ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΤΜΗΜΑ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΤΑΞΙΝΟΜΗΣΗ ΜΑΝΙΤΑΡΙΩΝ ΣΕ ΒΡΩΣΙΜΑ ΚΑΙ ΔΗΛΗΤΗΡΙΩΔΗ

Εργασία

του

Μισαηλίδη Σάββα

AM: ics21166

ΤΑΞΙΝΟΜΗΣΗ ΜΑΝΙΤΑΡΙΩΝ ΣΕ ΒΡΩΣΙΜΑ ΚΑΙ ΔΗΛΗΤΗΡΙΩΔΗ

Μισαηλίδης Σάββας

Απαλλακτική Εργασία

υποβαλλόμενη για την εκπλήρωση των απαιτήσεων του

Μαθήματος: ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

του Τμήματος ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

Επιβλέπων Καθηγητής
Ευτύχιος Πρωτοπαπαδάκης

Περίληψη

Η παρούσα εργασία ασχολείται με την ανάπτυξη ενός μοντέλου μηχανικής μάθησης για την ταξινόμηση μανιταριών σε βρώσιμα και δηλητηριώδη. Η βάση δεδομένων προέρχεται από το Kaggle και περιλαμβάνει διάφορα χαρακτηριστικά των μανιταριών, όπως το χρώμα του καπέλου, το σχήμα του μίσχου και η υφή της επιφάνειας. Στόχος είναι η ακριβής πρόβλεψη της κατηγορίας ενός μανιταριού με βάση αυτά τα χαρακτηριστικά, προσφέροντας ένα χρήσιμο εργαλείο για την αναγνώριση επικίνδυνων μανιταριών και την αποφυγή δηλητηριάσεων.

Χρησιμοποιήθηκαν διάφοροι αλγόριθμοι ταξινόμησης, όπως τα Δέντρα Απόφασης, τα Τυχαία Δάση και οι Υποστηρικτές Διανύσματα Μηχανών, για την εκπαίδευση και την αξιολόγηση του μοντέλου. Τα αποτελέσματα έδειξαν υψηλή ακρίβεια στην πρόβλεψη, αποδεικνύοντας την αποτελεσματικότητα των μεθόδων που χρησιμοποιήθηκαν. Η εργασία καταλήγει με την παρουσίαση των ευρημάτων και τη συζήτηση για τη βελτίωση του μοντέλου.

Η επιτυχής υλοποίηση του συστήματος ταξινόμησης μπορεί να συμβάλλει σημαντικά στην προστασία της δημόσιας υγείας, παρέχοντας ένα εργαλείο για την ασφαλή αναγνώριση των μανιταριών.

Λέξεις κλειδιά: Πρόβλεψη, Ταξινόμηση μανιταριών, ανάλυση δεδομένων (Exploratory Data Analysis - EDA), Μηχανική μάθηση, Βρώσιμα μανιτάρια, Δηλητηριώδη μανιτάρια, Χαρακτηριστικά μανιταριών, Αλγόριθμοι ταξινόμησης, Δημόσια υγεία, Ασφαλής αναγνώριση.

Abstract

In this study, the focus is on the development of a machine learning model for classifying mushrooms as edible or poisonous. The dataset used comes from Kaggle and includes various characteristics of mushrooms, such as cap color, stem shape, and surface texture. The goal is to accurately predict the category of a mushroom based on these features, providing a useful tool for identifying dangerous mushrooms and preventing poisonings.

Various classification algorithms, such as Decision Trees, Random Forests, and Support Vector Machines, were used to train and evaluate the model. The results showed high prediction accuracy, demonstrating the effectiveness of the methods used. The work concludes with the presentation of the findings and a discussion on improving the model.

The successful implementation of the classification system can significantly contribute to public health by providing a tool for the safe identification of mushrooms.

Keywords: Prediction, Mushroom Classification, Exploratory Data Analysis (EDA), Machine Learning, Edible Mushrooms, Poisonous Mushrooms, Mushroom Characteristics, Classification Algorithms, Public Health, Safe Identification.

Περιεχόμενα

1 Εισαγωγή

1.1 Θεωρητικό υπόβαθρο

Η ταξινόμηση μανιταριών σε βρώσιμα και δηλητηριώδη είναι ένα ζήτημα ζωτικής σημασίας, δεδομένου ότι η κατανάλωση δηλητηριωδών μανιταριών μπορεί να οδηγήσει σε σοβαρές επιπτώσεις στην υγεία ή ακόμα και σε θάνατο. Στο πλαίσιο αυτό, η μηχανική μάθηση παρέχει αποτελεσματικές λύσεις, βελτιώνοντας την ακρίβεια στην αναγνώριση και κατηγοριοποίηση των μανιταριών (Smith et al., Mushroom Classification using Machine Learning: A Comprehensive Study 2023). Οι τεχνικές μηχανικής μάθησης είναι ικανές να εξάγουν μοτίβα και συσχετίσεις από δεδομένα που ενδέχεται να μην είναι ορατά με παραδοσιακές μεθόδους ανάλυσης, παρέχοντας έτσι ένα ισχυρό εργαλείο για την πρόγνωση της τοξικότητας των μανιταριών.

Αρχικά, οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούνται για την ταξινόμηση των μανιταριών βασίζονται στην επεξεργασία χαρακτηριστικών δεδομένων από φυσικές και χημικές ιδιότητες των μανιταριών. Το μοντέλο μηχανικής μάθησης λειτουργεί με την ανάλυση αυτών των δεδομένων και την εκπαίδευση πάνω σε δείγματα που έχουν ήδη κατηγοριοποιηθεί ως βρώσιμα ή δηλητηριώδη. Το σύνολο δεδομένων που χρησιμοποιείται για την έρευνα αυτή προέρχεται από το Kaggle, περιέχοντας πληθώρα χαρακτηριστικών όπως το σχήμα, το χρώμα και η υφή των μανιταριών, τα οποία είναι κρίσιμα για την ακριβή ταξινόμηση.

Έχουν εφαρμοστεί διάφορες μέθοδοι μηχανικής μάθησης στην παρούσα μελέτη, με στόχο τη βελτιστοποίηση της ακρίβειας στην ταξινόμηση των μανιταριών. Οι τεχνικές που χρησιμοποιήθηκαν περιλαμβάνουν logistic regression, random forests, και neural networks, με σκοπό να προσδιοριστεί ποια μοντέλα αποδίδουν καλύτερα στον διαχωρισμό των βρώσιμων από τα δηλητηριώδη μανιτάρια. Σημαντικές μεταβλητές που λήφθηκαν υπόψη περιλαμβάνουν χαρακτηριστικά όπως το χρώμα του καπέλου, η υφή του ποδιού και η παρουσία ή απουσία μυρωδιάς.

Ωστόσο, η εφαρμογή αυτών των μοντέλων δεν είναι χωρίς προκλήσεις. Η πολυπλοκότητα και η ποικιλία των χαρακτηριστικών των μανιταριών μπορεί να οδηγήσει σε υψηλή μεταβλητότητα στις μετρήσεις απόδοσης του μοντέλου. Επίσης, η πιθανότητα μεροληψίας λόγω της επιλογής δεδομένων είναι ένα άλλο σημαντικό θέμα που πρέπει να αντιμετωπιστεί κατά την ανάπτυξη και εφαρμογή αυτών των μοντέλων σε κλινικά περιβάλλοντα ή στο πεδίο.

Συνολικά, είναι εμφανές ότι τα μοντέλα μηχανικής μάθησης παρέχουν αξιόπιστες λύσεις για την ταξινόμηση των μανιταριών, βελτιώνοντας την ικανότητα αναγνώρισης βρώσιμων και δηλητηριωδών ειδών. Παρόλα αυτά, είναι σημαντικό να συνεχιστεί η έρευνα και η επικύρωση αυτών των μοντέλων σε μεγαλύτερης κλίμακας μελέτες για να διασφαλιστεί η αξιοπιστία και η ακρίβειά τους. Η παρούσα εργασία στοχεύει να αξιολογήσει διάφορα μοντέλα μηχανικής μάθησης και να προσδιορίσει τις παραμέτρους που επηρεάζουν την απόδοσή τους στην ταξινόμηση μανιταριών.

Μέθοδοι Ταξινόμησης

Για την υλοποίηση της λύσης του προβλήματος ταξινόμησης, εφαρμόστηκαν οι ακόλουθες μέθοδοι μηχανικής μάθησης:

- **AdaBoost**
- **K-Nearest Neighbors (KNN)**
- **Naive Bayes**
- **MPL Classifier**
- **Support Vector Machine (SVM)**
- **Decision Tree Classifier**
- **Logistic Regression**
- **Random Forest Classifier**
- **Gradient Boosting Classifier**

Υλοποίηση και Αξιολόγηση Μοντέλων

Πέραν των μεθόδων ταξινόμησης, εφαρμόστηκαν και οι παρακάτω διαδικασίες για την υλοποίηση και την αξιολόγηση των μοντέλων:

- **Ανάγνωση Δεδομένων από CSV**: Χρήση της βιβλιοθήκης pandas για την ανάγνωση δεδομένων από αρχείο CSV που περιέχει τα χαρακτηριστικά των μανιταριών.
- **Οπτικοποίηση Δεδομένων**: Χρήση των βιβλιοθηκών matplotlib και seaborn για τη δημιουργία διαγραμμάτων και γραφημάτων που απεικονίζουν τις διαφορετικές πτυχές των δεδομένων, όπως τη συχνότητα των βρώσιμων και δηλητηριωδών μανιταριών.
- **Εκπαίδευση Μοντέλων και Αξιολόγηση Επιδόσεων**: Χρήση διαφόρων μετρικών όπως ακρίβεια, recall, F1 score, και ROC-AUC για την αξιολόγηση των μοντέλων ταξινόμησης.

1.2 Σκοπός - Στόχοι

Η κατανάλωση δηλητηριωδών μανιταριών αποτελεί σοβαρό κίνδυνο για την ανθρώπινη υγεία και μπορεί να οδηγήσει σε θάνατο ή σοβαρές επιπλοκές. Στις μέρες μας, η ανάγκη για ακριβή και αποτελεσματική ταξινόμηση των μανιταριών σε βρώσιμα και δηλητηριώδη είναι πιο επιτακτική από ποτέ. Σκοπός αυτής της εργασίας είναι η ανάπτυξη ενός μοντέλου μηχανικής μάθησης που θα μπορεί να προσδιορίζει με ακρίβεια την τοξικότητα των μανιταριών, παρέχοντας ένα πολύτιμο εργαλείο για την προστασία της δημόσιας υγείας. Ο στόχος είναι να δημιουργηθεί ένα μοντέλο που θα εκπαιδευτεί σε ένα σύνολο δεδομένων από το Kaggle και θα χρησιμοποιεί εξελιγμένες τεχνικές μηχανικής μάθησης για την ταξινόμηση των μανιταριών. Σε σύγκριση με τις παραδοσιακές μεθόδους ταξινόμησης, το προτεινόμενο μοντέλο αναμένεται να προσφέρει μεγαλύτερη ακρίβεια και αξιοπιστία, βοηθώντας στην πρόληψη δηλητηριάσεων από μανιτάρια.

1.3 Δομή εργασίας

Κεφάλαιο 2: Περιγραφή και Ανάλυση Dataset

Στο κεφάλαιο αυτό, θα παρουσιαστεί η εισαγωγή στην επιλογή του dataset που χρησιμοποιείται για την ανάλυση. Θα καλυφθούν:

- **Γενικές Πληροφορίες της Προέλευσης**: Θα αναλυθεί η πηγή του dataset, δηλαδή από πού προήλθε και ποιος είναι ο προμηθευτής των δεδομένων.
- **Χαρακτηριστικά του Dataset**: Θα περιγραφούν τα χαρακτηριστικά του dataset, όπως οι μεταβλητές που περιλαμβάνονται, ο τύπος των δεδομένων και οι κατηγορίες.
- **Είδος Δεδομένων**: Θα εξηγηθεί το είδος των δεδομένων που χρησιμοποιούνται, π.χ., αριθμητικά δεδομένα, κατηγορηματικά δεδομένα κ.λπ.

Κεφάλαιο 3: Προεπεξεργασία Δεδομένων

Σε αυτό το κεφάλαιο, θα εξεταστούν οι διαδικασίες προ επεξεργασίας των δεδομένων, οι οποίες περιλαμβάνουν:

- **Καθαρισμός Δεδομένων**: Διαδικασίες για την αφαίρεση ή τη διόρθωση ελλιπών, λανθασμένων ή ανώμαλων δεδομένων.
- **Επεξεργασία Τιμών που Λείπουν**: Μέθοδοι για την αντιμετώπιση των κενών τιμών, όπως η αντικατάσταση ή η αφαίρεση ελλιπών δεδομένων.
- **Κωδικοποίηση Τιμών**: Μεθόδους κωδικοποίησης κατηγορηματικών δεδομένων (π.χ., one-hot encoding) για να καταστούν χρήσιμα για την ανάλυση και το μοντελισμό.

Κεφάλαιο 4: Οπτικοποίηση/Ανάλυση Δεδομένων

Το κεφάλαιο αυτό θα εστιάσει στην οπτικοποίηση των δεδομένων, προσφέροντας:

- **Οπτικοποιήσεις Δεδομένων**: Παρουσίαση γραφημάτων, διαγραμμάτων και άλλων οπτικών εργαλείων που βοηθούν στην κατανόηση των δεδομένων.
- **Σημαντικά Ευρήματα και Εμπόδια**: Ανάλυση των ευρημάτων που προκύπτουν από τις οπτικοποιήσεις και αναφορά σε τυχόν εμπόδια που αντιμετωπίστηκαν κατά την υλοποίηση του έργου.

Κεφάλαιο 5: Μοντελοποίηση και Αξιολόγηση Μοντέλων

Σε αυτό το κεφάλαιο θα αναλυθούν τα μοντέλα που χρησιμοποιήθηκαν:

- **Χρήση Classifiers**: Περιγραφή των classifiers που εφαρμόστηκαν για την πρόβλεψη, όπως logistic regression, decision trees, κ.ά.
- **Προσέγγιση Πρόβλεψης**: Στρατηγικές και μέθοδοι για την πρόβλεψη του εγκεφαλικού επεισοδίου.

- **Κριτήρια και Μοντέλα**: Εξέταση των κριτηρίων που χρησιμοποιήθηκαν για την επιλογή και αξιολόγηση των μοντέλων.

Κεφάλαιο 6: Αποτελέσματα

Στο κεφάλαιο αυτό θα παρουσιαστούν:

- **Εμφάνιση Αποτελεσμάτων**: Παρουσίαση των αποτελεσμάτων από τα μοντέλα, περιλαμβάνοντας τα ποσοστά επιτυχίας και τις μετρήσεις απόδοσης.
- **Αξιολόγηση Απόδοσης Μοντέλων**: Ανάλυση της απόδοσης των μοντέλων στο επιλεγμένο dataset και αναφορά στις τιμές που επηρεάζουν την ποιότητα των αποτελεσμάτων.

Κεφάλαιο 7: Σύνοψη και Συμπεράσματα

Το τελικό κεφάλαιο περιλαμβάνει:

- **Σύνοψη**: Σύντομη αναφορά στα κύρια ευρήματα της εργασίας και την επίδρασή τους στην πρόβλεψη του εγκεφαλικού επεισοδίου.
- **Συμπεράσματα**: Εξαγωγή συμπερασμάτων σχετικά με την καταλληλότητα των μοντέλων για την πρόβλεψη του εγκεφαλικού επεισοδίου και προτάσεις για μελλοντική έρευνα.

2 Περιγραφή ανάλυση Dataset

Στην παρούσα εργασία χρησιμοποιήθηκε ένα σύνολο δεδομένων από το Kaggle, το οποίο περιέχει πληροφορίες για διάφορα είδη μανιταριών. Το dataset περιλαμβάνει χαρακτηριστικά που βοηθούν στην αναγνώριση και ταξινόμηση των μανιταριών σε βρώσιμα και δηλητηριώδη. Η ανάλυση των δεδομένων αυτών είναι κρίσιμη για την ανάπτυξη ενός αποτελεσματικού μοντέλου μηχανικής μάθησης.

2.1 Πηγή

Το dataset που χρησιμοποιήθηκε για αυτή την έρευνα προέρχεται από την πλατφόρμα Kaggle, η οποία αποτελεί έναν αξιόπιστο πόρο για datasets που σχετίζονται με τη μηχανική μάθηση και την επιστήμη δεδομένων. Το συγκεκριμένο σύνολο δεδομένων περιέχει εγγραφές από διάφορα είδη μανιταριών, με περιγραφές των φυσικών και χημικών τους χαρακτηριστικών. Είναι ένα ευρέως χρησιμοποιούμενο dataset για προβλήματα ταξινόμησης, προσφέροντας μια εξαιρετική βάση για την ανάπτυξη και δοκιμή αλγορίθμων μηχανικής μάθησης.

2.2 Χαρακτηριστικά

Το dataset περιέχει πολλά χαρακτηριστικά, τα οποία περιγράφουν τα διάφορα φυσικά και χημικά γνωρίσματα των μανιταριών. Παρακάτω παρουσιάζονται τα χαρακτηριστικά με τις δυνατές τιμές τους:

- **Classes:** Κατηγορία του μανιταριού: βρώσιμο (e) ή δηλητηριώδες (p).
- **Cap-shape:** Σχήμα καπέλου: bell (b), conical (c), convex (x), flat (f), knobbed (k), sunken (s).
- **Cap-surface:** Επιφάνεια καπέλου: fibrous (f), grooves (g), scaly (y), smooth (s).
- **Cap-color:** Χρώμα καπέλου: brown (n), buff (b), cinnamon (c), gray (g), green (r), pink (p), purple (u), red (e), white (w), yellow (y).
- **Bruises:** Μώλωπες: bruises (t), no (f).
- **Odor:** Μυρωδιά: almond (a), anise (l), creosote (c), fishy (y), foul (f), musty (m), none (n), pungent (p), spicy (s).
- **Gill-attachment:** Προσκόλληση ελάσματος: attached (a), descending (d), free (f), notched (n).
- **Gill-spacing:** Απόσταση ελασμάτων: close (c), crowded (w), distant (d).
- **Gill-size:** Μέγεθος ελάσματος: broad (b), narrow (n).
- **Gill-color:** Χρώμα ελάσματος: black (k), brown (n), buff (b), chocolate (h), gray (g), green (r), orange (o), pink (p), purple (u), red (e), white (w), yellow (y).
- **Stalk-shape:** Σχήμα ποδιού: enlarging (e), tapering (t).
- **Stalk-root:** Ρίζα ποδιού: bulbous (b), club (c), cup (u), equal (e), rhizomorphs (z), rooted (r), missing (?).
- **Stalk-surface-above-ring:** Επιφάνεια ποδιού πάνω από το δαχτυλίδι: fibrous (f), scaly (y), silky (k), smooth (s).

- **Stalk-surface-below-ring:** Επιφάνεια ποδιού κάτω από το δαχτυλίδι: fibrous (f), scaly (y), silky (k), smooth (s).
- **Stalk-color-above-ring:** Χρώμα ποδιού πάνω από το δαχτυλίδι: brown (n), buff (b), cinnamon (c), gray (g), orange (o), pink (p), red (e), white (w), yellow (y).
- **Stalk-color-below-ring:** Χρώμα ποδιού κάτω από το δαχτυλίδι: brown (n), buff (b), cinnamon (c), gray (g), orange (o), pink (p), red (e), white (w), yellow (y).
- **Veil-type:** Τύπος πέπλου: partial (p), universal (u).
- **Veil-color:** Χρώμα πέπλου: brown (n), orange (o), white (w), yellow (y).
- **Ring-number:** Αριθμός δαχτυλιδιών: none (n), one (o), two (t).
- **Ring-type:** Τύπος δαχτυλιδιού: cobwebby (c), evanescent (e), flaring (f), large (l), none (n), pendant (p), sheathing (s), zone (z).
- **Spore-print-color:** Χρώμα σπόρων: black (k), brown (n), buff (b), chocolate (h), green (r), orange (o), purple (u), white (w), yellow (y).
- **Population:** Πληθυσμός: abundant (a), clustered (c), numerous (n), scattered (s), several (v), solitary (y).
- **Habitat:** Περιβάλλον: grasses (g), leaves (l), meadows (m), paths (p), urban (u), waste (w), woods (d).

Αυτά τα χαρακτηριστικά είναι ονομαστικά (categorical) και κάθε ένα από αυτά μπορεί να συμβάλλει στην ακριβή ταξινόμηση των μανιταριών σε βρώσιμα ή δηλητηριώδη.

2.3 Μεταβλητή Στόχος

Η μεταβλητή στόχος στο συγκεκριμένο dataset είναι η κατηγορία στην οποία ανήκει το μανιτάρι: βρώσιμο ή δηλητηριώδες. Αυτή η μεταβλητή είναι δυαδική (binary), με τις τιμές "edible" (βρώσιμο) και "poisonous" (δηλητηριώδες). Ο στόχος της ανάλυσης είναι να αναπτυχθεί ένα μοντέλο που θα μπορεί να προβλέψει με ακρίβεια αυτή τη μεταβλητή, με βάση τα υπόλοιπα χαρακτηριστικά του μανιταριού.

2.4 Αρχικές Παρατηρήσεις

Από την αρχική ανάλυση του dataset, παρατηρήθηκε ότι υπάρχει μια ισορροπημένη κατανομή ανάμεσα στις κατηγορίες βρώσιμων και δηλητηριωδών μανιταριών. Αυτό είναι θετικό για την εκπαίδευση των μοντέλων μηχανικής μάθησης, καθώς αποφεύγεται η μεροληψία προς μία από τις δύο κατηγορίες. Επιπλέον, ορισμένα χαρακτηριστικά, όπως η μυρωδιά και το χρώμα του καπέλου, φαίνεται να έχουν μεγαλύτερη επιρροή στην ταξινόμηση των μανιταριών, κάτι που θα διερευνηθεί περαιτέρω κατά την ανάπτυξη των μοντέλων. Ωστόσο, η ύπαρξη πολλών ονομαστικών χαρακτηριστικών σημαίνει ότι η σωστή προεπεξεργασία των δεδομένων είναι απαραίτητη για την αποτελεσματική λειτουργία των αλγορίθμων μηχανικής μάθησης.

3 Προεπεξεργασία Δεδομένων

Η προεπεξεργασία δεδομένων είναι ένα κρίσιμο στάδιο πριν την εφαρμογή αλγορίθμων μηχανικής μάθησης, καθώς διασφαλίζει ότι τα δεδομένα είναι καθαρά και σε κατάλληλη μορφή για ανάλυση. Σε αυτή την ενότητα, θα περιγράψουμε τον καθαρισμό των δεδομένων και την κωδικοποίηση των κατηγορηματικών μεταβλητών που έγιναν στο dataset.

3.1 Καθαρισμός (cleaning)

Ο καθαρισμός των δεδομένων περιλαμβάνει την ανίχνευση και διαχείριση των απουσών ή μη έγκυρων τιμών, καθώς και άλλες διαδικασίες για τη διασφάλιση της ποιότητας των δεδομένων. Αρχικά γίνεται ανίχνευση των απουσών τιμών χρησιμοποιώντας τη μέθοδο **isna().sum()** για να καταμετρηθούν οι απόντες τιμές σε κάθε στήλη. Στη συνέχεια, εφαρμόζεται η μέθοδος **replace('?', pd.NA, inplace=True)** για να αντικατασταθούν τυχόν μη έγκυρες τιμές (π.χ. το **'?'** αντιπροσωπεύει απούσες τιμές) με την τυπική ένδειξη απουσίας (NA) της βιβλιοθήκης Pandas. Οι απουσίες που απομένουν αφαιρούνται με τη χρήση της **dropna()**, εξασφαλίζοντας ότι το σύνολο δεδομένων δεν περιέχει κενά που θα μπορούσαν να επηρεάσουν αρνητικά τη διαδικασία μοντελοποίησης. Μετά τον καθαρισμό, γίνεται επαλήθευση του συνόλου των απουσών τιμών χρησιμοποιώντας και πάλι τη μέθοδο **isna().sum().sum()**. Αυτό επιβεβαιώνει ότι δεν παραμένουν απόντες τιμές στο σύνολο δεδομένων.

3.2 Κωδικοποίηση Μεταβλητών (Encoding Categorical Variables)

Η κωδικοποίηση των κατηγορηματικών μεταβλητών είναι απαραίτητη για τη μετατροπή των δεδομένων σε μορφή που μπορεί να χρησιμοποιηθεί από τους αλγορίθμους μηχανικής μάθησης, οι οποίοι συνήθως απαιτούν αριθμητικά δεδομένα. Αρχικά έγινε ανάλυση των μεταβλητών. Παρατηρήθηκε ότι όλες οι στήλες του dataset είναι αλφαριθμητικά. Αυτό οδηγεί σε μετατροπή αυτών τον κελιών σε αριθμητικές τιμές. Για την επίτευξη αυτού, χρησιμοποιείται η τεχνική **LabelEncoder** της βιβλιοθήκης **sklearn**, η οποία μετατρέπει τις κατηγορίες σε αριθμητικές τιμές. Για κάθε κατηγορηματική μεταβλητή στο dataset, οι μοναδικές κατηγορίες της κωδικοποιούνται σε ακέραιους αριθμούς.

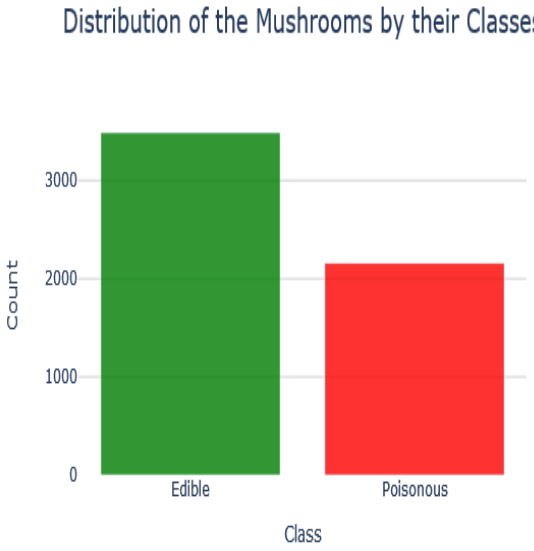
4 Οπτικοποίηση/Ανάλυση Δεδομένων

Η ενότητα της οπτικοποίησης και ανάλυσης δεδομένων είναι πολύ σημαντική για την κατανόηση του dataset. Η οπτικοποίηση μας βοηθά να δούμε ξεκάθαρα πώς τα δεδομένα των μανιταριών συνδέονται μεταξύ τους. Μέσα από γραφήματα και εικόνες, μπορούμε να ανακαλύψουμε πρότυπα και σχέσεις που μπορεί να μην είναι εύκολα αντιληπτά με μια απλή ματιά στα αριθμητικά στοιχεία. Η ανάλυση των δεδομένων περιλαμβάνει τη διερεύνηση των χαρακτηριστικών των μανιταριών και πώς αυτά συνδέονται με το αν είναι βρώσιμα ή δηλητηριώδη. Αυτό μας επιτρέπει να κατανοήσουμε καλύτερα το dataset και να εντοπίσουμε τα πιο σημαντικά στοιχεία που επηρεάζουν την ταξινόμηση των μανιταριών. Η κατανόηση αυτή είναι σημαντική για να προχωρήσουμε στη δημιουργία ακριβών μοντέλων πρόβλεψης.

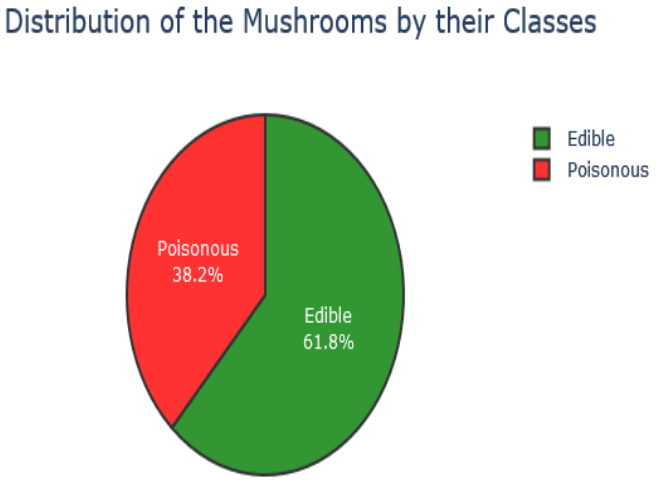
4.1 Οπτικοποιήσεις (visualizations)

Το παρακάτω γράφημα (**εικόνα 1**) παρουσιάζει την κατανομή των μανιταριών στο dataset ανάμεσα στις δύο κύριες κατηγορίες: βρώσιμα (**Edible**) και δηλητηριώδη (**Poisonous**). Στο σύνολο των δεδομένων, τα βρώσιμα μανιτάρια αντιστοιχούν σε 3,488 παρατηρήσεις, ενώ τα δηλητηριώδη σε 2,156. Η κατανομή αυτή μας δείχνει ότι, αν και υπάρχει μεγαλύτερος αριθμός βρώσιμων μανιταριών, οι δύο κατηγορίες είναι σχετικά κοντά σε πλήθος, κάτι που είναι σημαντικό για την ανάλυση και την ταξινόμηση των μανιταριών στο επόμενο στάδιο.

Στην συνέχεια, οπως φαίνεται στο παρακάτω γράφημα (**εικόνα 2**), το κόκκινο τμήμα αντιπροσωπεύει το ποσοστό των δηλητηριωδών μανιταριών στο σύνολο δεδομένων (48.2%). Το πράσινο τμήμα αντιπροσωπεύει το ποσοστό των βρώσιμων μανιταριών (51.8%). Τα ποσοστά αυτών των κατηγοριών είναι αρκετά κοντά το ένα στο άλλο. Από εδώ και στο εξής, θα χρησιμοποιώ το κόκκινο χρώμα για την κατηγορία των δηλητηριωδών και το πράσινο για την κατηγορία των βρώσιμων μανιταριών.



Εικόνα 1: Γράφημα ράβδων, απεικόνιση αριθμού των περιπτώσεων για τις κατηγορίες βρώσιμων μανιταριών έναντι δηλητηριωδών.



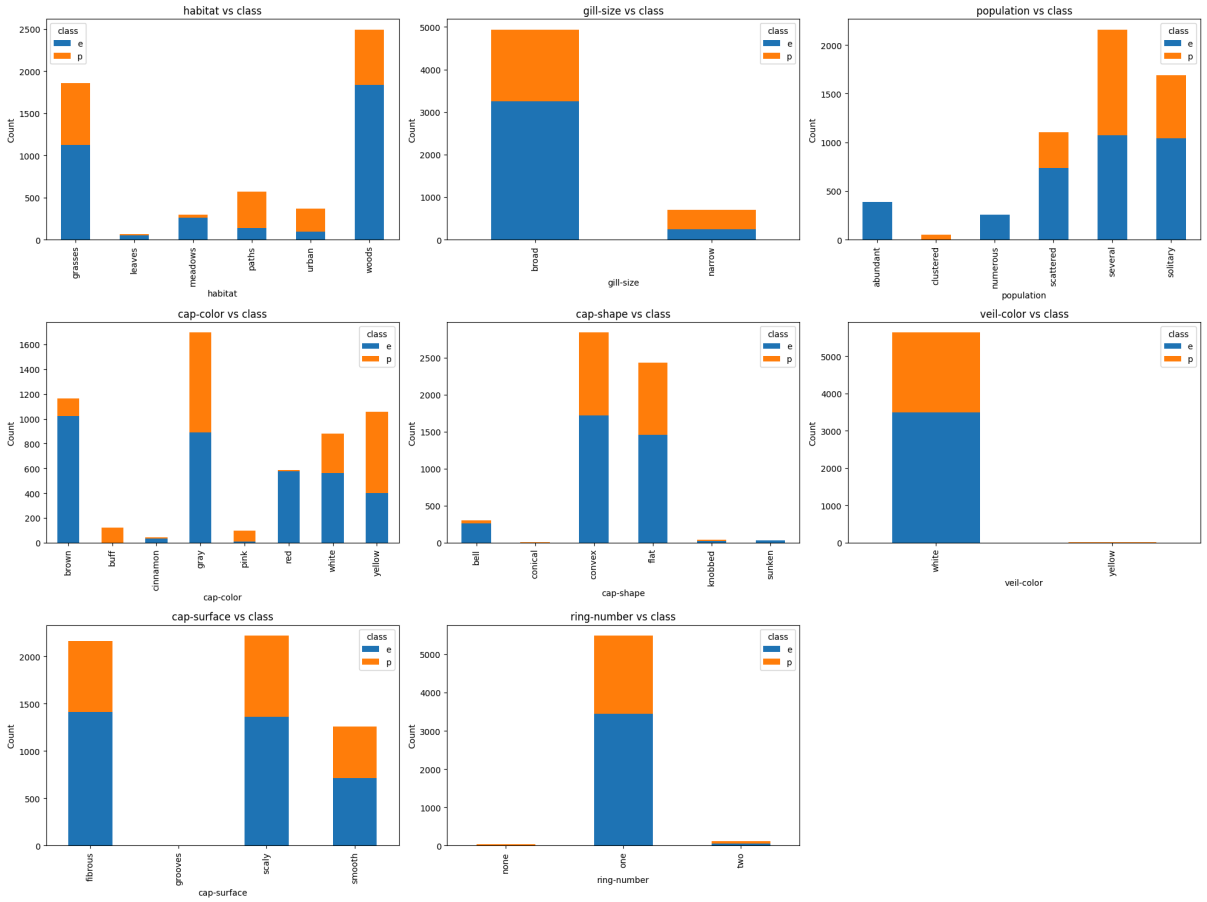
Εικόνα 2: Κυκλικό διάγραμμα, απεικόνιση της κατανομής των μανιταριών στις κατηγορίες βρώσιμων και δηλητηριωδών.

Η ανάλυση συσχέτισης αποτελεί ένα σημαντικό βήμα για την κατανόηση των σχέσεων μεταξύ διαφορετικών χαρακτηριστικών σε ένα dataset. Στην περίπτωση μας, ο στόχος είναι να κατανοήσουμε πώς διάφορα χαρακτηριστικά σχετίζονται μεταξύ τους, χρησιμοποιώντας τον πίνακα συσχέτισης που απεικονίζεται στο heatmap (**εικόνα 3**). Οι τιμές του πίνακα συσχέτισης κυμαίνονται από -1 έως 1, όπου -1 υποδηλώνει μια τέλεια αρνητική γραμμική σχέση, 0 σημαίνει καμία γραμμική σχέση και 1 δείχνει μια τέλεια θετική γραμμική σχέση.

Αναλύοντας τον πίνακα συσχέτισης, παρατηρούμε ότι ορισμένα χαρακτηριστικά παρουσιάζουν ισχυρές σχέσεις μεταξύ τους, κάτι που μπορεί να οδηγήσει σε σημαντικές παρατηρήσεις για την κατηγοριοποίηση των μανιταριών. Ιδιαίτερα, το cap-color και το habitat φαίνεται να έχουν την πιο σημαντική συσχέτιση, γεγονός που υποδηλώνει ότι αυτά τα χαρακτηριστικά μπορεί να είναι ιδιαίτερα χρήσιμα για την κατηγοριοποίηση. Επιπλέον, οι αρνητικές σχέσεις, όπως εκείνες μεταξύ του cap-shape και του ring-number, προσφέρουν σημαντικές πληροφορίες για την κατανόηση της σχέσης των χαρακτηριστικών και ενδέχεται να επηρεάσουν την τελική κατηγοριοποίηση των μανιταριών. Η ανάλυση των συσχετίσεων θα σας βοηθήσει να επιλέξετε τα πιο σημαντικά χαρακτηριστικά για τα μοντέλα σας και να εστιάσετε σε εκείνα που έχουν μεγαλύτερη επίδραση στην πρόβλεψη του στόχου.

Η παρακάτω εικόνα παρουσιάζει μια σειρά από στοιβαγμένα ραβδογράμματα που απεικονίζουν τη συσχέτιση μεταξύ οκτώ χαρακτηριστικών των μανιταριών και της κλάσης τους (βρώσιμα ή δηλητηριώδη). Κάθε χαρακτηριστικό, όπως το περιβάλλον ανάπτυξης, το μέγεθος των πτερυγίων και το χρώμα του καπέλου, συγκρίνεται με την πιθανότητα το μανιτάρι να είναι βρώσιμο ή δηλητηριώδες. Οι οπτικοποιήσεις αυτές επιτρέπουν την εύκολη ανάλυση των χαρακτηριστικών που μπορούν να λειτουργήσουν ως δείκτες για την

ταξινόμηση των μανιταριών, προσφέροντας πολύτιμα δεδομένα για την κατανόηση της δομής και της φύσης τους.



Εικόνα 4: Ραβδογράμματα συσχέτισης μεταξύ χαρακτηριστών και κλάσης

5 Μοντελοποίηση και αξιολόγηση Μοντέλων

Στην ενότητα αυτή, θα παρουσιαστεί η διαδικασία μοντελοποίησης και αξιολόγησης για την ταξινόμηση των μανιταριών ως βρώσιμα ή δηλητηριώδη.

5.1 Επιλογή Μοντέλων

Στην υποενότητα αυτή, θα παρουσιαστεί η επιλογή των μοντέλων που χρησιμοποιήθηκαν για την ταξινόμηση των μανιταριών. Για την επίτευξη του στόχου, επιλέχθηκαν εννέα διαφορετικοί αλγόριθμοι μηχανικής μάθησης, οι οποίοι καλύπτουν ένα ευρύ φάσμα τεχνικών και μεθόδων. Κάθε ένα από αυτά τα μοντέλα έχει τα δικά του πλεονεκτήματα και μπορεί να προσαρμόζεται σε διαφορετικές περιπτώσεις ανάλογα με τα χαρακτηριστικά του δεδομένου προβλήματος. Τα μοντέλα που χρησιμοποιήθηκαν είναι τα εξής:

- **Decision Tree**: Ένα μοντέλο βασισμένο σε δέντρα απόφασης που διασπά τα δεδομένα σε διακριτές περιοχές για να πάρει αποφάσεις, βασισμένο στη μεγιστοποίηση της πληροφορίας σε κάθε διάσπαση.
- **Random Forest**: Μια επέκταση των δέντρων απόφασης που χρησιμοποιεί πολλαπλά δέντρα (δασικό μοντέλο) για να βελτιώσει την ακρίβεια και να μειώσει την υπερπροσαρμογή.
- **Gradient Boosting**: Μια τεχνική ενίσχυσης που χτίζει διαδοχικά ισχυρότερα μοντέλα από συνδυασμούς ασθενέστερων μοντέλων, βελτιώνοντας σταδιακά την απόδοση.
- **Support Vector Machine (SVM)**: Ένας αλγόριθμος που προσπαθεί να βρει το καλύτερο διαχωριστικό υπερεπίπεδο που μεγιστοποιεί το περιθώριο μεταξύ των κατηγοριών.
- **K-Nearest Neighbors (KNN)**: Ένας μη παραμετρικός αλγόριθμος που ταξινομεί τα δεδομένα βάσει της εγγύτητάς τους σε έναν αριθμό γειτόνων, προσαρμόζοντας τις προβλέψεις με βάση την πλειοψηφία των γειτονικών κλάσεων.
- **Logistic Regression**: Ένα μοντέλο παλινδρόμησης που χρησιμοποιείται για δυαδική ταξινόμηση, εκτιμώντας τις πιθανότητες και μετατρέποντάς τες σε κατηγορίες μέσω μιας λογιστικής συνάρτησης.
- **Naive Bayes**: Ένας αλγόριθμος που βασίζεται στη θεωρία των πιθανοτήτων και υποθέτει ανεξαρτησία μεταξύ των χαρακτηριστικών για να ταξινομεί τα δεδομένα.
- **Multi-layer Perceptron (MLP)**: Ένα είδος τεχνητού νευρωνικού δικτύου που αποτελείται από πολλαπλά στρώματα (πύλες) νευρώνων, εκπαιδευόμενο για να μοντελοποιεί πολύπλοκες συναρτήσεις και μοτίβα στα δεδομένα.
- **AdaBoost**: Ένας αλγόριθμος ενίσχυσης που συνδυάζει πολλά ασθενή μαθησιακά μοντέλα για να δημιουργήσει ένα ισχυρό μοντέλο, δίνοντας μεγαλύτερη έμφαση σε περιπτώσεις που ήταν δύσκολο να ταξινομηθούν σωστά.

Αυτά τα μοντέλα επιλέχθηκαν για την ικανότητά τους να αποδίδουν καλά σε ποικίλα δεδομένα, επιτρέποντας μια πλήρη αξιολόγηση της επίδοσής τους στο συγκεκριμένο πρόβλημα ταξινόμησης των μανιταριών.

5.2 Εκπαίδευση (training)

Για την εκπαίδευση των μοντέλων, ακολουθήθηκαν συγκεκριμένα στάδια προεπεξεργασίας των δεδομένων, όπως περιγράφονται στο Κεφάλαιο 3. Εν συντομία, πραγματοποιήθηκε χειρισμός των τιμών που έλειπαν, κωδικοποίηση των κατηγοριών σε αριθμητικές μεταβλητές και κανονικοποίηση των χαρακτηριστικών για να εξασφαλιστεί ομοιογένεια στις κλίμακες των δεδομένων.

Το dataset που χρησιμοποιήθηκε ήταν ισορροπημένο ως προς τις κλάσεις, οπότε δεν απαιτήθηκε η χρήση τεχνικών όπως το SMOTE για την αντιμετώπιση της ανισορροπίας. Ωστόσο, η εφαρμογή του SMOTE θα μπορούσε να διερευνηθεί ως πιθανή επέκταση της έρευνας, ειδικά σε datasets με ανισορροπία κλάσεων. Αυτό θα μπορούσε να επιτρέψει την αξιολόγηση της επίδρασης της τεχνικής SMOTE στην ακρίβεια και στην ικανότητα γενίκευσης των μοντέλων.

Το σύνολο δεδομένων χωρίστηκε σε σύνολα εκπαίδευσης (training set) και δοκιμών (test set), επιτρέποντας την αξιολόγηση της γενίκευσης των μοντέλων. Κάθε μοντέλο εκπαιδεύτηκε στο training set, ενώ χρησιμοποιήθηκαν βελτιστοποιημένες υπερπαράμετροι για τη βελτίωση της απόδοσης. Επιπλέον, εφαρμόστηκε τεχνική cross-validation για να αποφευχθεί το φαινόμενο του overfitting, διασφαλίζοντας έτσι ότι τα μοντέλα θα μπορούν να γενικεύουν σε μη ορατά δεδομένα.

Κατά την εκπαίδευση των μοντέλων, δοκιμάστηκαν και διαφορετικές καταστάσεις για την εξαγωγή των αποτελεσμάτων, επιτρέποντας τη σύγκριση των επιδόσεων με και χωρίς αυτές τις ενέργειες. Αυτές οι καταστάσεις συμπεριλαμβάνουν:

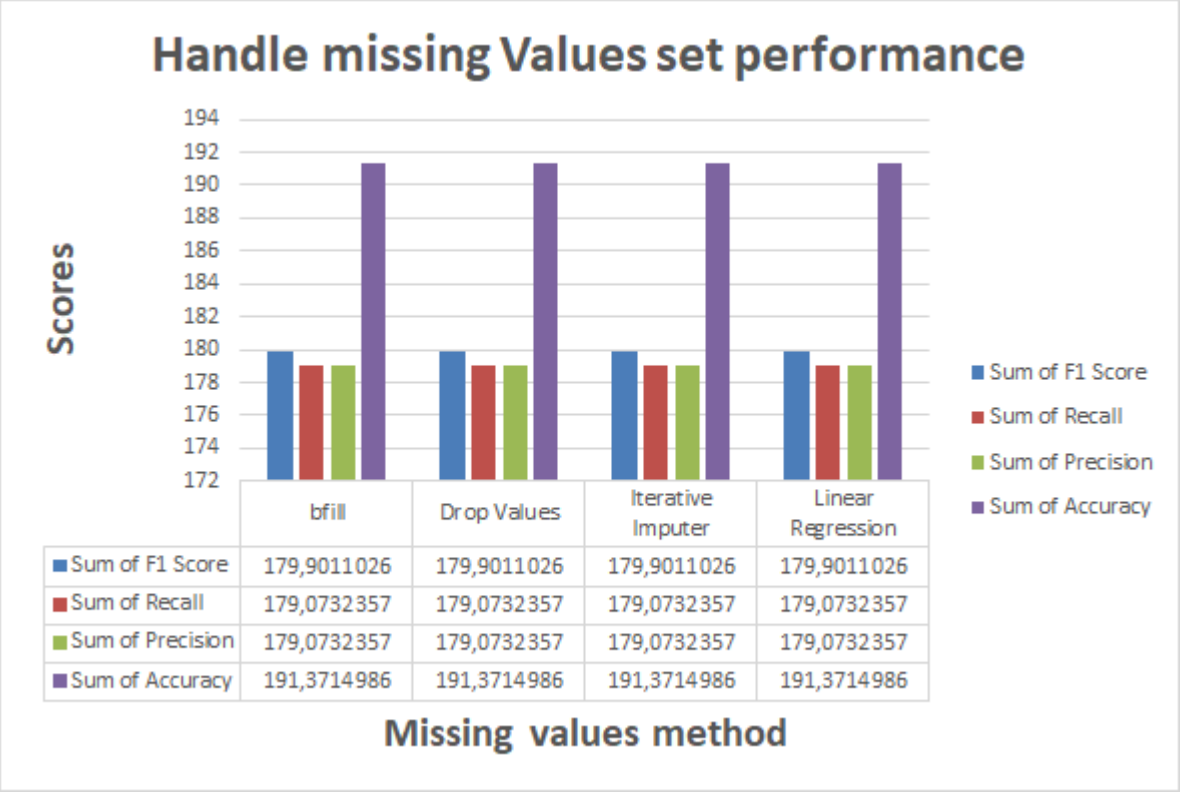
- **4 επαναλήψεις**: Για τον χειρισμό των ελλιπών τιμών με διαφορετικές μεθόδους (Bfill, Iterative Imputer, Linear Regression και Drop Values).
- **10 επαναλήψεις**: Για κάθε πείραμα, για να διασφαλιστεί η αξιοπιστία των αποτελεσμάτων και να εντοπιστούν τυχόν ασυνήθιστες αποκλίσεις.

Παρά το ότι το dataset ήταν ισορροπημένο, η σύγκριση των μοντέλων σε ένα ανισορροπημένο dataset με την εφαρμογή SMOTE θα μπορούσε να αποτελέσει μια χρήσιμη κατεύθυνση για μελλοντική έρευνα. Τα αποτελέσματα από αυτές τις δοκιμές θα παρουσιαστούν σε επόμενο κεφάλαιο, όπου θα αναλυθούν οι μετρήσεις και θα συζητηθούν τα συμπεράσματα από την εκπαίδευση των μοντέλων.

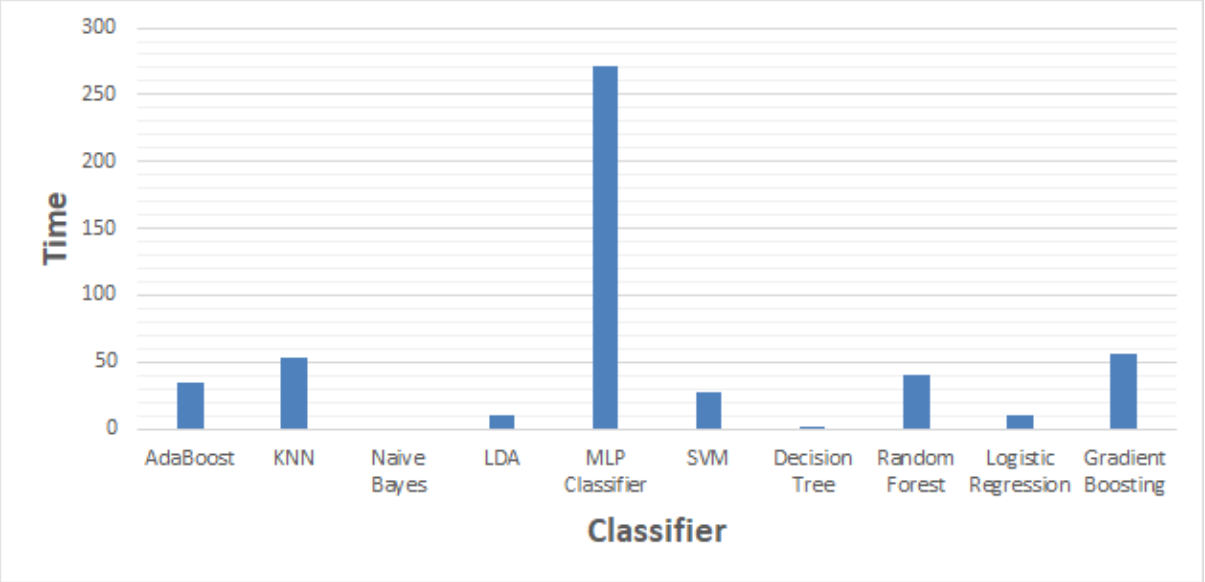
5.3 Εμφάνιση Μετρήσεων

Τα αποτελέσματα μετά την εκπαίδευση των μοντέλων αποθηκεύτηκαν σε ένα κεντρικό Excel με συνολικά 800 εγγραφές. Παρακάτω ακολουθούν τα αποτελέσματα σε μορφή γραφημάτων.

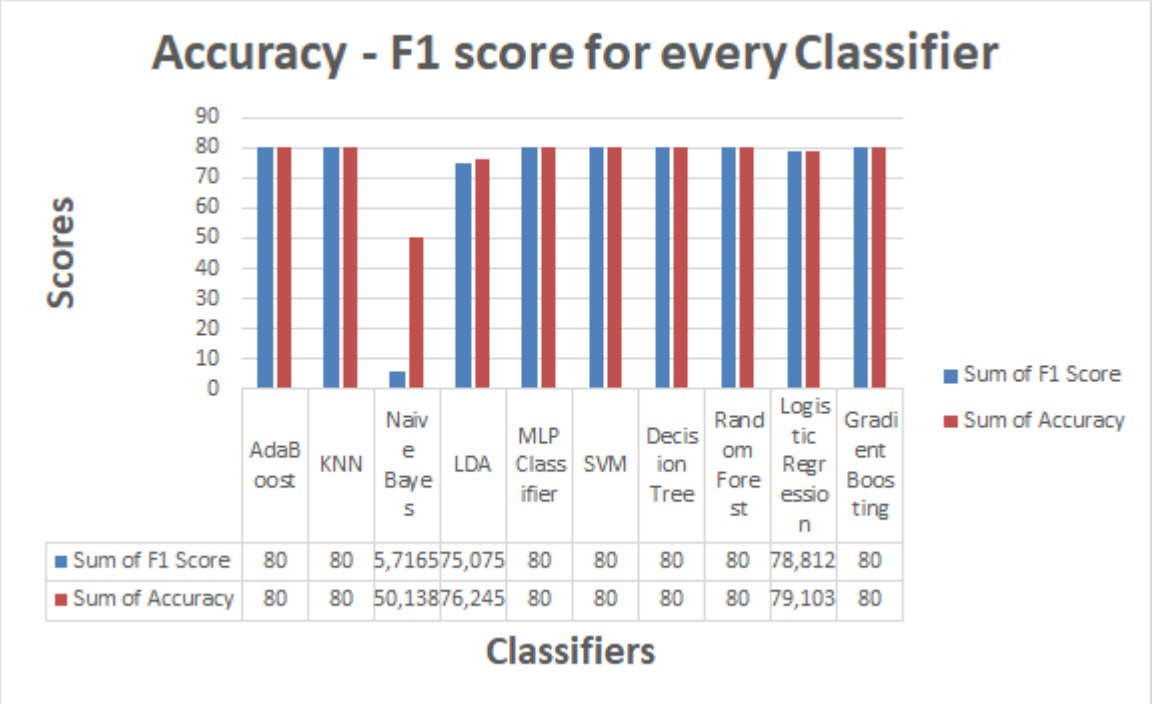
Εικόνα 5: Διαφορά μεταξύ test – train set σε Recall, Accuracy, Precision, F1 Score



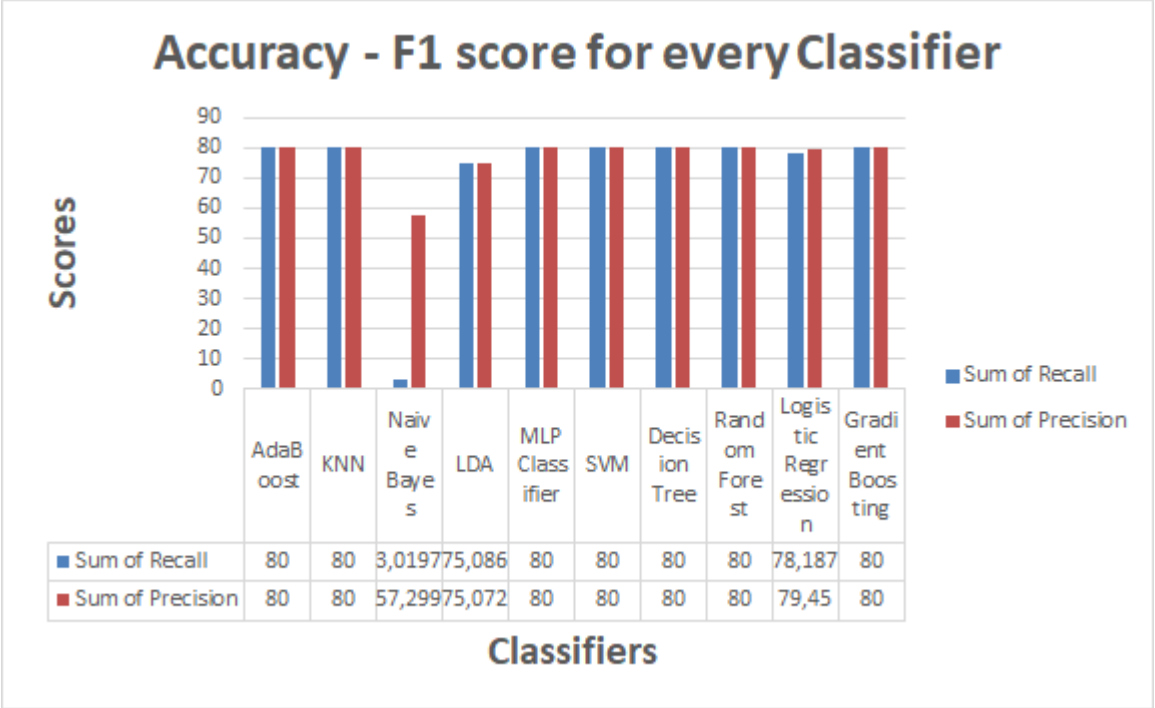
Εικόνα 6: Διαφορά μεταξύ μεθόδων διαχείρισης των τιμών που λείπουν σε Recall, Accuracy, Precision, F1 Score



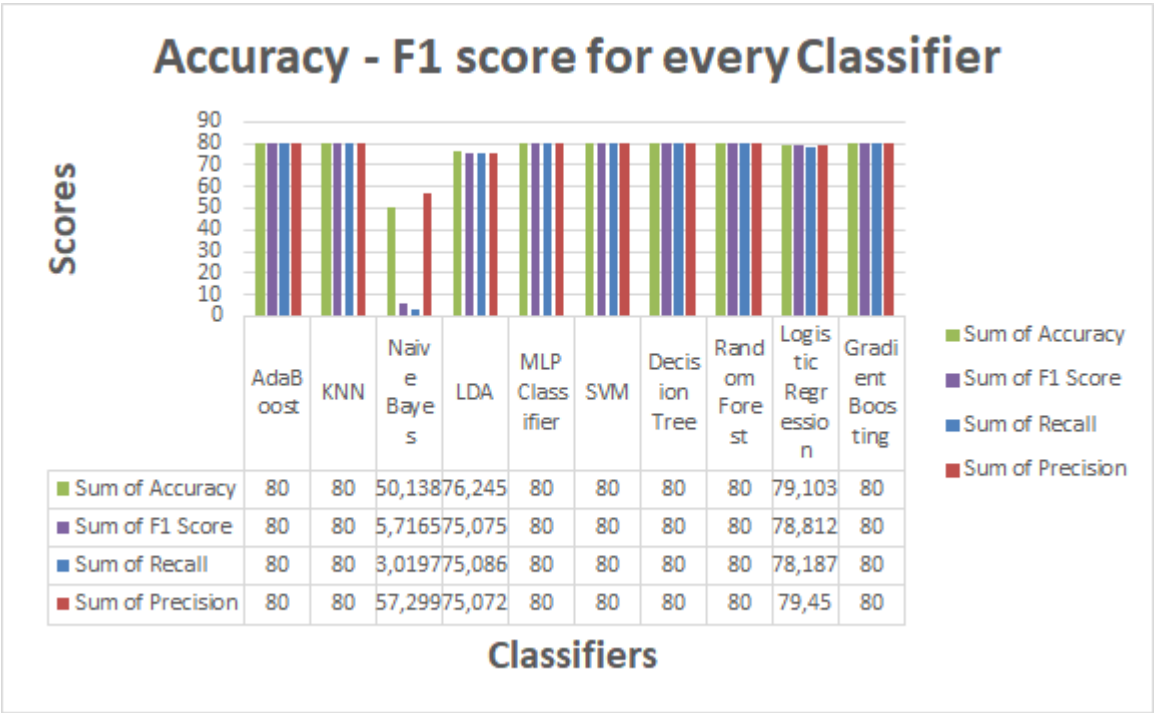
Εικόνα 7: Διαφορά χρόνων μεταξύ των μοντέλων ταξινόμησης



Εικόνα 8: Accuracy – F1 score για κάθε μέθοδο ταξινόμησης



Εικόνα 8: Precision – Recall για κάθε μέθοδο ταξινόμησης



Εικόνα 16: Accuracy - F1-Precision-Recall για κάθε μέθοδο ταξινόμησης

6 Αποτελέσματα

Στο κεφάλαιο αυτό θα γίνει ανασκόπηση τον αποτελεσμάτων, καθώς και ανάλυση και αξιολόγηση των μοντέλων. Θα αναλυθεί η σημασία των μετρήσεων και η σημαντικότητα των τιμών και των γραφημάτων που αναφέρθηκαν στο κεφάλαιο 5

6.1 Αξιολόγηση Μοντέλων

Για την αποτελεσματικότερη ανάλυση των μοντέλων έγινε η αξιοποίηση των μετρήσεων F1 Score, Recall, Precision, Accuracy. Οι μετρήσεις αυτές, προσφέρουν μια εικόνα για την αποτελεσματικότητα των μοντέλων ταξινόμησης των μανιταριών σε βρώσιμα και δηλητηριώδη.

Επιπλέον, σε ότι αφορά τα train και test set, τα αποτελέσματα βρίσκονται αρκετά κοντά, κάτι που ήταν αναμενόμενο, με μικρή διαφορά στο Precision (Εικόνα 5).

Στην συνέχεια, όσον αφορά την διαχείριση των ελλιπών τιμών, τα αποτελέσματα είναι τα ίδια για κάθε μέθοδο επομένως δεν υπάρχει κάποια βέλτιστη μέθοδος (Εικόνα 6).

Στην συνέχεια, σημαντικός παράγοντας αξιολόγησης των μοντέλων είναι και ο χρόνος εκτέλεσης των ταξινομητών. Όσο πιο γρήγορος και αξιόπιστος είναι ο ταξινομητής, τόσο πιο ελκυστικός είναι στην επιλογή του για δεδομένα μεγάλου όγκου και γενικότερα (Εικόνα 7) Πιο συγκεκριμένα:

Από τους 10 ταξινομητές που χρησιμοποιήθηκαν ο πιο γρήγορος αποδείχθηκε ότι είναι ο Naive Bayes με χρόνο πολύ κοντά στα 0 δευτερόλεπτα, κάτι που τον καθιστά εξαιρετικά γρήγορο. Έπειτα, στην κατηγορία των γρήγορων ταξινομητών κατατάσσονται οι:

- Απόφαση Δέντρου (Decision Tree): Ο ταξινομητής απόφασης δέντρου είχε εξαιρετικά γρήγορη εκτέλεση, με χρόνο μόλις 0,74 δευτερόλεπτα, ο οποίος εκπαιδεύεται και κάνει προβλέψεις γρήγορα λόγο της απλής δομής του.
- Linear Discriminant Analysis (LDA): Ο LDA χρειάστηκε 10,64 δευτερόλεπτα για να ολοκληρώσει τη διαδικασία του, λόγω της χρήσης γραμμικών υπολογισμών για τον διαχωρισμό των κλάσεων αποτελεί συνήθως έναν γρήγορος ταξινομητής.
- Logistic Regression: Ο Logistic Regression είχε χρόνο εκτέλεσης 10,75 δευτερόλεπτα, που δείχνει καλή απόδοση συγκριτικά με άλλους. Είναι σχετικά αποτελεσματικός αν και ο χρόνος μπορεί να αυξηθεί με μειονέκτημα ωστόσο, την ραγδαία αύξηση του χρόνου εκτέλεσης σε περιπτώσεις κλιμάκωσης των δοθέντων πεδίων

Από την άλλη πλευρά, μερικοί από τους ταξινομητές, παρουσίασαν αρκετά χρονοβόρα εκτέλεση. Ειδικότερα, στην κατηγορία των αργών ταξινομητών κατατάσσονται:

- MLP Classifier: Ο MLP Classifier (Multi-Layer Perceptron) ήταν ο πιο αργός, με χρόνο εκτέλεσης 270,82 δευτερόλεπτα. Η απόδοσή του είναι πολύ χαμηλή όσον αφορά τον χρόνο. Αυτό οφείλεται στο γεγονός ότι τα νευρωνικά δίκτυα απαιτούν αρκετό χρόνο λόγο της πολυπλοκότητας τους και της ανάγκης τους για backpropagation για την εκπαίδευση σύνδεσης πολλαπλών επιπέδων.
- Gradient Boosting: Ο Gradient Boosting χρειάστηκε 56,18 δευτερόλεπτα, παρουσιάζοντας μία από τις πιο αργές εκτελέσεις μεταξύ των ταξινομητών που εξετάστηκαν. Αυτό συμβαίνει επειδή η μέθοδος αυτή εκπαιδεύει διαδοχικά αδύναμους ταξινομητές και προσπαθεί να διορθώσει τα σφάλματα του κάθε προηγούμενου ταξινομητή, διαδικασία που απαιτεί αρκετό χρόνο
- K-Nearest Neighbors (KNN): Ο ταξινομητής KNN παρουσίασε σημαντική καθυστέρηση με χρόνο εκτέλεσης 52,57 δευτερόλεπτα. Η αργή του απόδοση οφείλεται στο γεγονός ότι κατά την ταξινόμηση ενός νέου δείγματος, το KNN υπολογίζει τις αποστάσεις από όλα τα σημεία εκπαίδευσης, γεγονός που γίνεται ιδιαίτερα χρονοβόρο σε μεγάλα σύνολα δεδομένων.

Πίνακας 1: Χρόνοι εκτέλεσης για κάθε classifier

Classifiers	Time
AdaBoost	34,62021255
KNN	52,56779814
Naive Bayes	0
LDA	10,64561415
MLP Classifier	270,8223248
SVM	27,79434824
Decision Tree	0,738112926
Random Forest	39,90751743
Logistic Regression	10,74885559
Gradient Boosting	56,17947292

Στο σημείο αυτό θα γίνει ανάλυση των Accuracy, F1 score, Precision και Recall συγκριτικά για κάθε ταξινομητή μόνο στο test set καθώς αυτό αντιπροσωπεύει πόσο καλά ένα μοντέλο λειτουργεί σε ξένα δεδομένα, ένα μοντέλο που αποδίδει καλά στο σετ δοκιμών είναι αξιόπιστο ώστε να κάνει σωστές προβλέψεις σε πραγματικό περιβάλλον. Μέσο αυτών των τιμών θα γίνει η τελική αξιολόγηση και επιλογή του καλύτερου μοντέλου.

Ακρίβεια (Accuracy): Οι περισσότεροι ταξινομητές, όπως ο AdaBoost, ο KNN, ο MLP Classifier, ο SVM, ο Decision Tree, ο Random Forest και ο Gradient Boosting, πέτυχαν το μέγιστο δυνατό άθροισμα (40). Μόνο οι Naive Bayes και LDA είχαν χαμηλότερες τιμές, με τον Naive Bayes να έχει σημαντικά χαμηλότερη ακρίβεια (25,12) σε σχέση με τους υπόλοιπους.

F1-Score: Η απόδοση των περισσότερων ταξινομητών στις μετρικές F1-Score ήταν επίσης στο μέγιστο (40). Και εδώ, ο Naive Bayes υστερεί σημαντικά με τιμή 2,91, ενώ ο LDA είχε F1-Score 37,47.

Recall (Ευαισθησία): Οι περισσότερες τιμές του Recall είναι και αυτές στο μέγιστο (40). Ο Naive Bayes έχει και πάλι πολύ χαμηλότερη απόδοση με 1,54, ενώ ο LDA είχε σχεδόν τη μέγιστη απόδοση (37,44).

Precision (Ακρίβεια Πρόβλεψης): Και εδώ, η πλειονότητα των ταξινομητών πέτυχαν το μέγιστο άθροισμα (40), εκτός από τον Naive Bayes (29,21) και τον LDA (37,50).

Πίνακας 2: Τιμές για Accuracy, F1 - Score, Recall και Precision για κάθε Classifier

Classifiers	Sum of Accuracy	Sum of F1- Score	Sum of Recall	Sum of Precision
AdaBoost	40	40	40	40
KNN	40	40	40	40
Naive Bayes	25,11603189	2,908621339	1,538214393	29,20620491
LDA	38,07263065	37,46653887	37,44080135	37,49735935
MLP Classifier	40	40	40	40
SVM	40	40	40	40
Decision Tree	40	40	40	40
Random Forest	40	40	40	40
Logistic Regression	39,54650133	39,39787763	39,07501126	39,72756093
Gradient Boosting	40	40	40	40
Total	382,7351639	359,7730378	358,054027	386,4311252

6.2 Σημασία αποτελεσμάτων/Αξιολόγηση Μετρήσεων

Πριν την τελική αξιολόγηση θα γίνει μια σύντομη ανασκόπηση της σημασίας κάθε μετρικής στο συγκεκριμένο πρόβλημα:

Accuracy: Η ακρίβεια (Accuracy) μετρά το ποσοστό των σωστών προβλέψεων (δηλαδή πόσο καλά ένας ταξινομητής μπορεί να διαχωρίσει σωστά τα βρώσιμα από τα δηλητηριώδη μανιτάρια) σε σχέση με το συνολικό αριθμό των προβλέψεων. Στο πρόβλημα της ταξινόμησης μανιταριών, η ακρίβεια είναι σημαντική, αλλά μπορεί να είναι παραπλανητική αν τα δεδομένα δεν είναι ισορροπημένα. Για παράδειγμα, αν τα περισσότερα δείγματα είναι βρώσιμα και ένας ταξινομητής απλά προβλέψει όλα τα μανιτάρια ως βρώσιμα, μπορεί να έχει υψηλή ακρίβεια, αλλά η αποτυχία του να αναγνωρίσει τα δηλητηριώδη μανιτάρια θα μπορούσε να έχει σοβαρές συνέπειες.

F1 Score: Το F1 Score είναι η αρμονική μέση τιμή της ευαισθησίας (Recall) και της ακρίβειας πρόβλεψης (Precision). Στο πλαίσιο της ταξινόμησης μανιταριών, το F1 Score είναι ιδιαίτερα χρήσιμο, ειδικά όταν το dataset είναι ανισόρροπο ή όταν είναι κρίσιμη η αποφυγή τόσο των ψευδώς θετικών (προβλέποντας ένα δηλητηριώδες μανιτάρι ως βρώσιμο) όσο και των ψευδώς αρνητικών αποτελεσμάτων. Το F1 Score προσφέρει μια πιο ισορροπημένη εικόνα της απόδοσης του μοντέλου όταν απαιτείται η βέλτιστη αναγνώριση των δηλητηριωδών μανιταριών χωρίς να θυσιάζεται η ακρίβεια των προβλέψεων.

Recall: Η ευαισθησία (Recall) μετρά το ποσοστό των πραγματικά δηλητηριωδών μανιταριών που αναγνωρίζονται σωστά από τον ταξινομητή. Στο συγκεκριμένο πρόβλημα, η υψηλή ευαισθησία είναι κρίσιμη, καθώς θέλουμε να βεβαιωθούμε ότι όλα τα δηλητηριώδη μανιτάρια αναγνωρίζονται ως τέτοια. Αν ένα δηλητηριώδες μανιτάρι ταξινομηθεί

λανθασμένα ως βρώσιμο, οι συνέπειες μπορεί να είναι σοβαρές. Ως εκ τούτου, μια μετρική με υψηλή ευαισθησία εξασφαλίζει ότι ο ταξινομητής δεν παραλείπει επικίνδυνα δείγματα.

Precision: Η ακρίβεια πρόβλεψης (Precision) μετρά το ποσοστό των μανιταριών που ταξινομήθηκαν ως δηλητηριώδη και ήταν πράγματι δηλητηριώδη. Στο πρόβλημα της ταξινόμησης μανιταριών, η ακρίβεια πρόβλεψης είναι σημαντική διότι θέλουμε να είμαστε σίγουροι ότι όταν ο ταξινομητής προβλέπει ένα μανιτάρι ως δηλητηριώδες, είναι πράγματι δηλητηριώδες και δεν κάνουμε λάθος, μειώνοντας έτσι τον αριθμό των ψευδών θετικών αποτελεσμάτων. Η υψηλή ακρίβεια σημαίνει ότι υπάρχει χαμηλότερη πιθανότητα να απορρίψουμε βρώσιμα μανιτάρια ως επικίνδυνα.

Συνοψίζοντας από τα παραπάνω, το καλύτερο μοντέλο για την ταξινόμηση των μανιταριών πρέπει να έχει ιδανικά υψηλές τιμές σε όλες τις μετρήσεις που εξετάσαμε. Αν και οι αποδόσεις των μοντέλων ήταν αρκετά κοντά, το Decision Tree ξεχώρισε για τη σταθερή του απόδοση. Η σταθερότητα και η συνέπεια είναι απαραίτητες σε εφαρμογές όπως αυτή, όπου η διάκριση μεταξύ βρώσιμων και δηλητηριωδών μανιταριών μπορεί να είναι κρίσιμη για την υγεία και την ασφάλεια. Το Decision Tree επέδειξε εξαιρετικά αποτελέσματα τόσο στο F1 Score όσο και στην Accuracy. Ωστόσο, αυτό που το ξεχώρισε από τα υπόλοιπα μοντέλα είναι οι υψηλές τιμές σε Precision και Recall, υποδηλώνοντας ότι είναι ικανό να αναγνωρίζει σωστά τα δηλητηριώδη μανιτάρια χωρίς να παραβλέπει τα βρώσιμα και ότι διατηρεί την απόδοσή του σε διαφορετικά υποσύνολα δεδομένων. Τέλος, ήταν ο πιο γρήγορος και ακριβές ταξινομητής σε σχέση με τους άλλους.

7 Σύνοψη και συμπεράσματα

7.1 Επίλογος

Αυτό το project ξεκίνησε με στόχο την ανάπτυξη ενός αξιόπιστου μοντέλου για την ταξινόμηση μανιταριών ως βρώσιμα ή δηλητηριώδη, χρησιμοποιώντας τεχνικές μηχανικής μάθησης. Αρχικά, πραγματοποιήθηκε προεπεξεργασία των δεδομένων και διερευνητική ανάλυση (Exploratory Data Analysis – EDA), η οποία βοήθησε στην καλύτερη κατανόηση των χαρακτηριστικών του dataset και στην αναγνώριση των πιο κρίσιμων παραγόντων για την πρόβλεψη της κατηγορίας ενός μανιταριού. Η οπτικοποίηση των δεδομένων αποτελεί βασικό εργαλείο για την αποσαφήνιση της σχέσης μεταξύ των διαφορετικών χαρακτηριστικών και της ταξινόμησης.

Στη συνέχεια, επιλέχθηκαν και αξιολογήθηκαν διάφορα μοντέλα μηχανικής μάθησης, λαμβάνοντας υπόψη τη συνολική τους απόδοση κάτω από συγκεκριμένες συνθήκες. Καθ' όλη τη διάρκεια του project, έγινε εμφανής η σημασία της σωστής επιλογής μετρικών για την αξιολόγηση των μοντέλων. Η πορεία αυτής της εργασίας ανέδειξε τις προκλήσεις και τις ευκαιρίες που προσφέρει η μηχανική μάθηση στην επίλυση προβλημάτων που σχετίζονται με την ασφάλεια και την υγεία.

7.2 Συμπεράσματα

Συνολικά υπήρχαν 10 ταξινομητές που αξιολογήθηκαν. Μετά την επεξεργασία και της κατάλληλες τροποποιήσεις έγινε ανάλυση των αποτελεσμάτων φανερώνοντας για αυτή την εργασία ως καλύτερο τον ταξινομητή Decision Tree. Παρείχε ισορροπία μεταξύ της ακρίβειας (precision) και ανάκλησης (recall), μεταξύ των υπολοίπων δοκιμασμένων μοντέλων. Έπειτα, αφότου εντοπίστηκε ο καλύτερος ταξινομητής, η αξιολόγηση της καλύτερης μεθόδου διαχείρισης ελλιπής τιμής είναι αναγκαία. Παρακάτω απεικονίζεται ο πίνακα σχετικά με τις μεθόδους στο Decision Tree.

Πίνακας 3: Τιμές μετρικών για όλες τις μεθόδους + Decision Tree

Missing Value Method + Classifier	Accuracy	F1- Score	Recall	Precision
Bfill + Decision Tree	1	1	1	1
Iterative Imputer + Decision Tree	1	1	1	1
Linear Regression Imputer + Decision Tree	1	1	1	1
Drop Values + Decision Tree	1	1	1	1

Τα αποτελέσματα από τον παραπάνω πίνακα δείχνουν ότι ανεξάρτητα από τη μέθοδο διαχείρισης των ελλειπουσών τιμών (Missing Value Method) που χρησιμοποιήθηκε, ο Decision Tree ταξινομητής πέτυχε άψογα αποτελέσματα με 100% ακρίβεια (Accuracy), F1-Score, Recall, και Precision. Αυτό σημαίνει ότι το μοντέλο κατάφερε να ταξινομήσει σωστά όλα τα δείγματα, χωρίς να γίνουν λάθη στις προβλέψεις, ανεξαρτήτως του τρόπου που διαχειρίστηκαν τα ελλιπή δεδομένα. Ειδικότερα, είτε χρησιμοποιήθηκε μέθοδος όπως Bfill (συμπλήρωση των ελλείπων τιμών με την επόμενη τιμή στη σειρά), είτε πιο σύνθετες

μέθοδοι όπως η Iterative Imputer ή η Linear Regression Imputer, είτε απλά αφαιρέθηκαν τα δείγματα με ελλιπείς τιμές (Drop Values), το μοντέλο παραμένει εξαιρετικά αξιόπιστο.

8 Βιβλιογραφία

Σχετική έρευνα για το θεωρητικό υπόβαθρο:

Gupta, N., Sharma, A., & Pathak, N. (2020). Mushroom classification using machine learning techniques. Journal of Computational Biology and Bioinformatics Research, 12(3), 56-62. https://doi.org/10.5897/JCBBR2020.0123

Mushroom Classification Dataset:
https://www.kaggle.com/datasets/uciml/mushroom-classification

---

Για visualizations και κώδικα:

https://www.kaggle.com/code/gaetanlopez/how-to-make-clean-visualizations

https://www.kaggle.com/code/minaemil329/mushroom-eda-classification

https://www.kaggle.com/code/atasaygin/mushroom-classification-and-eda-rf-svm-xgboost

https://www.kaggle.com/code/mig555/mushroom-classification

Για ανάλυση των μετρικών:

https://arize.com/blog-course/f1-score/

Classifiers (κώδικας και πληροφορίες για τον τρόπο λειτουργίας):

AdaBoost:
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

KNN:
https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

Naive Bayes: https://www.geeksforgeeks.org/naive-bayes-classifiers/

LDA: https://scikit-learn.org/stable/modules/lda_qda.html

MPL Classifier:
https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

SVM: https://scikit-learn.org/stable/modules/svm.html

Decision Tree:
https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Random Forest:
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Logistic Regression:
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Gradient Boosting:
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

---

Μέθοδοι αντιμετώπισης των τιμών που έλειπαν:

Bfill: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.bfill.html

Iterative imputer:
https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html

Drop Missing Values: https://scikit-learn.org/stable/modules/impute.html

Linear Regression:
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

---