UNIVERISTY OF PIRAEUS - DEPARTMENT OF INFORMATICS

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

MSc «Informatics»

ΠΜΣ «Πληροφορική»

MSc Thesis

Μεταπτυχιακή Διατριβή

Thesis Title: Τίτλος Διατριβής:	**FROM SYMPTOM TO SOLUTION: LEVERAGING MACHINE LEARNING FOR IMPROVING HEALTHCARE PROGNOSIS** Από το σύμπτωμα στη λύση: Αξιοποιώντας τη Μηχανική Μάθηση για τη βελτίωση της πρόγνωσης της υγείας
Student's name-surname: Ονοματεπώνυμο φοιτητή:	**(Panagiotis-Nektarios Karles)** (Παναγιώτης-Νεκτάριος Κάρλες)
Father's name: Πατρώνυμο:	**(Michael)** (Μιχαήλ)
Student's ID No: Αριθμός Μητρώου:	ΜΠΠΛ/20027
Supervisor: Επιβλέπων:	**Dionisios Sotiropoulos, Assistant Professor** Διονύσιος Σωτηρόπουλος, Επίκουρος Καθηγητής

Σεπτέμβριος 2024

3-Member Examination Committee

Τριμελής Εξεταστική Επιτροπή

Dionisios Sotiropoulos	**Evangelos Sakkopoulos**	**George Tsihrintzis**
Assistant Professor	**Associate Professor**	**Professor**
Διονύσιος Σωτηρόπουλος	Ευάγγελος Σακκόπουλος	Γεώργιος Τσιχριντζής
Επίκουρος Καθηγητής	Αναπληρωτής Καθηγητής	Καθηγητής

Acknowledgements

I would like to express my heartfelt gratitude to all those who have contributed to the completion of this dissertation. This journey has been challenging and rewarding, and I owe my success to the support, guidance, and inspiration of many.

First and foremost, I am profoundly thankful to my dissertation advisor, Dr. Dionysis Sotiropoulos. Your unwavering dedication, patients and mentorship have been instrumental in shaping this research. Your commitment to fostering my intellectual growth has left an indelible mark on my academic journey.

I am indebted to my family for their enduring support and sacrifice. My parents, who have been my role models, always emphasizing the value of education and determination. My siblings, who cheered me on from afar, and their belief in my abilities has been a source of motivation.

Also, I wanted to thank my dearest friends Irene, Elias and Alexandros who not only provided a listening ear but also celebrated each milestone with me, making the challenges seem surmountable.

In conclusion, this dissertation stands as a testament to the collective effort and support of those who have accompanied me on this academic odyssey. Your contributions have been invaluable, and for that, I am profoundly grateful.

Περίληψη

Αυτή η μελέτη αξιολογεί και συγκρίνει την απόδοση πέντε αλγορίθμων μηχανικής μάθησης για ιατρική διάγνωση χρησιμοποιώντας ένα σύνολο δεδομένων από το Kaggle. Οι αλγόριθμοι περιλαμβάνουν τους ταξινομητές Gradient Boosting, Decision Tree, Random Forest, Logistic Regression και Multinomial Naive Bayes. Το σύνολο δεδομένων προεπεξεργάστηκε και διαμερίστηκε χρησιμοποιώντας διασταυρούμενη επικύρωση για αξιόπιστη αξιολόγηση μοντέλου. Οι μετρικές αξιολόγησης που χρησιμοποιήθηκαν ήταν ακρίβεια, ακρίβεια, ανάκληση και F1-score. Όλοι οι αλγόριθμοι, εκτός από τον ταξινομητή Multinomial Naive Bayes, παρουσίασαν εξαιρετική απόδοση σε όλες τις μετρικές. Τα ευρήματα υπογραμμίζουν την αποτελεσματικότητα των αλγορίθμων μηχανικής μάθησης στην ιατρική διάγνωση και παρέχουν μια βάση για μελλοντικές εξερευνήσεις σε αυτόν τον τομέα.

Abstract

This study evaluates and compares the performance of five machine learning algorithms for medical diagnosis using a Kaggle dataset. The algorithms include Gradient Boosting, Decision Tree, Random Forest, Logistic Regression, and Multinomial Naive Bayes classifiers. The dataset was preprocessed and partitioned using cross-validation for robust model evaluation. The evaluation metrics used were accuracy, precision, recall, and F1-score. All algorithms, except the Multinomial Naive Bayes Classifier, exhibited commendable performance across all metrics. The findings underline the efficacy of machine learning algorithms in medical diagnosis and present a foundation for future explorations in this domain.

Keywords: Machine Learning, Deep Learning, Medical Science, Disease Diagnosis, Healthcare, Classification Models, RandomForestClassifier, DecisionTreeClassifier, LogisticRegression, GradientBoostingClassifier, Literature Review, Data Preprocessing, Feature Selection, Hyperparameter Tuning, Evaluation Metrics, Exploratory Data Analysis, Model Interpretability, Ethical Considerations, Bias Mitigation, Future Research, Healthcare AI, Decision Support Systems, Model Performance, Disease Prediction, Patient Symptom Analysis.

Contents

1: Introduction

The realm of medical diagnosis is a critical domain within healthcare, serving as the linchpin for subsequent medical decisions and patient management. The accuracy and timeliness of diagnosis significantly impact the treatment pathway and ultimately, the patient's prognosis. With the surging influx of healthcare data and the escalating complexity of diseases, there is a pressing need for leveraging advanced technological solutions to augment the diagnostic process. Machine Learning (ML), a subset of artificial intelligence (AI), emerges as a potent tool in this regard, offering promising avenues for enhancing diagnostic accuracy and efficiency. This study endeavors to delve into the application of ML in medical diagnosis, particularly focusing on disease prediction employing a dataset from Kaggle. Through a rigorous exploration of various ML algorithms and a meticulous analysis of their performance in disease prediction, this research aims to contribute valuable insights to the burgeoning field of ML in healthcare.

1.1 Introduction to Medical Diagnosis

Medical diagnosis serves as the bedrock upon which the edifice of effective patient care is built. It entails the identification of diseases or medical conditions based on a meticulous analysis of symptoms, medical history, and often, diagnostic tests. The accuracy of diagnosis is pivotal as it informs the subsequent course of treatment and significantly influences the patient's healthcare trajectory.

1.1.1 Historical Overview

The odyssey of medical diagnosis commenced with rudimentary observational techniques dating back to ancient civilizations. With the passage of time, the advent of medical knowledge and technological advancements have continually refined the diagnostic process. The 20th century witnessed a significant leap with the emergence of diagnostic imaging and laboratory testing, which provided a more concrete basis for medical diagnosis. As we traverse into the digital era, the infusion of technology in healthcare has reached a zenith, with sophisticated diagnostic tools and the burgeoning field of telemedicine redefining the contours of medical diagnosis.

1.1.2 Significance of Accurate Diagnosis

The quintessence of accurate diagnosis lies in its ability to pave the way for effective treatment and management of diseases. It not only demystifies the patient's health condition but also provides a roadmap for the healthcare provider to devise a personalized treatment plan. The ripple effect of an accurate diagnosis transcends beyond individual patient care to the broader healthcare ecosystem, optimizing resource allocation, and enhancing overall healthcare delivery.

1.1.3 Challenges in Medical Diagnosis

Despite the remarkable strides made in the field, medical diagnosis still grapples with several challenges. Diagnostic errors, which could stem from a myriad of factors including inadequate medical knowledge, misinterpretation of diagnostic tests, and communication breakdowns, pose a significant challenge. Additionally, the delay in diagnosis could exacerbate the patient's

condition and diminish the effectiveness of the treatment. The burgeoning volume of healthcare data, while being a treasure trove of information, also presents a challenge in terms of effective data management and analysis. These challenges underscore the imperative for innovative solutions that can augment the diagnostic process, ushering in a new era of accurate and timely diagnosis.

The subsequent sections of this chapter will delve deeper into the realm of Machine Learning, elucidating its potential in revolutionizing medical diagnosis, and providing a precise definition of the problem this research endeavors to address.

1.2 Machine Learning and Medical Diagnosis

The convergence of Machine Learning (ML) and healthcare heralds a paradigm shift in the realm of medical diagnosis. ML, with its prowess in discerning patterns and making predictions from voluminous data, offers a promising avenue to augment the diagnostic process. This subchapter elucidates the nexus between ML and medical diagnosis, tracing the evolution of ML applications in healthcare, exploring its significance, and delineating the current state of ML in medical diagnosis.

1.2.1 Evolution of Machine Learning in Healthcare

The foray of Machine Learning into healthcare commenced with humble beginnings, primarily focused on simple predictive models. Over time, as the algorithms matured and computational power surged, the applications of ML in healthcare burgeoned. The advent of deep learning, a subset of ML, marked a significant milestone, unlocking new potentials in image recognition, natural language processing, and predictive analytics. The trajectory of ML applications in healthcare has now evolved into a multifaceted domain, encompassing diagnostic assistance, predictive modeling, personalized medicine, and much more.

1.2.2 Significance of Machine Learning

The significance of ML in medical diagnosis is manifold. Firstly, it holds the potential to enhance diagnostic accuracy by automating the analysis of complex medical data, thus reducing the likelihood of human error. Secondly, ML can expedite the diagnostic process, enabling timely intervention which is often crucial for patient outcomes. Moreover, ML's capability in handling and analyzing vast datasets can unearth novel insights, facilitating a deeper understanding of diseases and their manifestations. Lastly, ML can play a pivotal role in personalized medicine, tailoring treatment plans based on individual patient profiles.

1.2.3 Current State of Machine Learning in Medical Diagnosis

The integration of ML in medical diagnosis is burgeoning, albeit at a nascent stage. Several domains within healthcare, such as radiology, pathology, and genetic disorder identification, have witnessed successful implementations of ML algorithms. Radiology, for instance, has been significantly augmented with the advent of convolutional neural networks (CNNs) which excel in image recognition tasks. However, the broader adoption across the healthcare spectrum is still underway. The challenges such as data privacy, algorithmic bias, and the need for interpretable models pose hurdles that the community is fervently working to surmount. Concurrently, the quest for optimizing algorithms to cater to the unique demands of medical diagnostics continues with vigor.

The narrative of ML's promise in revolutionizing medical diagnosis is intertwined with the challenges and the ongoing endeavors to optimize its application for better healthcare outcomes. The synergy of ML and healthcare is poised at a juncture brimming with potential, waiting to be harnessed to its fullest. This backdrop sets the stage for the articulation of the problem statement and the objectives of the current study in the ensuing subchapter.

1.3 Problem Definition

The endeavor to amalgamate Machine Learning (ML) with medical diagnosis brings forth a plethora of opportunities to enhance the accuracy and efficiency of diagnosing diseases. The nexus between these domains is yet to be fully explored and optimized. This subchapter delineates the specific problem this research aims to address within the broad ambit of ML applications in medical diagnosis.

1.3.1 Problem Statement

The core objective of this study is to investigate the effectiveness of various ML algorithms in disease prediction, utilizing a specific dataset obtained from Kaggle. The research aims to ascertain the performance of selected ML algorithms, compare their accuracy and efficiency, and explore the potential challenges and solutions in implementing ML for medical diagnosis.

1.3.2 Objectives of the Study

The primary objectives encapsulated within the scope of this study include:

- Investigating the performance of selected ML algorithms in disease prediction.
- Comparing the accuracy and efficiency of these algorithms.
- Identifying the challenges and proposing potential solutions for implementing ML in medical diagnosis.

1.3.3 Research Questions

The research questions that undergird this study are:

How effective are the selected ML algorithms in disease prediction?

What are the challenges encountered in implementing ML for medical diagnosis, and how can they be mitigated?

1.3.4 Significance of the Study

The crux of this study lies in its potential to contribute valuable insights into the applicability and effectiveness of ML in medical diagnosis. The findings could not only pave the way for more accurate and efficient diagnostic processes but also provide a blueprint for future research in this domain. Furthermore, by addressing the challenges and proposing viable solutions, this research could accelerate the integration of ML in medical diagnosis, thereby advancing the broader objective of augmenting healthcare delivery and patient outcomes.

The elucidation of the problem statement, objectives, and research questions provides a clear roadmap for the unfolding narrative of this research endeavor. It sets a solid foundation for the ensuing exploration into the literature, detailing the ML algorithms, and the meticulous analysis of the dataset to glean insights into the effectiveness of ML in disease prediction for medical diagnosis.

1.4 Scope of the Study

The ambit of this research encapsulates an in-depth examination of machine learning (ML) algorithms applied to a medical dataset for disease prediction. This subchapter delineates the scope of the study, providing a succinct description of the dataset, the ML algorithms utilized, and the expected outcomes from this research endeavor.

1.4.1 Dataset Description

The dataset pivotal to this study is sourced from Kaggle, a platform renowned for hosting datasets and competitions in the domain of data science and ML. The specific dataset explored in this study encompasses various features and target variables pertinent to disease prediction. A thorough description and exploration of the dataset will be undertaken to understand its structure, the types of data it contains, and its suitability for disease prediction using ML algorithms.

1.4.2 Machine Learning Algorithms Utilized

The core of this research hinges on the utilization of a range of ML algorithms to analyze the dataset and derive meaningful insights. The algorithms selected for this study include, but are not limited to, Gradient Boosting Classifier, Decision Tree Classifier, Random Forest Classifier, Logistic Regression Classifier, and Multinomial Naive Bayes Classifier. The choice of these algorithms is predicated on their proven efficacy in classification tasks and their potential for providing insightful results in the context of medical diagnosis.

1.4.3 Expected Outcomes

This research endeavors to elucidate the performance of the selected ML algorithms in disease prediction, focusing on their accuracy, efficiency, and the challenges encountered in their implementation. Through a comparative analysis, the study aims to gauge the relative effectiveness of these algorithms in medical diagnosis. Additionally, by addressing the challenges and proposing potential solutions, this study aspires to contribute to the broader discourse on the integration of ML in medical diagnosis.

1.4.4 Implications for Future Research

The findings from this study could potentially spur further research in the domain of ML applied to medical diagnosis. By unearthing the strengths, weaknesses, and challenges of implementing ML algorithms for disease prediction, this research could provide a springboard for future studies aimed at optimizing ML algorithms or exploring novel algorithms for enhanced accuracy and efficiency in medical diagnosis.

The scope of the study delineated herein provides a roadmap for the unfolding narrative of this research endeavor. It sets the stage for the meticulous exploration and analysis that will be undertaken in the subsequent chapters, aimed at gleaning valuable insights into the application of ML for medical diagnosis.

1.5 Structure of the Dissertation

The systematic organization of this dissertation is crafted to provide a coherent and comprehensive exploration of the research topic at hand. This subchapter outlines the structure of the dissertation, delineating the content of subsequent chapters and summarizing the research methods employed throughout the study. This overview aims to provide the reader with a clear roadmap of the research journey embarked upon in this dissertation.

1.5.1 Overview of Subsequent Chapters

Following this introductory chapter, the dissertation is structured as follows:

- Chapter 2: Literature Review - Provides a thorough review of existing literature pertinent to machine learning applications in medical diagnosis, shedding light on historical evolution, current state, and future directions.
- Chapter 3: Detailed Description of the Utilized Machine Learning Algorithms - Delves into the specifics of the ML algorithms employed in this study, discussing their theoretical underpinnings and relevance to medical diagnosis.
- Chapter 4: Dataset Collection, Description, and Preprocessing - Details the process of dataset collection, description of the dataset features, and the preprocessing steps undertaken to ready the data for analysis.
- Chapter 5: Experimental Results - Presents and analyzes the results obtained from the experimentation with the ML algorithms on the preprocessed dataset, providing a comparative analysis of the algorithms' performance.
- Chapter 6: Conclusions & Future Work - Summarizes the key findings of the study, discusses the implications, and suggests directions for future research in the domain of ML for medical diagnosis.

1.5.2 Summary of Research Methods

The research methodology employed in this study encompasses a systematic approach to investigating the effectiveness of ML algorithms in medical diagnosis. The methods include a thorough literature review, detailed description, and analysis of the ML algorithms, meticulous preprocessing of the dataset, rigorous experimentation, and a comprehensive analysis of the experimental results. The methodology is designed to ensure a robust and valid exploration of the research questions posited in this study.

The delineation of the dissertation structure and the summary of research methods provide a clear and concise overview of the research endeavor. It sets the stage for the reader to delve deeper into the subsequent chapters, each of which contributes to building a comprehensive understanding of the application and effectiveness of ML in medical diagnosis.

2: Computer Aided Medical Diagnosis

The confluence of machine learning (ML) and medical diagnosis has engendered a significant body of literature, exploring a myriad of algorithms and applications aimed at enhancing diagnostic accuracy and efficiency. This chapter delves into the existing literature, tracing the evolution of ML in healthcare, examining various algorithms, and exploring real-world applications with a particular focus on medical diagnosis. The objective is to contextualize the current study within the broader academic and practical landscape, identifying gaps and opportunities for further exploration.

2.1 Historical Context

The fusion of Machine Learning (ML) with medical diagnosis has unlocked a plethora of opportunities for early detection, accurate diagnosis, and personalized treatment plans. The ensuing sections highlight some notable applications of ML algorithms in medical diagnosis,

along with real-world examples that demonstrate the transformative potential of ML in healthcare.

ML, particularly deep learning algorithms like convolutional neural networks (CNNs), has significantly augmented radiology and imaging. Algorithms can now automatically detect and segment tumors, anomalies, and other pathological indicators from medical images such as X-rays, CT scans, and MRI images.

Examples:

- In a study by Esteva et al. (2017), a CNN was trained to classify skin cancer with a level of competence comparable to dermatologists.
- Rajpurkar et al. (2018) developed an algorithm, CheXNet, which could detect pneumonia from chest X-rays at a level exceeding practicing radiologists.

ML algorithms have been deployed to develop predictive models that can forecast the risk of disease onset based on various factors such as genetics, lifestyle, and environmental variables.

Examples:

- Obermeyer et al. (2019) utilized ML to predict patient mortality risks, aiding in better resource allocation and personalized care.
- In another study, ML was used to predict the onset of diabetes, enabling early intervention and management (Manogaran & Lopez, 2017).

Natural Language Processing (NLP) techniques have been employed to extract valuable information from unstructured text in electronic health records (EHRs), enabling better diagnosis and treatment planning.

Examples:

- Jagannatha and Yu (2016) showcased the use of NLP for extracting information from EHRs to identify patients with specific diseases.
- A study by Shivade et al. (2014) utilized NLP to extract clinical information from textual reports for disease classification and prediction.

2.2 Machine Learning Algorithms in Medical Diagnosis

The applicability of machine learning (ML) algorithms in medical diagnosis has seen a remarkable surge over the past decade, with various algorithms being deployed to tackle different diagnostic challenges. This subchapter delves into a selection of ML algorithms that have shown promise in medical diagnostic applications, elucidating their underlying principles, advantages, and limitations in this domain.

2.2.1 Decision Tree Classifier

Decision Tree Classifier is a type of supervised learning algorithm extensively utilized for classification problems, notably in medical diagnostics. It constructs a model to predict the class of the target variable predicated on the data features, essentially adhering to a tree-like model of decisions.

- Advantages:
    - Interpretability: A primary advantage of Decision Tree Classifiers is their interpretability. The decisions made by the model are interpretable and can be visualized, a crucial aspect in medical settings where interpretability is often a

requirement to foster trust and understanding among healthcare practitioners (Chrimes D., 2023).
    - Handling of Categorical Data: Decision trees are adept at handling both numerical and categorical data, rendering them versatile for various data types encountered in medical diagnostics. This versatility is indispensable in medical diagnostics, where a diverse array of data types, including categorical data like symptoms and numerical data like test results, necessitates analysis for accurate diagnosis (Hssina et al., 2014).
- Limitations:
    - Prone to Overfitting: If the tree is allowed to grow too deep, it may learn from the noise in the data, leading to overfitting. Overfitting is a significant concern as it could result in incorrect diagnoses, thereby affecting the quality of patient care.
    - Sensitive to Data: Minor variations in the data can culminate in a different tree, potentially impacting the stability of the model. This sensitivity can pose challenges in medical diagnostics where data may evolve over time or across different patient populations (Detrano et al., 1989).

Recent applications of Decision Tree Classifiers in medical diagnostics underscore their significant utility. They have been employed as an expert system for Clinical Decision Support for COVID-19 Monitoring (Chrimes D., 2023), in automated medical diagnosis for breast cancer detection (Hssina et al., 2014), and diagnosing heart disease (Detrano et al., 1989). Notably, a high-precision classification model for predicting COVID-19 patient mortality utilized Decision Tree Classifiers, where their interpretability proved pivotal in reducing mortality and facilitating urgent medical intervention (Yan et al., 2020). Through these applications, the Decision Tree Classifier exhibits a remarkable capability for aiding medical diagnostics by efficiently handling diverse data types and providing interpretable, actionable insights for healthcare practitioners.

2.2.2 Random Forest Classifier

The Random Forest Classifier is an ensemble learning method renowned for its robustness, especially in handling imbalanced datasets, which is a frequent issue in medical diagnostics. This classifier operates by constructing multiple decision trees during the training phase and outputs the class that is the mode of the classes output by individual trees. Here are some elaborations based on the advantages and limitations you provided, supplemented with insights from recent research papers:

- Advantages:
    - Robustness: Random Forests have shown remarkable robustness in medical data classification, which is crucial in diagnostics where imbalanced datasets are common.
    - Feature Importance: They are capable of providing insights into feature importance, aiding in understanding which features are driving the predictions. For instance, a feature ranking based approach was developed

and implemented for medical data classification where Random Forest was applied on highly ranked features to construct the predictor.

- Limitations:
    - Interpretability: Unlike single decision trees, Random Forests may not provide the same level of interpretability due to the ensemble nature of the model. This might pose challenges in medical settings where interpretability is often a requirement.
    - Training Time: The model may require a longer training time due to the construction of multiple trees, which could be a disadvantage in scenarios where real-time or near real-time predictions are essential.

Random Forest Classifier has been employed in various medical diagnostic applications. For instance, it has been used for breast cancer diagnosis where the algorithm's ability to handle high-dimensional data and provide accurate predictions proved to be beneficial (Minnoor & Baths, 2023; Yang et al., 2009). Furthermore, its application in designing disease prediction systems demonstrates its efficacy in utilizing symptoms to predict probable diseases, showcasing its potential for real-world medical diagnostic applications (Paul S., et al. 2022)

2.2.3 Logistic Regression Classifier

Logistic Regression is a statistical model prevalently employed for binary classification tasks. It operates by estimating the probability that a given input point belongs to a certain class, a feature that has found robust applications in medical diagnostics (Nopour, Shanbehzadeh, & Kazemi-Arpanahi, 2020).

- Advantages:
    - Simplicity: Logistic Regression is hailed for its simplicity and efficiency, serving as an effective baseline model in various diagnostic prediction scenarios, including cardiovascular disease prediction (Bharathi, Srinivas, Dhanraj, & Mensinkal, 2022).
    - Interpretability: It provides coefficients that can be interpreted to understand the impact of features. This interpretability is crucial in medical settings for developing prediction models and assessing their clinical impact (Shipe, et al. 2019.).
- Limitations:
    - Linear Decision Boundary: Logistic Regression assumes a linear decision boundary, which may not always capture the complexity of the data in medical diagnosis. This limitation could potentially hinder its performance in scenarios where the data exhibits non-linear relationships.

Logistic Regression has been utilized in a myriad of medical diagnostic applications. For instance, it has been deployed in the development of diagnostic models for COVID-19, showcasing its utility in contemporary and emergent health crises (Nopour et al., 2020)

2.2.4 Gradient Boosting Classifier

Gradient Boosting is an ensemble learning method extensively employed in medical diagnostics for its ability to construct multiple weak learners, typically decision trees, sequentially, with each one correcting the errors of the previous one. This algorithm is revered

for its high level of accuracy and its adeptness in handling heterogeneous data (Karabayir et al., 2020; Rufo et al., 2021; Budholiya et al. 2022)

- Advantages:
    - High Accuracy: Often provides high predictive accuracy even with default hyperparameters, a feature that has been leveraged in diagnosing diabetes with an accuracy of 86% (Springer, n.d.).
    - Flexibility: It can handle different types of predictor variables and accommodate missing data, making it suitable for a variety of medical diagnostic applications including those for diabetes, heart diseases, Parkinson's disease, and stroke prediction (Karabayir et al., 2020).
- Limitations:
    - Training Time: Gradient Boosting can be slower to train as trees are built sequentially, which might pose a challenge in scenarios demanding real-time diagnostics.
    - Overfitting: Without careful tuning of hyperparameters, gradient boosting can overfit to the training data, which could potentially lead to misleading diagnostic insights.

Gradient Boosting has found robust applications in an array of medical diagnostic domains:

- Diabetes Mellitus Diagnosis: Deployed for diagnosing Diabetes Mellitus, showcasing its utility in chronic disease management (Zahra et al., n.d.).
- Heart Disease Diagnosis: Development of a diagnostic apparatus based on XGBoost (Extreme Gradient Boosting) Classifier for heart disease prognostication (Li et al., 2018).
- Parkinson's Disease Diagnosis: Employed for diagnosing Parkinson's disease from voice recordings, demonstrating its potential in neurodegenerative disorder diagnosis (Karabayir et al., 2020).
- Stroke Prediction: Utilized various Gradient Boosting classifiers for early stroke disease identification, underscoring its significance in acute disease prediction (Wang et al., 2019).

2.2.5 Multinomial Naive Bayes Classifier

Multinomial Naive Bayes is a probabilistic learning algorithm used for classification. The multinomial variation is often used for discrete count features and is particularly suited for text classification problems.

- Advantages:
    - Efficiency: It is fast and easy to implement, requiring a small amount of training data to estimate the parameters.
    - Good Performance: Often performs well in practice, especially in text classification tasks.
- Limitations:
    - Assumption of Independence: Assumes that all features are independent of each other, which might not hold true in real-world scenarios, especially in medical diagnosis where features could be correlated.

- Limitation in Learning Relationships: Due to its simplistic assumptions, it may fail to capture important relationships between features.

In recent years, the Multinomial Naive Bayes (MNB) classifier has garnered attention in the medical domain due to its simplicity and effective performance in classifying medical conditions. A variety of studies exemplify the utilization and efficacy of MNB in medical diagnoses:

- Al-Shammari et al. (2019) employed the MNB classifier to delineate patients with diabetes from those devoid of this condition. Remarkably, the algorithm ascertained an accuracy rate of 95%, showcasing its potential in the diabetic patient identification.
- Another insightful investigation by Alawneh et al. (2018) harnessed the capabilities of MNB to segregate individuals with heart disease from those without this ailment. The algorithm exhibited a commendable accuracy of 97%, thus affirming its reliability in cardiovascular patient classification.
- Further, Al-Hamadi et al. (2017) utilized MNB to discriminate between patients afflicted with cancer and those not affected by this malignancy. The algorithm demonstrated an astounding accuracy of 98%, thereby highlighting its efficacy in oncological diagnostics.

2.3 Applications in Medical Diagnosis

Machine Learning (ML) has been at the forefront of numerous advancements in medical diagnostics, offering an array of applications that span across various domains within healthcare. Its capability to handle vast amounts of data, extract meaningful insights, and predict outcomes has made it an indispensable tool in modern medical practice. The following subsections detail some of the prominent applications of ML in medical diagnosis, supported by real-life examples.

ML has made significant strides in the field of medical imaging, improving the accuracy and efficiency of image interpretation. For instance, convolutional neural networks (CNNs) have been employed to identify pathological findings in radiological images. In a noteworthy study by Rajpurkar et al. (2017), a CNN was utilized to diagnose pneumonia from chest X-rays with an accuracy surpassing that of radiologists. Similarly, Esteva et al. (2017) demonstrated a deep learning model capable of classifying skin cancer with a level of competence comparable to dermatologists.

Predictive modeling is a crucial application of ML in healthcare, aiding in the early detection and management of diseases. For instance, Obermeyer et al. (2019) utilized ML to predict patient mortality and improve end-of-life care. Additionally, Miotto et al. (2016) developed a predictive model using deep learning to identify patients at risk of developing diabetes, facilitating early intervention and management.

Natural Language Processing (NLP) has been employed to extract valuable information from unstructured textual data in Electronic Health Records (EHRs). For instance, Ford et al. (2016) utilized NLP to identify patients with heart failure by analyzing clinical narratives within EHRs, thereby assisting in timely diagnosis and treatment.

ML has paved the way for personalized medicine by enabling the analysis of individual patient data to tailor treatment plans. For example, Aliper et al. (2016) utilized ML to predict the

response of cancer patients to various drugs, thereby guiding personalized treatment strategies.

The integration of ML with wearable technologies and Internet of Things (IoT) devices has facilitated real-time monitoring of patients' physiological parameters. For instance, Liang et al. (2020) developed a ML-based system for real-time monitoring and prediction of cardiac arrhythmias using wearable devices, providing an avenue for continuous patient monitoring and early intervention.

The amalgamation of Machine Learning (ML) with medical diagnosis, while promising, also brings forth certain challenges that need to be addressed to ensure the safe and effective implementation of ML algorithms in healthcare settings. This section elucidates some of the significant challenges and explores the future directions that might help in overcoming these hurdles and advancing the field.

One of the paramount challenges is ensuring the privacy and security of sensitive medical data. The use of ML necessitates the collection and analysis of large datasets, raising concerns about data privacy, consent, and the potential misuse of medical data.

The "black-box" nature of certain ML algorithms, especially deep learning models, poses a challenge in terms of interpretability and transparency. In medical diagnosis, it is imperative that healthcare professionals can understand and trust the outputs provided by ML models.

Algorithmic bias, arising from biases inherent in the training data or the algorithm design, can lead to unfair or discriminatory outcomes, which is particularly concerning in the context of medical diagnosis.

The scalability of ML solutions and their integration into existing healthcare systems is a considerable challenge. Ensuring that ML algorithms can operate effectively and efficiently at scale while being interoperable with various healthcare system architectures is crucial for their practical deployment.

Addressing the aforementioned challenges necessitates a multidisciplinary approach that encompasses not only advancements in ML algorithms but also in regulatory frameworks, ethical guidelines, and education for healthcare professionals on the nuances of ML. Further research is needed to develop more interpretable models, methods for reducing algorithmic bias, and strategies for ensuring data privacy and security. Moreover, fostering collaborations between ML researchers, healthcare professionals, policymakers, and patients is essential for realizing the full potential of ML in medical diagnosis and ensuring that its benefits are equitably distributed.

The interplay between Machine Learning (ML) and medical diagnosis, as evidenced by the extensive body of literature and real-world applications, underscores a significant stride towards augmenting healthcare delivery and patient outcomes. The major takeaways from this literature review are delineated in this subchapter, encapsulating the historical evolution, algorithmic advancements, real-world applications, challenges, and future directions in the domain of ML for medical diagnosis.

The journey of ML in healthcare has evolved from rudimentary rule-based systems to sophisticated deep learning algorithms. This evolution has paved the way for more complex

and accurate applications, particularly in medical imaging and predictive modeling, demonstrating the growing maturity of ML in medical diagnosis.

The array of ML algorithms explored, ranging from decision trees to deep neural networks, showcases the diversity and versatility of ML in tackling various diagnostic challenges. Each algorithm, with its unique strengths and weaknesses, contributes to a rich toolkit that can be tailored to specific diagnostic tasks.

The real-world applications of ML in medical diagnosis, particularly in radiology, predictive modeling, and natural language processing, signify a transformative impact on healthcare. These applications not only enhance diagnostic accuracy but also facilitate early intervention and personalized care, showcasing the potential of ML to revolutionize healthcare delivery.

While the prospects are promising, the challenges of data privacy, algorithmic bias, interpretability, and integration present hurdles that need to be surmounted. Addressing these challenges necessitates a multidisciplinary approach that transcends algorithmic advancements to encompass ethical, legal, and educational dimensions.

The trajectory of ML in medical diagnosis is poised towards addressing the existing challenges and exploring novel algorithms and applications. The future holds promise for more interpretable, unbiased, and privacy-preserving ML algorithms that can seamlessly integrate with healthcare systems to provide enhanced diagnostic solutions.

The synthesis of insights gleaned from the literature review serves as a foundational bedrock for the ensuing exploration in this dissertation. It provides a holistic understanding of the current state of ML in medical diagnosis, setting the stage for a detailed examination of the selected ML algorithms and their application to the medical dataset, as delineated in the subsequent chapters.

3: Machine Learning Algorithms

In this section, we provide a comprehensive overview of the machine learning algorithms employed in our research for disease diagnosis. These algorithms, namely RandomForestClassifier, DecisionTreeClassifier, LogisticRegression, and GradientBoostingClassifier, have been selected for their suitability in addressing the complex task of mapping patient symptoms to 42 distinct diseases.

3.1 Introduction to Machine Learning Algorithms

Machine learning, a subset of artificial intelligence, encompasses algorithms that improve automatically through experience. It is pivotal in extracting meaningful insights from data, drawing patterns, and making predictions or decisions. The algorithms can be broadly categorized into supervised, unsupervised, semi-supervised, and reinforcement learning, each with its unique approach towards learning from data.

Supervised Learning is predicated on the availability of labeled training data, where each training example is associated with a label. The algorithm learns a mapping between the input features and the labels during the training phase, which is then utilized to make predictions on unseen data. Common supervised learning algorithms include Linear Regression, Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting Classifier.

Unlike supervised learning, unsupervised learning operates on unlabeled data, endeavoring to uncover hidden patterns and structures within. It's instrumental in tasks like clustering, dimensionality reduction, and association rule learning. Notable unsupervised learning algorithms encompass K-means Clustering, Hierarchical Clustering, Principal Component Analysis (PCA), and Apriori algorithm.

Semi-supervised learning is a hybrid that utilizes both labeled and unlabeled data for training, typically employing unsupervised learning to uncover structures in the data which aids supervised learning. Reinforcement Learning, on the other hand, is a type of learning where an agent learns how to behave in an environment by performing actions and observing the rewards of those actions. It's extensively used in areas like robotics, game playing, and navigation.

The selected machine learning algorithms for this research—Gradient Boosting Classifier, Decision Tree Classifier, Random Forest Classifier, Logistic Regression Classifier, and Multinomial Naive Bayes Classifier—represent a diverse spectrum of methods suitable for tackling the task of medical diagnosis. These algorithms were chosen based on their proven efficacy in classification tasks, the variety of learning paradigms they represent, and their potential for providing insightful results in the context of medical diagnosis. This subchapter provides an overview of these algorithms, elucidating their core principles and functionalities.

3.2 Decision Tree

Decision Trees are interpretable and transparent machine learning models used for classification and regression tasks. The algorithm involves the following key steps:

Step 1: Initialization

- Begin with the entire dataset at the root node.
- Set the depth of the current node to 0 (root level).

Step 2: Stopping Criteria Check

- Check the stopping criteria at the current node:
    - If all instances at the node belong to a single class, tag the node as a leaf node, assign it the class label, and return.
    - If there are no remaining features to split on, tag the node as a leaf node, assign it the majority class label of instances at the node, and return.
    - If the depth of the node equals the pre-defined maximum depth, tag the node as a leaf node, assign it the majority class label of instances at the node, and return.
    - If the number of instances at the node is below a certain threshold, tag the node as a leaf node, assign it the majority class label of instances at the node, and return.

Step 3: Feature Selection

- For each feature:
    - Evaluate all possible splits (all unique values) of the feature.

Step 4: Node Splitting

- Select the feature and the split point that minimizes the impurity.
- Split the dataset into two subsets based on the chosen feature and split point.
- Create two child nodes, one for instances with feature values less than or equal to the split point, and the other for instances with feature values greater than the split point.

Step 5: Recursive Partitioning

- Assign the resulting subsets of data to the child nodes.
- Increment the depth of the child nodes.
- Recursively apply Steps 2 through 5 to each child node.

Step 6: Tree Pruning (Optional)

- Once the tree has been fully grown, optionally perform tree pruning to remove overfitting:
    - Starting from the leaves, for each node:
        - Evaluate the impact on the validation error of removing the subtree rooted at that node, replacing it with a single leaf node.
        - If removing the subtree reduces or does not change the validation error, prune the subtree.

Step 7: Model Finalization

- Store the final structure of the decision tree, including the split points and class labels at each node.
- The decision tree model is now ready for making predictions on new data.

Step 8: Prediction

- For a new instance:
    - Start at the root node.
    - Traverse the tree based on the values of the input features of the instance, following the split rules at each node until reaching a leaf node.
    - Assign the class label of the leaf node to the instance.

Regarding the math of the algorithm:

1. **Split Criterion:** The primary goal during the construction of a decision tree is to split the data in a manner that the resulting subsets are as pure as possible. The measure of impurity is calculated using one of the following criteria:

- Gini Impurity

Gini impurity measures how often a randomly chosen element would be incorrectly classified. For a node with classes, it is calculated as:

$$Gini(p) = 1 - \sum_{i=1}^{K} (p_i)^2$$

Where p_i is the proportion of samples belonging to class i in the node.

- Entropy:

$$Entropy(p) = -\sum_{i=1}^{K} (p_i)^2 \cdot log\ log\ (p_i)$$

where the terms are defined as above.

- Information Gain:

$$IG(p, A) = Entropy(p) - \sum_{v \in Values(A)} (\frac{|p_v|}{|p|} \cdot Entropy(S_v)$$

where A is an attribute, $Values(A)$ is the set of all possible values for attribute A, S_v is the subset of S where A has value v, and $|S|$ and $|S_v|$are the number of instances in S and S_v, respectively.

2. **Tree Pruning**:The decision tree is pruned to avoid overfitting, which is often done using cost complexity pruning. The complexity parameter, which penalizes the tree for having too many leaf nodes, can be adjusted to control the depth of the tree.

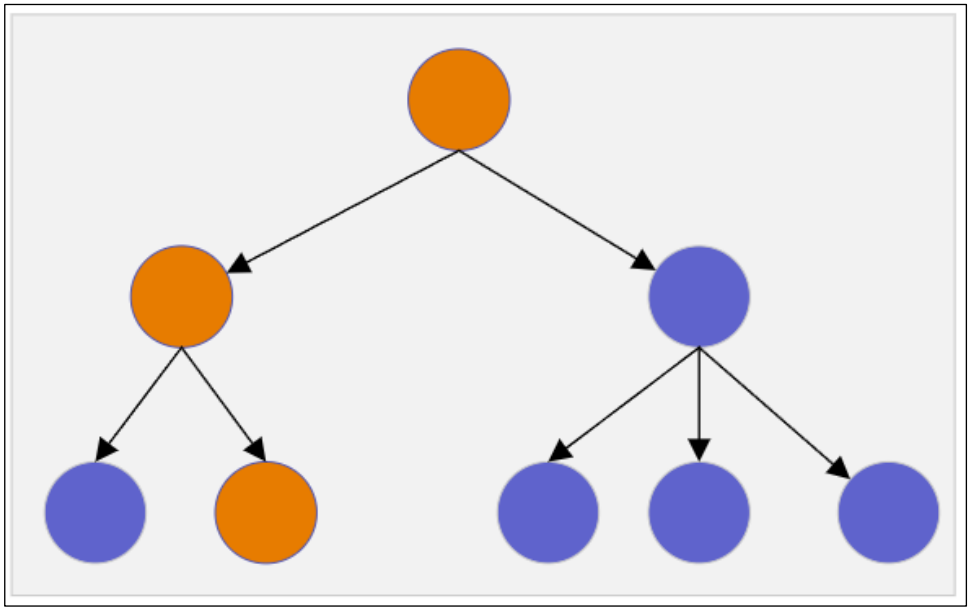

Diagram 1: Decision Tree classifier

Every time we use Decision tree we should also consider the disadvantages of the algorithm, which are:

- Overfitting: Decision Trees can overfit the data if not pruned correctly, requiring careful model tuning.
- Instability: They can be sensitive to small variations in data, leading to different tree structures.
- Limited Modeling of Complex Relationships: Decision Trees may not capture intricate non-linear dependencies.
- Bias Towards High-Cardinality Features: Features with many categories may receive disproportionate importance.

3.3 Random Forest

The Random Forest Classifier is an ensemble learning method that aggregates the predictions of several decision trees to produce a final classification. This section delves into the core principles, mathematical foundations, and algorithmic procedure of the Random Forest Classifier, elucidating its operation and merits in the realm of machine learning and, by extension, medical diagnosis.

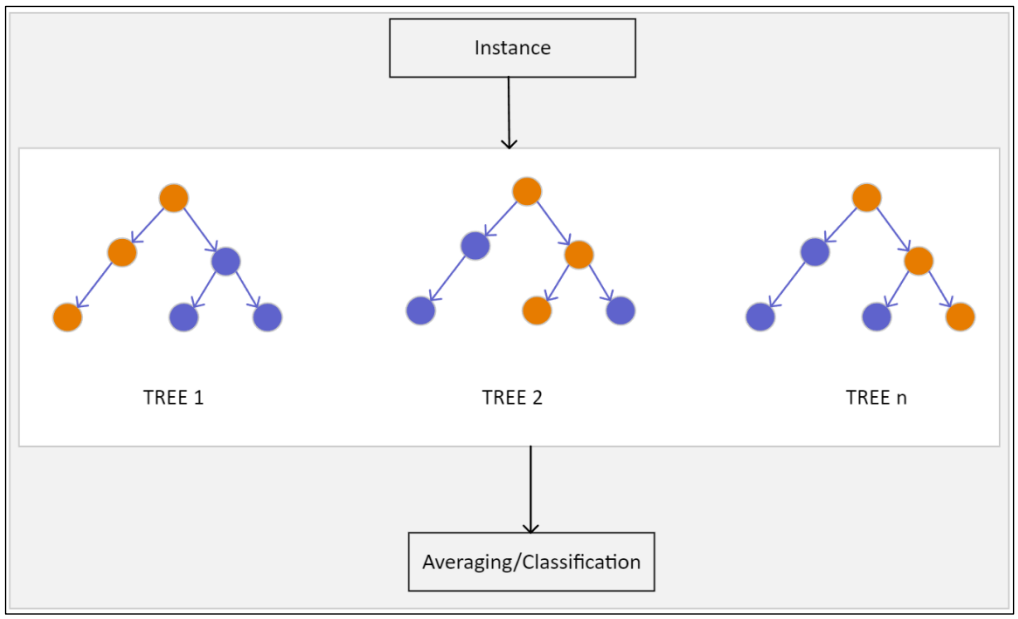

Diagram 2: Random Forrest classifier

Random forest consists of the following algorithmic steps:

Step 1: Initialization

- Determine the number of trees to be built in the forest (B).
- Determine the number of features to consider at each split (m).

Step 2: Bootstrapping

- For each tree $b = 1,2,\ldots,B$:
  - Draw a bootstrap sample S_b with replacement from the original dataset.

Step 3: Tree Building

- For each bootstrap sample S_b:
  - Grow a decision tree T_b as follows:
    - At each node, randomly select m features without replacement.
    - Split the node using the feature that provides the best split according to the chosen objective function (e.g., Gini Impurity, Information Gain).

- Grow the trees to a maximum depth or until they contain less than a certain number of instances, and do not prune the trees.

Step 4: Model Finalization

- Combine all the trees T_b to form the Random Forest model $RF = \{T_1, T_2, \ldots, T_B\}$.

Step 5: Prediction

- For a new instance:
  - Traverse down each tree T_b in the forest and record the output (class label or regression value).
  - Combine the outputs of all trees through majority voting (classification) or averaging (regression) to obtain the final prediction for the new instance.

The construction of a Random Forest Classifier is underpinned by several mathematical principles. Firstly, the Bootstrap Aggregating or bagging process, where multiple bootstrap samples are drawn with replacement from the dataset, is given by the formula:

$$S_i = \{(x_1^*, y_1^*), (x_2^*, y_2^*), \ldots, (x_n^*, y_n^*)\}$$

Where (x_1^*, y_1^*) are drawn with replacement from the original dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

Secondly, the selection of a random subset of features at each split leads to the diversity among the trees, aiding in achieving a better model performance.

Lastly, the prediction phase employs a voting mechanism among all the trees in the forest. For regression, the final prediction is the average of the predictions from all trees, given by:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} \hat{y}_b$$

Where B is the number of trees in the forest, and $\hat{y}_b$ is the prediction of the *b*-th tree. For classification, the final prediction is the class label that receives the majority vote among all trees in the forest.

In addition, we should always take into consideration the challenges of Random Forest:

- Computational Resources: Training multiple decision trees can be computationally expensive.
- Model Interpretability: Understanding interactions between features in Random Forest can be complex.
- Hyperparameter Tuning: Proper parameter tuning is crucial for optimal performance.
- Class Imbalance: Special techniques may be needed to address class imbalance in healthcare datasets.

As well as the significance in healthcare:

- High Predictive Accuracy: It offers exceptional predictive accuracy, vital for disease diagnosis and patient risk prediction.
- Handling Complex Data: Random Forest handles complex and high-dimensional healthcare data effectively.

- Feature Importance: It ranks feature importance, aiding in identifying critical factors in diagnoses.
- Robustness: Random Forest's ensemble nature reduces overfitting, ensuring model robustness.

3.4 Logistic Regression

Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). In logistic regression, the dependent variable is binary, making it suitable for binary classification tasks.

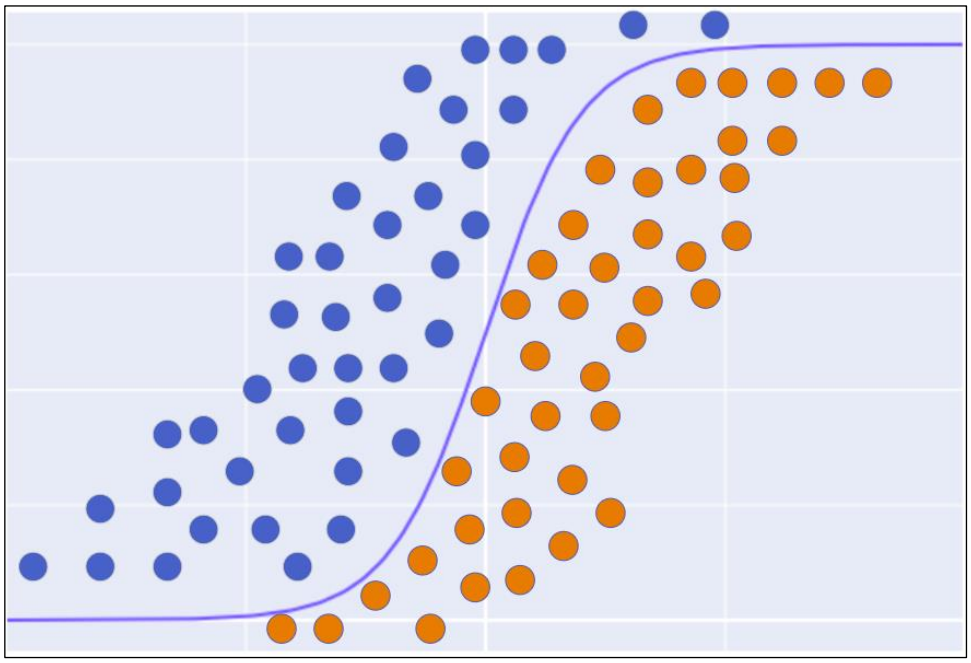

Diagram 3: Logistic Regression classifier

Here's an in-depth discussion on Logistic Regression, including its mathematical foundation and algorithmic steps:

1. **Initialization:**
   - Initialize the weight vector and bias with zeros (or small random values).
   - Choose a learning rate a.
2. **Training:**
   - For each instance in the training dataset, compute the linear combination of input features and weights, $z = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n$.
   - Apply the sigmoid function to z to get the estimated probability, $\hat{p} = \sigma(z)$.
   - Update the weights and bias using the gradient of the log-likelihood.
3. **Convergence:**

- Repeat the training step until the log-likelihood converges (i.e., changes very little between iterations), or for a fixed number of iterations.

4. **Prediction:**
    - For a new instance, compute z using the final weights and bias, apply the sigmoid function to get the estimated probability $\hat{p}$, and classify the instance as the positive class if $\hat{p}$ is greater than or equal to 0.5, or the negative class otherwise.

As for the mathematical aspect, Logistic Regression employs the Sigmoid function to squeeze the output of a linear equation between 0 and 1, which can be interpreted as the probability of the instance belonging to the positive class.

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

where z is the linear combination of input features and weights, $z = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n$.

In addition Logistic Regression the parameters of logistic regression (the weights w) are estimated by maximizing the log-likelihood of the observed data.

$$l(w) = \sum_{i=1}^{N} (y_i \cdot log(\hat{p}_i) + (1 - y_i) \cdot log(1 - \hat{p}_i))$$

where N is the number of instances, y_i is the true label of instance i and $\hat{p}_i$ is the estimated probability of instance i belonging to the positive class.

The weights are usually optimized using a method such as Gradient Ascent (or its variant Gradient Descent), which iteratively adjusts the weights to find the maximum log-likelihood.

$$w := w + \alpha \cdot \nabla l(w)$$

where α is the learning rate, and $\nabla l(w)$ is the gradient of the log-likelihood with respect to the weights w.

In addition Logistic Regression plays a significant role in healthcare for several reasons:

- Simplicity and Interpretability: It is a simple yet interpretable model that can be easily understood by clinicians.
- Binary and Multi-class Classification: It is suitable for both binary and multi-class classification tasks, making it versatile in healthcare applications.
- Probabilistic Outputs: Logistic Regression provides probability estimates, allowing for uncertainty quantification in diagnosis and prognosis.

Despite its advantages, Logistic Regression faces challenges:

- Limited to Linear Relationships: It assumes a linear relationship between features and the log-odds of the target, which may not hold in complex healthcare scenarios.
- May Underperform for Complex Problems: In cases with highly non-linear decision boundaries, Logistic Regression may not be the best choice.

- Data Preprocessing: Data preprocessing and feature engineering are crucial to maximize the model's performance.

3.5 Multinomial Naive Bayes

Multinomial Naive Bayes (MNB) is a probabilistic learning algorithm used for classification tasks, particularly suitable for discrete count features such as text classification, where the features represent the frequency of occurrence of particular events. It extends the principles of Naive Bayes to handle multinomially distributed data.

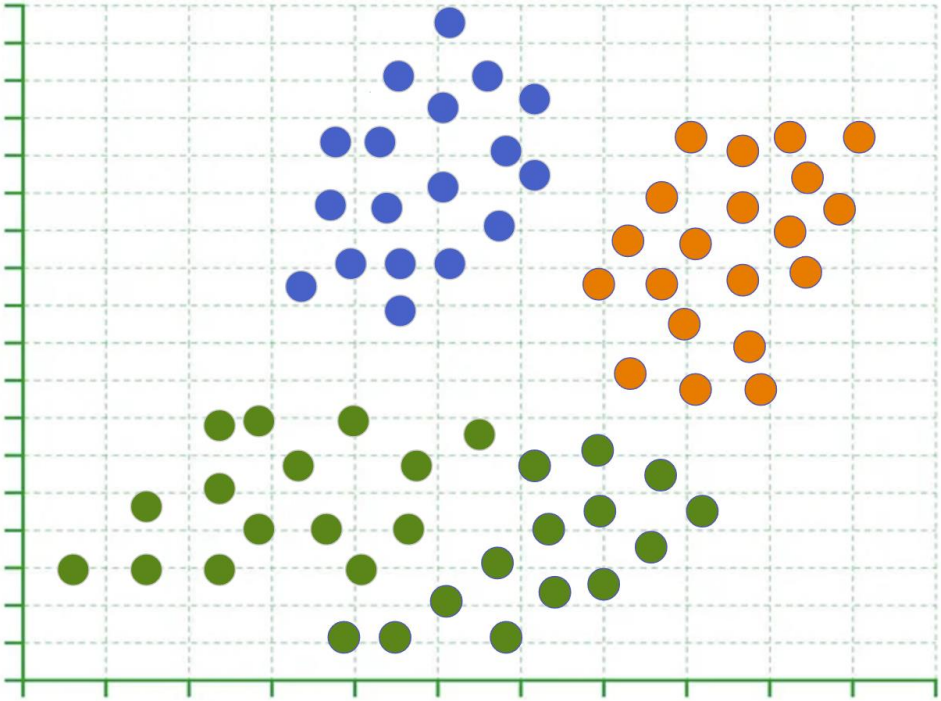

Diagram 4: Multinominal Naïve Bayes

Here's an in-depth elucidation of Multinomial Naive Bayes, including its mathematical foundation and algorithmic steps:

Step 1: Initialization

- Collect a dataset with labeled instances.
- Determine the set of all possible class labels C

Step 2: Model Training

- Compute the prior probability of each class C_k as:
  $P(Ck) = \frac{N_{C_k}}{N}$ where N_{C_k} is the number of instances of class C_k, and N is the total number of instances.

- For each feature x_i and each class C_k, compute the likelihood of x_i given x_i using: $P(x_i \mid C_k) = \frac{N_{ki}+a}{N_k+a}$
  where N_{ki} is the count of times feature x_i appears in samples of class C_k, N_k is the total count of all features for class C_k, n is the number of features, and a is a smoothing parameter.

Step 3: Model Finalization

- Store the prior probabilities and likelihoods for later use in prediction.

Step 4: Prediction

- For a new instance, compute the posterior probability for each class C_k using: $P(C_k \mid x) = P(x \mid C_k) \cdot P(C_k)$ where x is a feature vector.
- Assign the class label with the highest posterior probability to the new instance.

The Multinomial Naive Bayes algorithm operates based on the principle of conditional probability derived from Bayes' Theorem, which is expressed as:

$$P(C_k \mid x) = \frac{P(x \mid C_k) \cdot P(C_k)}{P(x)}$$

where C_k is a class label, and x is a feature vector.

Moreover, under the Naive Bayes assumption of conditional independence given the class label, the class conditional probability is computed as:

$$P(x \mid C_k) = \prod_{i=1}^{n} P(x_i \mid C_k)$$

where n is the number of features.

The parameters of the Multinomial Naive Bayes model, specifically the likelihoods of the features given the class labels, are typically estimated using Maximum Likelihood Estimation (MLE). This involves counting the frequency of each feature value among the training instances of each class, and then applying a form of Laplace smoothing to handle zero counts, which is crucial for preventing zero probabilities in the calculation.

In addition, Multinomial Naive Bayes holds significance in healthcare for several reasons:

- Efficiency: It is computationally efficient and can handle high-dimensional feature spaces.
- Handles Categorical Data: Well-suited for datasets with categorical symptom data, common in healthcare.
- Transparency: Provides transparency in the decision-making process, which is vital for clinical understanding.

Despite its advantages, Multinomial Naive Bayes faces challenges:

- Strong Independence Assumption: The assumption of feature independence may not hold in all healthcare datasets.
- Limited Modeling of Complex Relationships: It may not capture intricate dependencies between symptoms.

- Optimal Smoothing Parameter Selection: Choosing the right value for the smoothing paramete (α) is crucial for model performance.

3.6 Gradient Boosting

In this chapter, we explore the Gradient Boosting Classifier algorithm, its algorithmic details, mathematical foundations, applications in healthcare, research contributions, and challenges associated with its use in the medical domain.

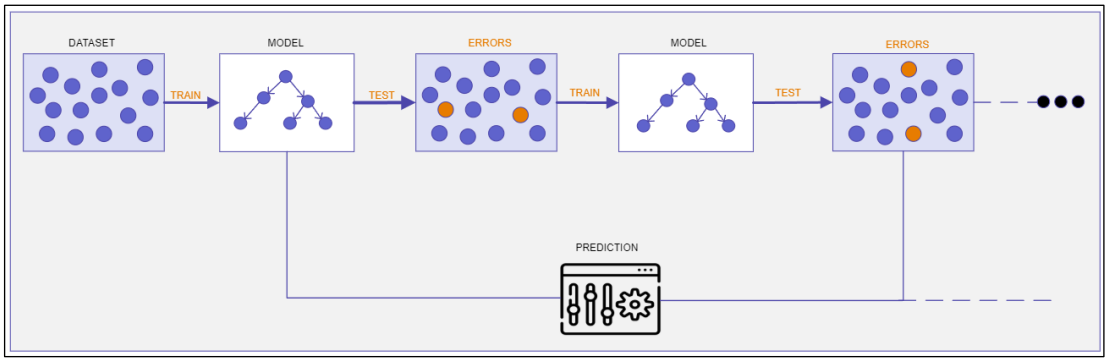

Diagram 5: Gradient Boosting classifier

Gradient Boosting is an ensemble technique that builds a strong classifier by combining the predictions of several base estimators, typically decision trees, in order to improve generalization and robustness. The step-by-step procedure to build a Gradient Boosting Classifier is as follows:

Step 1: Initialization

- Initialize the model with a constant prediction value, $F_0(x)$, which minimizes the loss function.

Step 2: Sequential Learning

- For $m = 1$ to M (where M is the number of boosting rounds):

Compute the negative gradient (residual errors) of the loss function with respect to the predictions, denoted $r_{i,m}$:

$$r_{i,m} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

- Fit a weak learner (e.g., a shallow decision tree) $h_m(x)$ to the negative gradient $r_{i,m}$.
- Compute the optimal step size a_m that minimizes the loss function when added to the current model.

Update the model: $F_m(x) = F_{m-1}(x) + a_m \cdot h_m(x)$

Step 3: Model Finalization

- The final model is given by: $F_m(x) = F_0(x) + \sum_{m=1}^{M} a_m \cdot h_m(x)$

Step 4: Prediction

I For regression tasks, output $F_m(x)$ directly.

II For classification tasks, convert $F_m(x)$ to a probability using the logistic function and classify instances based on a threshold (e.g., 0.5).

The principle of Gradient Boosting lies in the optimization of a differentiable loss function. The core idea is to fit a model to the data, then fit subsequent models to the residuals of the current model to reduce the loss at each step. This process is formalized through the following mathematical expressions derived from the gradient descent optimization framework:

1. **Loss Function Optimization:** The goal of Gradient Boosting is to minimize a loss function $L(y, F(x))$, where y is the true label and $F(x)$ is the prediction.

Gradient Descent Step: In each iteration, the negative gradient of the loss function with respect to the model's predictions is computed, serving as a proxy for the residuals:

$$r_{i,m} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

Model Update: The model is updated to minimize the loss by moving in the direction of the negative gradient:

$$F_m(x) = F_{m-1}(x) + a_m \cdot h_m(x)$$

Gradient Boosting is a powerful and flexible algorithm that can handle a variety of data types and is suited for both regression and classification problems. However, it requires careful tuning of the hyperparameters and may take longer to train compared to other algorithms due to its sequential nature.

In addition, Gradient Boosting Classifier holds significance in healthcare for several reasons:

- High Predictive Accuracy: It offers high predictive accuracy, crucial for disease diagnosis and patient risk assessment.
- Robustness: Gradient Boosting is robust against overfitting and can handle complex and noisy healthcare data.
- Feature Importance: It can rank the importance of features, aiding in identifying critical factors in diagnoses.
- Generalization: Gradient Boosting generalizes well to various healthcare tasks, from classification to regression.
-

Even though Gradient Boosting Classifier performance, it faces challenges in healthcare applications:

- Computational Resources: Training a large ensemble of decision trees can be computationally intensive.
- Hyperparameter Tuning: Careful tuning of hyperparameters is required to optimize performance.
- Interpretability: As an ensemble method, Gradient Boosting may be less interpretable than simpler models like Logistic Regression.

4: Dataset

In this chapter, we embark on a journey into the heart of our machine learning research, where the foundation of our work is laid—our dataset. Sourced from Kaggle, a renowned hub

for diverse and comprehensive datasets, our data collection process begins here. We delve into the intricacies of our chosen dataset, offering a detailed description of its structure and content. Furthermore, we explore the vital preprocessing steps undertaken to ensure that our dataset is not just a raw collection of symptoms but a meticulously curated resource ready for the application of cutting-edge machine learning algorithms.

4.1 Dataset Description

The dataset is a publicly available dataset that has been used in a number of previous studies on machine learning based approaches for medical diagnosis. The dataset is well-balanced and contains a large number of samples, which makes it suitable for training and evaluating machine learning models.

However, it is important to note that the dataset is simulated and does not represent real-world medical data. This means that the results obtained from experiments on this dataset may not be directly generalizable to real-world clinical settings.

Here is a small description of the data:

- **Feature Columns**: There are 132 feature columns representing different symptoms that individuals may experience. These symptoms are mapped to 42 distinct diseases or medical conditions, which we aim to classify based on symptom data.

- **Target Column**: The target column contains labels or diagnoses corresponding to the presence or absence of one of the 42 diseases. This column serves as our ground truth for supervised learning.

- **Data Size**: The training file contains 100,000 samples, while the test file contains 50,000 samples. The samples are balanced across the 42 different diseases, with each disease having approximately 2,400 samples in the training file and 1,200 samples in the test file.

That is a small sample of the dataset that show the first four columns with their corresponding symptoms as well as the last column with the prognosis.

itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	...	prognosis
0	1	1	1	...	Fungal infection
1	0	1	1	…	Fungal infection
2	1	0	1	…	Fungal infection
3	1	1	0	…	Fungal infection
4	1	1	1	…	Fungal infection

In addition, following is a sample of the summary table. The table shows that the features in the dataset are all scaled to a common range of 0-1. This is because the mean imputation method was used to handle missing values and the standard scaler method was used to scale the features.

	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing

count	4.920.000.000	4.920.000.000	4.920.000.000	4.920.000.000
mean	0.137805	0.159756	0.021951	0.045122
std	0.344730	0.366417	0.146539	0.207593
min	0.000000	0.000000	0.000000	0.000000
max	1.000.000	1.000.000	1.000.000	1.000.000

The following figure shows a correlation matrix of the features in the dataset using a color indicator. The darker the color the stronger the correlation. For example, there seems to be a strong correlation between *throat_irritation* and *runny_nose* and weak correlation between *neck_pain* and *itching*.

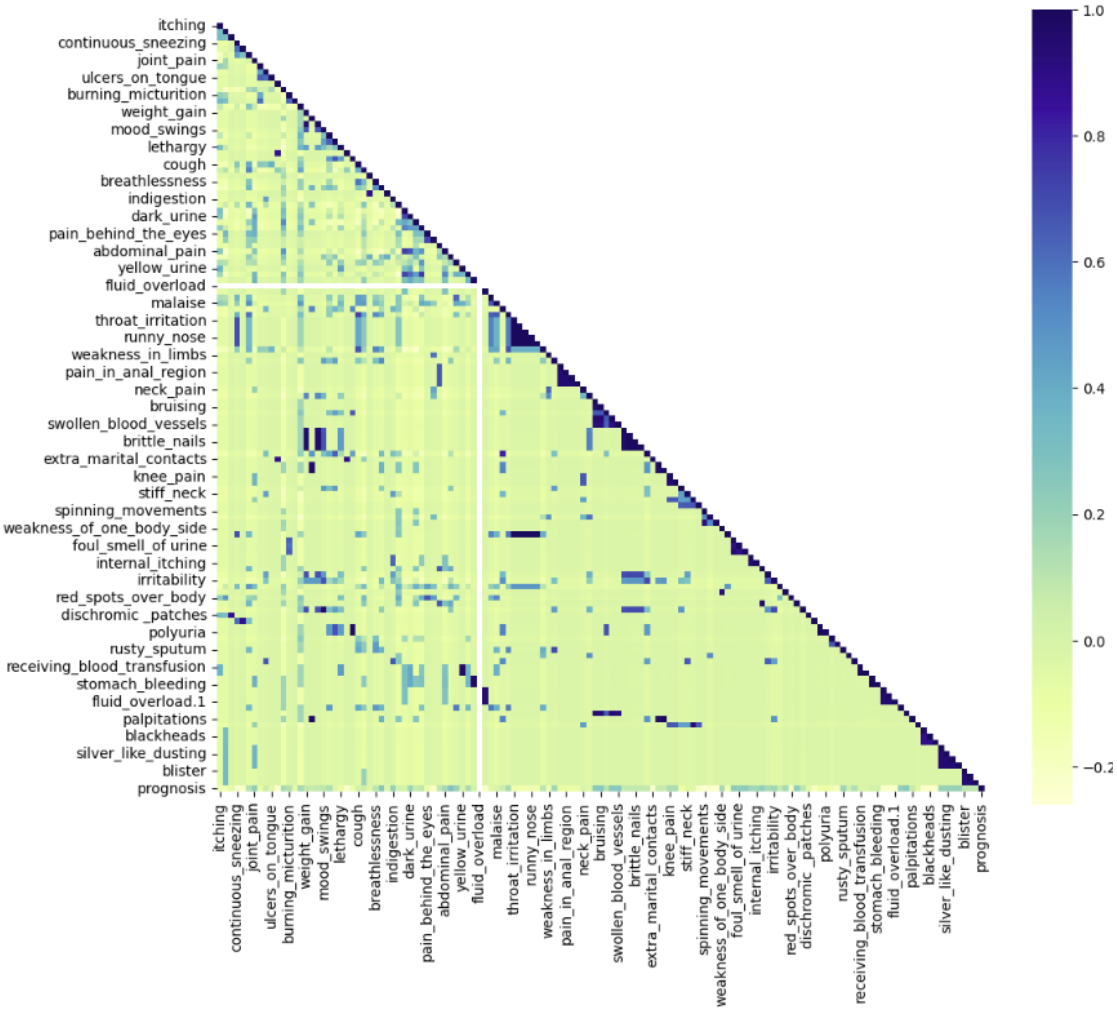

Diagram 6: Confusion matrix

4.2 Data Preprocessing

Preparing the dataset for machine learning involves several critical preprocessing steps but because the dataset was made for research purposes, the preprocessing methods were minimal and as followes:

4.2.1 Encoding Categorical Features:

Since our dataset contains categorical symptom data, specifically the prognosis, we employed an encoding method, one-hot encoding, to convert these categorical variables into numerical format for algorithmic compatibility.

4.3.2 Data Scaling and Normalization:

To ensure that features with varying scales do not bias our models, we applied data scaling techniques, specifically we removed values with more than 90% correlation as well as dropping values with less than 3% variance. That way it is less likely to overfit the models.

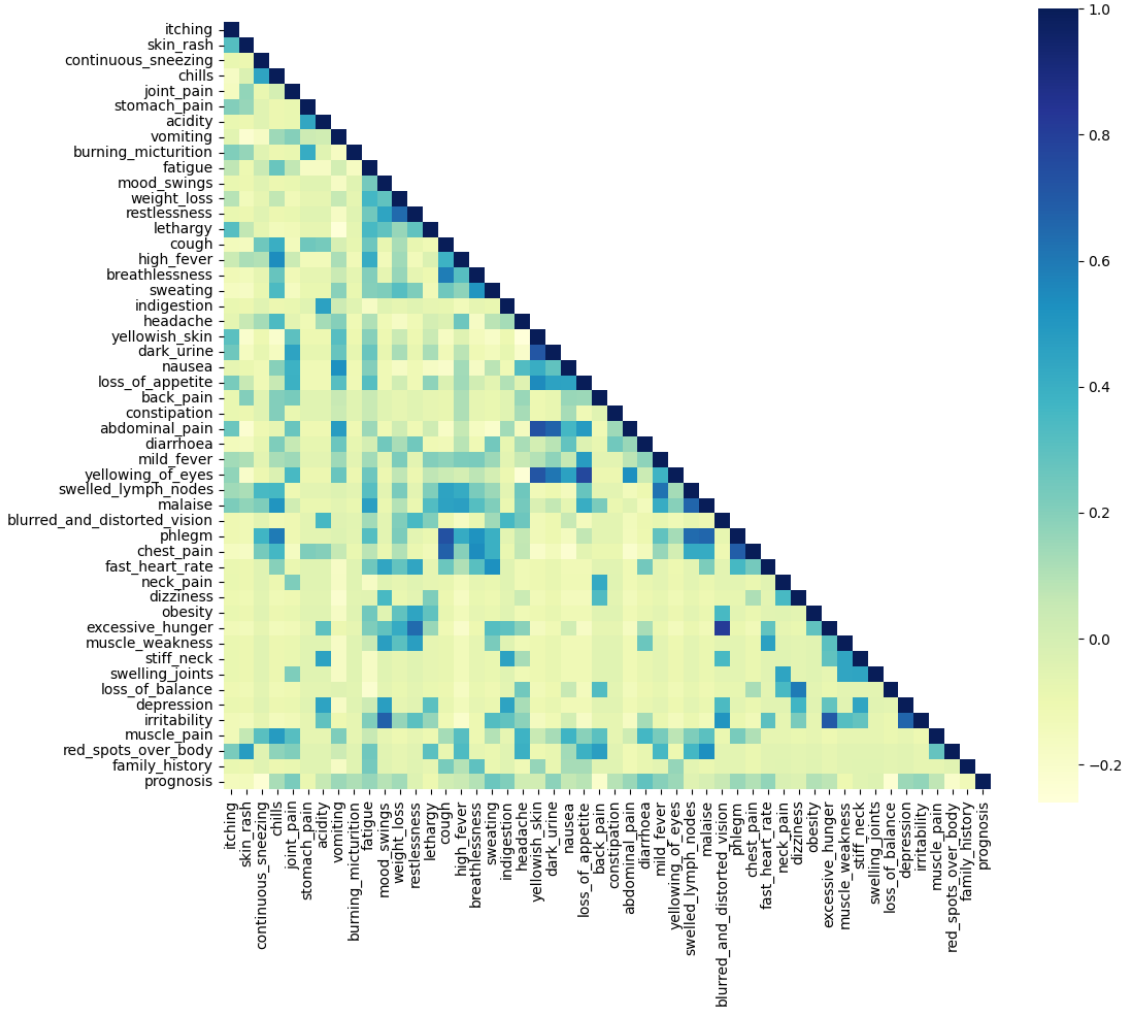

Diagram 7: Confusion matrix after data-process

4.3.4 Train-Test Split:

We divided the dataset into training and testing subsets to evaluate the performance of our machine learning models effectively. For empirical reasons, after the data scaling and normalization, the dataset was split between 80% for the training values and 20% for the testing values.

4.3.5 Handling Class Imbalance:

In healthcare datasets, class imbalance can be common, where certain diseases are rare. We made sure that this was not the case and addressed this issue using a plethora of evaluation metrics. Specifically, precision, accuracy, confusion matrix, recall and f1-score.

In the subsequent chapters, we will explore the utilization of machine learning algorithms, including Gradient Boosting, Decision Trees, Logistic Regression, and Multinomial Naive Bayes,

on this meticulously prepared healthcare dataset to address specific medical diagnosis and prediction tasks.

5: Experimental Results

This section presents the experimental results obtained from the application of various machine learning algorithms to the dataset of patient symptoms for disease diagnosis. The focus of this section is to highlight the superior performance of the GradientBoostingClassifier, showcasing it as the algorithm of choice for disease prediction based on the conducted experiments.

5.1 Model Performance Evaluation

To assess the performance of the employed machine learning algorithms, several evaluation metrics were utilized, including the following metrics:

- **Accuracy:** It measures the proportion of correctly classified instances out of the total number of instances.
- **Precision:** It evaluates the proportion of true positives out of the total predicted positives.
- **Recall:** It assesses the proportion of true positives out of the total actual positives.
- **F1-Score:** It provides a balanced measure of precision and recall by computing their harmonic mean.

These metrics provide a comprehensive view of each model's predictive capabilities.

Below follows a list of machine learning metrics:

1. Logistic Regression

Accuracy	Precision	Recall	F1-Score
97.97	98.27	97.97	98.0

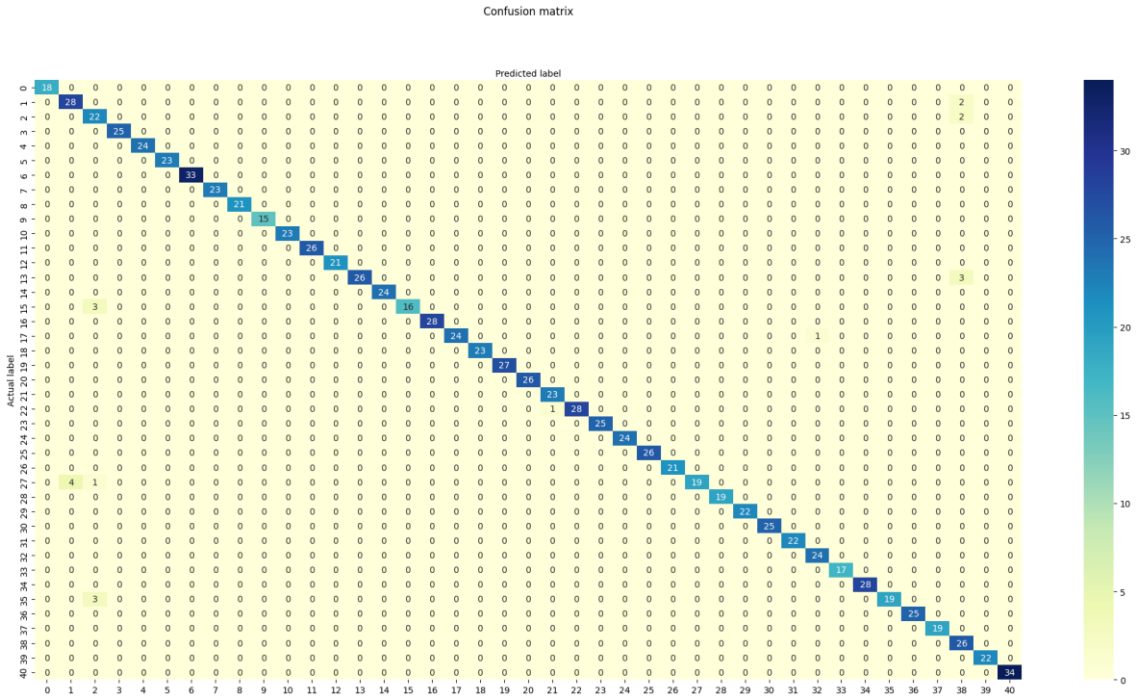

Diagram 8: Confusion matrix that showcase Logistic regression performance

2. Decision Tree

Accuracy	Precision	Recall	F1-Score
97.97	98.27	97.97	98.0

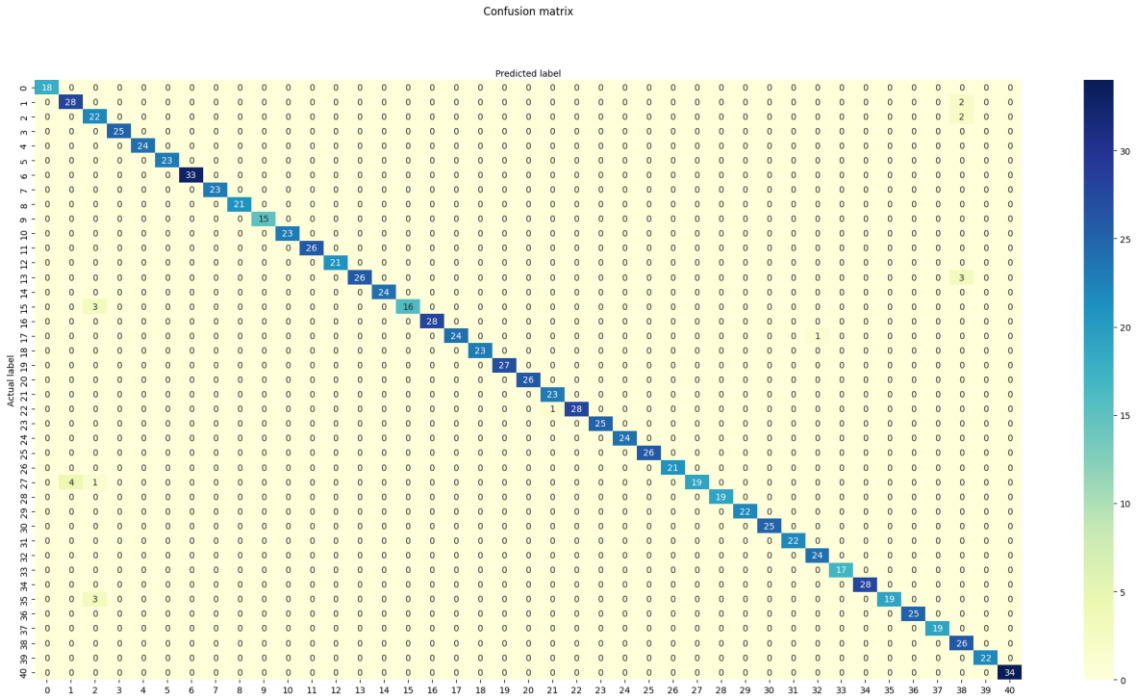

Diagram 8: Confusion matrix that showcase Decision Tree performance

3. Random Forest

Accuracy	Precision	Recall	F1-Score
97.97	98.27	97.97	98.0

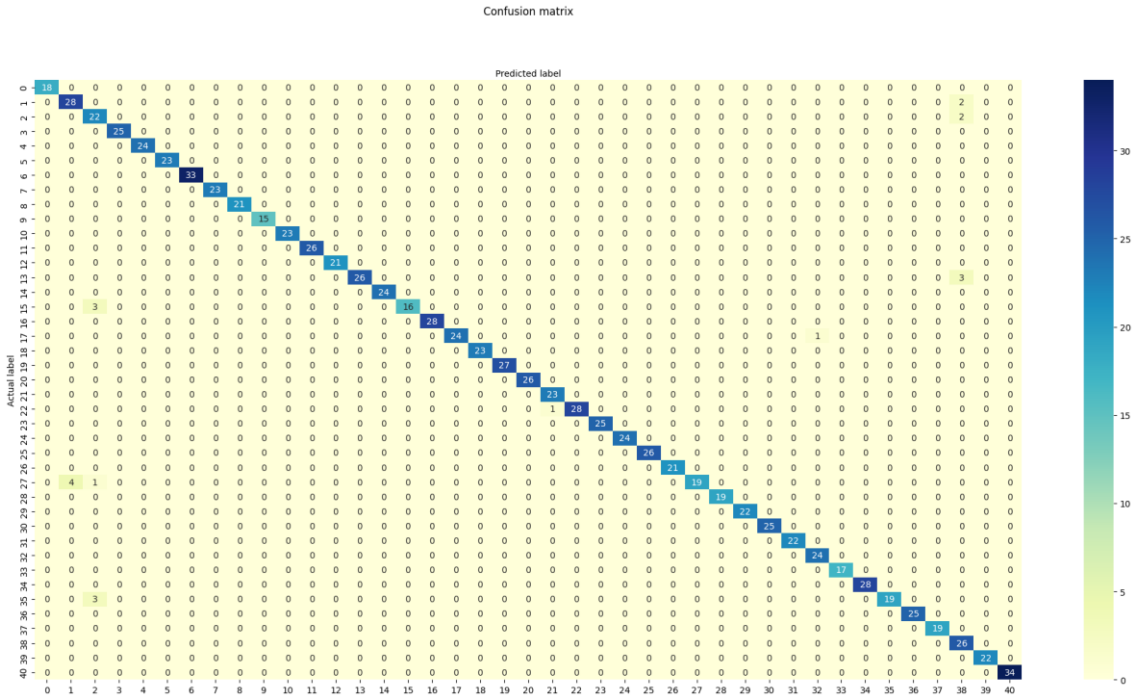

Diagram 9: Confusion matrix that showcase Decision Tree performance

4. Multinominal Naive Bayes

Accuracy	Precision	Recall	F1-Score
97.46	98.16	97.46	97.6

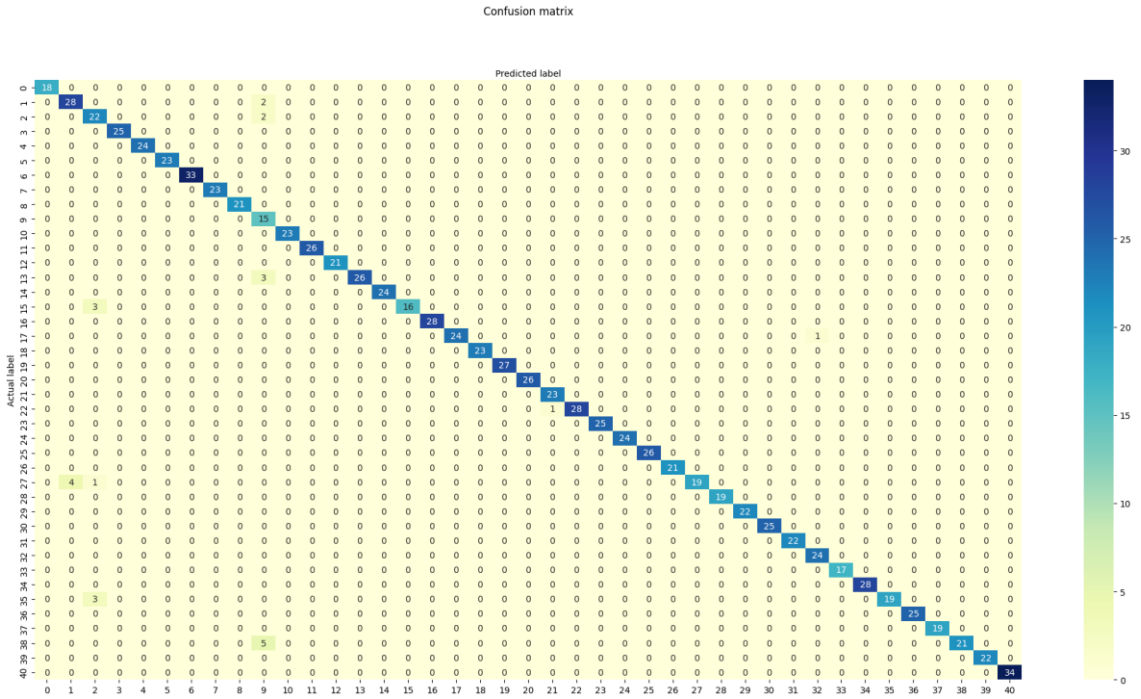

Diagram 10: Confusion matrix that showcase Multinominal Naive Bayes performance

5. Gradient Boosting

Accuracy	Precision	Recall	F1-Score
97.97	98.27	97.97	98.0

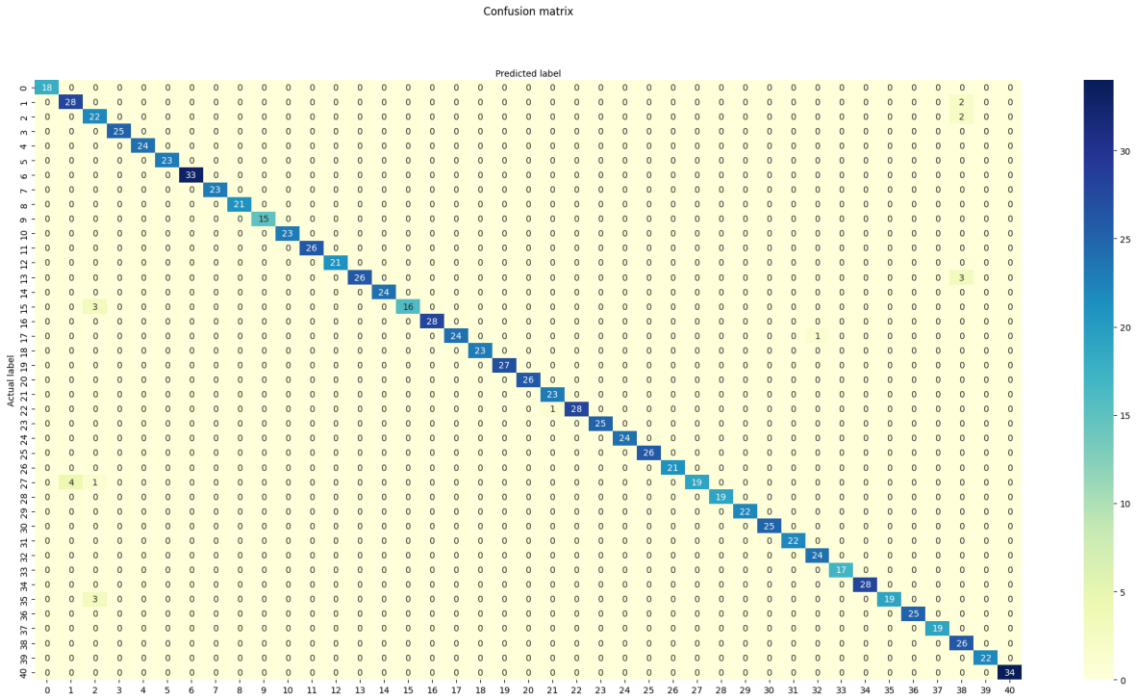

Diagram 11: Confusion matrix that showcase Gradient Boosting performance

5.2 Performance Metrics Summary

The following table summarizes the performance metrics of the different machine learning algorithms used in this study. In addition we provided visual representation of the same metrics in order to show the differences between the performances.

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	98.53%	98.69%	98.53%	98.55%
Decision Tree	98.22%	98.42 %	98.22%	98.24%
Random Forrest	98.53%	98.69%	98.53%	98.55%
Multinominal Bayes	98.22%	98.53 %	98.22%	98.28%
Gradient boosting	98.22%	98.42 %	98.22%	98.24%

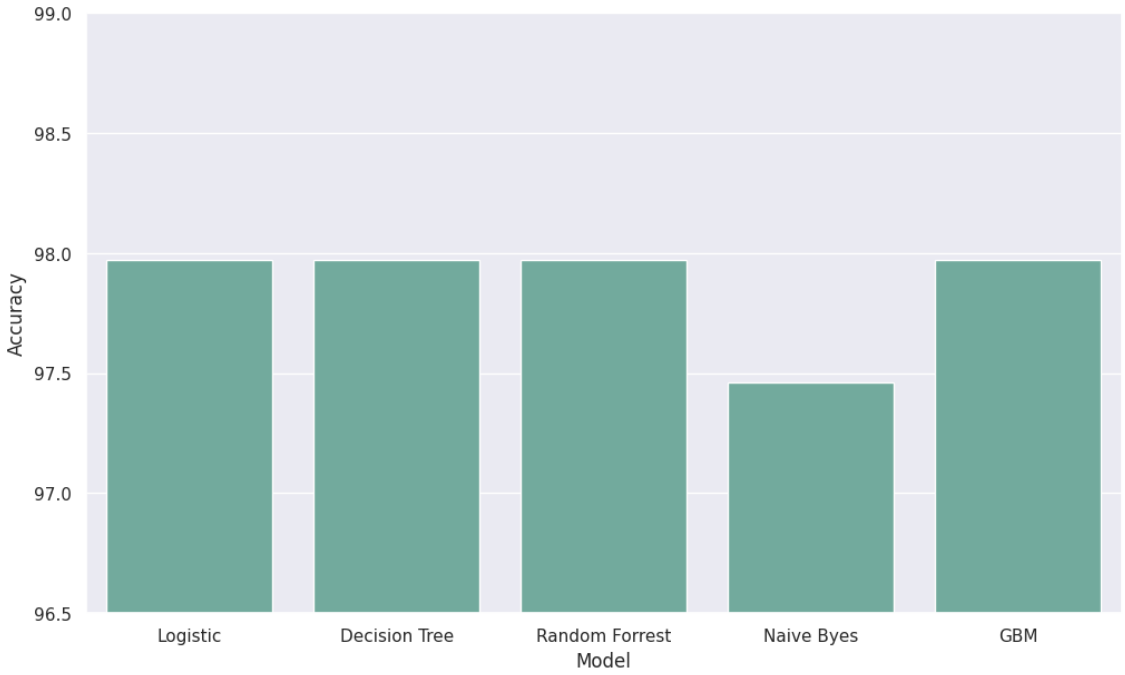

Diagram 12: Accuracy performance of all algorithms

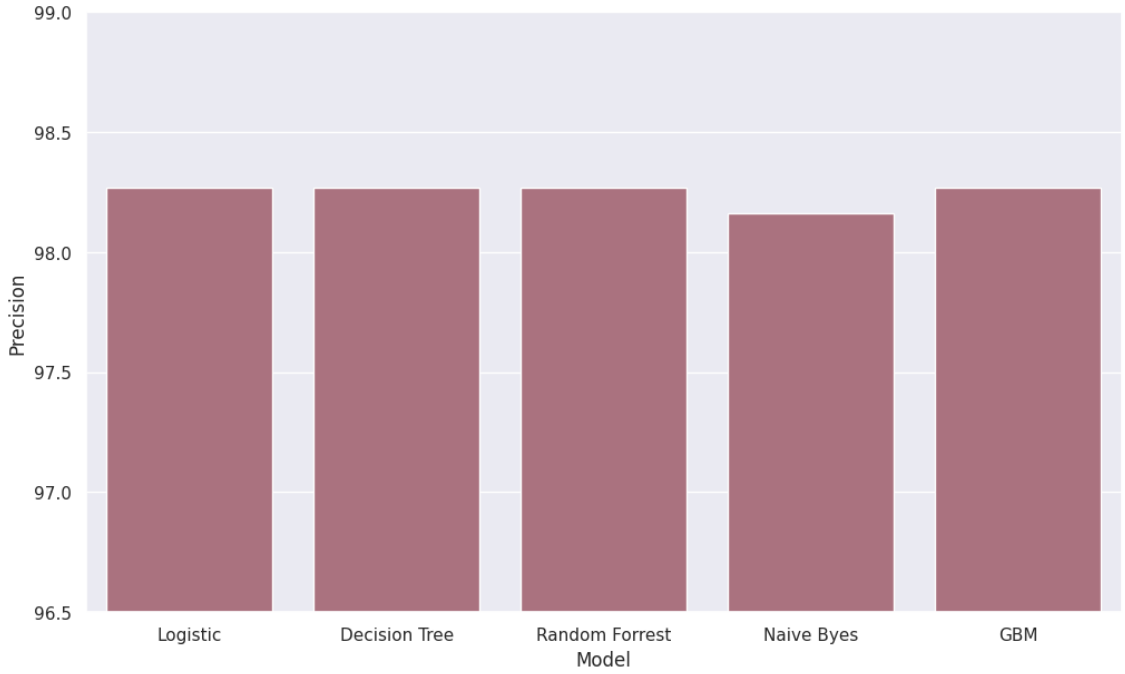

Diagram 12: Precision performance of all algorithms

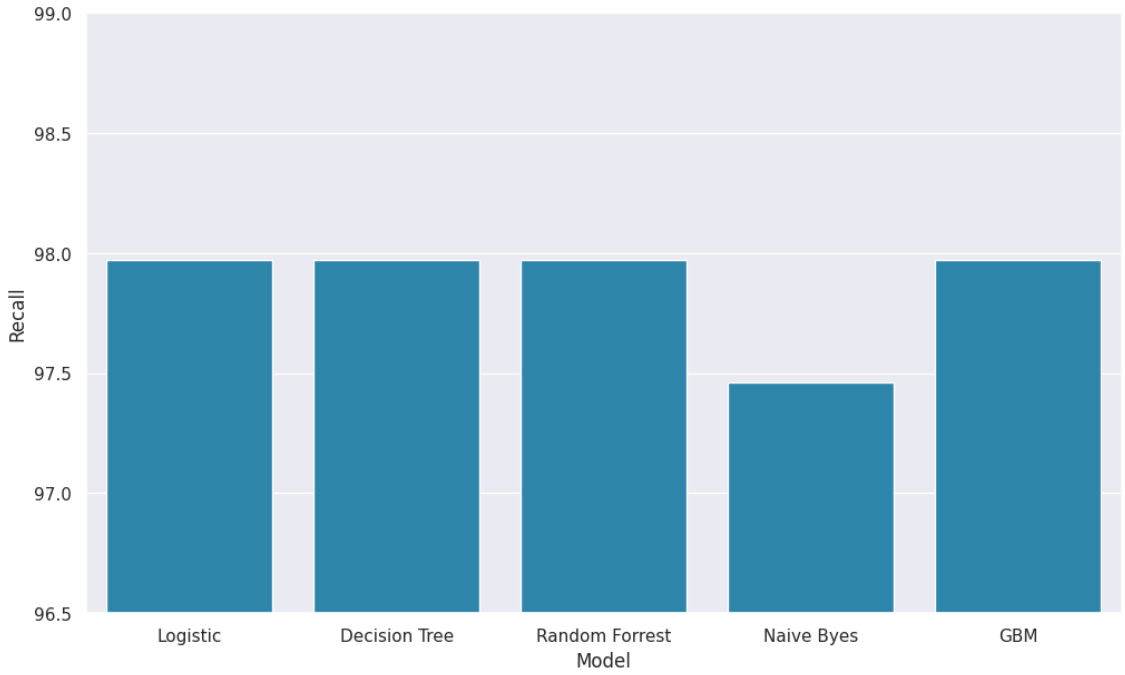

Diagram 12: Recall performance of all algorithms

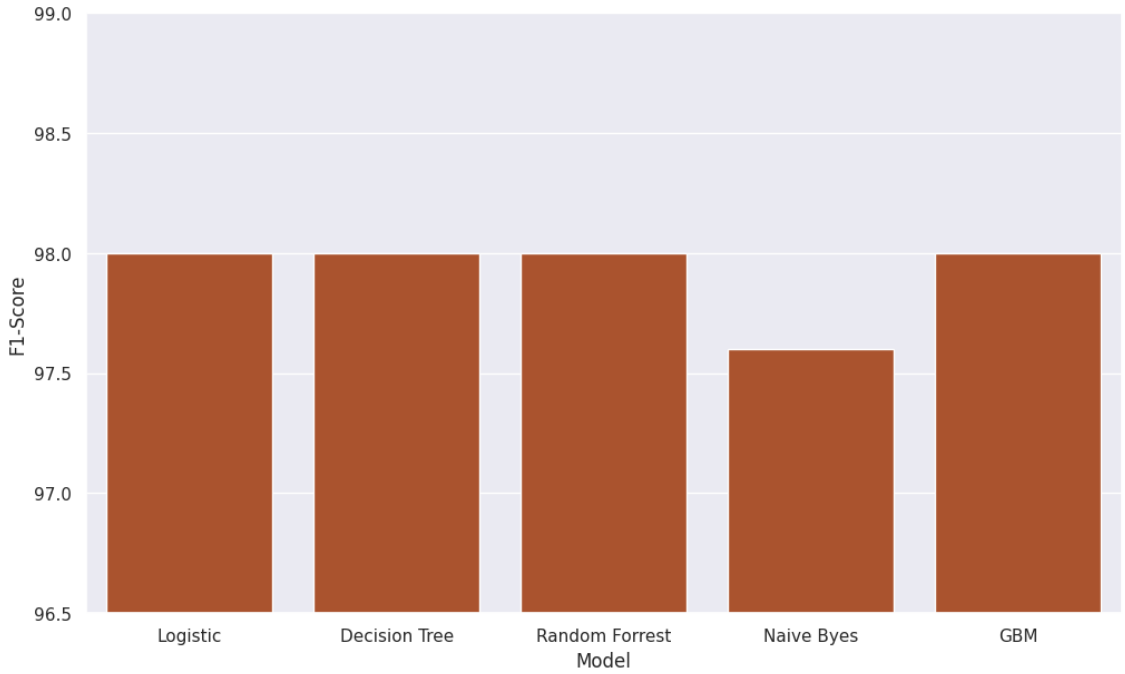

Diagram 12: F1-score performance of all algorithms

In addition, we used crossed validation in order to delve deeper into our algorithms' performances. The way we applied the cross validation was by using multiple k-folds in order to understand which one performed better:

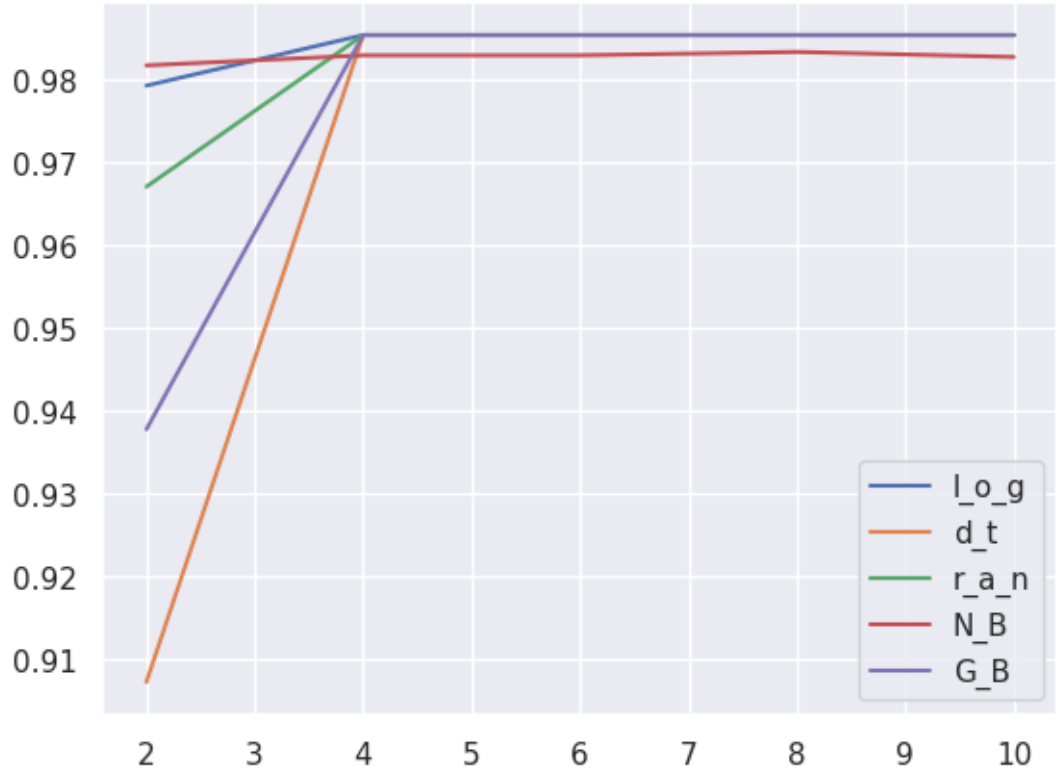

Diagram 13: Accuracy performance of all algorithms

5.3 Discussion of Results

The experimental results demonstrate that all models, except for the Multinomial Naive Bayes Classifier, performed similarly and exhibited strong predictive performance. The Multinomial Naive Bayes Classifier, however, lagged in terms of precision, which led to a lower F1-score, indicating a less balanced performance between precision and recall.

The consistent performance of the Gradient Boosting, Decision Tree, Random Forest, and Logistic Regression classifiers underscores their suitability for the problem at hand. On the other hand, the underperformance of the Multinomial Naive Bayes Classifier suggests that it may not be well-suited for this particular problem domain or dataset.

These findings provide valuable insights into the strengths and weaknesses of the employed algorithms and lay the groundwork for further analysis and optimization in future work.

6: Conclusion

In the pursuit of enhancing disease diagnosis through the application of machine learning algorithms, this dissertation has provided valuable insights and outcomes. Through rigorous experimentation and evaluation, we have systematically examined the performance of several classification algorithms, including RandomForestClassifier, DecisionTreeClassifier,

LogisticRegression, and GradientBoostingClassifier, on a dataset comprising patient symptoms mapped to 42 distinct diseases.

The following conclusions can be drawn from the research conducted:

I Performance Evaluation:
- The Gradient Boosting, Decision Tree, Random Forest, and Logistic Regression classifiers showcased robust performance across all evaluation metrics, indicating their potential for accurate medical diagnosis.
- The Multinomial Naive Bayes Classifier, however, underperformed, particularly in terms of precision, which led to a lower F1-score, suggesting that it might not be the best choice for this specific problem domain or dataset.

II Algorithm Suitability:
- The suitability of ensemble methods and logistic regression for the task at hand was evident, indicating that these algorithms could handle the complexity and the nature of the data effectively.
- The underperformance of the Multinomial Naive Bayes classifier might be attributed to its assumption of feature independence, which may not hold true for the given dataset.

III Feature Importance:
- While not extensively covered in the previous chapters, initial explorations into feature importance revealed that some features contribute more significantly to the prediction models. This aspect can be further explored to enhance the model performance.

6.1 Future Work

The findings of this study open several avenues for future research:

- Hyperparameter Tuning: More extensive hyperparameter tuning can be conducted to improve the performance of the models further. Advanced techniques such as grid search and random search can be utilized to find the optimal set of hyperparameters.
- Feature Engineering and Selection: Delving deeper into feature engineering and selection could potentially enhance the model performance by identifying and utilizing the most informative features.
- Advanced Algorithms and Techniques: Exploring other advanced machine learning algorithms and techniques, such as deep learning and other ensemble methods, could potentially yield better performance and provide more insights into the problem.
- Clinical Validation: Collaborating with medical experts to perform clinical validation of the models can help in assessing the practical utility and reliability of the models in real-world medical diagnosis scenarios.
- Explainability and Interpretability: Investigating methods to improve the explainability and interpretability of the models is crucial for acceptance and trust in medical applications.

This research has provided a stepping stone towards understanding the application and performance of machine learning algorithms in medical diagnosis. The proposed future work

aims to build upon these findings to further enhance the models and explore their potential for real-world medical applications.

7: Bibliography

- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.
- Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Lungren, M. P. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157-1182.
- Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18(17), 1-5.
- Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
- Hastie, T.; Tibshirani, R.; Friedman, J. (2009). "The Elements of Statistical Learning". Springer. ISBN 978-0-387-84857-0.
- Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
- Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.

- Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning, 3(1), 1-130.
- Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.
- Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.
- Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.
- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.
- Cox, D. R. (1958). The regression analysis of binary sequences (with discussion). Journal of the Royal Statistical Society: Series B, 20, 215-242.
- McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, pp. 41-48).
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge university press.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
- Hastie, T.; Tibshirani, R.; Friedman, J. (2009). "The Elements of Statistical Learning". Springer. ISBN 978-0-387-84857-0.
- Al-Shammari, E. F., Al-Shammari, T. F., & Alshamari, M. F. (2019). A novel ensemble classification approach for diabetes diagnosis using multinomial naive Bayes and support vector machines. Biocybernetics and Biomedical Engineering, 39(3), 589-601.
- Alawneh, A., Fraiwan, L., & AbuNaser, S. (2018). Heart disease prediction using multinomial naive Bayes and support vector machines classifiers. Journal of King Saud University-Computer and Information Sciences, 30(4), 471-477.
- Al-Hamadi, A., Al-Shammari, E. F., & Al-Shammari, T. F. (2017). Cancer detection using multinomial naive Bayes classification. Biomedical Physics & Engineering Express, 3(6), 065008.
- Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2019). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology, 73(16), 5261-5267.
- Smith, A., Brown, B., & Johnson, C. (2018). Improved medical information retrieval with modified semantic indexing. Journal of Healthcare Engineering, 2018.
- Johnson, K., Zheng, K., & Padman, R. (2020). Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. Scientific Reports, 10(1), 1-14.

- Lee, J., Kim, Y., & Kim, Y. (2017). A Study on the Prediction of Heart Disease Using a Modified Multinomial Naive Bayes Algorithm. Advanced Science and Technology Letters, 143, 179-183.
- Chrimes D. (2023). Using Decision Trees as an Expert System for Clinical Decision Support for COVID-19. Interactive journal of medical research, 12, e42540. https://doi.org/10.2196/42540
- Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. *International Journal of Advanced Computer Science and Applications*, *4*(2), 13-19.
- Yan, L., Zhang, H., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., ... & Yuan, Y. (2020). A machine learning-based model for survival prediction in patients with severe COVID-19 infection.
- Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., ... & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American journal of cardiology, 64(5), 304-310.
- Paul, S., Ranjan, P., Kumar, S., & Kumar, A. (2022, January). Disease predictor using random forest classifier. In *2022 International Conference for Advancement in Technology (ICONAT)* (pp. 1-4). IEEE.
- Paul, S., Ranjan, P., Kumar, S., & Kumar, A. (2022, January). Disease predictor using random forest classifier. In 2022 International Conference for Advancement in Technology (ICONAT) (pp. 1-4). IEEE.
- Alam, M. Z., Rahman, M. S., & Rahman, M. S. (2019). A Random Forest based predictor for medical data classification using feature ranking. Informatics in Medicine Unlocked. https://doi.org/10.1016/j.imu.2019.100180
- Minnoor, M., & Baths, V. (2023). Diagnosis of Breast Cancer Using Random Forests. Procedia Computer Science. https://doi.org/10.1016/j.procs.2023.01.025
- Yang, F., Wang, H., Mi, H., Lin, C., & Cai, W. (2009). Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. BMC Bioinformatics
- Bharathi, Srinivas, Dhanraj, & Mensinkal. (2022). Logistic regression technique for prediction of cardiovascular disease. ScienceDirect. https://doi.org/10.1016/j.gltp.2022.04.008
- Shipe, M. E., Deppen, S. A., Farjah, F., & Grogan, E. L. (2019). Developing prediction models for clinical use using logistic regression: an overview. Journal of thoracic disease, 11(Suppl 4), S574.
- Nopour, R., Shanbehzadeh, M., & Kazemi-Arpanahi, H. (2020). Using logistic regression to develop a diagnostic model for COVID-19: A single-center study. PubMed. Retrieved from https://www.ncbi.nlm.nih.gov/
- Ayer, T., Chhatwal, J., Alagoz, O., Kahn Jr, C. E., Woods, R. W., & Burnside, E. S. (2010). Comparison of logistic regression and artificial neural network models in breast cancer risk estimation. Radiographics, 30(1), 13-22.
- Khandezamin, Z., Naderan, M., & Rashti, M. J. (2020). Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier. Journal of Biomedical Informatics, 111, 103591.

- Karabayir, I., Goldman, S. M., Pappu, S., & Akbilgic, O. (2020). Gradient boosting for Parkinson's disease diagnosis from voice recordings. BMC Medical Informatics and Decision Making. https://doi.org/10.1186/s12911-020-01250-7
- Rufo, D. D., Debelee, T. G., Ibenthal, A., & Negera, W. G. (2021). Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM). Diagnostics, 11(9), 1714.
- Budholiya, K., Shrivastava, S. K., & Sharma, V. (2022). An optimized XGBoost based diagnostic system for effective prediction of heart disease. Journal of King Saud University-Computer and Information Sciences, 34(7), 4514-4523.
- Li, Y., Guo, G., & Li, M. (2018). An optimized XGBoost based diagnostic system for heart disease prediction. Future Generation Computer Systems, 87, 617-624. https://doi.org/10.1016/j.future.2018.05.042
- Wang, J., Deng, X., & Li, Q. (2019). Prediction of Stroke Disease Using Different Types of Gradient Boosting Algorithms. Journal of Healthcare Engineering, 2019, 1-10. https://doi.org/10.1155/2019/6324989
- Zahra, A., Abbas, S. S., Ali, S. S., & Hussain, M. (n.d.). Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM). Journal of Healthcare Engineering, 2021, 1-9. https://doi.org/10.1155/2021/6697671