3	3.8	16.5	1.2	4.8	0.5	not fire

3.2.4 Στατιστικά

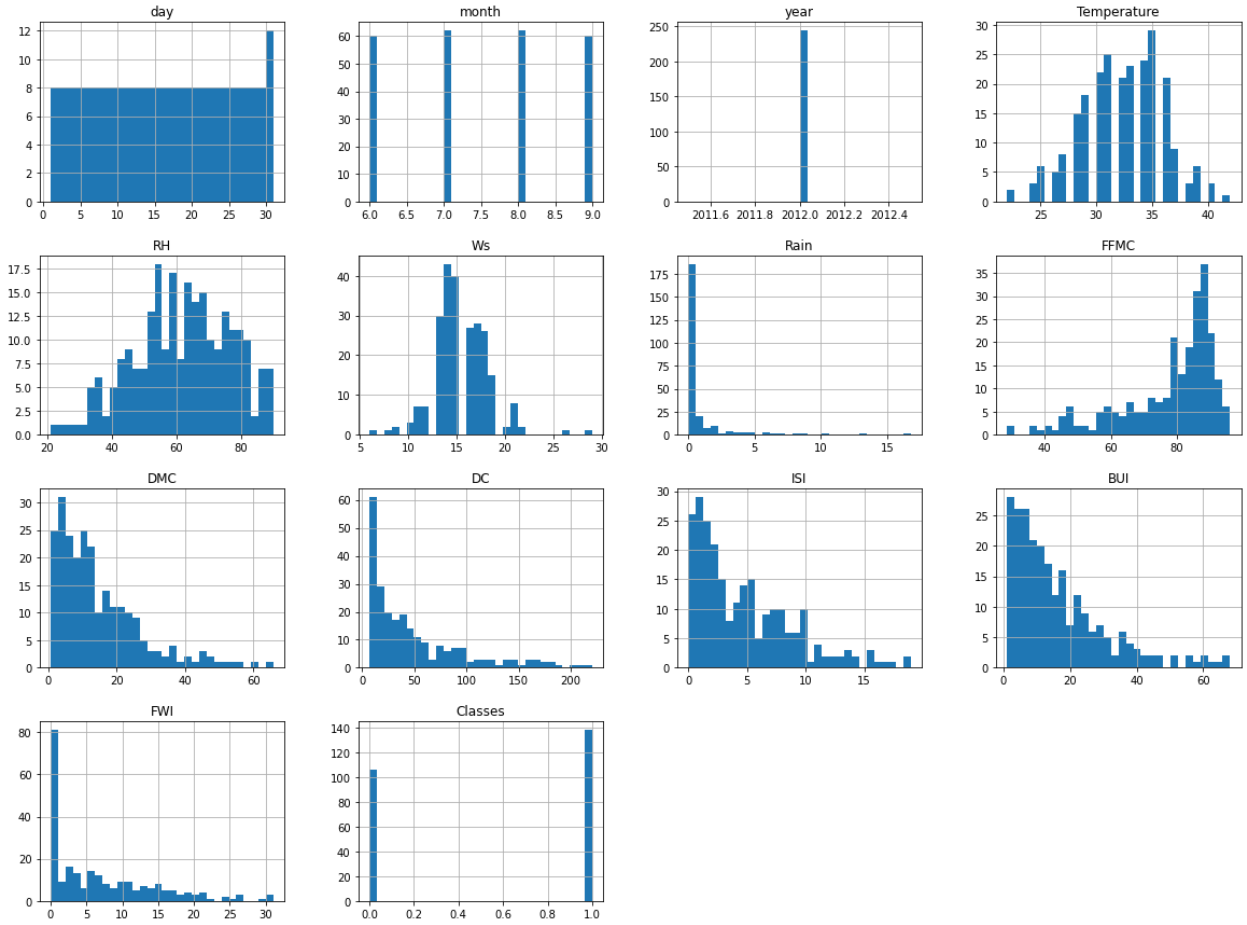
Στον παρακάτω πίνακα θα παρατηρηθούν μερικά στατιστικά του συνόλου δεδομένων Algerian Forest Fires Dataset μέσω της συνάρτησης describe() της βιβλιοθήκης Pandas. Αυτά τα στατιστικά των 244 συνολικών εγγραφών είναι η μέση τιμή, η τυπική απόκλιση, η ελάχιστη τιμή και η μέγιστη τιμή. Παρακάτω παρουσιάζεται ο κάθε τύπος της μέσης τιμής και της τυπικής απόκλισης των αποτελεσμάτων που παρουσιάστηκαν στον παρακάτω πίνακα.

Πίνακας 7: Στατιστικά συνόλου δεδομένων Algerian Forest Fires Dataset

Χαρακτηριστικά	**mean**	**std**	**min**	**max**
day	15.75	8.83	1	31
month	7.5	1.11	6	9
year	2012	0	2012	2012
Temperature	32.17	3.63	22	42
RH	61.94	14.88	21	90
Ws	15.5	2.81	6	29
Rain	0.76	2	0	16.8
FFMC	77.89	14.34	28.6	96
DMC	14.67	12.37	0.7	65.9
DC	49.29	47.62	6.9	220.4
ISI	4.76	4.15	0	19
BUI	16.67	14.2	1.1	68
FWI	7.05	7.43	0	31.1
Classes	0.57	0.5	0	1

3.2.5 Ιστόγραμμα

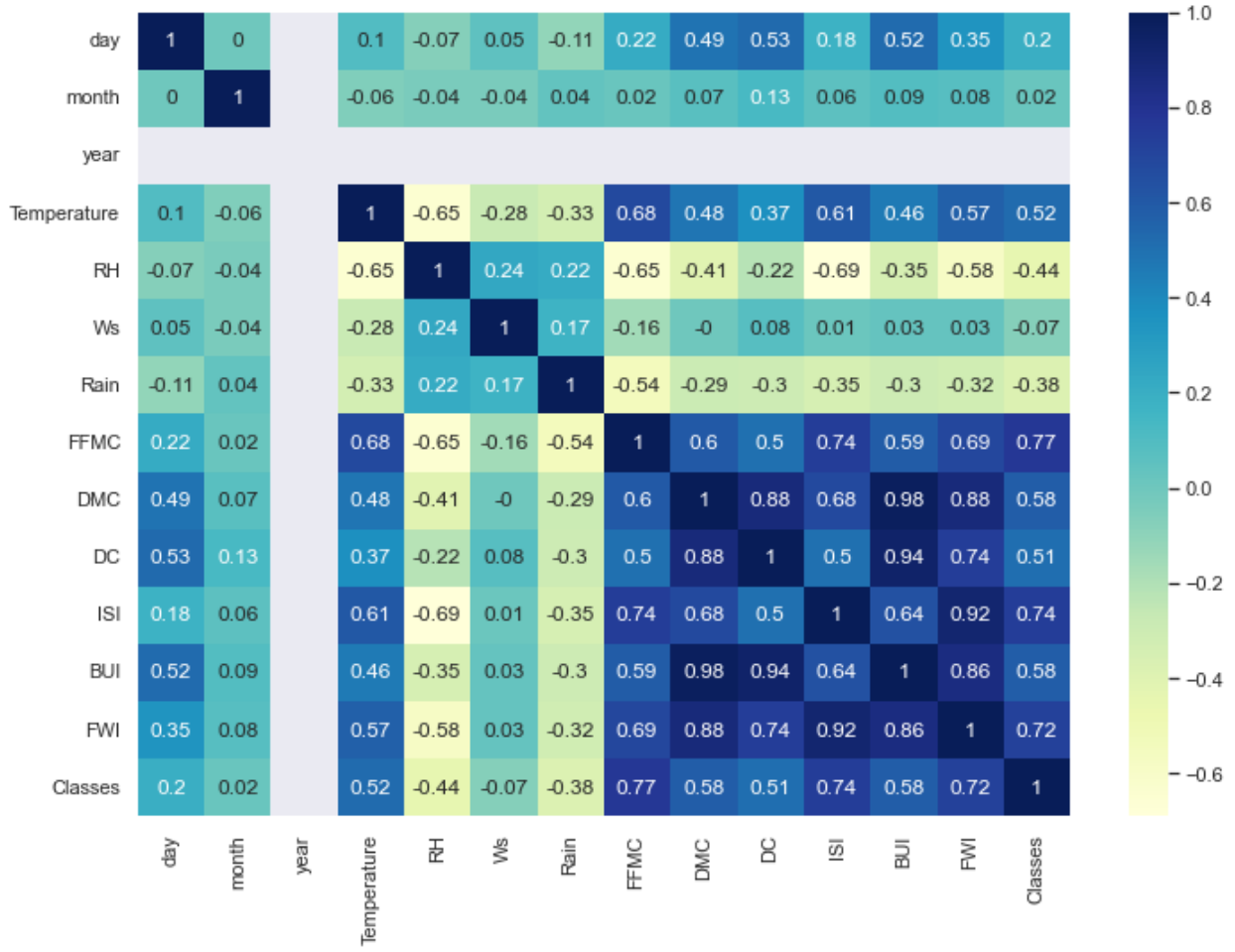
Παρακάτω παρουσιάζεται το ιστόγραμμα του συνόλου δεδομένων Algerian Forest Fires Dataset μέσω της συνάρτησης hist() της βιβλιοθήκης Matplotlib. Με την χρήση του ιστογράμματος δίνεται μια πρώτη όψη στα τα δεδομένα του συνόλου για κάθε χαρακτηριστικό.



Εικόνα 21: *Ιστογράμματα συνόλου δεδομένων Algerian Forest Fires Dataset*

3.2.6 Πίνακας συσχέτισης

Παρακάτω παρατηρείται ο πίνακας συσχέτισης του συνόλου δεδομένων Algerian Forest Fires Dataset χρησιμοποιώντας σε συνδυασμό την βιβλιοθήκη Pandas για τον υπολογισμό των τιμών του πίνακα συσχέτισης και τις βιβλιοθήκες Matplotlib και Seanborn για οπτικοποίηση του. Με τον πίνακα συσχέτισης θα εμφανιστεί στην παρακάτω εικόνα η συσχέτιση που υπάρχει στις μεταβλητές του συνόλου δεδομένων.



Εικόνα 22: *Πίνακας συσχέτισης συνόλου δεδομένων Algerian Forest Fires Dataset*

Όσο πιο ελαφρύ είναι το μπλε χρώμα τόσο πιο μεγαλύτερη είναι η συσχέτιση. Όσο πιο σκούρο είναι το μπλε χρώμα τόσο τόσο πιο λιγότερη είναι η συσχέτιση. Παρατηρείται ότι η στήλη year δεν έχει τιμή. Αυτό οφείλεται στο γεγονός ότι όλες οι τιμές στην στήλη year είναι 2012 και δεν υπάρχει κάποια διαφορετική.

3.3 Επιλογή αλγορίθμων

Οι αλγόριθμοι μηχανικής μάθησης για κατηγοριοποίηση ή και αλλιώς κατηγοριοποιητές (classifiers) που θα χρησιμοποιηθούν για την πρόβλεψη των δασικών πυρκαγιών είναι υλοποιημένοι στην βιβλιοθήκη Scikit-Learn της Python. Οι συγκεκριμένοι αλγόριθμοι επιλέχθηκαν με σκοπό να μελετηθεί η λειτουργία τους στην παρούσα διπλωματική εργασία και να μελετηθούν τα αποτελέσματά τους. Οι περισσότεροι από αυτούς είναι από τους πιο γνωστούς αλγόριθμους κατηγοριοποίησης όπως για παράδειγμα ο K-κοντινότεροι-γείτονες. Ακόμη η έρευνα

[8] που χρησιμοποιεί το ίδιο σύνολο δεδομένων χρησιμοποίησε αλγόριθμους που θα χρησιμοποιηθούν με αρκετά καλά αποτελέσματα. Άλλοι αλγόριθμοι όπως η μηχανή διανυσματικής υποστήριξης (Support Vector Machine) δεν χρησιμοποιήθηκαν καθώς η συγκεκριμένη διπλωματική εργασία έχει σαν στόχο να χρησιμοποιήσει ένα συγκεκριμένο εύρος αλγόριθμων μηχανικής μάθησης. Παρακάτω παρουσιάζονται ορισμένοι λόγοι που επιλέχθηκαν οι συγκεκριμένοι αλγόριθμοι μηχανικής μάθησης για κατηγοριοποίηση:

- **Κ-κοντινότεροι-γείτονες:** είναι από τους πιο δημοφιλείς και ακριβείς αλγόριθμους μηχανικής μάθησης για κατηγοριοποίηση. Επεξηγήθηκε στο κεφάλαιο 2 και τα αποτελέσματα του διαφέρουν ανάλογα με τις παραμέτρους που θα επιλεχθούν. Οι βασικοί παράμετροι του είναι η επιλογή της τιμής Κ και του μέτρου απόστασης. Ο αλγόριθμος που χρησιμοποιήθηκε είναι ο KNeighborsClassifier().
- **Δέντρα απόφασης:** είναι και αυτός ένας από τους πιο δημοφιλείς αλγόριθμους μηχανικής μάθησης για κατηγοριοποίηση. Είναι κατάλληλος για κατηγοριοποίηση διακριτών τιμών 0 ή 1 (δηλαδή not fire ή fire) και επεξηγήθηκε στο κεφάλαιο 2. Ο αλγόριθμος που χρησιμοποιήθηκε είναι ο DecisionTreeClassifier().
- **Τυχαία δάση:** ανήκει στην συλλογιστική μάθηση και ουσιαστικά είναι μια συλλογή δέντρων απόφασης. Χρησιμοποιείται για κατηγοριοποίηση και η έξοδος του είναι το δέντρο το οποίο έχει τις περισσότερες ψήφους από όλα τα δέντρα του δάσους. Ο αλγόριθμος που χρησιμοποιήθηκε είναι ο RandomForestClassifier().
- **AdaBoost:** ανήκει στην συλλογιστική μάθηση και χρησιμοποιεί την μέθοδο της ενδυνάμωσης. Χρησιμοποιείται για κατηγοριοποίηση και μπορεί να χρησιμοποιηθεί για την ενίσχυση της απόδοσης και μπορεί να παρέχει ακριβείς προβλέψεις (predictions). Ο αλγόριθμος που χρησιμοποιήθηκε είναι ο AdaBoostClassifier ().
- **Gradient tree bosting:** ανήκει στην συλλογιστική μάθηση και χρησιμοποιεί την μέθοδο της ενδυνάμωσης. Χρησιμοποιείται για κατηγοριοποίηση και συγκρίνοντας με τα τυχαία δάση προσπαθεί να διορθώσει τα λάθη του. Ο αλγόριθμος που χρησιμοποιήθηκε είναι ο GradientBoostingClassifier().
- **Λογιστική παλινδρόμηση:** είναι ένας από τους πιο κατάλληλους κατηγοριοποιητές σε προβλήματα κατηγοριοποίησης. Η λειτουργία του επεξηγήθηκε στο κεφάλαιο 2 και είναι κατάλληλος για την συγκεκριμένη διπλωματική εργασία καθώς μπορεί να προβλέψει

διακριτές τιμές 0 ή 1 (δηλαδή not fire ή fire). Ο αλγόριθμος που χρησιμοποιήθηκε είναι ο LogisticRegression().

- **Νευρωνικό δίκτυο πολλών επιπέδων:** χρησιμοποιείται σε πολλά προβλήματα τεχνητής νοημοσύνης όπου ένα από αυτά είναι και η κατηγοριοποίηση. Έχουν την δυνατότητα να μαθαίνουν από μόνα τους καθώς η λειτουργία τους προσπαθεί να αντιγράψει την λειτουργία του ανθρώπινου εγκεφάλου. Ο αλγόριθμος που χρησιμοποιήθηκε είναι ο MLPClassifier().
- **Απλοϊκός Bayes:** είναι ένας από τους πιο γνωστούς κατηγοριοποιητές (classifiers) και εφαρμόζει το θεώρημα Bayes. Έχει αρκετούς τύπους αλλά αυτός που θα χρησιμοποιηθεί για την συγκεκριμένη διπλωματική εργασία είναι Bernoulli απλοϊκός Bayes ο οποίος εφαρμόζει την κατανομή του Bernoulli. Επεξηγήθηκε στο κεφάλαιο 2 και είναι κατάλληλος για την πρόβλεψη δυαδικών αποτελεσμάτων 0 ή 1 (δηλαδή not fire ή fire). Ο αλγόριθμος που χρησιμοποιήθηκε είναι ο BernoulliNB().
- **Κλιμάκωση χαρακτηριστικών:** για την κλιμάκωση χαρακτηριστικών (feature scaling) θα χρησιμοποιηθεί ο αλγόριθμος StandardScaler().

3.4 Παράμετροι αλγορίθμων

Τώρα θα παρουσιαστούν οι παράμετροι των αλγορίθμων μηχανικής μάθησης ή αλλιώς κατηγοριοποιητών που παρουσιάστηκαν παραπάνω. Κάθε αλγόριθμος έχει διαφορετικές παραμέτρους και παρουσιάζονται παρακάτω:

- **KNeighborsClassifier():** θα χρησιμοποιηθεί 3 φορές με 3 διαφορετικές παραμέτρους. Ο αριθμός των n_neighbors είναι το πιο σημαντικό κριτήριο και θα πάρει 3 διαφορετικές τιμές 3, 5 και 7. Επίσης θα χρησιμοποιηθούν 3 διαφορετικά metric δηλαδή μέτρο απόστασης επιλέχθηκαν οι τιμές euclidean, manhattan και minkowski. Τέλος η παράμετρος weights θα πάρει την τιμή distance και στις 3 διαφορετικές περιπτώσεις δηλαδή οι πιο κοντινοί γείτονες ενός σημείου θα έχουν μεγαλύτερη επιρροή από τους γείτονες που βρίσκονται πιο μακριά.

Πίνακας 8: Παράμετροι KNeighborsClassifier

Παράμετρος	Τιμή

n_neighbors	3, 5, 7
metric	euclidean, manhattan, minkowski
weights	distance

- **DecisionTreeClassifier():** θα χρησιμοποιηθεί με 3 παραμέτρους. Ως criterion επιλέχθηκε η τιμή entropy δηλαδή η εντροπία. Η παράμετρος splitter θα πάρει την τιμή best δηλαδή την καλύτερη τιμή για τον διαχωρισμό των κόμβων. Τέλος η παράμετρος max_depth δηλαδή το μέγιστο βάθος του δέντρου θα πάρει την τιμή 3.

Πίνακας 9: *Παράμετρος DecisionTreeClassifier*

Παράμετρος	**Τιμή**
criterion	entropy
splitter	best
max_depth	3

- **RandomForestClassifier():** θα χρησιμοποιηθεί με 3 παραμέτρους. Η παράμετρος n_estimators δηλαδή ο αριθμός των συνολικών δέντρων στο δάσος θα πάρει την τιμή 100. Ως criterion επιλέχθηκε η τιμή entropy δηλαδή η εντροπία. Τέλος η παράμετρος max_depth δηλαδή το μέγιστο βάθος του κάθε δέντρου μέσα στο δάσος θα πάρει την τιμή 3.

Πίνακας 10: *Παράμετροι RandomForestClassifier*

Παράμετρος	**Τιμή**
n_estimators	100
criterion	entropy
max_depth	3

- **AdaBoostClassifier():** θα χρησιμοποιηθεί με 2 παραμέτρους. Η παράμετρος n_estimators δηλαδή ο αριθμός των συνολικών δέντρων στο δάσος θα πάρει την τιμή 50. Τέλος η παράμετρος learning_rate δηλαδή βάρος που εφαρμόζεται σε κάθε κατηγοριοποιητή σε κάθε επανάληψη ενίσχυσης θα πάρει την τιμή 1.0.

Πίνακας 11: Παράμετροι AdaBoostClassifier

Παράμετρος	Τιμή
n_estimators	50
learning_rate	1.0

- **GradientBoostingClassifier():** θα χρησιμοποιηθεί με 3 παραμέτρους. Η παράμετρος n_estimators δηλαδή ο αριθμός των συνολικών δέντρων στο δάσος θα πάρει την τιμή 100. Η παράμετρος learning_rate δηλαδή βάρος που εφαρμόζεται σε κάθε κατηγοριοποιητή σε κάθε επανάληψη ενίσχυσης θα πάρει την τιμή 0.1. Τέλος η παράμετρος loss δηλαδή η συνάρτηση απώλειας θα πάρει την τιμή deviance δηλαδή αναφέρεται στην λογιστική παλινδρόμηση.

Πίνακας 12: Παράμετροι GradientBoostingClassifier

Παράμετρος	Τιμή
n_estimators	100
learning_rate	0.1
loss	deviance

- **LogisticRegression():** θα χρησιμοποιηθεί με 5 παραμέτρους. Η παράμετρος penalty δηλαδή η τιμωρία θα πάρει την τιμή l2 και η παράμετρος dual θα πάρει την τιμή False η οποία εφαρμόζεται μόνο όταν η παράμετρος penalty έχει την τιμή l2 και η παράμετρος solver έχει την τιμή liblinear. Ακόμη η παράμετρος tol δηλαδή η ανοχή στα κριτήρια παύσης θα πάρει την τιμή 1e-4 και η παράμετρος C δηλαδή η αντίστροφη δύναμη κανονικοποίησης θα πάρει την τιμή 1. Τέλος ως solver επιλέχθηκε η τιμή liblinear.

Πίνακας 13: Παράμετροι LogisticRegression

Παράμετρος	Τιμή
penalty	l2
dual	False
tol	1e-4

C	1
solver	liblinear

- **MLPClassifier():** θα χρησιμοποιηθεί με 4 παραμέτρους. Η παράμετρος hidden_layer_sizes δηλαδή το μέγεθος του κάθε κρυφού επίπεδου θα πάρει την τιμή (50, 50, 50) δηλαδή 3 κρυφά επίπεδα 50 νευρώνων. Η συνάρτηση ενεργοποίησης για τα κρυφά επίπεδα δηλαδή η παράμετρος activation θα πάρει την τιμή relu που σημαίνει διορθωμένη συνάρτηση γραμμικής μονάδας. Ως solver επιλέχθηκε η τιμή adam δηλαδή ένας στοχαστικός βελτιστοποιητής που βασίζεται σε κλίση. Τέλος ως solver δηλαδή το πρόγραμμα ρυθμού εκμάθησης για ενημερώσεις των βαρών επιλέχθηκε η τιμή constant.

Πίνακας 14: *Παράμετροι MLPClassifier*

Παράμετρος	**Τιμή**
hidden_layer_sizes	(50, 50, 50)
activation	relu
solver	adam
learning_rate	constant

- **BernoulliNB():** θα χρησιμοποιηθεί με 2 παραμέτρους. Η παράμετρος alpha θα πάρει την τιμή 1.0. Τέλος η παράμετρος binarize δηλαδή το κατώφλι για δυαδοποίηση των χαρακτηριστικών του συνόλου δεδομένων θα πάρει την τιμή 0.0. Οι τιμές των παραμέτρων είναι ήδη οι προκαθορισμένες τιμές της sklearn.

Πίνακας 15: *Παράμετροι BernoulliNB*

Παράμετρος	**Τιμή**
alpha	1.0
binarize	0.0

Συνοπτικά στον παρακάτω πίνακα παρουσιάζονται οι κατηγοριοποιητές δηλαδή οι αλγόριθμοι της μηχανικής μάθησης που θα χρησιμοποιηθούν με τις παραμέτρους και τις τιμές τους.

Πίνακας 16: Συνοπτικός πίνακας παραμέτρων όλων των κατηγοριοποιητών

Κατηγοριοποιητής	**Παράμετροι και τιμές**
KNeighborsClassifier	n_neighbors=3, metric='euclidean', weights='distance'
KNeighborsClassifier	n_neighbors=5, metric='manhattan', weights='distance'
KNeighborsClassifier	n_neighbors=7, metric='minkowski', weights='distance'
DecisionTreeClassifier	criterion='entropy', splitter='best', max_depth=3
RandomForestClassifier	n_estimators=100, criterion='entropy', max_depth=3
AdaBoostClassifier	n_estimators=50, learning_rate=1.0
GradientBoostingClassifier	n_estimators=100, learning_rate=0.1, loss='deviance'
LogisticRegression	penalty='l2', dual=False, tol=1e-4, C=1, solver='liblinear'
MLPClassifier	hidden_layer_sizes=(50, 50, 50), activation='relu', solver='adam', learning_rate='constant'
BernoulliNB	alpha = 1.0, binarize = 0.0

3.5 Βήματα υλοποίησης

Στο περιβάλλον Spyder χρησιμοποιήθηκε η Python όπου μέσω αυτής χρησιμοποιήθηκαν οι παραπάνω αλγόριθμοι μηχανικής μάθησης ή αλλιώς κατηγοριοποιητές. Στόχος της παρούσας διπλωματικής εργασίας είναι η εφαρμογή αυτών των αλγόριθμων μηχανικής μάθησης για την συγκριτική τους αξιολόγηση για την πρόβλεψη δασικών πυρκαγιών. Τα βήματα υλοποίησης περιγράφονται παρακάτω:

1. Άνοιγμα του αρχείου Algerian_Forest_Fires_Dataset.csv και αποθήκευση σε ένα πλαίσιο δεδομένων (dataframe).
2. Αντικατάσταση των τιμών της κλάσης δηλαδή η τιμή not fire γίνεται 0 και η τιμή fire γίνεται 1.
3. Εμφάνιση των πρώτων και των τελευταίων 5 δεδομένων του πλαισίου δεδομένων.
4. Εμφάνιση περιγραφής του πλαισίου δεδομένων.
5. Εμφάνιση διαγράμματος του ιστογράμματος.
6. Εμφάνιση διαγράμματος του πίνακα συσχέτισης.
7. Διαγραφή των στηλών day, month και year.

8. Αποθήκευση πλαισίου δεδομένων σε σύνολο δεδομένων X και y.
9. Εκπαίδευση του συνόλου δεδομένων.
10. Προεπεξεργασία του συνόλου δεδομένων με κλιμάκωση χαρακτηριστικών.
11. Αρχικοποίηση κατηγοριοποιητών και των πινάκων των αξιολογήσεων τους δηλαδή της ορθότητας (accuracy), της ακρίβειας (precision), της ανάκλησης (recall) και του αρμονικού μέσου (f1 score).
12. Εφαρμογή κατηγοριοποιητών (classifiers) και αξιολόγησή τους.
13. Εμφάνιση συνοπτικού και αναλυτικών διαγραμμάτων του πίνακας σύγχυσης (confusion matrix).
14. Εμφάνιση συνοπτικού και αναλυτικών διαγραμμάτων των αξιολογητών δηλαδή της ορθότητας (accuracy), της ακρίβειας (precision), της ανάκλησης (recall) και του αρμονικού μέσου (f1 score).
15. Εμφάνιση ξεχωριστών διαγραμμάτων των αξιολογητών δηλαδή της ορθότητας (accuracy), της ακρίβειας (precision), της ανάκλησης (recall) και του αρμονικού μέσου (f1 score).

Κεφάλαιο 4: Αποτελέσματα

Σε αυτό το κεφάλαιο θα παρουσιαστούν οι προβλέψεις των αλγόριθμων μηχανικής μάθησης δηλαδή, θα παρουσιαστεί ο πίνακας σύγχυσης (confusion matrix), η ορθότητα (accuracy), η ακρίβεια (precision), η ανάκληση (recall) και τέλος ο αρμονικός μέσος (f1 score) του κάθε αλγόριθμου μηχανικής μάθησης. Στην συνέχεια θα γίνει συγκριτική αξιολόγηση όλων των αλγόριθμων μηχανικής μάθησης δηλαδή θα συγκριθεί ο πίνακας σύγχυσης (confusion matrix), η ορθότητα (accuracy), η ακρίβεια (precision), η ανάκληση (recall) και τέλος ο αρμονικός μέσος (f1 score) του κάθε αλγόριθμου μηχανικής μάθησης. Υπενθυμίζεται ότι κάθε αλγόριθμος χρησιμοποιεί τις μεθόδους test_train_split και StandardScaler. Τέλος αφού γίνει συγκριτική αξιολόγηση θα επιλεχθεί ο βέλτιστος αλγόριθμος μηχανικής μάθησης δηλαδή αυτός που θα έχει τα υψηλότερα ποσοστά και θα είναι ο πιο βέλτιστος από τους υπόλοιπους αλγόριθμους μηχανικής μάθησης για την πρόβλεψη δασικών πυρκαγιών του συνόλου δεδομένων Algerian Forest Fires Dataset Data Set.

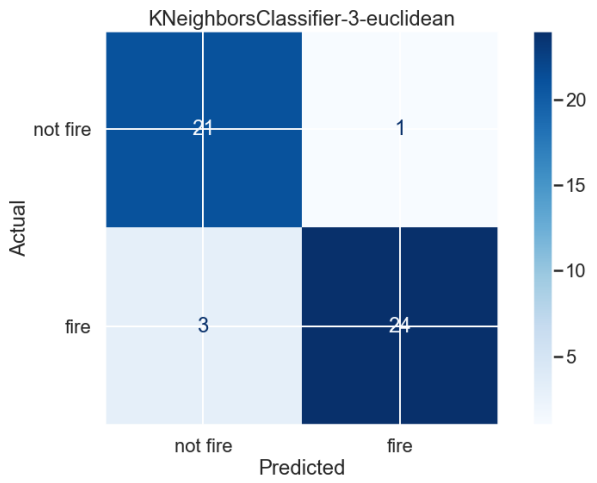
4.1 Προβλέψεις αλγόριθμων μηχανικής μάθησης

Τώρα θα παρουσιαστούν οι προβλέψεις των αλγόριθμων μηχανικής μάθησης ή αλλιώς των κατηγοριοποιητών (classifiers). Οι αλγόριθμοι μηχανικής μάθησης είναι οι Κ-κοντινότεροι γείτονες (K-nearest-neighboors - kNN) όπου θα χρησιμοποιηθούν 3 διαφορετικοί παράμετροι, τα δέντρα απόφασης (decision trees), τα τυχαία δάση (random forests), ο AdaBoost, ο Gradient tree boosting, η λογιστική παλινδρόμηση (logistic regression), τα νευρωνικά δίκτυα πολλών επιπέδων και ο Bernoulli απλοϊκός Bayes.

4.1.1 Κ-κοντινότεροι-γείτονες

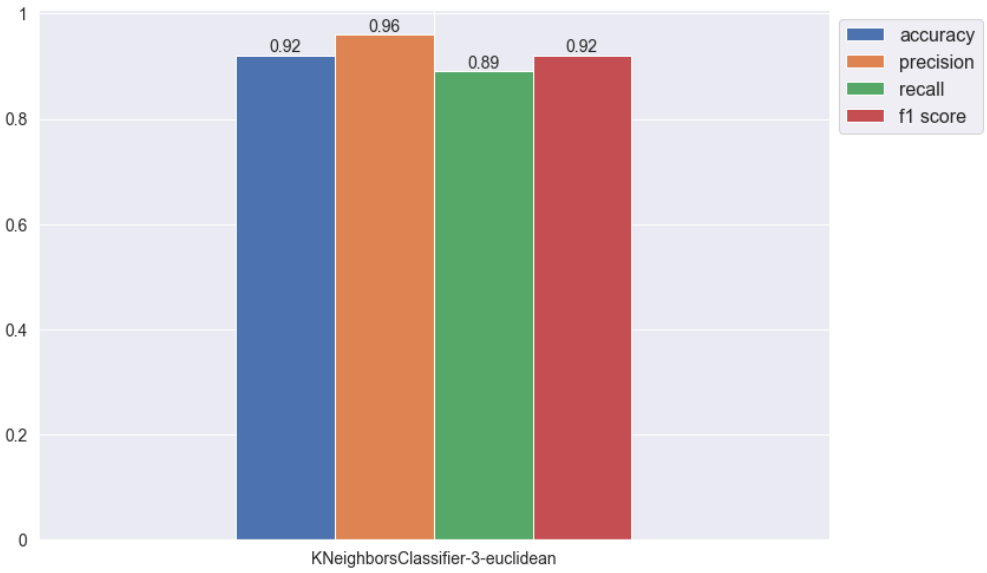
Ο κατηγοριοποιητής Κ-κοντινότεροι-γείτονες ή αλλιώς KNeighborsClassifier χρησιμοποιήθηκε 3 φορές με 3 διαφορετικές παραμέτρους.

Παρακάτω παρουσιάζονται τα αποτελέσματα του 1^{ου} κατηγοριοποιητή KNeighborsClassifier ο οποίος έχει την τιμή 3 στην παράμετρο n_neighbors και την τιμή euclidean στην παράμετρο metric.



Εικόνα 23: *Πίνακας σύγχυσης 1ου KNeighborsClassifier*

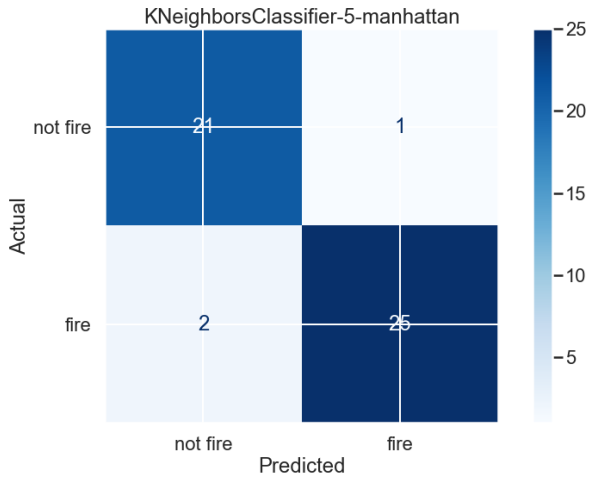
Παραπάνω στον πίνακα σύγχυση (confusion matrix) παρατηρείται ότι ο 1ος κατηγοριοποιητής KNeighborsClassifier πρόβλεψε 24 αληθινές θετικές πυρκαγιές (True Positive), 21 αληθινές αρνητικές μη πυρκαγιές (True Negative), 1α ψευδή θετική πυρκαγιά (False Positive), και 3εις ψευδής αρνητικές μη πυρκαγιές (False Negative).



Γραφική Παράσταση 1: *Αποτελέσματα 1ου KNeighborsClassifier*

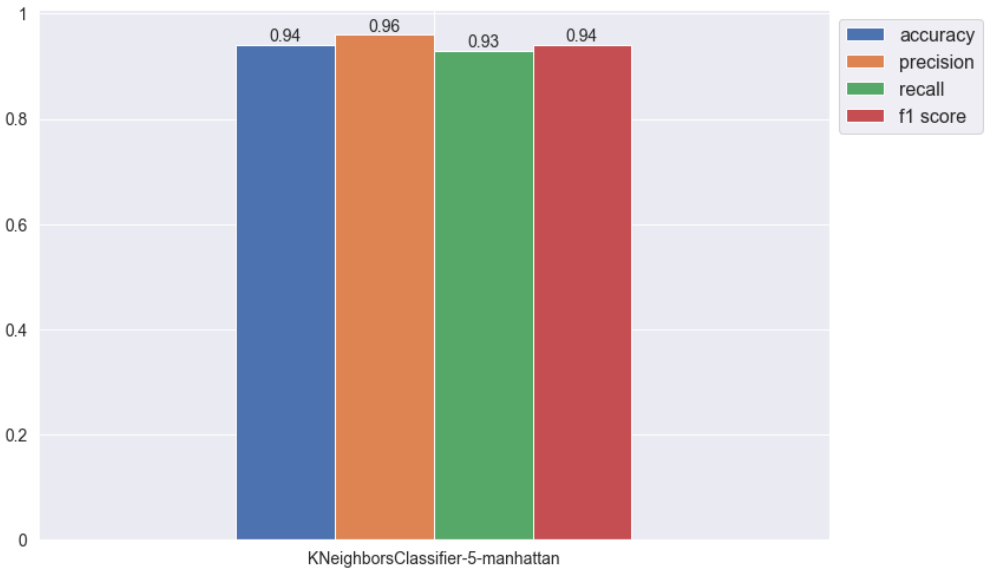
Παραπάνω παρατηρείται ότι ο 1ος κατηγοριοποιητής KNeighborsClassifier έχει ορθότητα (accuracy) 0.92, έχει ακρίβεια (precision) 0.96, έχει ανάκληση (recall) 0.89 και έχει αρμονικό μέσο (f1 score) 0.92.

Συνεχίζοντας παρακάτω παρουσιάζονται τα αποτελέσματα του 2ου κατηγοριοποιητή KNeighborsClassifier ο οποίος έχει την τιμή 5 στην παράμετρο n_neighbors και την τιμή manhattan στην παράμετρο metric.



Εικόνα 24: *Πίνακας σύγχυσης 2ου KNeighborsClassifier*

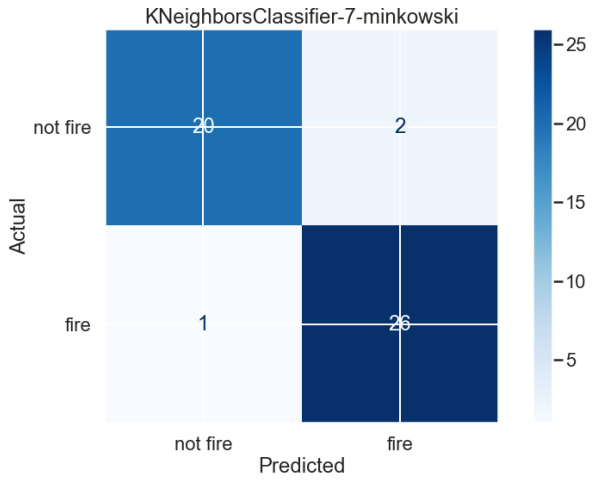
Παραπάνω στον πίνακα σύγχυση (confusion matrix) παρατηρείται ότι ο 2ος κατηγοριοποιητής KNeighborsClassifier πρόβλεψε 25 αληθινές θετικές πυρκαγιές (True Positive), 21 αληθινές αρνητικές μη πυρκαγιές (True Negative), 1α ψευδή θετική πυρκαγιά (False Positive), και 2 ψευδής αρνητικές μη πυρκαγιές (False Negative).



Γραφική Παράσταση 2: *Αποτελέσματα 2ου KNeighborsClassifier*

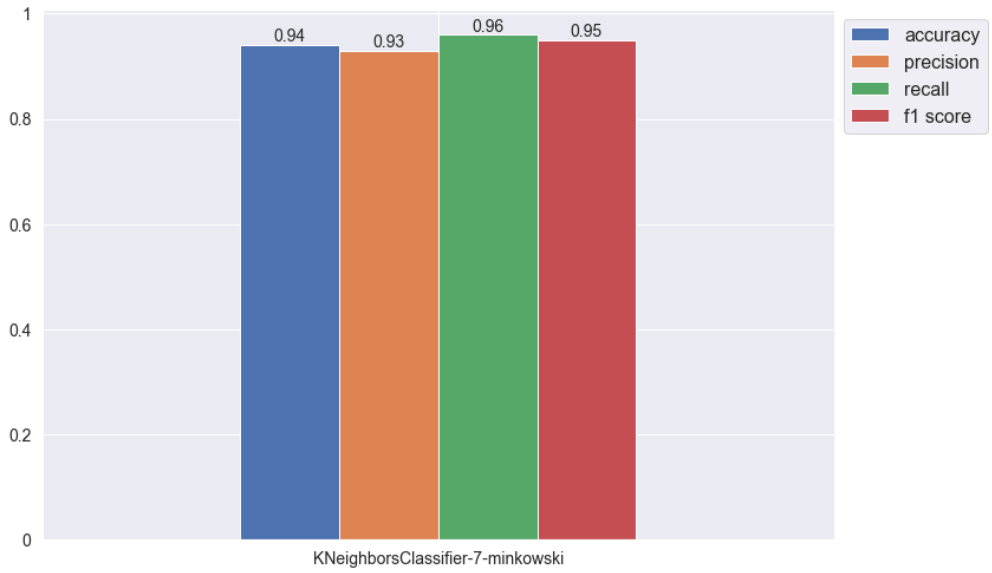
Παραπάνω παρατηρείται ότι ο 2ος κατηγοριοποιητής KNeighborsClassifier έχει ορθότητα (accuracy) 0.94, έχει ακρίβεια (precision) 0.96, έχει ανάκληση (recall) 0.93 και έχει αρμονικό μέσο (f1 score) 0.94.

Τέλος παρακάτω παρουσιάζονται τα αποτελέσματα του 3ου κατηγοριοποιητή KNeighborsClassifier ο οποίος έχει την τιμή 7 στην παράμετρο n_neighbors και την τιμή minkowski στην παράμετρο metric.



Εικόνα 25: *Πίνακας σύγχυσης 3ου KNeighborsClassifier*

Παραπάνω στον πίνακα σύγχυση (confusion matrix) παρατηρείται ότι ο 3ος κατηγοριοποιητής KNeighborsClassifier πρόβλεψε 26 αληθινές θετικές πυρκαγιές (True Positive), 20 αληθινές αρνητικές μη πυρκαγιές (True Negative), 2 ψευδής θετική πυρκαγιά (False Positive), και 1α ψευδή αρνητική μη πυρκαγιά (False Negative).

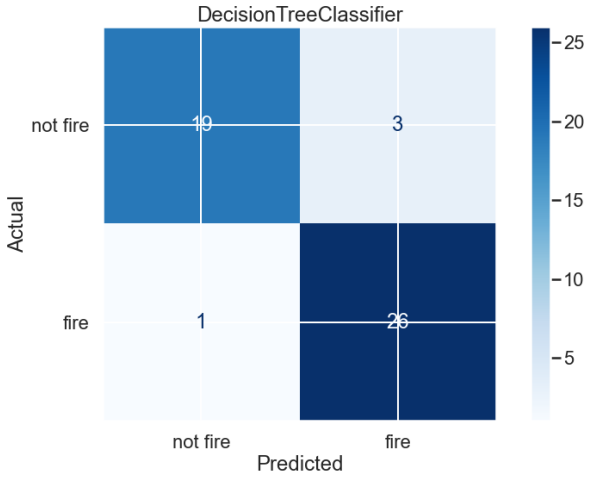


Γραφική Παράσταση 3: *Αποτελέσματα 3ου KNeighborsClassifier*

Παραπάνω παρατηρείται ότι ο 3ος κατηγοριοποιητής KNeighborsClassifier έχει ορθότητα (accuracy) 0.94, έχει ακρίβεια (precision) 0.93, έχει ανάκληση (recall) 0.96 και έχει αρμονικό μέσο (f1 score) 0.95.

4.1.2 Δέντρα απόφασης

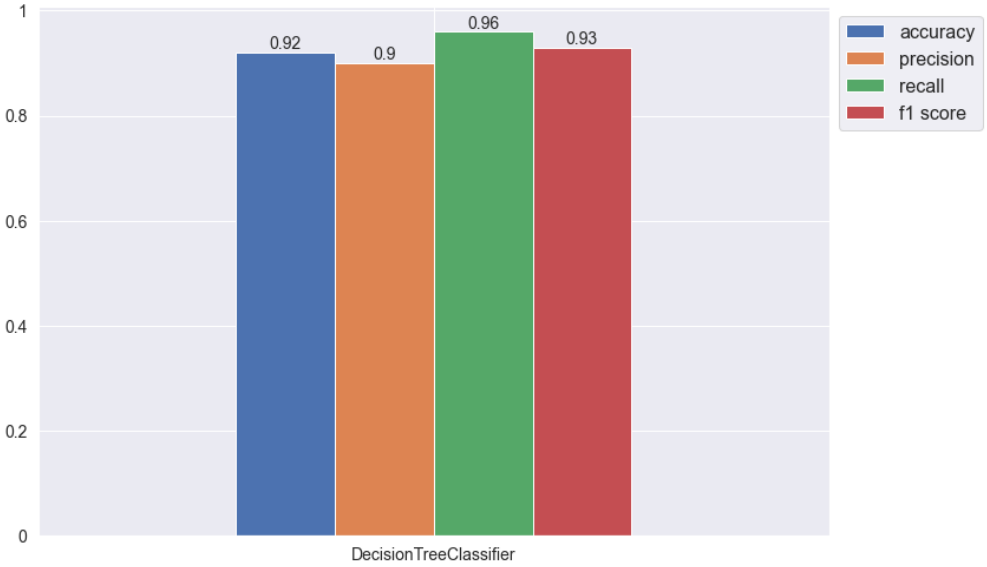
Παρακάτω παρουσιάζονται τα αποτελέσματα του κατηγοριοποιητή DecisionTreeClassifier.



Εικόνα 26: *Πίνακας σύγχυσης DecisionTreeClassifier*

Παραπάνω στον πίνακα σύγχυση (confusion matrix) παρατηρείται ότι ο κατηγοριοποιητής DecisionTreeClassifier πρόβλεψε 26 αληθινές θετικές πυρκαγιές (True Positive), 19 αληθινές

αρνητικές μη πυρκαγιές (True Negative), 3ες ψευδής θετικές πυρκαγιές (False Positive), και 1α ψευδή αρνητική μη πυρκαγιά (False Negative).

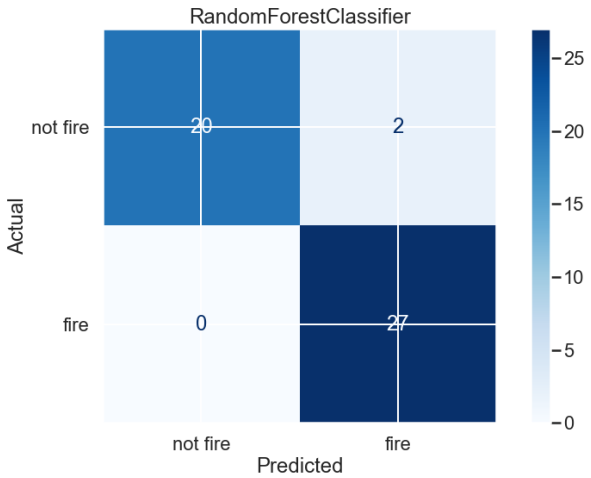


Γραφική Παράσταση 4: *Αποτελέσματα DecisionTreeClassifier*

Παραπάνω παρατηρείται ότι ο κατηγοριοποιητής DecisionTreeClassifier έχει ορθότητα (accuracy) 0.92, έχει ακρίβεια (precision) 0.90, έχει ανάκληση (recall) 0.96 και έχει αρμονικό μέσο (f1 score) 0.93.

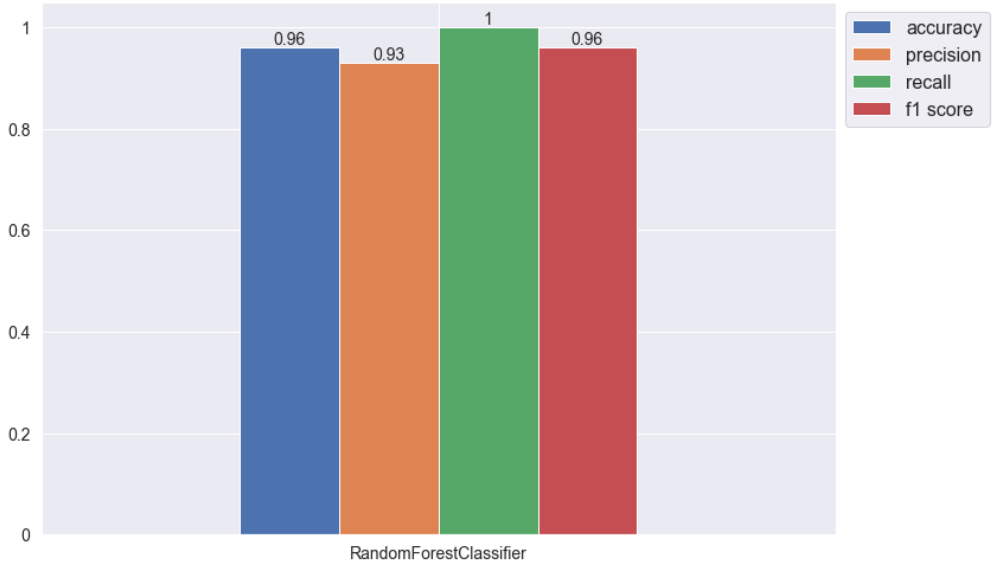
4.1.3 Τυχαία δάση

Παρακάτω παρουσιάζονται τα αποτελέσματα του κατηγοριοποιητή RandomForestClassifier.



Εικόνα 27: *Πίνακας σύγχυσης RandomForestClassifier*

Παραπάνω στον πίνακα σύγχυση (confusion matrix) παρατηρείται ότι ο κατηγοριοποιητής RandomForestClassifier πρόβλεψε 27 αληθινές θετικές πυρκαγιές (True Positive), 20 αληθινές αρνητικές μη πυρκαγιές (True Negative), 2 ψευδής θετικές πυρκαγιές (False Positive), και 0 ψευδής αρνητικές μη πυρκαγιές (False Negative).

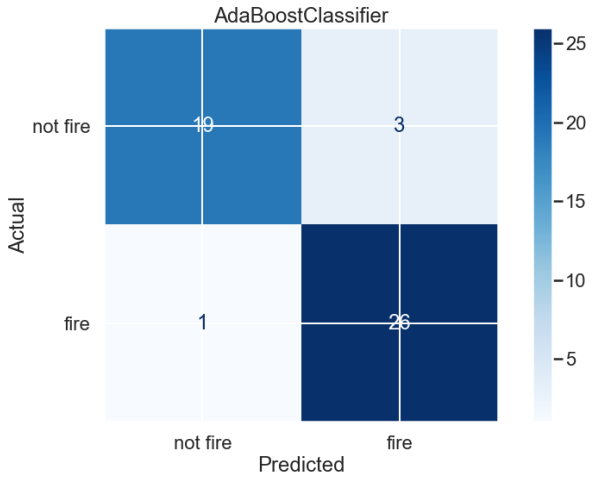


Γραφική Παράσταση 5: *Αποτελέσματα RandomForestClassifier*

Παραπάνω παρατηρείται ότι ο κατηγοριοποιητής RandomForestClassifier έχει ορθότητα (accuracy) 0.96, έχει ακρίβεια (precision) 0.93, έχει ανάκληση (recall) 1 και έχει αρμονικό μέσο (f1 score) 0.96.

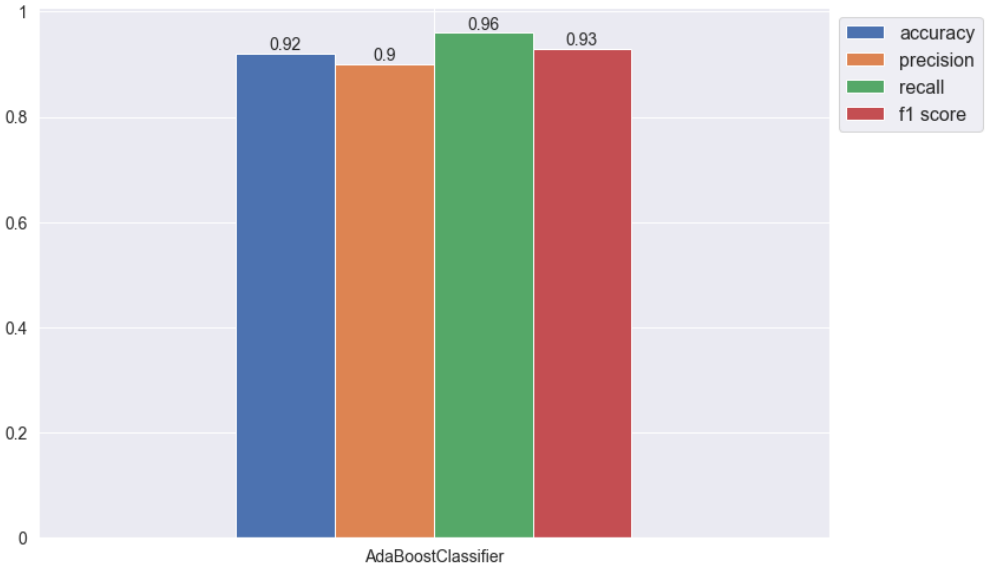
4.1.4 AdaBoost

Παρακάτω παρουσιάζονται τα αποτελέσματα του κατηγοριοποιητή AdaBoostClassifier.



Εικόνα 28: *Πίνακας σύγχυσης AdaBoostClassifier*

Παραπάνω στον πίνακα σύγχυση (confusion matrix) παρατηρείται ότι ο κατηγοριοποιητής AdaBoostClassifier πρόβλεψε 26 αληθινές θετικές πυρκαγιές (True Positive), 19 αληθινές αρνητικές μη πυρκαγιές (True Negative), 3 ψευδής θετικές πυρκαγιές (False Positive), και 1_{α} ψευδή αρνητική μη πυρκαγιά (False Negative).

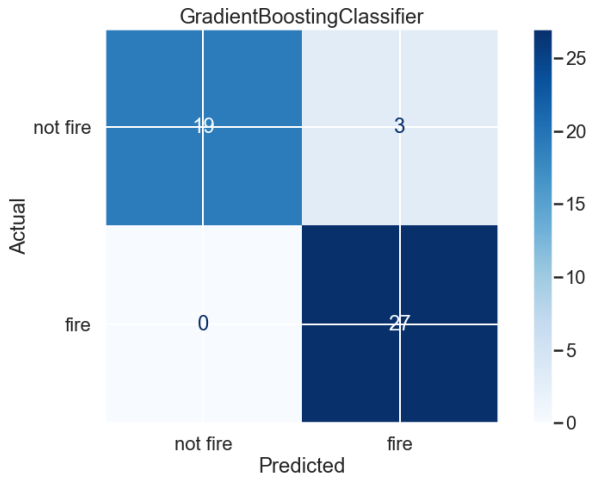


Γραφική Παράσταση 6: *Αποτελέσματα AdaBoostClassifier*

Παραπάνω παρατηρείται ότι ο κατηγοριοποιητής AdaBoostClassifier έχει ορθότητα (accuracy) 0.92, έχει ακρίβεια (precision) 0.90, έχει ανάκληση (recall) 0.96 και έχει αρμονικό μέσο (f1 score) 0.93.

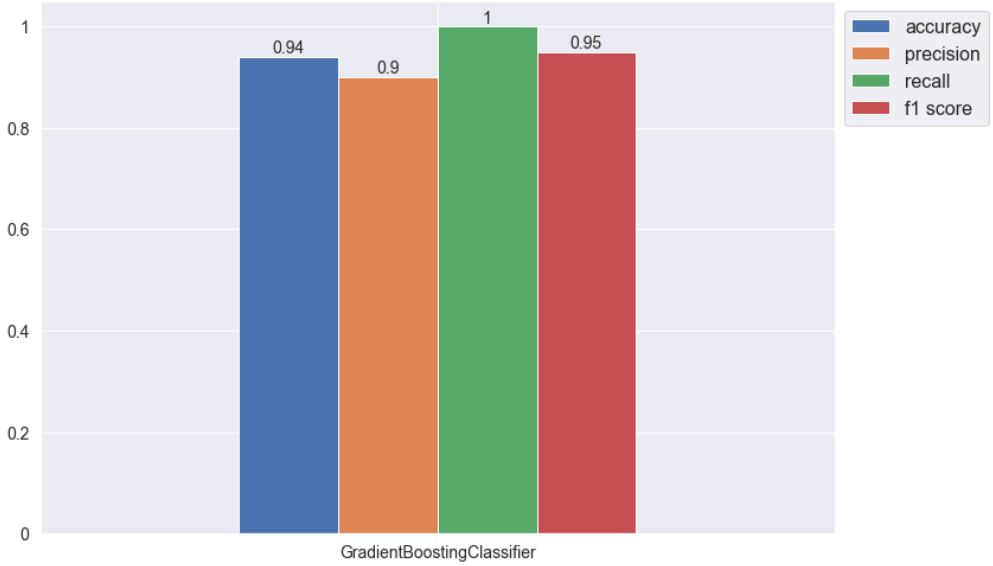
4.1.5 Gradient tree boosting

Παρακάτω παρουσιάζονται τα αποτελέσματα του κατηγοριοποιητή GradientBoostingClassifier.



Εικόνα 29: *Πίνακας σύγχυσης GradientBoostingClassifier*

Παραπάνω στον πίνακα σύγχυση (confusion matrix) παρατηρείται ότι ο κατηγοριοποιητής GradientBoostingClassifier πρόβλεψε 27 αληθινές θετικές πυρκαγιές (True Positive), 19 αληθινές αρνητικές μη πυρκαγιές (True Negative), 3 ψευδής θετικές πυρκαγιές (False Positive), και 0 ψευδής αρνητικές μη πυρκαγιές (False Negative).

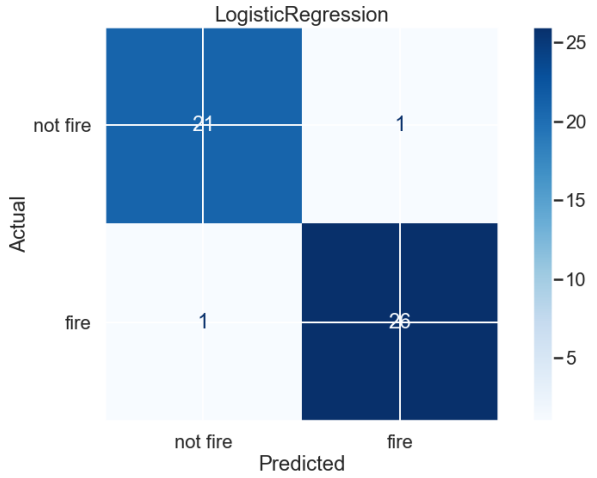


Γραφική Παράσταση 7: *Αποτελέσματα GradientBoostingClassifier*

Παραπάνω παρατηρείται ότι ο κατηγοριοποιητής GradientBoostingClassifier έχει ορθότητα (accuracy) 0.94, έχει ακρίβεια (precision) 0.90, έχει ανάκληση (recall) 1 και έχει αρμονικό μέσο (f1 score) 0.95.

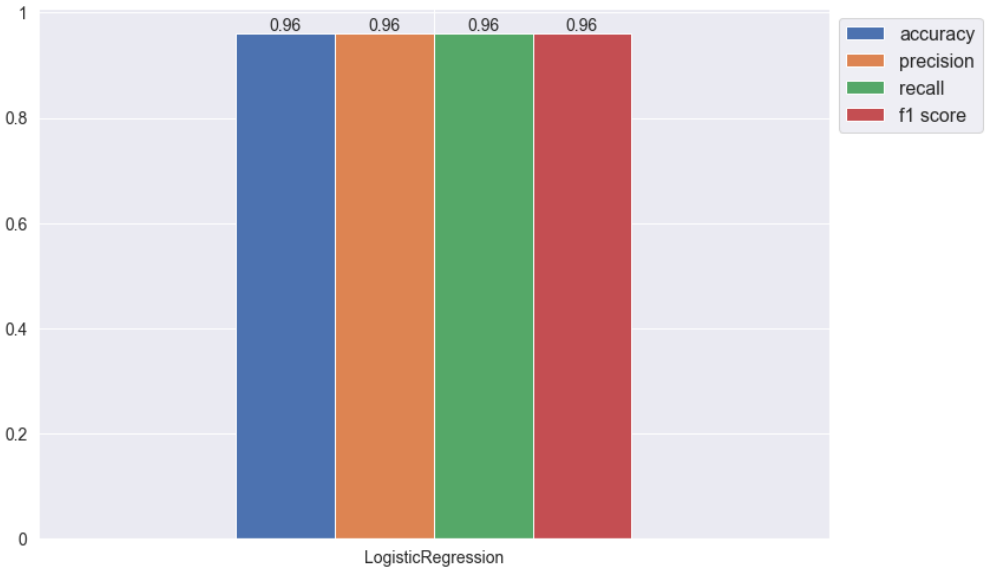
4.1.6 Λογιστική παλινδρόμηση

Παρακάτω παρουσιάζονται τα αποτελέσματα του κατηγοριοποιητή LogisticRegression.



Εικόνα 30: *Πίνακας σύγχυσης LogisticRegression*

Παραπάνω στον πίνακα σύγχυση (confusion matrix) παρατηρείται ότι ο κατηγοριοποιητής LogisticRegression πρόβλεψε 26 αληθινές θετικές πυρκαγιές (True Positive), 21 αληθινές αρνητικές μη πυρκαγιές (True Negative), 1_{α} ψευδή θετική πυρκαγιά (False Positive), και 1_{α} ψευδή αρνητική μη πυρκαγιά (False Negative).

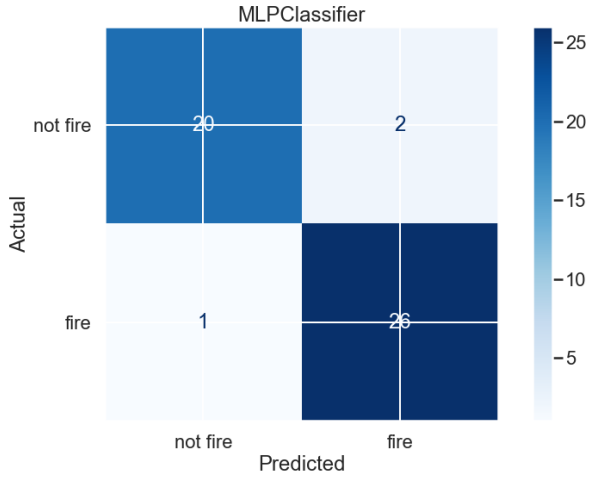


Γραφική Παράσταση 8: *Αποτελέσματα LogisticRegression*

Παραπάνω παρατηρείται ότι ο κατηγοριοποιητής LogisticRegression έχει ορθότητα (accuracy) 0.96, έχει ακρίβεια (precision) 0.96, έχει ανάκληση (recall) 0.96 και έχει αρμονικό μέσο (f1 score) 0.96.

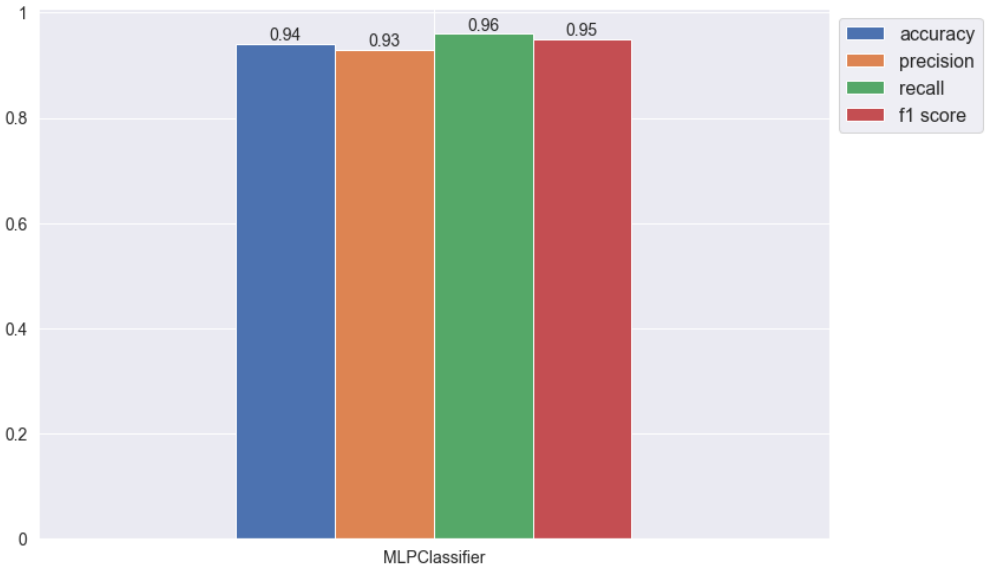
4.1.7 Νευρωνικά δίκτυα πολλών επιπέδων

Παρακάτω παρουσιάζονται τα αποτελέσματα του κατηγοριοποιητή MLPClassifier.



Εικόνα 31: *Πίνακας σύγχυσης MLPClassifier*

Παραπάνω στον πίνακα σύγχυση (confusion matrix) παρατηρείται ότι ο κατηγοριοποιητής MLPClassifier πρόβλεψε 26 αληθινές θετικές πυρκαγιές (True Positive), 20 αληθινές αρνητικές μη πυρκαγιές (True Negative), 2 ψευδής θετικές πυρκαγιές (False Positive), και 1_{α} ψευδή αρνητική μη πυρκαγιά (False Negative).

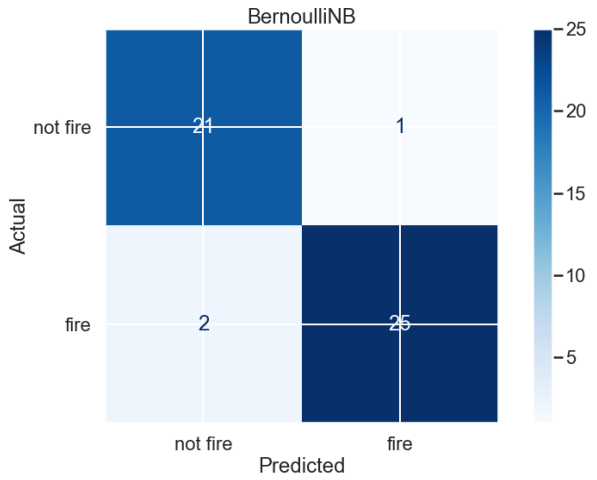


Γραφική Παράσταση 9: *Αποτελέσματα MLPClassifier*

Παραπάνω παρατηρείται ότι ο κατηγοριοποιητής MLPClassifier έχει ορθότητα (accuracy) 0.94, έχει ακρίβεια (precision) 0.93, έχει ανάκληση (recall) 0.96 και έχει αρμονικό μέσο (f1 score) 0.95.

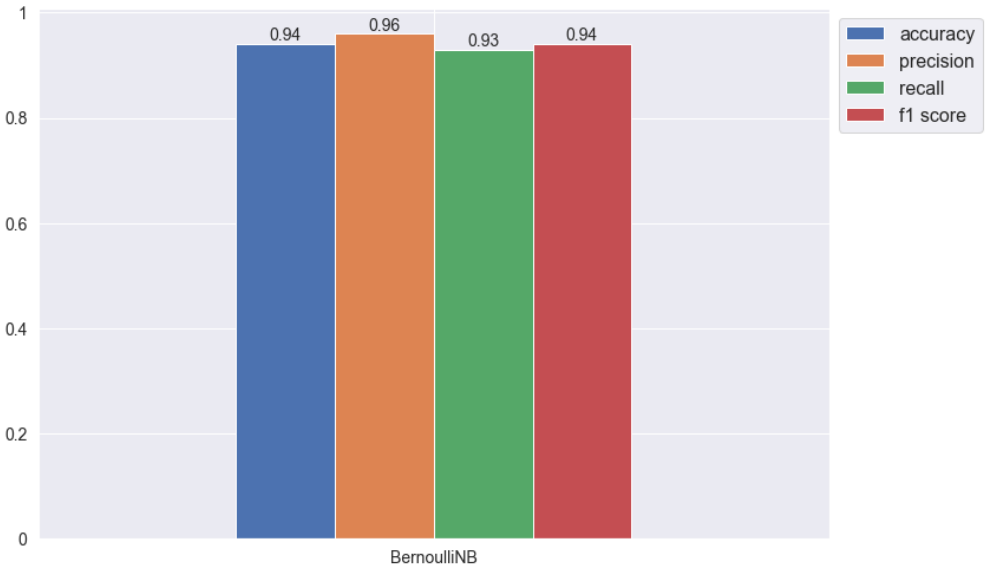
4.1.8 Bernoulli απλοϊκός Bayes

Παρακάτω παρουσιάζονται τα αποτελέσματα του κατηγοριοποιητή BernoulliNB.



Εικόνα 32: *Πίνακας σύγχυσης BernoulliNB*

Παραπάνω στον πίνακα σύγχυση (confusion matrix) παρατηρείται ότι ο κατηγοριοποιητής BernoulliNB πρόβλεψε 25 αληθινές θετικές πυρκαγιές (True Positive), 21 αληθινές αρνητικές μη πυρκαγιές (True Negative), 1α ψευδή θετική πυρκαγιά (False Positive), και 2 ψευδής αρνητικές μη πυρκαγιές (False Negative).



Γραφική Παράσταση 10: *Αποτελέσματα BernoulliNB*

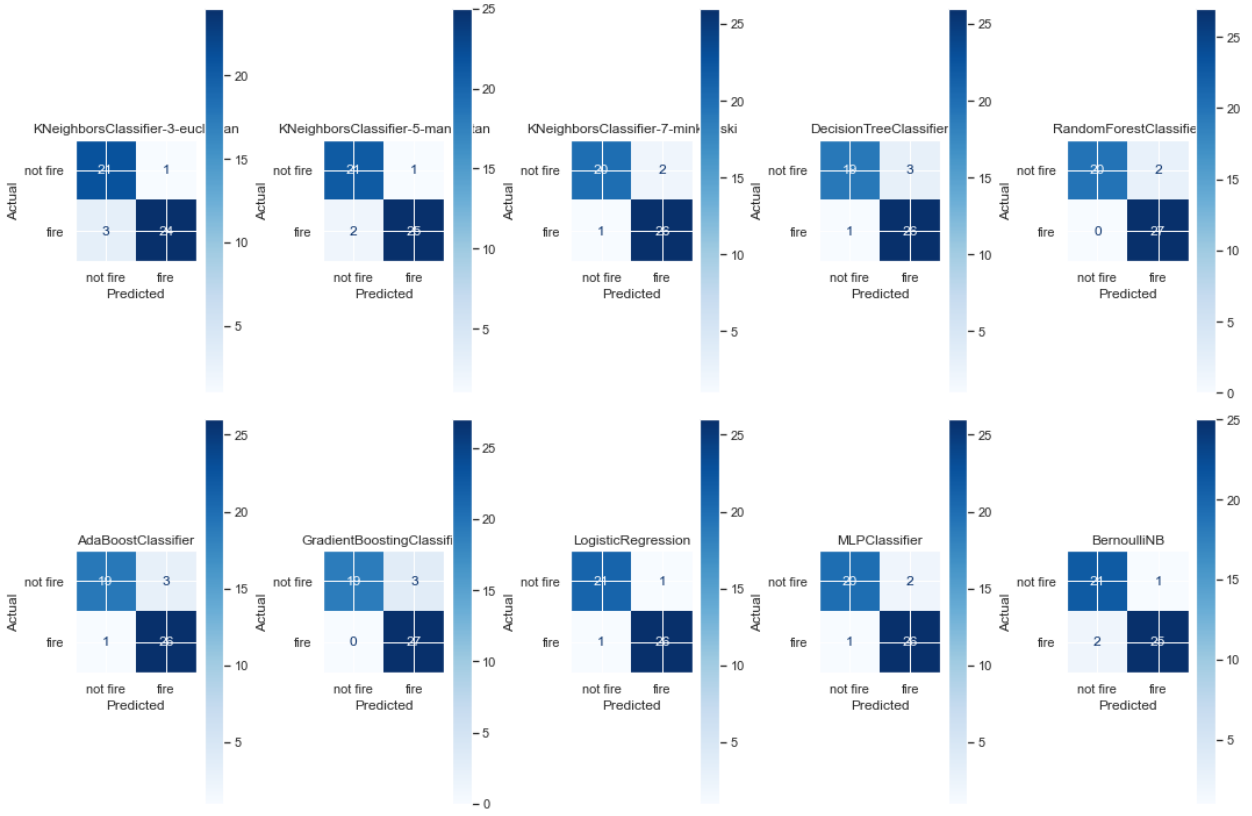
Παραπάνω παρατηρείται ότι ο κατηγοριοποιητής BernoulliNB έχει ορθότητα (accuracy) 0.94, έχει ακρίβεια (precision) 0.96, έχει ανάκληση (recall) 0.93 και έχει αρμονικό μέσο (f1 score) 0.94.

4.2 Συγκριτική αξιολόγηση αλγόριθμων μηχανικής μάθησης

Τώρα θα γίνει συγκριτική αξιολόγηση των αλγόριθμων μηχανικής μάθησης ή αλλιώς των κατηγοριοποιητών (classfiers). Αρχικά θα γίνει συγκριτική αξιολόγηση των πινάκων σύγχυσης (confusion matrices) και στην συνέχεια θα γίνει συγκριτική αξιολόγηση της ορθότητας (accuracy), της ακρίβειας (precision), της ανάκλησης (recall) και του αρμονικού μέσου (f1 score) όλων των κατηγοριοποιητών (classifiers). Τέλος θα γίνει σύνοψη όλων των αποτελεσμάτων όλων των κατηγοριοποιητών (classifiers).

4.2.1 Συγκριτική αξιολόγηση πίνακα σύγχυσης

Παρακάτω παρουσιάζονται όλοι οι πίνακες σύγχυσης (confusion matrices) των κατηγοριοποιητών (classifiers).



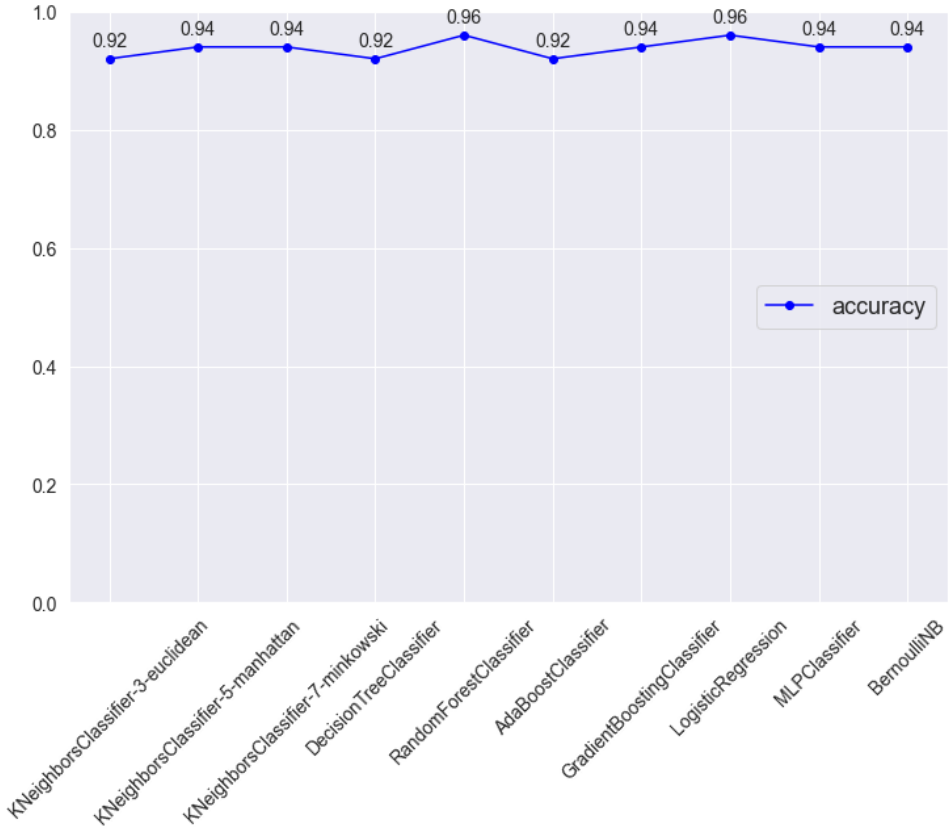
Εικόνα 33: *Πίνακες σύγχυσης όλων των κατηγοριοποιητών*

Παραπάνω παρατηρείται ότι όλοι οι κατηγοριοποιητές (classifiers) έχουν χαμηλό συνολικό αριθμό ψευδών (false) προβλέψεων. Οι κατηγοριοποιητές RandomForestClassifier και LogisticRegression έχουν τις λιγότερες ψευδές (false) προβλέψεις. Πιο συγκεκριμένα ο

κατηγοριοποιητής RandomForestClassifier πρόβλεψε 2 ψευδής θετικές πυρκαγιές (False Positive) και 0 ψευδής αρνητικές μη πυρκαγιές (False Negative). Ενώ ο κατηγοριοποιητής LogisticRegression πρόβλεψε 1_{α} ψευδή θετική πυρκαγιά (False Positive) και 1_{α} ψευδή αρνητική μη πυρκαγιά (False Negative).

4.2.2 Συγκριτική αξιολόγηση ορθότητας

Παρακάτω παρουσιάζεται η ορθότητα (accuracy) όλων των κατηγοριοποιητών (classifiers).



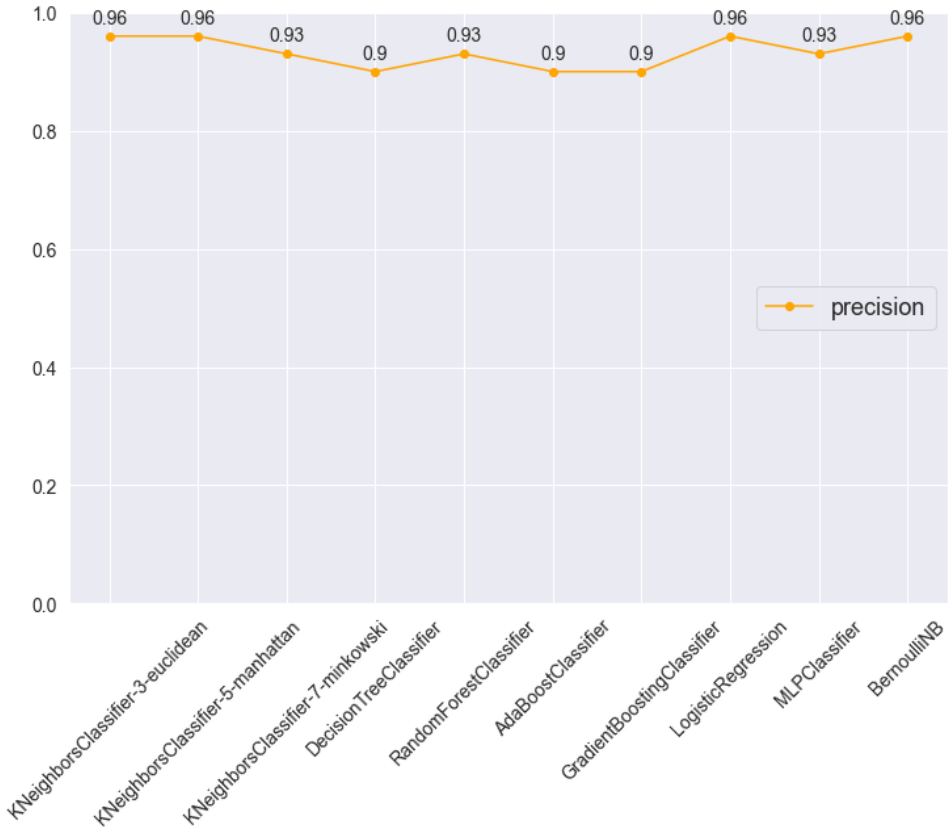
Γραφική Παράσταση 11: *Συγκριτική αξιολόγηση ορθότητας κατηγοριοποιητών*

Παραπάνω παρατηρείται ότι όλοι οι κατηγοριοποιητές (classifiers) έχουν υψηλή ορθότητα (accuracy). Οι κατηγοριοποιητές RandomForestClassifier και LogisticRegression έχουν την υψηλότερη ορθότητα (accuracy) σε σύγκριση με τους υπόλοιπους κατηγοριοποιητές (classifiers). Αυτό οφείλεται στο γεγονός ότι οι κατηγοριοποιητές RandomForestClassifier και LogisticRegression στον πίνακα σύγχυσης στην εικόνα 33 έχουν τις λιγότερο συνολικές ψευδές προβλέψεις. Πιο συγκεκριμένα οι κατηγοριοποιητές RandomForestClassifier και LogisticRegression έχουν 2 συνολικά ψευδείς προβλέψεις ενώ οι υπόλοιποι κατηγοριοποιητές

έχουν περισσότερες από 2. Υπενθυμίζεται ότι ο τύπος της ορθότητας (εξίσωση 28) διαιρεί τον συνολικό αριθμό προβλέψεων συνεπώς όσο πιο χαμηλός είναι αυτός ο αριθμός τόσο πιο υψηλή θα είναι η ορθότητα. Για αυτόν τον λόγω οι υπόλοιποι κατηγοριοποιητές έχουν χαμηλότερη ορθότητα ενώ ο RandomForestClassifier και LogisticRegression έχουν την υψηλότερη ορθότητα (accuracy).

4.2.3 Συγκριτική αξιολόγηση ακρίβειας

Παρακάτω παρουσιάζεται η ακρίβεια (precision) όλων των κατηγοριοποιητών (classifiers).



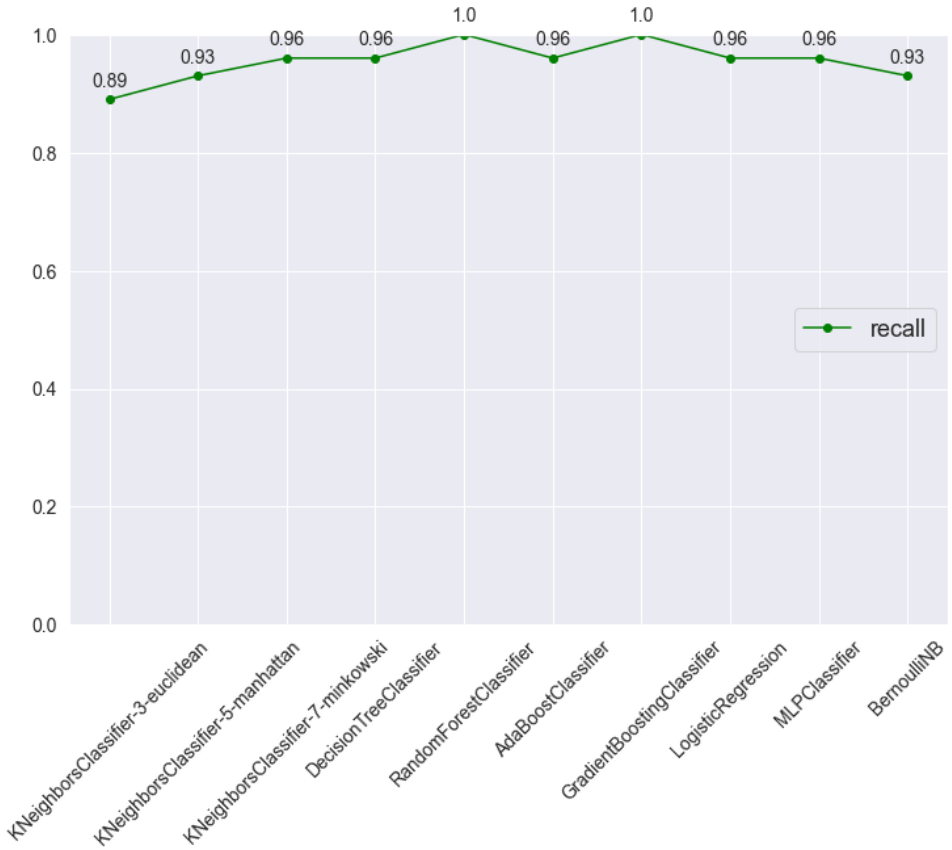
Γραφική Παράσταση 12: *Συγκριτική αξιολόγηση ακρίβειας κατηγοριοποιητών*

Παραπάνω παρατηρείται ότι όλοι οι κατηγοριοποιητές (classifiers) έχουν υψηλή ακρίβεια (precision). Οι κατηγοριοποιητές KNeighborsClassifier-3-euclidean, KNeighborsClassifier-5-manhattan, LogisticRegression και BernoulliNB έχουν την υψηλότερη ακρίβεια (precision) σε σύγκριση με τους υπόλοιπους κατηγοριοποιητές (classifiers). Αυτό οφείλεται στο γεγονός ότι οι συγκεκριμένοι κατηγοριοποιητές έχουν μόνο 1_{α} ψευδή θετική πρόβλεψη ενώ οι υπόλοιποι κατηγοριοποιητές έχουν περισσότερες από 1_{α}. Υπενθυμίζεται ότι ο τύπος της ακρίβειας (εξίσωση

29) διαιρεί τις ψευδείς θετικές προβλέψεις συνεπώς όσο πιο χαμηλός είναι αυτός ο αριθμός τόσο πιο υψηλή θα είναι η ακρίβεια. Για αυτόν τον λόγω οι υπόλοιποι κατηγοριοποιητές έχουν χαμηλότερη ακρίβεια ενώ ο KNeighborsClassifier-3-euclidean, KNeighborsClassifier-5-manhattan, LogisticRegression και BernoulliNB έχουν την υψηλότερη ακρίβεια (precision).

4.2.4 Συγκριτική αξιολόγηση ανάκλησης

Παρακάτω παρουσιάζεται η ανάκληση (recall) όλων των κατηγοριοποιητών (classifiers).



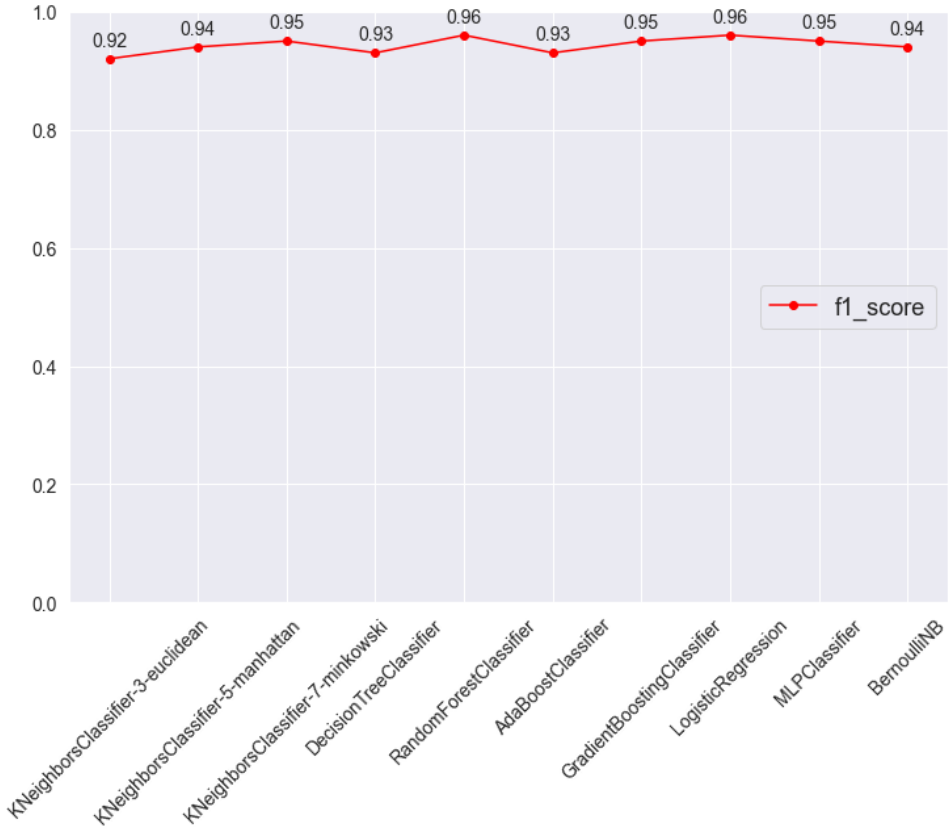
Γραφική Παράσταση 13: *Συγκριτική αξιολόγηση ανάκλησης κατηγοριοποιητών*

Παραπάνω παρατηρείται ότι όλοι οι κατηγοριοποιητές (classifiers) έχουν υψηλή ανάκληση (recall). Οι κατηγοριοποιητές RandomForestClassifier και GradientBoostingClassifier έχουν την υψηλότερη ανάκληση (recall) σε σύγκριση με τους υπόλοιπους κατηγοριοποιητές (classifiers). Αυτό οφείλεται στο γεγονός ότι οι συγκεκριμένοι κατηγοριοποιητές έχουν 0 ψευδής αρνητικές προβλέψεις ενώ οι υπόλοιποι κατηγοριοποιητές έχουν περισσότερες. Υπενθυμίζεται ότι ο τύπος της ανάκλησης (εξίσωση 30) διαιρεί τις ψευδείς αρνητικές προβλέψεις συνεπώς όσο πιο χαμηλός είναι αυτός ο αριθμός τόσο πιο υψηλή θα είναι η ανάκληση. Για αυτόν τον λόγω οι υπόλοιποι

κατηγοριοποιητές έχουν χαμηλότερη ανάκληση ενώ οι κατηγοριοποιητές RandomForestClassifier και GradientBoostingClassifier έχουν την υψηλότερη ανάκληση (recall).

4.2.5 Συγκριτική αξιολόγηση αρμονικού μέσου

Παρακάτω παρουσιάζεται ο αρμονικός μέσος (f1 score) όλων των κατηγοριοποιητών (classifiers).



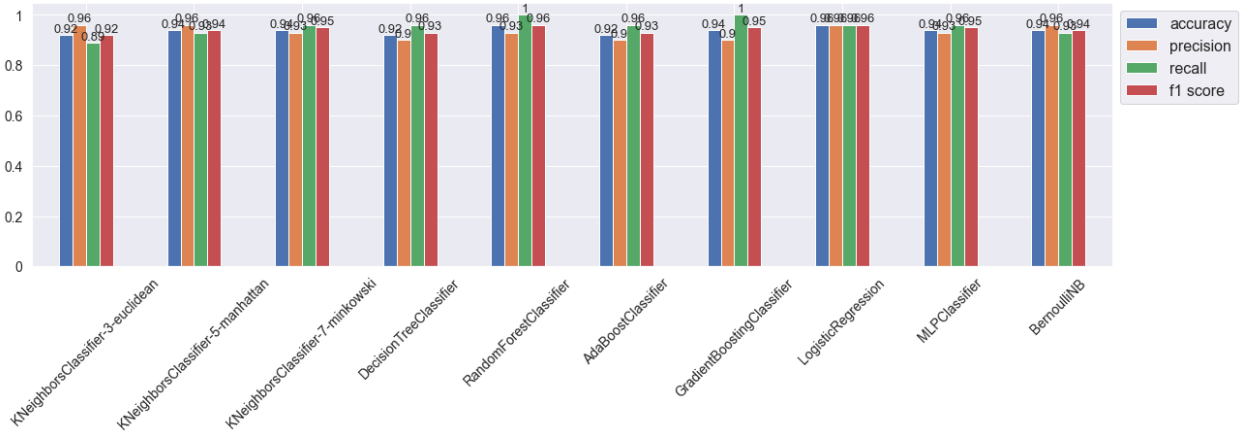
Γραφική Παράσταση 14: *Συγκριτική αξιολόγηση αρμονικού μέσου κατηγοριοποιητών*

Παραπάνω παρατηρείται ότι όλοι οι κατηγοριοποιητές (classifiers) έχουν υψηλό αρμονικό μέσο (f1 score). Οι κατηγοριοποιητές RandomForestClassifier και LogisticRegression έχουν την υψηλότερο αρμονικό μέσο (f1 score) σε σύγκριση με τους υπόλοιπους κατηγοριοποιητές (classifiers). Αυτό οφείλεται στο γεγονός ότι RandomForestClassifier είναι ένας από τους κατηγοριοποιητές με υψηλότερη ανάκληση, ενώ ο LogisticRegression είναι ένας από τους κατηγοριοποιητές με ακρίβεια. Υπενθυμίζεται ότι ο τύπος του αρμονικού μέσου (εξίσωση 31) χρησιμοποιεί την ανάκληση και την ακρίβεια συνεπώς κοντά στο 2 είναι η πρόσθεση της ακρίβειας με την ανάκληση τόσο πιο υψηλά αποτελέσματα θα έχουν. Για αυτόν τον λόγω οι

υπόλοιποι κατηγοριοποιητές έχουν χαμηλότερο αρμονικό μέσο ενώ οι κατηγοριοποιητές RandomForestClassifier και LogisticRegression έχουν τον υψηλότερο αρμονικό μέσο (f1 score).

4.2.6 Σύνοψη αποτελεσμάτων

Παρακάτω θα παρουσιαστεί η σύνοψη των αποτελεσμάτων των κατηγοριοποιητών (classifiers) δηλαδή θα παρουσιαστούν η ορθότητα (accuracy), η ακρίβεια (precision), η ανάκληση (recall) και ο αρμονικός μέσος (f1 score) μαζί.



Γραφική Παράσταση 15: *Συνοπτικά αποτελέσματα κατηγοριοποιητών*

Από μια γενική εικόνα παρατηρείται ότι γενικά όλοι οι κατηγοριοποιητές (classifiers) έχουν υψηλά αποτελέσματα. Αυτό οφείλεται στο γεγονός ότι όλοι οι κατηγοριοποιητές έχουν χαμηλό συνολικό αριθμό ψευδών (false) προβλέψεων όπως παρατηρείται στο κεφάλαιο 4.2.1.

4.3 Επιλογή βέλτιστου αλγόριθμου μηχανικής μάθησης

Όλοι οι κατηγοριοποιητές (classifiers) έχουν υψηλά αποτελέσματα όπως παρατηρείται στο κεφάλαιο 4.2.6 και έχουν πολύ μικρές διαφορές μεταξύ τους. Όπως παρατηρείται στο κεφάλαιο 4.2.1 στην συγκριτική αξιολόγηση των πινάκων σύγχυσης (confusion matrices) όλοι οι κατηγοριοποιητές (classifiers) έχουν χαμηλό συνολικό αριθμό ψευδών (false) προβλέψεων. Αυτό είναι εμφανές και στα κεφάλαια 4.2.2. έως 4.2.5 όπου τα αποτελέσματα της ορθότητας (accuracy), της ακρίβεια (precision), της ανάκληση (recall) και του αρμονικού μέσου (f1 score) ήταν αρκετά κοντά μεταξύ τους.

Για την επιλογή του βέλτιστου αλγόριθμου μηχανικής μάθησης θα δούμε 2 κριτήρια. Την ορθότητα (accuracy) και τον αρμονικό μέσο (f1 score) ο οποίος συνδυάζει την ακρίβεια (precision) και την ανάκληση (recall).

Οι κατηγοριοποιητές με την υψηλότερη ορθότητα (accuracy) και αρμονικό μέσο (f1 score) είναι οι RandomForestClassifier και LogisticRegression. Ο RandomForestClassifier έχει ορθότητα (accuracy) και αρμονικό μέσο (f1 score) 0.96 ενώ ο LogisticRegression έχει ακριβός το ίδιο αποτέλεσμα με ορθότητα (accuracy) και αρμονικό μέσο (f1 score) 0.96.

Οι δύο αυτοί κατηγοριοποιητές (classifiers) μπορεί να φαίνονται ότι είναι ίσοι αλλά υπάρχει μια πολύ μικρή διαφορά μεταξύ τους στα δεκαδικά ψηφία. Ο κατηγοριοποιτής RandomForestClassifier έχει ακρίβεια (precision) 0.93 και ανάκληση (recall) 1. Αυτό σημαίνει ότι από τον τύπο του αρμονικού μέσου (f1 score) το αναλυτικό αποτέλεσμά του είναι 0.9637. Ενώ ο κατηγοριοποιητής LogisticRegression έχει ακρίβεια (precision) 0.96 και ανάκληση (recall) 0.96 που αυτό σημαίνει ότι το αποτέλεσμα του αρμονικού μέσου (f1 score) είναι 0.96.

Αυτό που παρατηρείται είναι ότι υπάρχει μια πολύ μικρή διαφορά 0.0037 μεταξύ τους στον αρμονικό μέσο (f1 score) κάνοντας τον κατηγοριοποιητή RandomForestClassifier να υπερισχύει έναντι του κατηγοριοποιτή LogisticRegression. Συνεπώς ο κατηγοριοποιητής RandomForestClassifier είναι ο πιο βέλτιστος ενώ ο κατηγοριοποιητής LogisticRegression είναι ο $2_{ος}$ πιο βέλτιστος.

Κεφάλαιο 5: Συμπεράσματα και προτάσεις για μελλοντικές κατευθύνσεις

Στο παρόν κεφάλαιο θα παρουσιαστεί το συμπεράσματα της παρούσας διπλωματικής εργασίας καθώς και προτάσεις για μελλοντικές κατευθύνσεις. Θα υπάρξει μια μικρή σύνοψη της διπλωματικής εργασίας και θα παρουσιαστεί το συμπεράσματα. Τέλος θα παρουσιαστούν ορισμένες προτάσεις για το πώς μπορεί να αναπτυχθεί η διπλωματική εργασία στο μέλλον.

5.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία έγινε συγκριτική αξιολόγηση αλγόριθμων μηχανικής μάθησης για την πρόβλεψη δασικών πυρκαγιών. Η μηχανική μάθηση είναι ένα πολύ ισχυρό μέρος της τεχνητής νοημοσύνης και μπορεί να είναι κατάλληλη σε προβλήματα που αφορούν κατηγοριοποιήση. Με την άνοδο των χρόνων η μηχανική μάθηση χρησιμοποιείται όλο και περισσότερο και θα μπορέσει να δώσει λύσεις σε πολλά προβλήματα στο μέλλον [31-47]. Ένα από αυτά τα προβλήματα είναι η πρόβλεψη των δασικών πυρκαγιών. Οι δασικές πυρκαγιές αποτελούν ένα από τα σημαντικότερα προβλήματα στον πλανήτη μας. Μπορούν να προκαλέσουν οικολογικές ζημιές όπως την καταστροφή δασών, την μόλυνση του περιβάλλοντος και την αύξηση της κλιματικής αλλαγής. Εκτός από οικολογικές καταστροφές μπορούν να προκαλέσουν και οικονομικές καταστροφικές καθώς επίσης και απειλή προς την ανθρώπινη ζωή. Για την πρόβλεψη των δασικών πυρκαγιών χρησιμοποιήθηκαν δασικά δεδομένα από δύο περιοχές της Αλγερίας, την περιοχή Bejaia Region και την περιοχή Sidi-Bel Abbes Region. Μέσω της κατηγοριοποίησης οι αλγόριθμοι τις μηχανικής μάθησης μπόρεσαν να προβλέψουν δασικές πυρκαγιές και στην συνέχεια να αξιολογηθούν. Συνεπώς η μηχανική μάθηση μπορεί να εφαρμοστεί και να δώσεις λύσεις σε προβλήματα πραγματικού κόσμου καθώς και να αξιολογηθεί έτσι ώστε οι προβλέψεις να είναι πιο ακριβείς.

5.2 Προτάσεις για μελλοντικές κατευθύνσεις

Για την πρόβλεψη και την αξιολόγηση των αποτελεσμάτων της παρούσας διπλωματικής εργασίας χρησιμοποιήθηκαν οι τεχνικές test_train_split και StandardScaler της βιβλιοθήκης scikit-learn. Η συνάρτηση test_train_split χρησιμοποιήθηκε για την διαχώριση των δεδομένων εκπαίδευσης και δοκιμής με την παράμετρο 0.2 δηλαδή χρησιμοποίησε το 20% των δεδομένων για δοκιμή. Η συνάρτηση StandardScaler χρησιμοποιήθηκε για την κλιμάκωση των χαρακτηριστικών. Για την

ανάπτυξη της διπλωματικής εργασίας μπορούν να χρησιμοποιηθούν και άλλες τεχνικές. Μία από αυτές είναι η k-Fold Cross-Validation όπου διαχωρίζει τα δεδομένα σε k υποσύνολα και κάθε σύνολο μπορεί να αξιολογηθεί k φορές έτσι ώστε να βρεθεί το υποσύνολο με την μέγιστη αξιολόγηση. Επίσης για την κλιμάκωση των χαρακτηριστικών μπορεί να χρησιμοποιηθεί η τεχνική MinMaxScaler όπου είναι διαφορετική από την StandardScaler και εξηγήθηκε στο κεφάλαιο 2.4.2.1. Ακόμη μπορούν να χρησιμοποιηθούν και οι τεχνικές PCA και LDA για την μείωση διαστάσεων. Αυτές οι τεχνικές μείωσης διαστάσεων μπορούν να χρησιμοποιηθούν και με συνδιασμό κλιμάκωσης χαρακτηριστικών όπως StandardScaler και MinMaxScaler. Ένα μειονέκτημα της μείωσης διαστάσεων είναι να υπάρξει απώλεια πληροφοριών έχοντας ως συνέπεια να υπάρξει μείωση ορθότητας, ακρίβειας, ανάκλησης και αρμονικού μέσου. Με αυτές τις παραπάνω τεχνικές μπορεί η ορθότητα, η ακρίβεια, η ανάκληση και ο αρμονικός μέσος να έχουν υψηλότερα αποτελέσματα από τα αποτελέσματα που παρουσιάστηκαν στην παρούσα διπλωματική.

Βιβλιογραφία

1. Kontellis, E., Troussas, C., Krouska, A., & Sgouropoulou, C. (2021). *Real-time face mask detector using convolutional neural networks amidst COVID-19 pandemic. In Novelties in Intelligent Digital Systems*: Proceedings of the 1st International Conference (NIDS 2021), Athens, Greece, September 30-October 1, 2021 (Vol. 338, p. 247-255). IOS Press. doi:10.3233/FAIA210102
2. Christos Troussas, Akrivi Krouska, Cleo Sgouropoulou: *Collaboration and fuzzy-modeled personalization for mobile game-based learning in higher education*, Computers & Education, Volume 144, 2020, 103698, https://doi.org/10.1016/j.compedu.2019.103698.
3. Kanetaki, Z., Stergiou, C., Bekas, G., Troussas, C., & Sgouropoulou, C. (2021). *Data Mining for Improving Online Higher Education Amidst COVID-19 Pandemic: A Case Study in the Assessment of Engineering Students. Novelties in Intelligent Digital Systems: Proceedings of the 1st International Conference* (NIDS 2021), Athens, Greece, September 30-October 1, 2021 (Vol. 338, p. 157-165). doi:10.3233/FAIA210088.
4. Kapetanaki, A., Krouska, A., Troussas, C., & Sgouropoulou, C. (2021). *A Novel Framework Incorporating Augmented Reality and Pedagogy for Improving Reading Comprehension in Special Education. In Novelties in Intelligent Digital Systems: Proceedings of the 1st International Conference* (NIDS 2021), Athens, Greece, September 30-October 1, 2021 (Vol. 338, p. 105-110). IOS Press. doi:10.3233/FAIA210081.
5. Akrivi KROUSKA, Christos TROUSSAS, Filippos GIANNAKAS, Cleo SGOUROPOULOU, and Ioannis VOYIATZIS, *Enhancing the Effectiveness of Intelligent Tutoring Systems Using Adaptation and Cognitive Diagnosis Modeling, Novelties in Intelligent Digital Systems*: Proceedings of the 1st International Conference (NIDS 2021), Athens, Greece, September 30-October 1, 2021 (Vol. 338, p. 40-45). IOS Press, doi:10.3233/FAIA210073.
6. Krouska A, Troussas C, Virvou M (2017) *Comparative Evaluation of Algorithms for Sentiment Analysis over Social Networking Services*. JUCS - Journal of Universal Computer Science 23(8): 755-768. https://doi.org/10.3217/jucs-023-08-0755
7. C. Troussas, M. Virvou and S. Mesaretzidis, *"Comparative analysis of algorithms for student characteristics classification using a methodological framework,"* 2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA), 2015, pp. 1-5, doi: 10.1109/IISA.2015.7388038.
8. F. Abid and N. Izeboudjen, 2020. *Predicting Forest Fire in Algeria Using Data Mining Techniques: Case Study of the Decision Tree Algorithm*, [online] Available at: https://www.researchgate.net/publication/339062373_Predicting_Forest_Fire_in_Algeria_Using_Data_Mining_Techniques_Case_Study_of_the_Decision_Tree_Algorithm
9. Bentekhici, N., Bellal, S. and Zegrar, A., 2020. *Contribution of remote sensing and GIS to mapping the fire risk of Mediterranean forest case of the forest massif of Tlemcen (North-West Algeria).* Natural Hazards, [online] Available at: https://link.springer.com/article/10.1007/s11069-020-04191-6

10. Albini, F., 1984. *Wildland Fires: Predicting the behavior of wildland fires—among nature's most potent forces—can save lives, money, and natural resources.* [online] jstor.org. Available at: https://www.jstor.org/stable/27852969?casa_token=vqVwXS9bhoAAAAA%3AzY7pnegNeo4aDwGBFq57YcBYexFB1LSvkpBlKYD1f2eznHETkV7cJ_fT7HqrHpMhie4j6YaRut73qd3HBWIw2mjulZEjJttnVZa3MeUBa0ej9_hC4&seq=1#metadata_info_tab_contents
11. Géron, A., 2019. *Hands-on machine learning with Scikit-Learn and TensorFlow*. 2nd ed. Sebastopol (Clif.) [etc.]: O'Reilly.
12. Mitchell, T., 1997. *Machine Learning*. McGraw-Hill Science/Engineering/Math
13. Covington, P., Adams, J. and Sargin, E., 2016. *Deep Neural Networks for YouTube Recommendations*. [online] ACM Conferences. Available at: https://dl.acm.org/doi/abs/10.1145/2959100.2959190
14. Amatriain, X., 2013. *Big & personal | Proceedings of the 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*. [online] dl.acm.org. Available at: https://dl.acm.org/doi/abs/10.1145/2501221.2501222?casa_token=lFhnyrXOi34AAAAA:y6TNrJLh3jSYJSWX2yrb6SJL20uGOwIo8u-GEqdLryqaY0uHPLUvTfAscNoeZvmG09nlvbZqKN-gmQ
15. Khanzode, C. and Sarode, R., 2020. *Advantages and disadvantages of artificial intelligence and machine learning: A literature review*. *IAEME Publication*, [online] Available at: https://www.academia.edu/44895767/ADVANTAGES_AND_DISADVANTAGES_OF_ARTIFICIAL_INTELLIGENCE_AND_MACHINE_LEARNING_A_LITERATURE_REVIEW?from=cover_page
16. Georgouli, A. (2015). *Τεχνητή νοημοσύνη* [Undergraduate textbook]. Athens: Kallipos, Open Academic Editions. http://hdl.handle.net/11419/3381
17. Nayak, R., Jiwani, S. and Rajitha, B., 2021. *Spam email detection using machine learning algorithm*. *elsevier*
18. Τήλλυρος, Χ., 2019. *Συγκριτική αξιολόγηση αλγορίθμων μηχανικής μάθησης σε δεδομένα ασθενών με διαβήτη*. Μεταπτυχιακή Διπλωματική. Πανεπιστήμιο Πειραιώς. Πειραιάς
19. Barnadas M., 2016. *Machine Learning applied to crime prediction.* Degree Thesis. Polytechnic University of Catalonia. Barcelona
20. I. Hendrickx, A. Van den Bosch. *Hybrid algorithms with Instance-Based Classification*. Machine Learning: ECML2005.
21. Παπαστεργίου, Κ., 2019. *Μαθαίνω Python & Tkinter*.
22. Raschka, S. 2015. *Unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics.*

23. Taunk, K., De, S., Verma, S. and Swetapadma, A., 2019. *A Brief Review of Nearest Neighbor Algorithm for Learning and Classification*. [online] Available at: https://ieeexplore.ieee.org/document/9065747
24. Sathyadevan, S. and Nair, R., 2014. *Comparative Analysis of Decision Tree Algorithms: ID3, C4.5 and Random Forest*. [online] Available at: https://link.springer.com/chapter/10.1007/978-81-322-2205-7_51
25. Shalev-Shwartz, S. and Ben-David, S., 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press. https://www.cs.huji.ac.il/w~shais/UnderstandingMachineLearning
26. Vezhnevets, A. and Vezhnevets, V., 2005. *'Modest AdaBoost' – Teaching AdaBoost to Generalize Better*. [online] Available at: https://www.researchgate.net/publication/239542136_%27Modest_AdaBoost%27_-_Teaching_AdaBoost_to_Generalize_Better
27. F. Giannakas, C. Troussas, I. Voyiatzis, C. Sgouropoulou, A deep learning classification framework for early prediction of team-based academic performance, Applied Soft Computing, Volume 106, 2021, 107355. https://doi.org/10.1016/j.asoc.2021.107355.
28. Troussas, C., Krouska, A. & Virvou, M. A multilayer inference engine for individualized tutoring model: adapting learning material and its granularity. Neural Comput & Applic (2021). https://doi.org/10.1007/s00521-021-05740-1.
29. Giannakas, F., Troussas, C., Krouska, A., Sgouropoulou, C., Voyiatzis, I. (2021). XGBoost and Deep Neural Network Comparison: The Case of Teams' Performance. In: Cristea, A.I., Troussas, C. (eds) Intelligent Tutoring Systems. ITS 2021. Lecture Notes in Computer Science, vol 12677. Springer, Cham. https://doi.org/10.1007/978-3-030-80421-3_37.
30. C. Troussas, F. Giannakas, C. Sgouropoulou & I. Voyiatzis (2020). Collaborative activities recommendation based on students' collaborative learning styles using ANN and WSM, Interactive Learning Environments, DOI: 10.1080/10494820.2020.1761835.
31. Krouska, A., Troussas, C., Virvou, M. (2019). Computerized Adaptive Assessment Using Accumulative Learning Activities Based on Revised Bloom's Taxonomy. In: Virvou, M., Kumeno, F., Oikonomou, K. (eds) Knowledge-Based Software Engineering: 2018. JCKBSE 2018. Smart Innovation, Systems and Technologies, vol 108. Springer, Cham. https://doi.org/10.1007/978-3-319-97679-2_26
32. K. Chrysafiadi, C. Troussas and M. Virvou, "A Framework for Creating Automated Online Adaptive Tests Using Multiple-Criteria Decision Analysis," 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2018, pp. 226-231, doi: 10.1109/SMC.2018.00049.
33. Troussas, C., Virvou, M. & Alepis, E. Comulang: towards a collaborative e-learning system that supports student group modeling. SpringerPlus 2, 387 (2013). https://doi.org/10.1186/2193-1801-2-387

34. Virvou, M., Troussas, C., Caro, J., Espinosa, K.J. (2012). User Modeling for Language Learning in Facebook. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds) Text, Speech and Dialogue. TSD 2012. Lecture Notes in Computer Science(), vol 7499. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-32790-2_42
35. Papakostas C., Troussas C., Krouska A., Sgouropoulou C. Measuring User Experience, Usability and Interactivity of a Personalized Mobile Augmented Reality Training System. Sensors. 2021; 21(11):3888. https://doi.org/10.3390/s21113888
36. C. Troussas, M. Virvou, and K. J. Espinosa, "Using visualization algorithms for discovering patterns in groups of users for tutoring multiple languages through Social Networking", Journal of Networks, vol. 10, no. 12, pp. 668-674, 2015.
37. Troussas, C., Virvou, M., Caro, J., & Espinosa, K. J. (2013). Language Learning Assisted by Group Profiling in Social Networks. International Journal of Emerging Technologies in Learning (iJET), 8(3), pp. 35–38. https://doi.org/10.3991/ijet.v8i3.2684.
38. Troussas C., Krouska A., Sgouropoulou C. Improving Learner-Computer Interaction through Intelligent Learning Material Delivery Using Instructional Design Modeling. Entropy. 2021; 23(6):668. https://doi.org/10.3390/e23060668
39. Krouska, A., Troussas, C., Sgouropoulou, C. (2020). A Personalized Brain-Based Quiz Game for Improving Students' Cognitive Functions. In: Frasson, C., Bamidis, P., Vlamos, P. (eds) Brain Function Assessment in Learning. BFAL 2020. Lecture Notes in Computer Science(), vol 12462. Springer, Cham. https://doi.org/10.1007/978-3-030-60735-7_11.
40. Krouska, A., Troussas, C., Sgouropoulou, C. (2020). Applying Genetic Algorithms for Recommending Adequate Competitors in Mobile Game-Based Learning Environments. In: Kumar, V., Troussas, C. (eds) Intelligent Tutoring Systems. ITS 2020. Lecture Notes in Computer Science(), vol 12149. Springer, Cham. https://doi.org/10.1007/978-3-030-49663-0_23.
41. Krouska, A., Troussas, C. and Sgouropoulou, C. 2019. Fuzzy Logic for Refining the Evaluation of Learners' Performance in Online Engineering Education. European Journal of Engineering and Technology Research. 4, 6 (Jun. 2019), 50–56. DOI: https://doi.org/10.24018/ejeng.2019.4.6.1369.
42. Troussas, C., Krouska, A., Sgouropoulou, C. (2020). Dynamic Detection of Learning Modalities Using Fuzzy Logic in Students' Interaction Activities. In: Kumar, V., Troussas, C. (eds) Intelligent Tutoring Systems. ITS 2020. Lecture Notes in Computer Science(), vol 12149. Springer, Cham. https://doi.org/10.1007/978-3-030-49663-0_24.
43. C. Troussas, A. Krouska, E. Alepis & M. Virvou (2020) Intelligent and adaptive tutoring through a social network for higher education, New Review of Hypermedia and Multimedia, 26:3-4, 138-167, DOI: 10.1080/13614568.2021.1908436
44. C. Troussas, A. Krouska, F. Giannakas, C. Sgouropoulou, and I. Voyiatzis. Automated reasoning of learners' cognitive states using classification analysis. In 24th Pan-Hellenic

Conference on Informatics, pp. 103–106, 2020. https://doi.org/10.1145/3437120.3437285.
45. Kanetaki, Z., Stergiou, C., Bekas, G., Troussas, C., & Sgouropoulou, C. (2022). A Hybrid Machine Learning Model for Grade Prediction in Online Engineering Education. International Journal of Engineering Pedagogy (iJEP), 12(3), pp. 4–24. https://doi.org/10.3991/ijep.v12i3.23873.
46. A. Krouska, C. Troussas, A. Voulodimos, C. Sgouropoulou, A 2-tier fuzzy control system for grade adjustment based on students' social interactions, Expert Systems with Applications, Volume 203, 2022, 117503, https://doi.org/10.1016/j.eswa.2022.117503.
47. Giannakas, F., Troussas, C., Krouska, A. et al. Multi-technique comparative analysis of machine learning algorithms for improving the prediction of teams' performance. Educ Inf Technol (2022). https://doi.org/10.1007/s10639-022-10900-4.