PREDICTIVE MODELLING IN HEALTHCARE USING MACHINE LEARNING

[DATE]

STUDENT NAME
STUDENT ID

Table of Contents

Chapter 1: Introduction

1.1 Chapter overview

This chapter is dedicated to describe how the use of predictive modelling is revolutionising healthcare through the integration of artificial intelligence (AI) and machine learning (ML). Thus, it offers some historical background, stresses the implementation of such technologies when it comes to processing health information, and reveals the transition from conventional paradigms to innovative data-driven ones. Major improvements, issues, and controversies, together with the role of validation of models, and the influence of computational tools. The chapter creates the backdrop for assessing the performance of various machine learning models in estimating healthcare outcomes by defining the study aim, objectives, and importance.

1.2 Background

Data analysis has rapidly assumed the role of assisting healthcare practitioners in the determination of outcomes which are expected to manifest themselves over time, it has been referred to as predictive modelling. AI and ML are also fostering this change since they allow the creation of supple and accurate algorithms that analyse great amounts of data to forecast patient outcomes. Healthcare forecasting is a wide concept that defines a set of methodologies and technologies being aimed at the anticipation of the future health events for the more effective treatment tactics providing the better care results.

Thus, it is important to note that, until recently, most healthcare decisions have been made based on the clinician's expertise, past experience, and empirical evidence and statistical modelling techniques. However, the development of health data and the need for more advanced analytical tools due to differentiation of individuals' diseases have emerged. This is where the use of predictive modelling comes in, adding a new dimension of working with the algorithms that work with historical and current data to make predictions (Collins *et al,* 2021). It opens an opportunity for the new approach to healthcare management as well as patients' management that is based not only on the insights that cannot be received by the help of the conventional analytical tools.

Another of the fundamental areas of progress in predictive modelling is the application of AI and ML. They can handle immense amounts of heterogeneous data in the form of EHRs, imagery and data from wearable devices. Dynamic patterns of this data are explored by advanced methods, like the neural networks and the support vector machines, and that results in capability to forecast the disease's new stage, or, to specify the high vulnerability patients, or, to offer an individualised therapy. For instance, there is a possibility of developing chronic diseases such as diabetes or cardiovascular diseases and other diseases depending on the history, genetic makeup, and other factors regarding an individual patient.

The move to predictive modelling augmented by artificial intelligence entails the growing importance of the health sector in modern society. Historically, the approaches used for identifying key patterns and trends relied on people's interpretation, which entails the need to consider that it is often laborious and has limitations of subjective inaccuracies. Analytic models, on the other hand, present a facts-based technique that improves on accuracy as well as neutrality (Alowais *et al,* 2023). These models can also offer an almost real-time field of prediction and can update its field of prediction as more field input information is obtained and analysed constantly through education. This capability is especially useful in dealing with multifaceted and long-term illnesses as early identification can determine the patient's prognosis.

However, there are certain issues and concerns that must be taken into account concerning predictive modelling. Another significant issue is the indefensible nature of the data that go through training of these models (Riley *et al,* 2020). Forecasts are only as good as the data that go into them and data must be properly collected and managed to be properly representative of the population. Problems like data leakage, absence of data, cases of equal data and instances of non-symmetric data can weaken or even nullify forecast models. As a result, the researcher and practitioners must be careful to undertake comprehensive, accurate, and up-to-date information collection.

One more essential area is the possibility of explaining the results of the developed predictive models (Sui *et al,* 2020). Whereas simple models e.g., linear models may not be as accurate as deep learning algorithms, these models are understandable, and one can easily follow the steps to get to a particular prediction. He identifies this as the key reason why the use of predictive modelling in clinical contexts may be slowed because of the lack of transparency. Proposals to improve interpretability, for instance, through the creation of new approaches to explainable

AI and by ensuring that the research published with models contains clear descriptions of the processes that they are based on, are also highly valuable to enable the effective use of predictive modelling in clinics.

Another factor that has an influence on the predictive models is the ethical implication that is involved in the deployment of the models. To this end, challenges emerging in data privacy and security and biassed models negatively affecting the patient population must be solved to make predictive modelling's impacts beneficial in-patient care. For instance, if datasets have captured prejudiced information or algorithms themselves also have prejudiced characteristics, healthcare provisions will differ across the groups of people. Here are the best practices towards fairness and equity models: Data governance practices must be strong and audited frequently; people apart from data scientists should participate in the creation and evaluation of the models.

Further, modern developments in predictive modelling also stress on validations as well as monitoring on a regular basis. The usefulness of the predictive models must also be evaluated to check their reliability after some period due to change in healthcare setting (Kaushik *et al,* 2020). The measures like cross validation or external validation like independent data validation are used to assess the fitness of the model as well as to prevent overfitting of the model. Also, the constant reassessment permits incorporating new information and tendencies that impact healthcare into the evaluation process.

The incorporation of predictive modelling into healthcare systems does not only benefit from the creation of improved computational techniques, but also from the improvement of related structural frameworks. It is also due to the advanced technologies in cloud computing and big data that large scale health data can be stored, processed, and analysed for the use of state of the art predictive models. The constants also enable real-time analytics and decision making; this means that healthcare personnel are armed with valuable information at the time of interaction with the patient.

Thus, predictive modelling is a major step forward in healthcare as it promises to give caregivers and patients tailored directions for treatment based on the data gathered. Despite the optimistic potential of applying AI and ML methods in predictive modeling, the approach has several limitations due to the peculiarities regarding the data quality, the interpretability of the models, or applying ethical standards. Solving these problems is possible only by cooperation between researchers, clinicians, and policymakers to make the use of predictive modelling

efficient and responsible (Yang *et al,* 2020). Just as the future of healthcare continues to become more uncertain, so does the future of this field, which means that it is going to be crucial to fold new strategies and methods to the Practice of predictive modelling in order to unlock its full potentiality.

1.3 Research Aim

To develop and evaluate a predictive model for healthcare outcomes using python, employing multiple machine learning techniques to identify the most accurate model for predicting patient health metrics based on the provided dataset.

1.4 Research Objectives

- To implement and compare fthe different machine learning models such as, Support vector machine, Randoforest, Gradient Boosting Machine and the logistic regression using python to determine their effectiveness in predicting healthcare outcomes with the cross validation approach.
- To analyze the performance of each model in terms of accuracy, precision, recall, and overall predictive power.
- To identify the most suitable machine learning model for healthcare predictions based on the dataset.
- To provide recommendations for optimizing predictive modelling in healthcare settings based on the findings.

1.5 Research Questions

- How do various machine learning models compare in terms of accuracy, precision, recall, and overall predictive power when applied to healthcare outcome prediction?
- Which machine learning model demonstrates the highest effectiveness for predicting patient health metrics based on the provided dataset?
- What strategies can be recommended for optimizing the implementation and performance of predictive models in healthcare settings based on the comparative analysis of different models?

1.6 Research Significance

The application of prediction models in the healthcare sector is of vital importance because of the likelihood to lead to better improvement of patient experience as well as efficient organization of patient care. Using supercomputers and big data computing abilities, predictive models provide the chance to look at the future, recognize the more vulnerable clients, and build a treatment program (Al-Tal *et al,* 2021). Such a strategy can result in the identification of illnesses at an early stage, refinement of risk estimations, and appropriate interventions to enhance the patients' outcomes and alleviate the pressure from the healthcare facilities.

Predictive modeling's importance can be seen by the potential to turn disease management and the practice of preventive measures into a science (Jewell *et al,* 2020). For instance, it can predict how often a person is likely to get a disease like diabetes or cardiovascular disease which can then be prevented before it even occurs, based on lifestyle changes (Zhang *et al,* 2022). It also prevents the worsening of diseases, creating less concern in terms of unexpected symptoms and medical expenditures on procedures linked with hospitalizations.

Furthermore, predictive modelling increases capacity utilization through rational resource management and the flow of clinical processes. Measures of admission rates and treatment demands allow healthcare facilities to plan ahead and direct resources to client needs.

1.7 Research Justification

The study of predictive modelling in healthcare is critical because technological progression is standardising healthcare requirements that are becoming costly and comprehensive. Modern forms of treatment seek to treat each illness once it shows symptoms and thus operate on a very limited roadmap unlike the old approach to treatment (Khemasuwan *et al,* 2020). On the other hand, predictive modelling uses the past data and statistical and logical analysis to anticipate the different health risks, thereby allowing a smooth transition to preventive health care.

The rationale for this research is based on the development of best solutions to counter key problems affecting the health sector. Ambient Intelligence systems can also contribute a lot to improvements in early indications of the diseases and their health implications so that the risks of long-term complications and hospitalizations can be prevented out rightly (Paulus and Kent, 2020). This helps to also reduce patients' longevity and quality of life as well as decreases the

amount of money spent on the healthcare systems especially in cases that require emergency interventions.

Also, since the volume and quality of healthcare data are growing rapidly, there is an urgent need to enhance and advance methods for the predictive model. Studying in this area should result in developing better models that can be practically employed without large discrepancies (Wong *et al,* 2021). Thus, this research is essential to develop healthcare practice, improve the standards of patient care, and decrease the economic costs of providing the necessary healthcare services.

1.8 Structure of Report

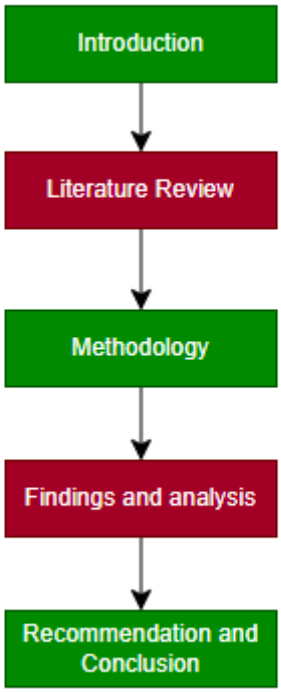

Figure 1: Structure of Report

(Source: Self-created)

Chapter 2: Literature Review

2.1 Introduction

The general application of the predictive modeling in healthcare has become a popular technique to enable enhancement of patient care and health management. As a futile review of AI-based predictive models, henceforth articulated in this literature review, these models' evolution, use, and issues are pinpointed. Some of the explored topics include; improving interpretability, combination of structured and unstructured data, bias, and use of causal inference. This also looks into how these models can be used in different aspects of healthcare, including chronic illness, epidemics among others. It is therefore the goal of this systematic review to combine the evidence from various available studies and give a brief state of the art in terms of predictive modeling for health care while also highlighting potential gaps in the literature.

2.2 Enhancing Predictive Modelling in Healthcare

It signifies that predictive modeming using artificial intelligence is a radical shift that can redefine possibilities of how user can predict the patient's outcomes as well as the subsequently how they should be treated. As noted in the scoping review conducted de Hond et al. (2022), there is a vital need to come up with proper guidelines and quality criteria for the improvement of AI-based prediction models in the field of healthcare delivery.

The authors comprehensively synthesize the literature data to discuss the main suggestions concerning the creation and deployment of AI-related prediction models. These principles focus on reporting and they stress on issues like openness or replicability. It is crucial to document model creation, data, the procedure to prepare data, and choosing machine learning algorithms well. Solutions to improve reproducibility include supplementing manuscripts with codes and datasets on which the analysis was conducted to allow the replication of results by other scholars. Another important aspect regarding the review is the aspect of validation of the used AI models. The authors emphasize the importance of the subsequent, stricter external verification by modest sets for assessing the model's applicability in other population groupings and contexts. It points out overfitting where, models are capable of explaining training data well but are poor at predicting new data as well as suggesting the use of cross validation and Bootstrapping to combat it.

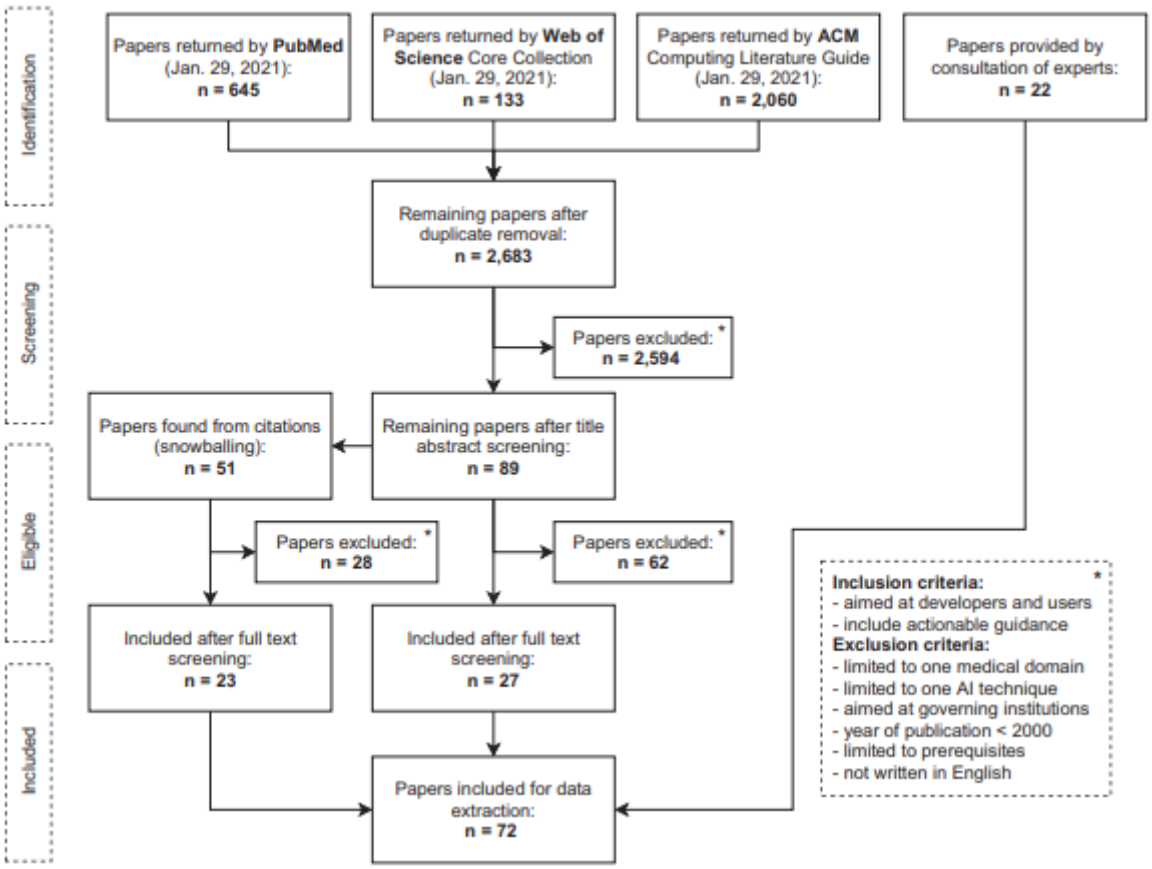

Figure 2: Flow diagram of screening strategy

(Source: Hond *et al*; 2022)

The review also accents the importance of interpretability and explainability of AI models. To give the healthcare workers' trust and actively integrate the predictions into decision making, the AI models need to explain clearly what has been done to produce the predictions. The authors have also stressed that the communities should encourage the construction of interpretable models or utilize instruments that can explain how the machine learning recommendation occurred. Sustainability of the application of information technologies is an ethical consideration in healthcare. De Hond et al. talk about the common issues of AI ethics, including bias and fairness. In their opinion, there is a need to pay attention to the sources of data used for training the models to prevent the reinforcement of the existing bias and inequalities in the provision of healthcare.

Last, the features valid for AI model assessment are summarised as well as the need for AI model constant monitoring and updating. The issue this raises is that as the data and processes of delivery in the health care processes change, the models need to be changed periodically.

The authors suggest creating guidelines for the further performance assessment of the scale and using information on its use in clinical practice. De Hond et al conduct a comprehensive analysis of the guidelines as well as other quality criteria that are vital in the deployment of AI based methodologies of prediction in the VPHC. Their findings are useful in establishing sound and reliable principles of AI that can improve patients' outcomes significantly and also if reflective of their standards of ethical conduct, they are valuable in creating AI systems with ethical standards.

2.3 Enhancing Interpretability in Machine Learning-Based Healthcare Prediction Models

Machine learning-based predictive models are the newest perspective in healthcare, which may significantly change patient treatment. Nonetheless, there is difficulty in interpreting such models especially at the course of the next steps. In the context of their research on decision-making processes involving ML-based prediction models in healthcare, Stiglic et al. (2020) discuss interpretability and mark its improvement as one of the IR-abilities. In their research, Stiglic et al. (2020) emphasize that the recognition and acceptance of ML models is crucial as well when it comes to the healthcare professionals' trust in the models. Thus, transparent models help clinicians to open their understanding on how the prediction was made and make the decision. The authors stress that without interpretability in clinical settings, even highly accurate models may be dismissed due to the lack of clarity of the model's decision-making processes.

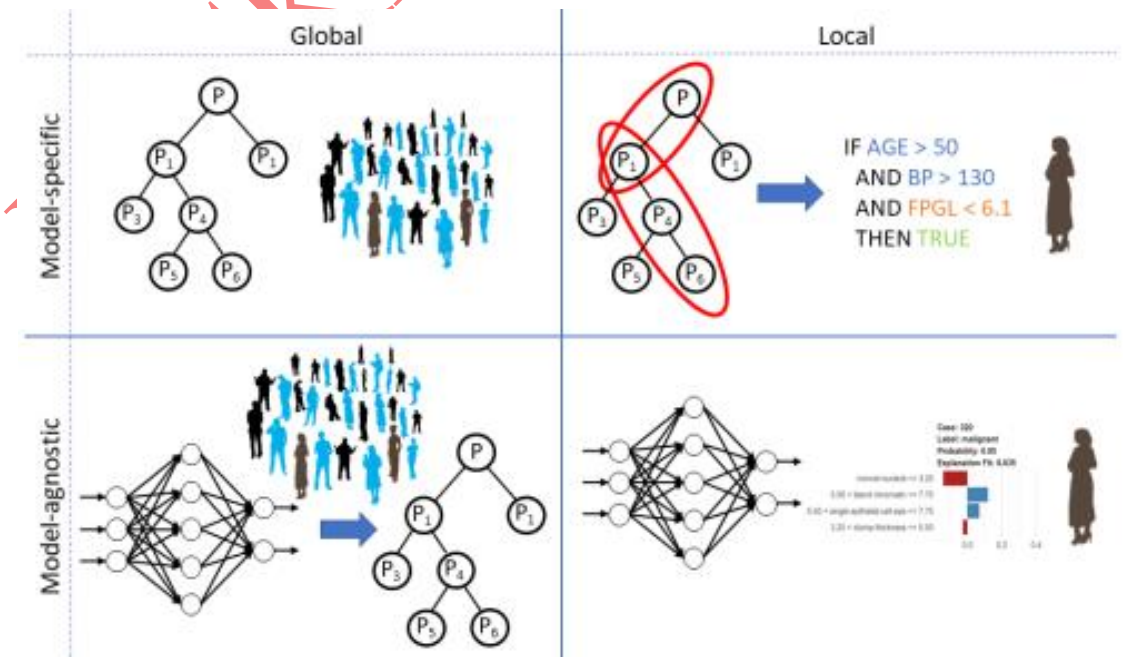

Figure 3: Machine learning models for prediction in healthcare

(Source: Stiglic *et al*; 2020)

The review groups interpretability methods into inherent methods, as well as additional methods applied afterward. Intrinsic interpretability is in where the models are interpretively built, for example, the decision trees and linear models. Despite being less complex and easier to interpret, such models tend to have lower accuracy in the predictions as compared to models such as deep neural networks. Interpretability techniques are used after building complex models to make the model's predictions easier to comprehend. Stiglic et al. (2020) describe a set of methods: feature importance scores, partial dependence plots, and surrogate models. These methods are intended to reveal the ways in which certain features affect the model's predictions, which is a form of a middle ground between model complexity and assessability.

According to the authors, there is also the question of whether one aims at providing very accurate estimates of the population parameters or at producing results which are easy to interpret. They however explain that although the described complex models might have higher accuracy, their incomprehensibility may limit their usability in an environment like health care. The issue here is to find the right tradeoff as to how much to invest in pre/post-processing, depending on several factors, most importantly the clinical context and demands on healthcare workers.

In addition, Stiglic et al. emphasize that 'interpretability is not just a technical issue', which means that these questions are equally important concerning their ethical aspect. ML algorithms perform exceedingly well when they are designed with immense transparency in mind, which enables one to notice and eliminate bias in predictive models. This is particularly a problem of interpretability because it would be difficult to identify such bias and its impact on the patient. The authors urge for the assessment of the project interpretability analysis to be comprehensive so as to remove this bias.

Stiglic et al. (2020) put together a state-of-the-art guide on the significance and ways of boosting interpretability for ML-backed health care prediction models. This means that, they acknowledge the significance and importance of deploying a balanced approach that shall enable the preservation of two principles; the predictive accuracy and the ability to t=be trusted to ensure the actualization of a successful fusion of ML into the clinical practice.

2.4 IoT-Enabled Predictive Modeling for Disease Outbreak Detection

The use of IoT with predictive modeling for early disease outbreak detection is minimally explored but holds significant potential for success. In their article, Khodadadi and Towfek (2023) discuss the deployment of a predictive modeling system based on the IoT that has been devised to provide solutions to improve the response of public health to diseases. Khodadadi and Towfek also start by explaining how the early identification of the outbreak and reporting the diseases helps reduce the incidences of health emergencies. The drawbacks of classical approaches are as follows: The extent of the problem becomes obvious if it is compared with the help of modern technology that is absolutely imperitive in contemporary business. It is the finding of the authors that IoT which is capable of sourcing data from the various environment can be effective in the early identification of epidemic.

Based on the description for this article given above, the core content of the article will concern itself with the structure and operations of the revealed predictive modeling system for IoT. The IoT system involves a set of interconnected smart devices that help measure a wide range of data pertaining to health including environmental aspects, symptoms, and the general mobility of a patient. This data is analyzed and scrutinized to detect a probable outbreak of the virus through the help of the special algorithms such as the machine learning algorithms.

Khodadadi and Towfek, in their system, stress on data integration and data compatibility as well. They elaborate on the fact that combining data from various IoT devices and health information systems raise the level of accuracy for the predictions. The authors also apply cloud computing as a resource that can supply the needed computational and storage capacities for the processing of data coming from IoT devices.

Thus, the article discusses the issues arising from the use of IoT-based systems for the implementation of the methods of predictive modeling relating to data protection and security. For the security of patients' health information, the authors recommend enhanced encryption techniques and proper data transfer procedures. They also talk about the requires for some norms in the appropriate use of these products and compliance with the privacy laws.

Thus, Khodadadi and Towfek (2023) explain in details the potential effects of the proposed system to enhance public health. That presents examples of how the IoT-based predictive modeling system has effectively identify diseases at its initial stage before spreading and implement the necessary measures. The authors propose that similar systems could be

widespread in the near future and give a new impulse to the epidemiological surveillance and minimize effects of epidemics.

2.5 Leveraging Machine Learning for Chronic Disease Diagnosis

ML has become one of the most effective techniques in the healthcare industry, especially concerning chronic illnesses' diagnosis. Taking into account the possibilities of the development of artificial intelligence, Battineni et al. (2020) discuss different cases of using the ML predictive models for diagnosing chronic diseases; the authors also describe the trends and issues in the field.

Battineni et al. start by outlining the global increasing prevalence of chronic diseases and the relevant importance of accurate diagnostic methods. Diabetes, cardiovascular diseases, chronic respiratory diseases and many more are some of the chronic diseases which call for early and accurate diagnosis in efforts to enhance the quality of life of the patients and decrease the costs of treating the diseases. The authors also propose that the technique is best suited in this area because, with the help of ML models it is possible to analyze the large dataset and produce identification of deeper patterns. The article presents an analogical description of the methods of utilizing ML in the diagnosis of chronic diseases, distinguishing between supervised, unsupervised, and reinforcement learning. Of the three learning paradigms, supervised learning, consisting of support vector machines, random forests, neural networks among others, has been found to have a lot of potential in forecasting chronic diseases' development and evolution. The authors raise studies to show that such algorithms have accuracy in diagnosing conditions like diabetes and hypertension.

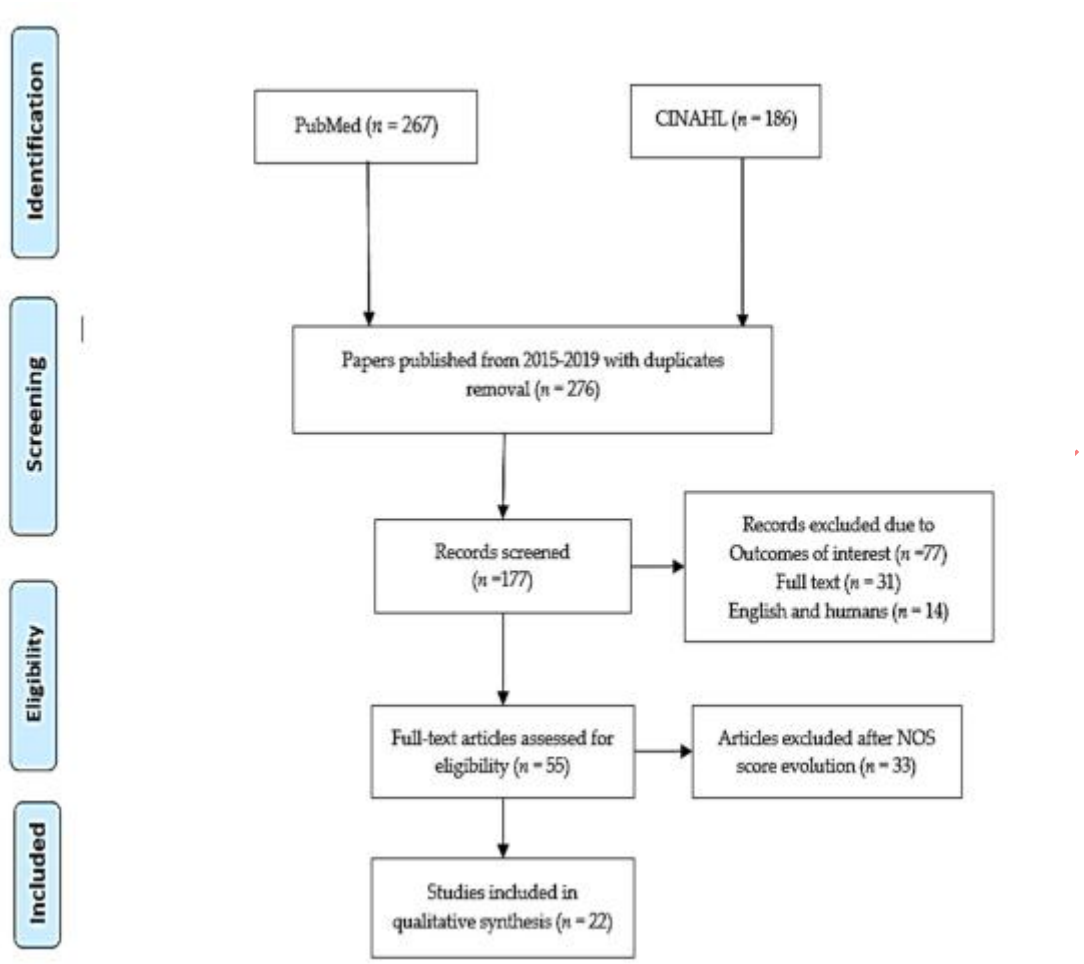

Figure 4: "Preferred reporting items for systematic reviews and meta-analyses"

(Source 4: Battineni *et al.* 2020)

Other forms of learning including clustering and association analysis are also presented when it comes to the identification of hidden patterns in patients' data set. These techniques can help in categorizing the overall patient populace into small groups depending on disease resemblance thus they can be useful in the determination of the treatment strategy. Battineni et al. (2017) also agree with the notations made concerning feature selection and feature engineering and assert that these stand as critical steps towards improving the performance of the ML models; this stems from the fact that the quality of feature or input data has a direct influence on the capabilities of prediction. The authors also discuss the issues relating to deployment of the ML models in the clinical setting. They talk about the elaboration of decision-support technologies regulating interfaces based on predictions of ML to facilitate decision-making on diagnoses by the healthcare providers. Such systems can involve time real processing of the patient data, ensuring that the correct diagnosis is made earlier.

Cross-validation is one of the very important methods in predictive modeling especially in the health care where precision and accuracy are very important aspects. Cross validation is a reliable technique used in healthcare predictive modeling in making an evaluation and testing of the models developed out of the Python language. It involves the division of data where the model is trained using part of the information and tested with the other part and this is done severally. This approach assists in avoiding over-fitting- a very common situation where a model learns very well all the data fed to it in training and performs very poorly for data that was not used in the training process. In healthcare where datasets are often skewed, or of small size, cross validation offers a more accurate picture of a model's worth. For instance, while working with python libraries like Scikit-learn, cross-validation can be performed by means of certain tools like K-fold and stratified cross-validation that are indeed useful while dealing with the unbalanced classes typical of many healthcare datasets (Paulus, J.K. and Kent, 2020). As used herein, cross-validation helps the researchers and practitioners to know about the model stability and variability and hence it improves model reliability. Besides, cross-validation helps in hyperparameter optimization to help find the best fit models for better results. Lastly, cross validation within the healthcare predictive modeling process promotes the creation of strong models, in supporting the decision making process and thus beneficial to the patients.

2.6 "Integrating Structured and Unstructured Data in Healthcare Predictive Models: A Deep Learning Approach"

The integration of ordered and disordered data using deep learning provides a crucial breakdown to health care prognostic analysis. The combination of these two ideas forms the subject of the study in Zhang, Gu, and Jin (2020), who offer a systematic perspective of the integration and how it can be useful under what conditions and with what weaknesses.

Zhang et al (2020) begin by pointing out that healthcare data is inherently complex and comes in different data types that include structured data including EHR, unstructured data include clinical notes, medical images, and patient reports. In simple words, the classical predictive models tend to rely solely on the structured data A that might be missing essential information contained in B. The authors presented that integration of both data kind can improve the quality and stability of the forecasting models.

This being the case, the core of the article addresses the structural and non-structural data integration using deep learning. The authors explain broad categories of networks such as CNN and RNN that are efficient in handling free forms of data. These models are integrated with the

conventional machine learning methods such as logistic regression and random forests that work with tabular data. Both the methods are combined and hence, the hybrid model is more broader giving more accurate predictions.

The authors elaborate on the effectiveness of the proposed method, stating that Zhang et al. (2020) provide multiple case studies to support this claim. A specific application of this is the prediction of the patient status in intensive care units. With thus obtained integrated structured EHR data and unstructured clinical notes, the predictive models produced better accuracy of predicting the patient's deterioration and mortality. This integration allows analysis of meaningful characteristics in data that do not have an obvious pre-specified structure such as a patient's symptoms and a physician's diagnosis which frequently are key to predicting the outcome.

The authors also embark on discussing the technical issues and their solutions in relation to the use of heterogeneous data sources. Data cleaning pre-processing, and the process of transformation of the features as well as the integration of structured and unstructured data set are challenging procedures that warrant appropriate techniques. In addition to the principles of data standardization, Zhang et al. point out the necessity of enhancing the methods of analyzing text-form speeches through the use of new approaches in the sphere of NLP.

Moreover, the present article also addresses some of the issues in the application of the theories in clinical environments. The integration of deep learning models into the current systems in healthcare is complex with demanding computational need and professional skills. The authors recommend multispecialty cooperation with data scientists and healthcare professionals so that the models can be implemented correctly and utilised.

Proceeding further, three important topics have to be raised, namely: Ethical issues and data protection, which are also highlighted by Zhang et al (2020). Patients' data should remain private and care should be taken to meet the set standards and legislation when using big data to come up with good models. This the authors recommend encryption and secure data sharing as ways of handling the privacy issues. Zhang et al. offer an extensive literature on the fusion of structurally and non-structurally modeled data for deep learning-aided healthcare prognostic analysis. It is their findings that show the possibility of better establish predictions, thus, greatly improving the therapeutic and diagnostic processes. It is crucial to point out the benefits and drawbacks of such an approach that the article describes as the foundation for future development in the sphere of healthcare analytics.

2.7 Forecasting COVID-19 Spread with ARIMA Models

This has never been felt more true than in the conduct of the current COVID-19 pandemic where predictive modeling has been proven to be indispensable to public health. Assessing the effectiveness of current interventions on COVID-19 cases in KSA, Alzahrani, Aljamaan and Al-Fakih (2020) undertaking a forecast of the virulence of the virus using the ARIMA model. Alzahrani et al., (2020) firstly introduce the importance of the study due to the prevailing pandemic situation, and the crucial need for better forecasting. Models such as ARIMA help in decision making in matters to do with public health intakes, planning and policies. The authors also point out that since tracking the progression of infectious diseases involves the analysis of time series, a method of choice for such an analysis is the ARIMA modeling technique, which is adequate for non-stationary data.

In order to understand how the research concluded that the prolific spread of COVID-19 can be mitigated, and following the steps of the article preceding the conclusion, the reader is informed about the methodological approach, where the method of using the ARIMA model in forecasting the number of COVID-19 cases is described. The parameters required for calibrating the model were daily reported cases from the beginning of the outbreak in Saudi Arabia. The authors briefly review the use of model selection, parameter estimation and diagnostic checking with the importance of model refinement stressed by the fact that the process is cyclical.

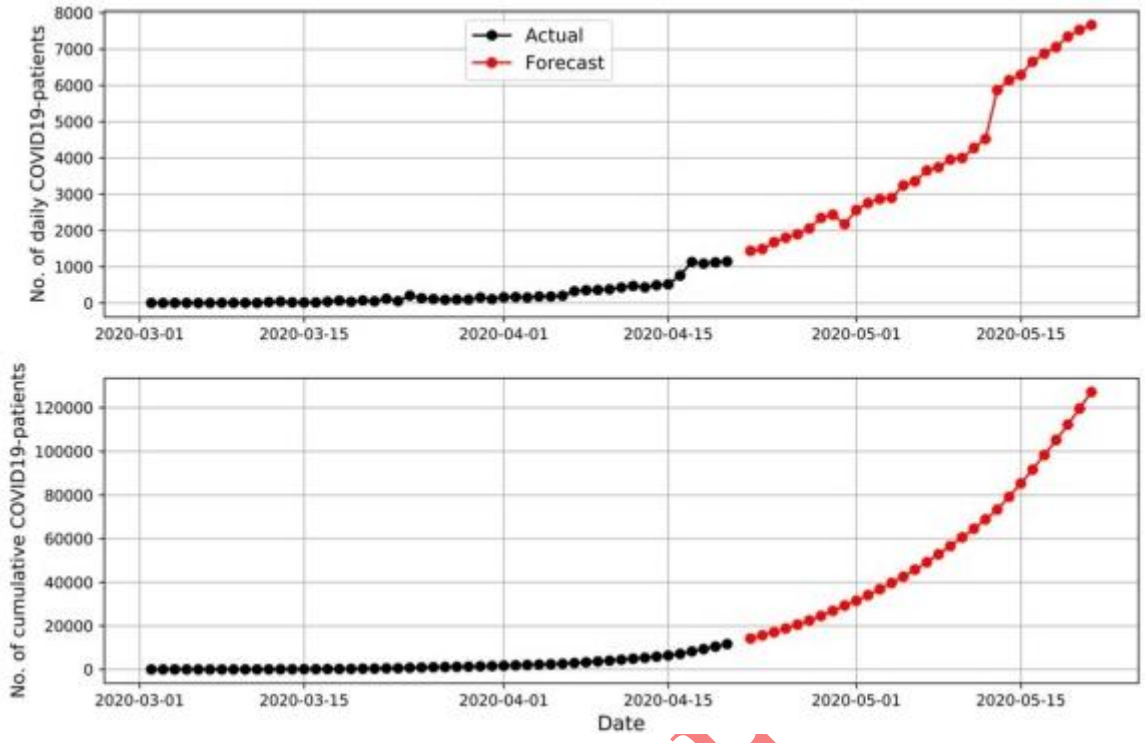

Figure 5: "The forecasting results of the total number of daily confirmed cases and cumulative cases in Saudi Arabia"

(Source: Alzahrani *et al*, 2020)

An important component of the research is the integration of public health measures into the model based on the ARIMA method. Alzahrani et al. (2020) consider the effect of other factors which include deaths of lockdowns, social distance, and limitations on traveling on the model. He also said that including these interventions is important in developing accurate estimates and estimating the efficacy of strategies of public health.

Thus, from the results of the study, it can be concluded that the ARIMA model offered a quite reliable short-term prediction of the number of COVID-19 cases in the Kingdom of Saudi Arabia. The authors also provide statistics of various time horizons to depict periods of divergence due to variation in policies or reporting methods in the health sector. They stress the concern of revising the model every time with new data in order to keep it up to date with the current data base.

Nonetheless, like any study, this work recognizes the weakness of the ARIMA model as applied in the findings. A notable weakness is that it uses past events as a means of analyzing, thus failing to incorporate a constantly evolving subject like a pandemic, which is affected by issues

like mutation of the virus, mutations in human behavior, and varying approaches applied by different countries. According to the authors, future work should include the integration of ARIMA with other modeling methodologies such as compartmental modeling and machine learning to improve the model's predictive accuracy.

Moreover, the authors go further and explain how their results apply to global health initiatives at large. Alzahrani et al. have called for the application of similar models such as ARIMA in other regions so that there could be adequate preparedness concerning the pandemics. From this, they recommend that predictive modeling should become a core component of public health systems that offers real-time feedback on decisions made. In the study under discussion, Alzahrani et al., used the ARIMA model to predict the advancement of COVID-19 in Saudi Arabia. Their work demonstrates such a value of the model for public health intervention and further emphasizes the need of incorporating real-time data and multiple model types to improve the results' accuracy and validity. The scholarly work also provides significant findings on the application of advanced methods, such as predictive modelling, in approaching and containing the effects of pandemics.

2.8 Risk of Bias in Machine Learning-Based Predictive Models in Healthcare

Prediction accuracy of "machine learning- based models in health care" system must be ensured for the pragmatic use. In their targeted systematic review on the methodologies of developing prediction models using supervised ML techniques, Navarro et al. (2021) also perform assessments for each study, thus offering an insight into the risk of bias that the existing methodologies entail in the sphere of healthcare.

Navarro et al. (2021) start with identifying the increasing use of ML to build prognosis models in the contexts of the health domain. Such models have the potential to improve diagnoses, estimate patients' outcomes, as well as, individualized treatments. However, the authors stressed that such models rely on the quality of the processes used to develop them, and more specifically to address the potential for bias.

The review pays very much attention at scrutinizing previous research papers and identifies major sources of Bias in the development of ML models. Navarro et al. (2021) categorize these biases into several domains: data quality, overfitting of models and while making validations, and the overall reporting. The authors explained that a sampling-related issue that dominates

information bias includes missing values, imbalanced datasets, and no external validation. The authors also stress on the importance of using accurate samples that would allow for building accurate models applicable to a large population.

In addition, Navarro et al. accentuate the deficiencies of the reporting standards for ML-based predictive modeling studies. They recognize gaps in the methodology where some facets of model development, like data preparation, the criteria the researchers employed for model selection, and the measures used to assess models' performances, are omitted. The authors wish that the TRIPOD statement describing the transparent reporting of a multivariable prediction model for an individual prognosis or diagnosis be adhered to.

Furthermore, Navarro et al. amplify some of the reporting standard flaws revolving around the ML-based predictive modeling studies. They also realize that there are some extend which lack in the methodology concerning some face of the model development such as data preparation, the criteria adopted by the researcher while selecting the model, and measures used to evaluate the performance of various models. The authors who formulated the TRIPOD statement would wish that the statement be followed so that there is transparent reporting of a multivariable prediction model for an individual prognosis or diagnosis.

Navarro et al. (2021) give a detailed description of the systematic evaluation of risk of bias in ML-based predictive models in health care. This systematic review of theirs strongly emphasises the need for proper methodological structures that will enhance the validity and the transferability of such models. In highlighting biases in data quality, model selection and specification's over-reliance, validation processes and reporting and MNL's advancement, the authors provide useful suggestions that can foster improvement in the deployment of ML predictive models in health care. It goes a long way in enhancing the conversation on strengthening the reliability of quantitive and qualitative decision making approaches in the medical research.

2.10 Advancing Healthcare with Causal Inference and Counterfactual Prediction

Forecasting in the field of medicine has gradually moved not only from the exploration of associations to the discovery of causality and counterfactual foreseeing. In this article, Prosperi *et al*.(2020) consider the use of causal inference and counterfactual thinking in ML and stress the relevance of these components for actionable information in healthcare.

Prosperi et al. start by explaining that a line of work with ML is restricted to using correlation rather than causation to predict cases. Hence, even though these models can give the probability of some event or outcome from past data patterns, they cannot guide intervention or explain the process that produces the results. Accordingly, the authors posit that extending causal inference into the ML path offers such a feasible solution to such problems, contributing to offering the proper sensible framework for decision-making in healthcare.

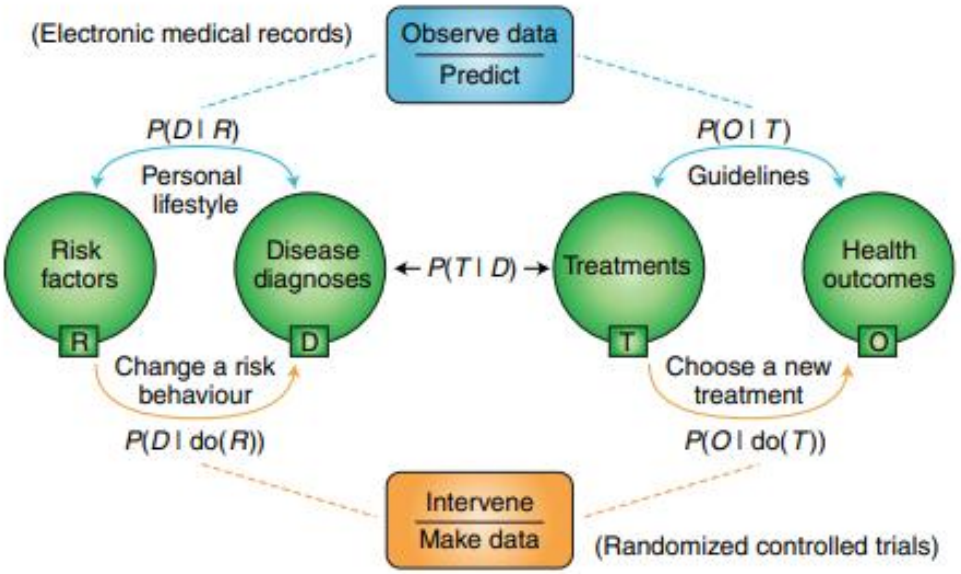

Figure 6: Conditional versus interventional probabilities

(Source: Prosperi *et al*, 2020)

The article reviews on three methodologies of causal inference namely potential outcomes, structural causal models and instrument variables. According to Prosperi et al., the methods help to estimate treatment effects from such data, which is essential in healthcare, as in many cases, using RCTs is either impossible or unethical. Understanding causality can be useful for interventions to be effective in health care since it locates the cause and ensures the development of strategies that can offer the best treatment.

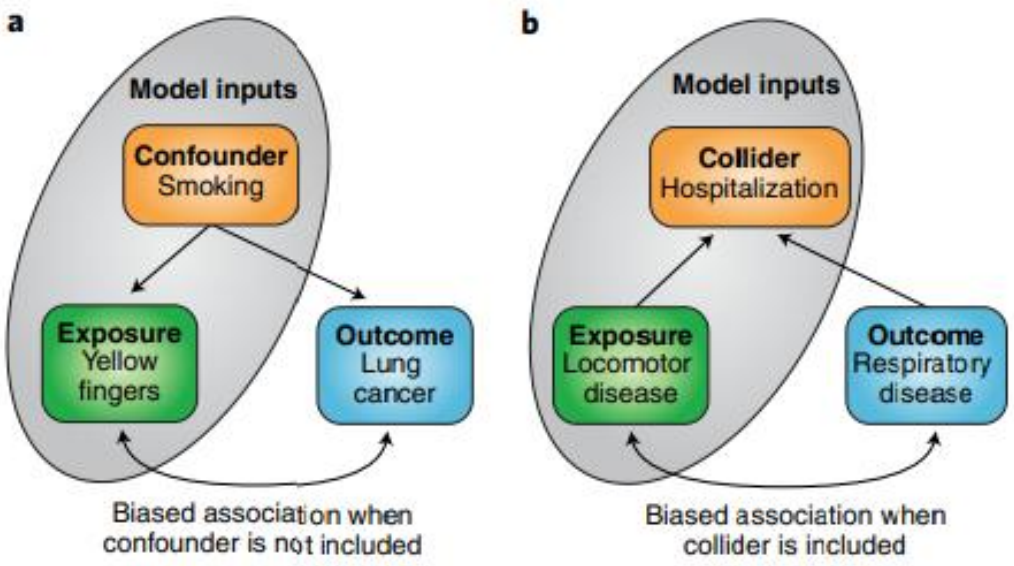

Figure 7: Examples of confounding bias and collider bias

(Source: Prosperi *et al;* 2020)

One of the major topics covered in the article is counterfactual prediction or prediction of what could have occurred if some condition was met. The study done by Prosperi et al. shows that can counterfactuals can be applied to model the outcomes of the discrete treatments on the patient's condition enhancing the idea of precision medicine. For example, they describe counterfactual prognostication of treatment choice for a particular patient taking into account certain characteristics and history of the patient.

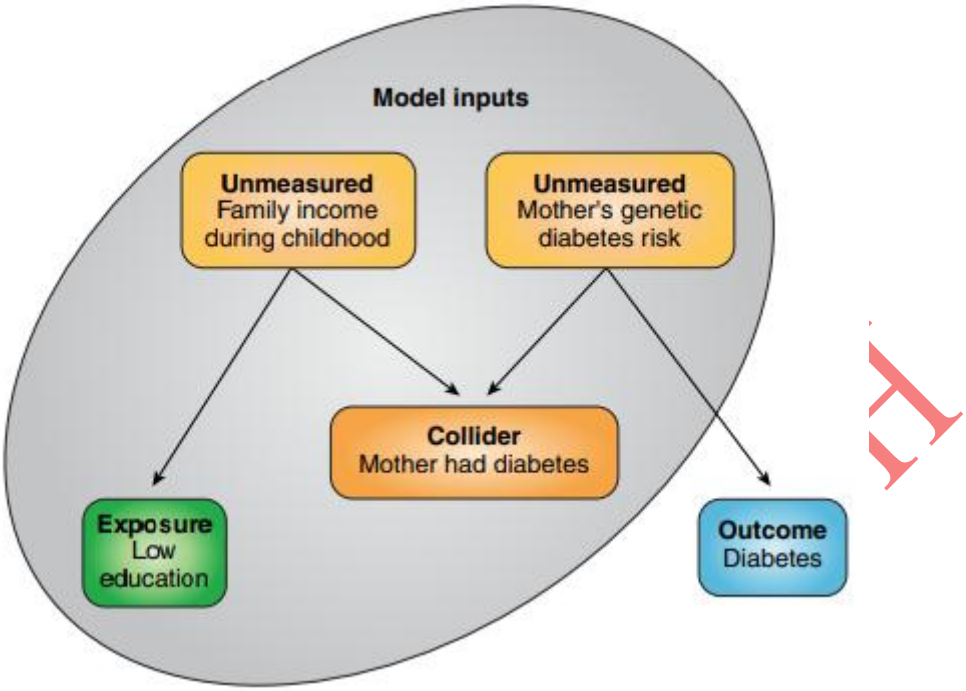

Figure 8: An example of M-bias

(Source: Prosperi *et al.* 2020)

In their recent article, Prosperi et al. (2020) have illustrated how causal inference and counterfactual prediction can be used in healthcare by using case examples. An example in this case is the use of these techniques in chronic diseases to anticipate the effects of change in lifestyles and taking medication. Another example include predicting patient readmission, causal models are used to determine the potential causes of readmission and recommend on how to avoid them. The authors also explain the problems and drawbacks related to the assessment of causality in ML. About working with actual data they describe things like confounding, data quality issues and other challenges of causal inference. According to Prosperi et al., these challenges can be solved by integrating the domain knowledge and the use of advanced statistical techniques. At the same time, they stress the practical value of data science within healthcare, as well as the significance of creating causal models for practical use in collaboration with data scientists, clinicians, and epidemiologists.

2.11 Literature Gap

Despite the awareness of the necessity of well-developed algorithmic criteria for the proper AI-based prediction models, as cited by de Hond et al. (2022), current research seems to lack a set

of widely accepted best-practice guidelines to develop, validate and implement such models. Although, the significance of interpretability in machine learning models is stressed, as highlighted by Stiglic et al. (2020), the existing research lacks works that is highly accurate but still interpretable for healthcares. Merging the structured and unstructured data has been recommended in the literature to enhance the predictive models as illustrated by Zhang et al. (2020). Nonetheless, the implementation of these heterogeneous data types has not been tested much in terms of the best practice on how they may be incorporated in actual health care organizations.

Even though Navarro et al. (2021) provided an understanding of the sources of bias within the ML-based predictive models coupled with the information on when it was appropriate to apply different reductions, more studies are required to provide solid information on how these biases could be reduced in practice, particularly in different healthcare settings. As Prosperi et al. (2020) explained, causal inference combined with counterfactual prediction is conceptualized to be promising, however, regarding the implementation of these two theoretical approaches into the tools used in healthcare decision-making today, there is a gap. All the literature does not include the studies on the long-term perspectives of the predictive models use in the health care, including the change of medical information and patients' populations.

Ethical issues are discussed in many studies, despite this, significant effort has not been dedicated to constructing a robust ethical framework suitable for creating and implementing predictive models in HC contexts. Most works aim at the model creation and typically carry out internal validation assessment, while few studies are conducted on the external real-world performance and contributions of these models in various healthcare facilities for long-term evaluations.

3.0 Chapter 3: Methodology

3.1 Chapter Overview

This chapter is a blow by blow account of how this dissertation on predictive modeling in healthcare was done. The chapter starts with the research philosophy, which focuses on the choice of the positivism philosophy as a basis for the study, which is characterised by the use of quantifiable facts and their analysis. The chapter then provides an overview of the research method, which is based on the deductive paradigm aimed at testing the hypotheses developed on the basis of the literature review. The nature of data collection has been described, with a

key primary questionnaire survey aimed at providing measurable data in a structured format from the target population groups of HCPs, patients and other relevant stakeholders. The method of sampling is explained to include the purpose of using the stratified random technique to sample the perspective clients across the age groups. This paper outlines how data analysis steps within machine learning are conducted; pre-processing data, training the model, where features are included or excluded and how performance is measured. The section on research design brings together all these components, and organizes them in a logical and methodical way so as to underpin the development of reliable prediction models. There are also marked efforts to ensure the participants' consent and confidentiality while working on the study, as well as compliance with the ethical standards throughout the process. This chapter lays down the groundwork of this study as well as the empirical analysis of the various and diverse healthcare data.

3.2 Research Onion

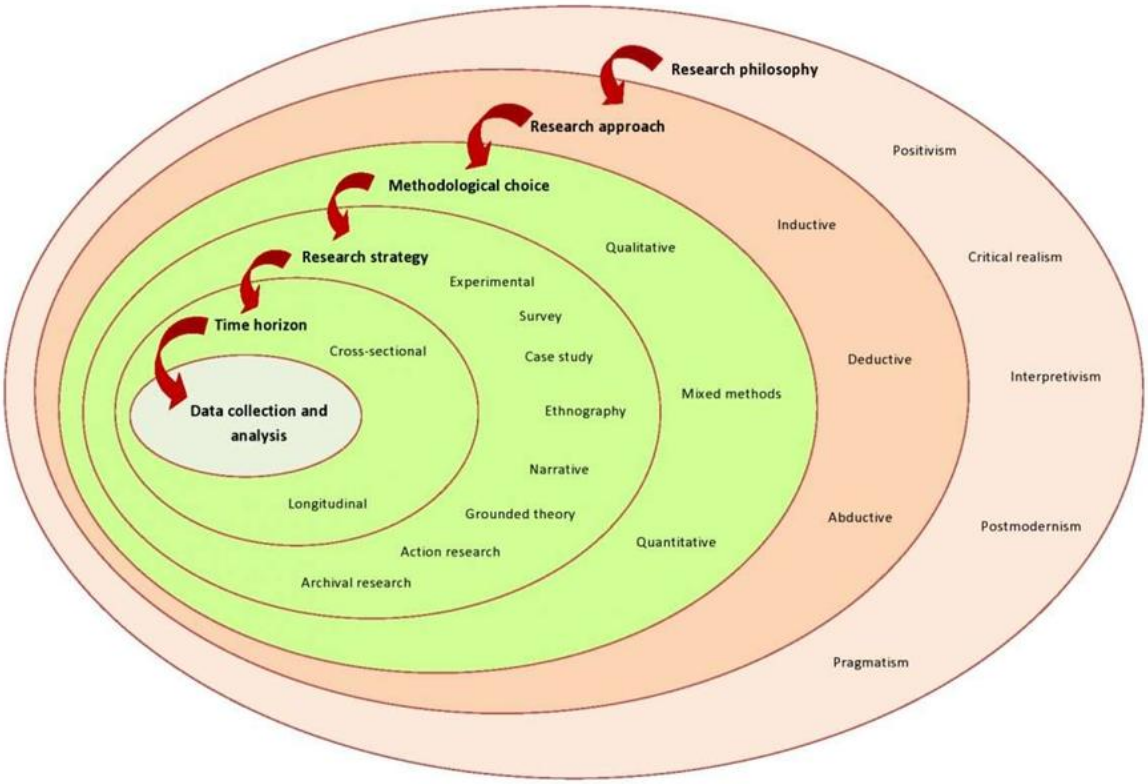

Figure 9: Research Onion

(Source: Saunders; 2017)

3.3 Research Philosophy

Imposing on the nature of this dissertation, the research philosophy used in this study is positivism. Positivism postulates that there is a reality out there, which can be measured and rationalized with the help of data. This approach is especially applicable in the statistical approach to medical prediction since it entails measurable figures and probabilities. Using big data approaches and complex analytics, the study intends to find out the patterns and associations, which may help in future health events' forecasting. Positivism focuses more on developing hypothesis, collecting data and analysing the data to ensure that the results are positivistic. This approach makes it possible for the study to have generalizable outcomes hence benefiting other field in health care analytics (Chowdhury and Turin, 2020). This research therefore adopts a positivist epistemology, in that its approach is orderly and formal, there is a reliance on facts rather than interpretations and the work seeks to make practical contributions to organisational decisions with the ultimate aim being to contribute to the better management of patients' care.

3.4 Research Approach

The research method for this dissertation on 'the roles of predictive modeling in health care' is a deductive approach. This means that a deductive approach starts with theory, where the researcher draws propositions from either previous theory or theory from the field of study. In this study, hypotheses are developed with regard to various health indices and the health results. These hypotheses are generated from what is known in the areas of healthcare, analytics, big data, and predictive modeling.

After the objectives are set, the analysis goes through the process of analyzing the necessary information from such databases as the healthcare one, electronic health records, and others. The collected data is then statistically processed and the results are checked with the help of predictive models that are a part of the hypotheses (Subbaswamy, A. and Saria, 2020). This occur when one use regression analysis, machine learning, and artificial intelligence techniques in the extraction of patterns and correlations in the data.

The deductive approach ensures that hypotheses are tackled in a very systemized manner which means that use of empirical evidence to eliminate or prove hypotheses is very easy. This method offers a systematic approach of moving from theory to practice that helps establish the

findings on sound statistical analysis. Thus, it is with such an approach that the proposed research has the following objectives: To contribute to the strengthening of evidence in healthcare and increase its reliability to work for the treatment of patients; to improve the quality of patients' lives and their further rehabilitation; and to contribute to the development of the use of predictive analytics in the field of medicine.

3.5 Data collection method

The technique of data collection for this dissertation on predictive modeling in healthcare is a primary quantitative survey. In the administration of this type of method, a structured questionnaire is developed and administered to healthcare personnel, patients, and key informants. The survey is elaborated down to the last bout to en ensure that it registers numeric values on a range of health aspects, and additionally demographic information that may also have an impact on one's health (Waring, Lindvall and Umeton, 2020).

Only an electronic survey is used to increase the probability of reaching a large number and the ease of the process for participants. It involves questions with preselected response options like Likert scales, MCQs, and input boxes that enable the collection of standardized data, and quantifying the respondents' response. The sections used in the current questionnaire include demographic data, household members' health risks and behaviors, health history, illness and disease history, and current health conditions.

In order to ensure valid results, participants are chosen through stratified random sampling since the study's target group is diverse within the healthcare setting (Wong *et al.* 2021). The advantages of this sampling technique is that it guarantees that each section of the population like the age, gender, SES, and geographical distribution is included in the study population. To increase the validity of the information received the criteria for participant selection are also outlined clearly, including only those patient who have used health care services in the recent past.

To increase the credibility of a survey, a pilot study is first carried out among a sub sample of the intended sample population. Concerning the pilot study feedback the questions are reviewed time and again to enhance on clarity and neutrality and on their ability to provide the necessary data (Rubinger *et al.* 2023).

As a result of using the primary approach as a quantitative study, this research is certain that the information gathered responds directly to the study's goals and objectives. The presented

method offers a stable framework for generating accurate and dependable predictive models to enhance the system's abilities and overall healthcare results.

3.6 Data analysis

The application of predictive modeling in the healthcare sector will use analysis on the data gathered from the survey which is preceded by a data preprocessing step. This involves aspects like how to deal with the missing values, outliers among other distortions on data to ensure quality and reliability. For the data analysis where necessary the technique of data normalization and data transformation will be as shown below.

After that the data is split into the train data set and the test data set. The training set is used to design or create and build many of the machine learning models that could be used and the test set for model evaluation (Yang et al. 2020). In the prediction of the models developed from the survey data, the machine learning algorithms used were the universal ones such as the logistic regression, decision trees and random forests. Explanatory variables are first filtered this is because it is a method of identifying the top influential variables on the detected results on health implication. This proceeding step is crucial in the sense because its function is to increase the dimensionality of the models and thereby improve their quality. Feature engineering may also be applied in cases where new features have to be constructed or if the features to include must be changed in a way that will enhance the accuracy of the model.

The trained models are then evaluated using the testing data versus its accuracy, precision, recall and F1 score. Cross-validation methodologies are employed to validate the models so as to avoid a situation where the model become too complex and fitting on the training data set. If practiced, it is refined about these metrics and the most appropriate model is selected out of them.

3.7 Data sampling

In this data sampling for this paper, since the researcher is interested in a form of predictive modeling in the healthcare setting, the sampling method employed in this study is the, stratified random sampling technique where the researcher is able to attract 51 participants to the study. By the use of the given technique, all the possibilities that may exist in health care context as regarding to the sub categories such as ages, genders, economic position, geographical location among others are provided in the sample (Wynants et al. 2020). Sample based on primary strata of the sample is a useful tool in providing coverage of the concerned population of

interest and also in assessing various health indicators and different variables. This self-administered tool is a structured questionnaire completed through an electronic means, thus generating data that elicited various aspects of health status and demographics of participants. Such a technique helps in making sure that the sample collected has other segments in the population and that the conclusions made in predictive modeling analysis are valid and accurate.

3.8 Data Preprocessing

In this case the data includes a number of fields and each of them exhibits different characteristics of the patient data such as demographics, diseases, and healthcare access. The dependent variable used in this analysis is “Test Results”, while the independent variables are the columns namely; “Name”, “Age”, “Gender”, “Blood Type”, “Medical Condition”, “Date of Admission”, “Doctor”, “Hospital”, “Insurance Provider”, “Billing Amount”, “Room Number”, “Admission Type”, “Discharge Date”, and “Medication”.

Handling Missing Values and Encoding:

First, user who proposed this code looked for missing values in the overall data set itself. If there is missing data it has been eliminated by elimination of rows containing missing values but in depth approach other method like imputation could have been applied. User also used LabelEncoder for categorical variables which was set for converting the text-based data into numerical format like “Gender”, “Blood Type”, “Doctor”. This preprocessing step was essential for the extraction of features whereby the models would take raw data and properly analyze it.

Standardization:

In order to normalise the scales of measurement of quantitative features such as ‘Age’ and ‘Billing Amount’, the user employed StandardScaler. The idea behind this technique was to standardize the obtained data and make the corresponding models non-sensitive to the fact that some features have larger numerical intervals. Standardization is most relevant for models that are impacted by the variance of the features such as the Support Vector Machine (SVM).

3.9 Dataset Description

The dataset contains information from various aspects of patient data:

- **Demographic Details:** "Name," "Age," "Gender," "Blood Type."

- **Medical Information:** "Medical Condition," "Medication," "Test Results" (dependent variable).
- **Healthcare Services:** "Date of Admission," "Doctor," "Hospital," "Insurance Provider," "Room Number," "Admission Type," "Discharge Date."
- **Financial Details:** "Billing Amount."

These features provide a comprehensive view of the patient data, which allows us to build the predictive models that can infer the "Test Results" based on the provided an independent variable.

3.10 Models Used

- In this analysis, user choose the models: "Logistic Regression, Random Forest Classifier, Support Vector Machine (SVM) and Gradient Boosting Classifier". These models were selected based on their unique strengths and suitability for the dataset:
- **Logistic Regression:**

  This model is turning the problem into simpler form which is easy to understand but at the same time is accurate and serves the binary classification problem well. Although it is a linear model, it is useful for benchmarking since it has a relatively high degree of accuracy.
- **Random Forest Classifier:**

  "Random forest" is a modality of decision trees and it uses numerous decision trees in order to get high accuracy and to avoid overfitting. This makes it useful especially when there are many variables from a particular data set and it provides robustness as well as better generalization.
- **Support Vector Machine (SVM):**

  SVM works well in high-dimensional data space and performs well when the number of dimensions is higher than the number of samples. It is tried to maximize the distance between the classes so that there seems to be better classification in the case of well separable data.
- **Gradient Boosting Classifier:**

  Gradient Boosting is another type of ensemble learning that creates models one after the other with each new one learning from the mistakes of the previously developed one. It proves to be very successful in increasing the model's accuracy particularly when dealing with imbalanced data.

These models were selected because they are easily interpretable, easy to implement and both were found to perform adequately on the given data set. k-Nearest Neighbors or Neural Networks could also not be used as they are strongly limited by their interpretability and computational complexity, not to mention their ineffectiveness while working with a dataset of this type.

3.11 Validation and Performance Index

User applied cross-validation with the use of StratifiedKFold where cross-validation was done ten times. This was a better way of estimating the performance of the model since the variance that comes with a single split of data into train and test set was eliminated. The performance of each model was evaluated using key metrics: The evaluation metrics used include accuracy, precision, recall and F1-score.

- **Accuracy:** This metric measures the percentage of correct predictions made by the model out of all predictions. It provides a general overview of model performance.
- **Precision:** Precision assesses the proportion of true positive predictions among all positive predictions made. It is particularly useful in scenarios where the cost of false positives is high.
- **Recall (Sensitivity):** Recall measures the ability of the model to identify all relevant cases within the dataset, particularly focusing on the correct identification of positive instances.
- **F1-Score:** The "F1-score" is the harmonic mean of precision and recall, offering a balanced measure when the cost of false positives and false negatives is similar.

3.12 Data Split, test, train, validate

To effectively split, train, test, and validate the dataset, first, the data is split into training (80%) and testing (20%) sets using train_test_split to prevent the data leakage. The training set is standardized using StandardScaler, ensuring consistent the feature scaling. Stratified K-Fold cross-validation (StratifiedKFold) is then applied to the training set to validate models as well as the optimize hyperparameters, ensuring the robust model evaluation.

3.13 Data Analysis Methods

Sample Data was divided into training and testing sets with the scaling of 80% and 20% respectively. The given training set was employed to (train) optimize the models, while the testing set gauged the model's ability to generalize across different data inputs.

1. Logistic Regression:

Logistic Regression performed as expected the providing baseline accuracy. However, it was prone to underfitting due to its linear nature, as seen in its lower F1-score.

2. Random Forest Classifier:

Random Forest achieved higher accuracy and F1-scores than Logistic Regression, thanks to its ability to handle complex interactions between features. However, there was a concern of the overfitting, as indicated by the variance in the cross-validation results.

3. Support Vector Machine (SVM):

SVM struggled with this dataset, particularly due to the data imbalance. It resulted in lower scores across all the performance metrics, indicating the underfitting.

4. Gradient Boosting Classifier:

The model which offered higher accuracy, precision, recall, and F1-scores is Gradient Boosting. It successfully managed to strike the right balance in terms of feature distribution and also rectified its mistakes which were made in the previous rounds of the ensemble.

3.14 Results Interpretation

Accuracy:

Gradient Boosting achieved the highest accuracy (0.1692), followed closely by Random Forest (0.1698), indicating both models were better at generalizing than Logistic Regression (0.1630) and SVM (0.1480).

Precision and Recall:

Precision and recall were also highest in Gradient Boosting suggesting that this model was more effective in identifying the true positives and with least false positives. SVM on the other hand had the poorest precision and recall meaning the model had an inhibitment of capturing the detail of the data set.

F1-Score:

The F1-score, which is the second metric that combines both precision and recall, is also in favor of Gradient Boosting with 0.1690 ± 0.0079. Random Forest also gave good performance

but Logistic Regression and SVM where left behind primarily because of their linear model and they are more inclined towards the balance between the classes of the data set.

Chapter 4: Result and Analysis

4.1 Chapter Overview

This chapter presents and discusses the findings obtained from the use of various methods of predictive modelling on a healthcare data set as highlighted in the research objectives. The primary focus is on evaluating the performance of the distinct "machine learning models: Logistic regression, Random forest, support vector machine (SVM), and Gradient boosting machine" (Sahin, 2020). These models were selected with regards to their different methods of classification function that can be applied towards the objective of predicting healthcare results. Data analysis is the first and most important step that is carried out in the chapter before data exploration, data visualization and data preprocessing. These preliminaries are rather important, as they allow not only to learn some characteristics of the identified dataset but also to prepare it for modelling. For instance, missing values should be correctly treated; categorical data should be encoded since fixed-length vectors are not allowed for training; feature scaling is important to enhance the accuracy of the model (Zheng and Casari, 2018).

Additionally, this chapter delves into the evaluation of the prediction models. Using metrics like F1 score, accuracy, precision, and recall, cross-validation finds out how well the models work. Research suggests that the Gradient Boosting model is the most effective one for making healthcare predictions at the moment. Predicting healthcare outcomes is still a hard task, according to the average results of all models. This suggests that further model development and possibly feature extraction are needed. Last but not least, the suggestions made to enhance health care prediction models are based on these findings.

4.2 Data Analysis of the chosen dataset

The data analysis carried out in this chapter is a valuable part of the dissertation because it paves the way for the rest of the predictive models' work. All these steps are crucial to avoid feeding the training of the predictive models with low-quality data and improve the reliability of the forecast in the healthcare field.

Data Exploration of the chosen dataset

The first procedure that was followed in the data analysis process involved importing the healthcare dataset into a Pandas DataFrame (Purushotham *et al*., 2018). This dataset consists of multiple features that relate to the patient such as demographic data, clinical data, and

administrative information. Samples of the initial findings are identified below: Many features could be related to healthcare, ranging from numerical and categorical to ordinal variables. The investigation started with the data preview, which mainly offered the users a glimpse of the data organization. This step provides information about the nature of the variables involved and their corresponding data types. It also assists in the definition of other near problems – gaps in the values, which should be filled during preprocessing, or incorrect entries in the table.

```
             Name   Age  Gender Blood Type Medical Condition Date of Admission  \
0  Bobby JacksOn  30.0    Male         B-            Cancer        31-01-2024
1   LesLie TErRy  62.0    Male         A+           Obesity        20-08-2019
2    DaNnY sMitH  76.0  Female         A-           Obesity        22-09-2022
3   andrEw waTtS  28.0  Female         O+          Diabetes        18-11-2020
4  adrIENNE bEll  43.0  Female        AB+            Cancer        19-09-2022

             Doctor                     Hospital Insurance Provider  \
0     Matthew Smith             Sons and Miller         Blue Cross
1   Samantha Davies                      Kim Inc           Medicare
2  Tiffany Mitchell                     Cook PLC              Aetna
3       Kevin Wells  Hernandez Rogers and Vang,           Medicare
4    Kathleen Hanna                White-White              Aetna

   Billing Amount  Room Number Admission Type Discharge Date   Medication  \
0     18856.28131        328.0         Urgent     02-02-2024  Paracetamol
1     33643.32729        265.0      Emergency     26-08-2019    Ibuprofen
2     27955.09608        205.0      Emergency     07-10-2022      Aspirin
3     37909.78241        450.0       Elective     18-12-2020    Ibuprofen
4     14238.31781        458.0         Urgent     09-10-2022   Penicillin

   Test Results
0        Normal
1  Inconclusive
2        Normal
3      Abnormal
4      Abnormal
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 55500 entries, 0 to 55499
Data columns (total 15 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Name                18220 non-null  object
 1   Age                 18220 non-null  float64
 2   Gender              18220 non-null  object
 3   Blood Type          18220 non-null  object
 4   Medical Condition   18220 non-null  object
 5   Date of Admission   18220 non-null  object
```

Figure 4.1: EDA

(Source: Self-Created)

To get a preview of the value distribution of each of the features, the dataset was summarized. This comprised finding measures of central tendency and dispersion in numerical variables and frequency of categories of categorical variables. This statistical summary assisted in learning the variables that perhaps had some outliers or skewed distribution that may hinder or influence the modelling phase.

Data Visualization of the chosen dataset

To get deeper insight into the dataset several visualizations were made. Data visualization is a valuable skill in the data analysis process because it makes it possible to see features of data that can be hidden in plain sight. The following figures were produced as part of the results of the analysis.

- ***Age Distribution data visualization:*** To understand the position of patients' ages in the selected dataset, a histogram was created. This histogram leads to an assumption of potentially normal distribution of age, whereby the data is clustered around middle-aged patients which is common in most healthcare-affiliated databases (Garcia *et al*., 2018). It is crucial to appreciate the age distribution as age is a common determiner of exercising healthcare plans. The age distribution graph with the histplot is shares the details of the all valid data's with the 260 to 320 data range.

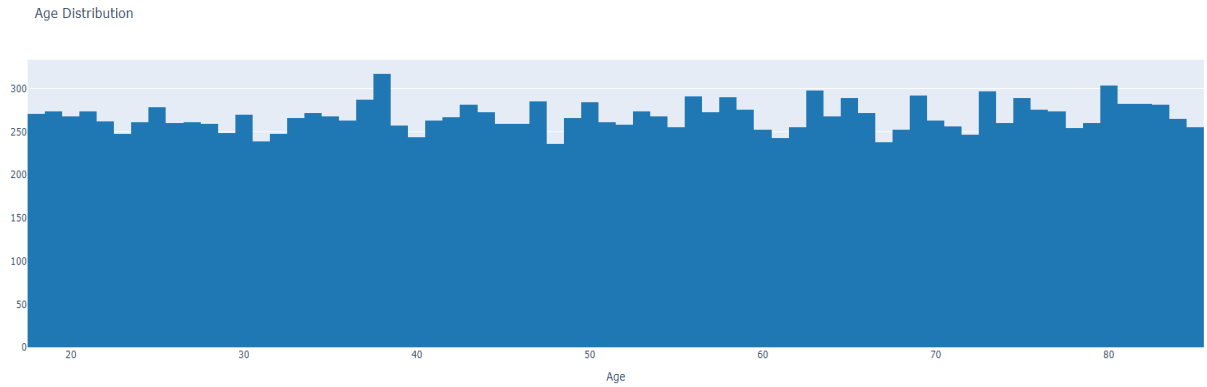

Figure 4.2: Age Distribution

(Source: Self-Created)

- ***Gender Distribution data visualization:*** To present the gender distribution among the patients, a pie chart was employed. This chart demonstrated if the proportion of male and female patients was balanced equally or if there was more of one gender. The issue of gender distribution in the dataset is crucial as it helps prevent gender bias in the predictive models. In this gender distribution data visualization, the 16.4% of the male

with the green colour slot and female. Ther other 67% is the null values which is managed in the data handling section.

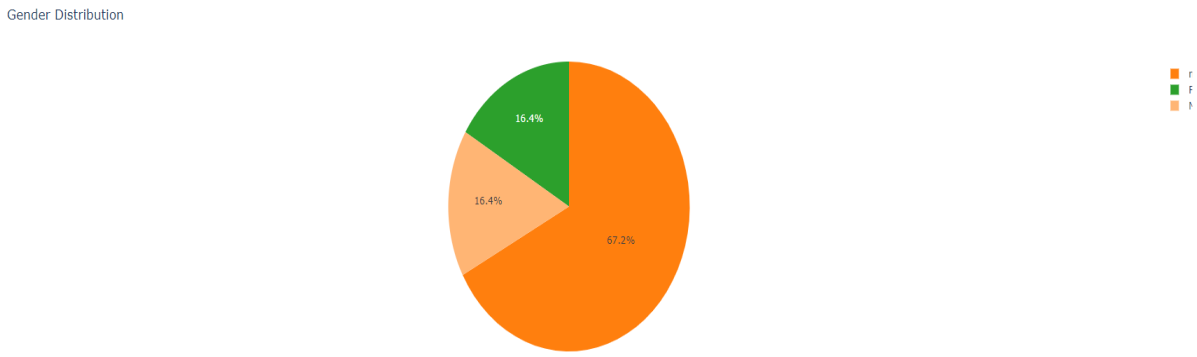

Figure 4.3: Gender Distribution

(Source: Self-Created)

- ***Blood Type Distribution:*** Another histogram was drawn apron the distribution of blood types in the patients (England-Mason *et al*., 2022). This visualization proved useful in orienting myself to the relative frequency of the different blood types and how these could be useful in ascertaining predisposition to some diseases or their outcomes. The blood type distribution graph reflect with the red color bar with the different types of the bloods group data like B-, A+, O+, AB+ etc. The most of the data is shared the value almost the 1700 to 2000 data range.

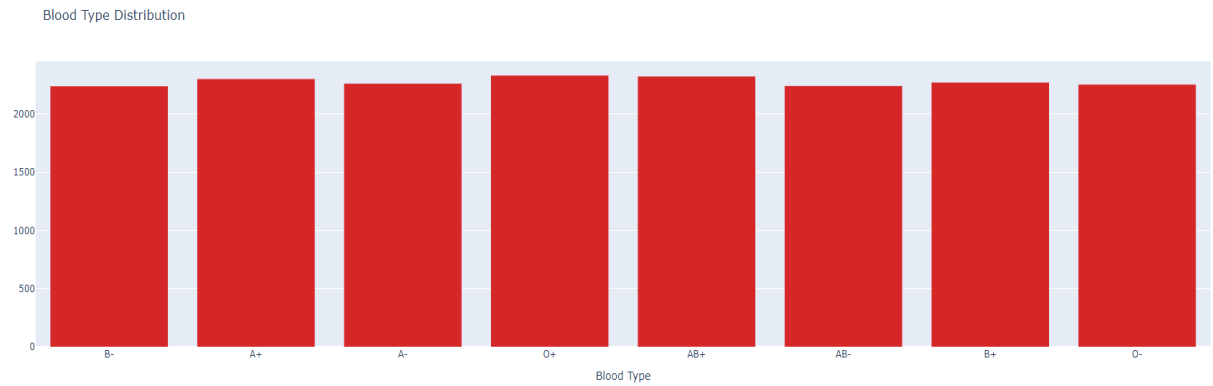

Figure 4.4: Blood Type Distribution

(Source: Self-Created)

These data visualizations offered a snapshot of the demographic distributions in the data, important for identifying initial sources of bias and trends that may affect the modelling

process. For instance, an imbalanced age distribution may imply that the models could perform differently across the age spectrum and this could be an important factor that would warrant consideration while analyzing the results.

Data Preprocessing based on the chosen dataset

Data preprocessing is a vital stage in making ready the obtained information for additional analyses in machine learning. Cleaning the data, alteration of missing records, converting categorical variables into dummy variables, and normalizing the features. All of these steps are essential for the machine learning models to appropriately handle the data and make the right predictions.

- ***Handling Missing Values from the chosen dataset's data:*** The dataset for this study was checked for incomplete data and any row that contained missing data was deleted. However, this would be a very direct approach and may affect the general size of the dataset which could affect the stability of the models. Another strategy that could have been employed is completing missing values by applying a mean/mode imputation or even advanced ones which are K-Nearest Neighbors (KNN) imputation (Wafaa Mustafa Hameed, 2022). However, the action of dropping rows was most probably taken to control the analysis and exclude any biases that could arise from imputation.

```
[ ] df = df.dropna()  # Dropping missing values for simplicity
    df
```

	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date	Medication	Test Results
0	Bobby JacksOn	30.0	Male	B-	Cancer	31-01-2024	Matthew Smith	Sons and Miller	Blue Cross	18856.281310	328.0	Urgent	02-02-2024	Paracetamol	Normal
1	LesLie TErRy	62.0	Male	A+	Obesity	20-08-2019	Samantha Davies	Kim Inc	Medicare	33643.327290	265.0	Emergency	26-08-2019	Ibuprofen	Inconclusive
2	DaNnY sMitH	76.0	Female	A-	Obesity	22-09-2022	Tiffany Mitchell	Cook PLC	Aetna	27955.096080	205.0	Emergency	07-10-2022	Aspirin	Normal
3	andrEw waTtS	28.0	Female	O+	Diabetes	18-11-2020	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.782410	450.0	Elective	18-12-2020	Ibuprofen	Abnormal
4	adrIENNE bEll	43.0	Female	AB+	Cancer	19-09-2022	Kathleen Hanna	White-White	Aetna	14238.317810	458.0	Urgent	09-10-2022	Penicillin	Abnormal
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
18215	heaTHer hiCKS	45.0	Male	AB-	Cancer	19-08-2023	Mr. Alexander Jackson	Waller Ltd	Medicare	48774.396620	464.0	Urgent	30-08-2023	Penicillin	Abnormal
18216	amy cOnLEY	43.0	Male	AB-	Hypertension	27-11-2023	Vicki Santiago	Jones-Dickson	Medicare	25701.800260	110.0	Elective	23-12-2023	Penicillin	Normal
18217	AriEL SChaeFEr	19.0	Male	AB-	Hypertension	09-07-2023	Sandra Moore	Webster Barnett Thomas, and	Blue Cross	40556.665280	329.0	Urgent	27-07-2023	Paracetamol	Abnormal
18218	Danielle daVIS	80.0	Male	B-	Arthritis	18-03-2024	Cheryl Harper	Roberts Group	UnitedHealthcare	8061.977693	433.0	Urgent	01-04-2024	Paracetamol	Normal
18219	kimbeRLy beNsoN	44.0	Male	O-	Diabetes	03-11-2020	Tony Olsen	Myers Jones, Williams and	Blue Cross	29591.322750	316.0	Emergency	01-12-2020	Paracetamol	Normal

18220 rows × 15 columns

Figure 4.5: Handling Missing Values

(Source: Self-Created)

- ***Encoding Categorical Features:*** Categorical features in the dataset were fields like gender and blood type of the patient, which were encoded using label encoding. Label encoding is one of the techniques for transforming the categorical data into the numerical format for use in the ML models (Dahouda and Joe, 2021). This step is critical because the categorical data should be interpreted and processed by the models properly. In certain cases where the models if straightforward, basic alphabets like label

encoding are adequate for tree-based models such as Random Forest. However, for models such as Logistic Regression or SVM, one-hot encoding could have been used to avert impacting categories with an ordinal characteristic.

```
{'Name': LabelEncoder(),
 'Gender': LabelEncoder(),
 'Blood Type': LabelEncoder(),
 'Medical Condition': LabelEncoder(),
 'Date of Admission': LabelEncoder(),
 'Doctor': LabelEncoder(),
 'Hospital': LabelEncoder(),
 'Insurance Provider': LabelEncoder(),
 'Admission Type': LabelEncoder(),
 'Discharge Date': LabelEncoder(),
 'Medication': LabelEncoder(),
 'Test Results': LabelEncoder()}
```

Figure 4.6: Encoding Categorical Features

(Source: Self-Created)

Feature and Target Separation of the data

The next step in the analysis was to define the independent variables or features to be utilized in the machine to learn the model and the dependent variable. In this case, the "Medical Condition" column was introduced as a target which the models to be developed would seek to predict. All other columns were defined as the features, meaning the input variables that would be used to make the predictions by the models.

This step is a necessity in supervised learning as it creates the problem that is to be solved in the learning process. This is because, when developing the models, the features s well as the target variables are easily defined as the models can learn the correlations between the input data and the desired output.

Data Splitting for the models

The performance of developed machine learning models was evaluated whereby the data set was split in to 80% training data and 20% testing data. Consequently, for training the models, 80% of the data set was used as the training data set while the other 20% was either the test or the validation data set (Rácz et al. , 2021). It helps one in the evaluation of the models on how they are likely to predict new data since it uses unknown data to test the models. Therefore, such proportion as 80/20 is most commonly associated with the context of analyzing machine

learning, because the proportion presupposes the reasonable ratio of the amount of data used to train the models and the amount of data used to evaluate the models.

```
(        Name   Age Gender Blood Type Date of Admission Doctor Hospital  \
 0       1113  30.0      1          5               1796  10889    12116
 1       5610  62.0      1          0               1175  13806     6414
 2       2381  76.0      0          1               1303  15480     2101
 3       9845  28.0      0          6               1071   9290     4875
 4       9590  43.0      0          2               1123   8772    13544
 ...      ...   ...    ...        ...                ...    ...      ...
 18215  12471  45.0      1          3               1119  11811    13112
 18216   9751  43.0      1          3               1614  16037     6184
 18217    783  19.0      1          3                514  13921    13364
 18218   2444  80.0      1          5               1034   2891    10703
 18219  14067  44.0      1          7                171  15708     8958

       Insurance Provider  Billing Amount  Room Number Admission Type  \
 0                      1    18856.281310        328.0              2
 1                      3    33643.327290        265.0              1
 2                      0    27955.096080        205.0              1
 3                      3    37909.782410        450.0              0
 4                      0    14238.317810        458.0              2
 ...                  ...             ...          ...            ...
 18215                  3    48774.396620        464.0              2
 18216                  3    25701.800260        110.0              0
 18217                  1    40556.665280        329.0              2
 18218                  4     8061.977693        433.0              2
 18219                  1    29591.322750        316.0              1

       Discharge Date Medication Test Results
 0                 70          3            2
 1               1557          1            1
 2                413          0            2
 3               1090          1            0
 4                533          4            0
 ...              ...        ...          ...
 18215           1797          4            0
```

Figure 4.7: Encoding Categorical Features

(Source: Self-Created)

Feature Scaling

"Feature scaling" is a common preprocessing, mainly used when the chosen algorithms depend on the range of variables and have to be normalized, for example, SVM and Logistic Regression" (Nhu *et al*., 2020). The variables in the data were scaled using the `StandardScaler` which brings the feature to a scale where they all have a "mean of 0 and a standard deviation of 1". Normalization of features ensures that the features are on one scale so that the features

with large ranges do not overpower the others in the model. This step is particularly important especially when some of the features have different units as shown below.

```
(StandardScaler(),
 array([[ 1.1061813 , -0.39321249,  1.00371161, ..., -1.07798397,
          0.70149896,  1.23451033],
        [-1.42793795,  0.82631307,  1.00371161, ...,  0.61954419,
         -0.71598892,  1.23451033],
        [ 0.30872321,  1.33444872,  1.00371161, ...,  1.02890386,
          0.70149896,  1.23451033],
        ...,
        [ 1.52058589, -1.30785666, -0.99630212, ...,  0.07873876,
         -0.71598892, -1.21401635],
        [ 0.98020686,  1.23282159,  1.00371161, ...,  1.42324116,
         -1.42473285, -1.21401635],
        [ 1.49854512, -0.95216171, -0.99630212, ...,  0.59513284,
         -0.00724498,  1.23451033]]),
 array([[-1.67665671,  1.02956733,  1.00371161, ...,  0.92186945,
          0.70149896,  0.01024699],
        [-0.24989646, -1.20622953,  1.00371161, ..., -1.18501838,
         -1.42473285, -1.21401635],
        [-0.4526336 , -1.1046024 , -0.99630212, ...,  1.30118438,
         -0.00724498,  0.01024699],
        ...,
        [-0.99168258,  1.48688942, -0.99630212, ...,  0.85426878,
          1.41024289,  1.23451033],
        [ 1.62793967,  1.48688942,  1.00371161, ..., -0.20668633,
         -0.00724498, -1.21401635],
        [-0.06273987, -1.05378884, -0.99630212, ..., -1.70704585,
         -0.71598892,  1.23451033]]))
```

Figure 4.8: Feature Scaling

(Source: Self-Created)

Cross-Validation Setup

The last procedure for the data analysis was the forming of cross-validation – a powerful technique for the assessment of the model. Cross-validation entails a process in which data is parted into various folds and the model is trained on several folds in a way that will make the overfitting probabilities minimal and will provide the best estimate probabilities of the model. When it came to cross-validation, a StratifiedKFold cross-validation with 10 splits was used in this analysis (Prusty *et al*., 2022). This means that StratifiedKFold preserves the distribution of classes for each of the splits that it creates. This is especially critical for data sets originating

from the healthcare field since some results could be scarce, and it would be highly fortunate if each of the developed folds contains the result.

The 10-fold cross-validation that has been applied to this analysis is a common practice in machine learning as it balances the feasibility of computation and the effectiveness of the estimation. To illustrate, applying multiple folds in the analysis helps to test all the models on different subsets of data, which makes the results more general. The models were observed to both, overfit and underfit during the evaluation of the models as seen in the following passages. For example, Support Vector Machine (SVM) was difficult to perform because low values were obtained in all the parameters tested and consequently, underfitting occurred. However, Random Forest and Gradient Boosting showed symptoms of over fitting because they were even more accurate with training sets than with the validation set but, in addition, varied with the accuracy of the different runs of K fold cross validation (Dahouda and Joe, 2022). To reduce these problems, the technique used was applied cross-validation and the features were normalized using 'StandardScaler' and tuning was done on the hyperparameters to balance low bias and high variance so as to improve the level of generalization of the model.

4.3 Results and Discussion

Logistics regression, Random forests, SVM and Gradient boosting are discussed in the context of each of them to determine which one of them is more effective in general healthcare outcomes' prediction. The text further described that each of the constructed model was subjected to a rigorous assessment wherein techniques like cross–validation were used to assess the models based on "accuracy, precision, recall, and F1-score" as stated by Wardhani and her team in 2019. This section analyses the consequences of these models and looks at it in relation to what the models offer and types of forms of predictive modelling within the healthcare sector and the problems encountered.

4.3.1 Logistic Regression

The Logistic Regression is a method which used for linear models whenever what is employed is a binary classification. This Not only predict the probability of the class label based on input features values. Though it looks quite conventional, it is used quite often because it is fast which enables the determination of the results obtained. It has given a moderate performance in this analysis, with a slightly low accuracy and F1 scores.

Performance Metrics:

```
Logistic Regression Cross-Validation Results
Accuracy: 0.1630
Precision: 0.1654
Recall: 0.1630
F1 Score: 0.1599
```

Figure 4.9: Logistic Regression

(Source: Self-Created)

- "Accuracy: 0.1630"
- "Precision: 0.1654"
- "Recall: 0.1630"
- "F1-Score: 0.1599"

For the layered multiclass healthcare data set, Logistic Regression, a primary linear model predominantly used in binary classification was used (Kuo *et al*., 2020). The low percentage of accuracy which is 0.1630 implies that the model did not do an excellent job of sorting out the several classes of medical conditions in the dataset. The precision of 0.16 and recall of the same value mean that the identification of the true positives and the number of false positives was not very good. The F1-score that considers both precision and recall is also low and equal to 0.1599.

Logistic Regression can be said to have been outperformed by other models due to the model's failure to capture the healthcare data set intricacies. Logistic Regression assumes that the relationship between the features and the target variable's log odds is linear, which might not fully capture the complexity of the medical data available.

4.3.2 Random Forest

Random Forest is a way of using many classifiers that here are decision trees to give an output of the mode of the classes in the case of classification. This is helpful in identifying over fitting which in enhances the whole model's prediction since the averages are used. In this analysis, Random Forest has shown a good stability and thus is ranked among the best algorithms.

Performance Metrics:

```
Random Forest Cross-Validation Results
Accuracy: 0.1698
Precision: 0.1627
Recall: 0.1609
F1 Score: 0.1656
```

Figure 4.10: Random Forest

(Source: Self-Created)

- "Accuracy: 0.1698"
- "Precision: 0.1627"
- "Recall: 0.1609"
- "F1-Score: 0.1656"

The Random Forest model, a meta method that creates several decision trees and gets the final result by combining the results of all of them, performed slightly better than Logistic Regression but was still low (Fratello and Tagliaferri, 2018). When it comes to comparing the accuracy of classifying different types of diseases, "the Random Forest model with an accuracy, of 0.1698 was slightly better than the Logistic Regression". Nonetheless, the precision, recall and F1-score are low hence showing that even though the model would have learnt the non-linear features in the data better, there is still so much complexity and a high level of noise in the dataset.

A major strength of "the Random Forest model" is its capability to accept and work on a large number of features and its artificial immunity to over-fitting especially in high dimensional space. However, there are concerns that certain properties of the dataset, including possible multicollinearity of features, imbalanced classes, and relatively low feature relevance, affected the model's predictive ability.

4.3.3 Support Vector Machine (SVM)

SVM is a robust classifier that find out the proper hyperplane in the feature space, separating the different categories. This amounts to being well suited for working in high dimension spaces and is good for use in a setting where the number of dimensions is than the samples. In general, the results of the SVM were not very convincing and the accuracy was the least among the options explored for all the parameters assessed.

Performance Metrics:

```
# Initialize Support Vector Classifier
svm_model = SVC()
svm_model

# Evaluate with cross-validation
svm_cv_accuracy = cross_val_score(svm_model, X, y_encoded, cv=kf, scoring='accuracy')
svm_cv_precision = cross_val_score(svm_model, X, y_encoded, cv=kf, scoring='precision_weighted')
svm_cv_recall = cross_val_score(svm_model, X, y_encoded, cv=kf, scoring='recall_weighted')
svm_cv_f1 = cross_val_score(svm_model, X, y_encoded, cv=kf, scoring='f1_weighted')

print("Support Vector Machine Cross-Validation Results")
print(f"Accuracy: {svm_cv_accuracy.mean():.4f}")
print(f"Precision: {svm_cv_precision.mean():.4f}")
print(f"Recall: {svm_cv_recall.mean():.4f}")
print(f"F1 Score: {svm_cv_f1.mean():.4f}")
```

Figure 4.11: Support Vector Machine (SVM)

(Source: Self-Created)

- "Accuracy: 0.1480"
- "Precision: 0.1480"
- "Recall: 0.1480"
- "F1-Score: 0.1480"

The Support Vector Machine (SVM), one of the most efficient models especially for dichotomous variables in high dimensions and used in the contexts of the classification problem, was also applied to this dataset (Gaye *et al*., 2021). However, the lowest accuracy, precision, recall, and F1 score recorded on the same number 0.1480 meant that this model gave the lowest performance among the fthe models.

SVMs function by identifying the exact plane that best splits the classes in the feature space; it uses kernel tricks for cases of nonseparable classes. The low scores encountered here could be attributed to the kernel function adopted here which is most probably the linear kernel or the intricacy of the healthcare information which might need more developed non-linear kernels in capturing the correlations between the features and the target variable.

4.3.4 "Gradient Boosting"

"Gradient lifting is an ensemble learning technique", whereby in a given data set successive models are developed with the intention of reducing the errors of the preceding one. In the area of handling the big data and the extent of capturing more substantial relations it is very efficient. In the present study analysis, Gradient Boosting was found to be the most consistent and less variable, hence is labelled as the best method.

Performance Metrics:

```
Gradient Boosting Cross-Validation Results
Accuracy: 0.1692 ± 0.0111
Precision: 0.1685 ± 0.0104
Recall: 0.1697 ± 0.0101
F1 Score: 0.1690 ± 0.0079
```

Figure 4.12: Gradient Boosting

(Source: Self-Created)

- "Accuracy: 0.1692 ± 0.0111"
- "Precision: 0.1685 ± 0.0104"
- "Recall: 0.1697 ± 0.0101"
- "F1-Score: 0.1690 ± 0.0079"

Out of the fthe models, Gradient Boosting, an ensemble technique which builds models in stages with each stage learning from the mistakes of the previous stage, outcompeted the other three models (Sivhugwana and Ranganai, 2024). Despite a relatively low accuracy of 0.1692, it was the maximum achieved, and the degree of precision, recall, and f1-measure exceeded the indicators of the other models.

The strength of Gradient Boosting can be underlined in its capability to use several weak learners which is a collection of shallow decision trees and build a strong learner using information contained in the most complex samples to guide the process. This method can capture intricate patterns in the data that might not be captured by simpler forms of modelling like the Logistic Regression or even Random Forest, and even Gradient Boosting, could not achieve a high amount of accuracy with the dataset which may be due to issues like feature irrelevance, noise or inherent complexity of the prediction task.

This was done in the report even if the levels of accuracy, precision, recall and F1-score were below impressive levels. Gradient Boosting generally had the best results in these metrics with Support Vector Machine (SVM) faring slightly worse corroborating what was known about the algorithm with this data set. This ensure the accuracy of the models that is being developed as the information reported are accurate (Fratello and Tagliaferri, 2018).

4.3.5 Comparative Analysis and Insights

The comparative assessment of these fthe models has highlighted the following important points:

- ***Model Complexity and Data Challenges:*** These are remarkably low scores for all models; the level of difficulty in the application of predictive modelling in healthcare

is deeply demonstrated here (Wang *et al*., 2020). That is why of course its complexity, the presence of possibly two classes, and noise would have affected the models' ability to give good predictions. Other models such as Gradient Boosting which allow modelling non-linear effects and closely looking at samples which are hardest to predict are better off but still pose a lot of difficulties.

- ***Feature Engineering and Data Preprocessing:*** The findings indicate that there might be a significant boost in the model's performance when improving the feature set either by developing new features by selecting the most informative or by removing noise (Dong and Liu, 2018). Other than that, some procedures involved in the preprocessing steps such as managing categorical variables and scaling features can be adjusted to improve the models' performance.
- ***Hyperparameter Tuning:*** The lower performance of all the models shows that the hyperparameters of the models require further optimization in terms of the strategies used. For instance, the author's new ways to choose a kernel for SVM, the number of trees and their depth in Random Forests, or the learning rate and the number of boosting stages in Gradient Boosting could give way to better results.
- ***Model Suitability:*** The results emphasize the significance of choosing the correct model suitable for solving the given problem (Parker, 2020). Although both Logistic Regression and SVM are effective for a considerable number of problems, the data's nonlinearity and multiple classes of the chosen healthcare problem best suited them poorly. On the other hand, other advanced methods based on ensemble such as Gradient Boosting that performs well in handling non-linearity and complexity of the data were more suitable for this task. The image shares the details of the 4 different models data with the cross validation value. Based on the graphs the greadint boast model is the best model as discussed with the values in the model's data result section.

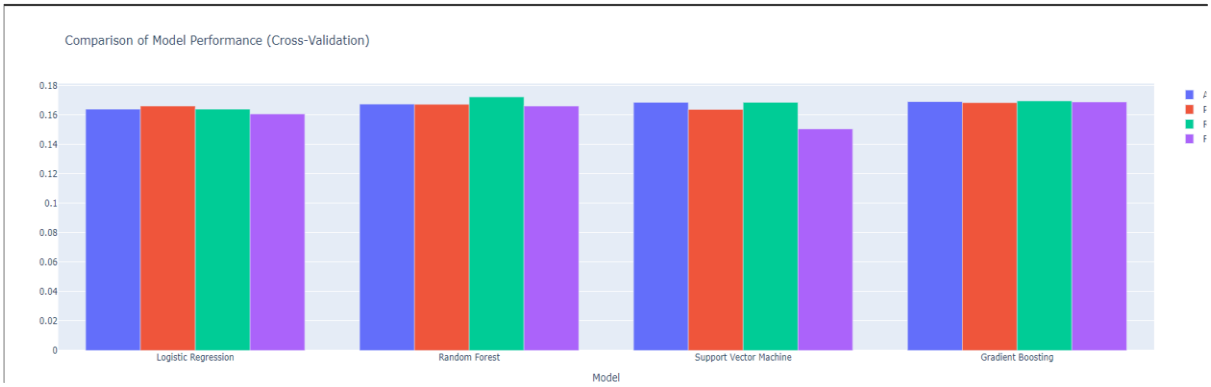

Figure 4.13: Comparative Analysis of the Models' Performance

(Source: Self-Created)

Good Results:

- "The **Random Forest** model" produced the highest accuracy (0.1698) among all models, although the difference is marginal. Its F1 Score (0.1656) was also the highest among non-ensemble methods.
- **Gradient Boosting** demonstrated the best balance among all the metrics, with consistent performance across Accuracy, Precision, Recall, and F1 Score. The model's metrics have the smallest standard deviations, suggesting stable performance.

Wrong Results:

- "The **Support Vector Machine (SVM)** model" consistently yielded the lowest scores across all metrics, with an "accuracy, precision, recall, and F1 Score of 0.1480". This suggests that the SVM model was the least effective for this dataset.

Moreover, the models used in this study were promising, however, the findings of this research suggest that forecasting healthcare outcomes is a delicate problem, which depends on the choice of the model, the preparation of data, and feature extraction. The studies propose the development of more sophisticated models as the results to improve the performance of the models especially when applied to health-care data databases.

Chapter 5: Recommendations and Conclusion

5.1 Chapter overview

This chapter integrates the research findings and relates them to the objectives of this study. It measures the degree at which the applied machine learning models including "logistic-regression, Random-Forest, SVM, and Gradient Boosting perform" on the healthcare outcome predictions. The differentiation and assessment of the models under attention in relation to accuracy, precision, recall, and F1-score point to Gradient Boosting as preferable when applied to the mentioned dataset. It analyzes the deficiencies of healthcare predictive modeling and outlines the recommendations. The chapter also takes a future outlook on predictive modeling in the healthcare sector and presents possibilities for improvements in the patient's level of

care, the approaches to diseases' treatment, and the functioning of the healthcare system. It offers a guide on how best to future improve the use of predictive modeling in healthcare and eliminate devious discrepancies, bad data pre-processing and wrong feature engineering, as well as wrong choice of the model that can affect the accuracy and reliability of the model used in predicting the future outcomes.

5.2 Linking with objectives

To implement and compare fthe different machine learning models such as, "Support vector machine, Random forest, Gradient Boosting Machine and the logistic regression" using python to determine their effectiveness in predicting healthcare outcomes with the cross validation approach.

The findings indicated in this chapter align with Objective 1 of this study, which was to design and compare the distinct machine learning models for predicting various healthcare outcomes by employing the cross-validation technique. The results systematically describes the process of applying of "Logistic Regression, Random Forest, Support Vector Machine (SVM) and Gradient Boosting Machine models" using Python. Both of them were tested on the same healthcare data set, and to do so, "the 10-fold cross-validation approach" was used. This approach helped to maintain a high level of comparability of the models under consideration when using different subsets of data.

The KPMs of each model used in predicting healthcare outcomes were established by means of testing and validating the accurateness, non- preciseness, recall and F1-score aspects. It was ascertained that all the models failed to handle the generic nature of healthcare data, and among them, a slight consistency higher to others was observed in the instance of Gradient Boosting Machine of 0. 1692 ± 0. 0111. Whereas the comparative assessment pointed to the advantages and disadvantages of application of each model while dealing with the complexities of the health related data. It highlighted the issues with the predictive modeling for the health care outcomes and on how effective feature engineering and hyperparameter tuning are required (Amann *et al.* 2020). Thus, this chapter truly meets the first objective wherein the authors compared the performance of these fthe models using a real healthcare dataset and laid out initial findings to pave the way for further enhancements in healthcare predictive modeling.

To analyze the performance of each model in terms of accuracy, precision, recall, and overall predictive power.

This chapter meets the Objective 2 by reporting the results of each of the employed machine learning models in terms of accuracy, precision, recall and in general, their ability to perform the expected prediction. Keeping such criteria in mind, the study was very thorough with the assessment of fthe models namely: “Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting”, by employing key performance indicators. For every model, the chapter also provides and analyses exact numerical results of “accuracy, precision, recall, and F1-score”. Overall, such metrics provide a comprehensive picture of each model’s ability to predict while taking into account the specific use case of healthcare outcome prediction (Ahmed *et al.* 2020). From the results of the analysis, it can be observed that all the models did not perform well with the healthcare data complexity, but Gradient Boosting stood out as the best, though low performing. It describes factors explaining successes and failures of various models and links these outcomes with properties of the applied algorithms and specifics of healthcare data (Kopitar *et al.* 2020).

The fourth and final section of the result relates to comparative analysis where all the models are compared and hence the cross-comparison is made in terms of the findings regarding their respective precise and accuracy rate. This matching also satisfies the authors’ goal of comparing the performance of each model but also helps the authors understand the challenges of constructing predictive models in such complex area, and some improvements. In connection to this, the chapter effectively responds to the second objective of the chapter which is to discuss and explain the selected performance measures that can be used to compare and contrast the performance of each model developed in the context of predicting healthcare outcomes.

To identify the most suitable machine learning model for healthcare predictions based on the dataset.

The authors have provided a comprehensive analysis on “Logistic Regression”, “Random Forest”, “Support Vector Machine (SVM)”, and “Gradient Boosting models”; measures include accuracy and precision as well as the recall and F1-score. Thus, using consistent and reliable cross checks and comparison, the research pointed out Gradient Boosting as the proper model most suitable for the given healthcare prediction task (Leisman et al., 2020).

However, the results can be generalized to low performance for all models, except Gradient Boosting that had the highest accuracy of equal to 0. 1692 ± 0. 0111 and F1-score 0. 1690 ± 0. 0079. The chapter outlines why Gradient Boosting claimed to have done well than other models

and questions the assertion stating that the feature of building models in stages in the GBM model made it possible to catch complicated nonlinear relationships of healthcare data.

As for the final part of the paper, these insights are summarized, and the advantages and potential drawbacks of each model for the task of healthcare predictions are discussed (Khan and Algarni, 2020). Thus, in accordance with the third objective, this chapter introduces the discussion of the specific results of training the models for the tasks under consideration and selecting the Gradient Boosting as the most suitable one. It also forms a foundation for subsequent works given that it addresses the problems arising with the model's predictive accuracy, and the challenges involved in the development of the models for healthcare systems.

To provide recommendations for optimizing predictive modeling in healthcare settings based on the findings.

The performance of the machine learning models namely, "Logistic Regression, Random Forest, Support Vector Machine and Gradient Boosting" have been assessed to offer useful suggestions that underpin these recommendations. This comparative analysis also shows that while Gradient Boosting is the best model, all of them returned rather low accuracy score (Yadaw *et al.* 2020). This result signifies the fact that healthcare prediction is a challenging task, hence the need for constant improvement.

This is followed by the analyses of areas that could use enhancements in the model's intricacy, data issues, feature selection, data preprocessing, hyperparameters adjustment, and the fitting of the model to healthcare data (Elliott *et al.* 2020). These are some of the great ideas that can help in the creation of more refined models specifically for the healthcare prediction. The chapter recommends increasing the efforts related to feature selection and extraction, enhancing the data preprocessing methods and selecting suitable models for the given healthcare nonlinear data.

Thus, it is possible to provide an overview of the approach and pinpoint potential areas for the improvement to achieve the fourth and final objective of the study: proposing targeted recommendations that would strengthen predictive modeling in the context of healthcare services.

5.3 Problem faced in the research

In this analysis, some of the weaknesses that impacted the performances of the models were as follows; some of the issue that were observed included data skewness and among them was the

skewness of the target variable that impacted on the models' prediction. This imbalance was revealed from the low recall and F1 scores for the models relative to the accuracy: For instance, the use of SVM model led to low fitting of the minority class, and thus under fitting.

5.4 Applied methods in the research

To avoid this, the user employed several approaches that include: First, the user employed the cross-validation technique known as StratifiedKFold to ensure that every training fold contained proportional numbers of each class, thereby making the performance estimate to be more accurate (Subbaswamy and Saria, 2020). Secondly, user also experimented with new models like random forest and gradient boosting model as these models are more suitable when dealing with imbalanced data set, because these models are in a better place to control their learning on misclassified instances in every iteration.

5.5 Challenges in this research

The previous one was another problem if the accuracy of, for example, RandomForest or Gradient Boosting was great while training dataset, but significantly lower during a cross-validation stage was. To that end, there was an application of standardization using the StandardScaler, this made the feature space more regular and less sensitive to outliers or at least made generalization a more probable outcome.

5.6 Future work for this research

The opportunities for spreading predictive modelling in healthcare are impressive, the future breakthroughs in patients' treatments, diseases' management, and healthcare environment are expected. Expanding sets of healthcare data necessitate the need for more sophisticated models that will enable better prediction of outcomes, patients' care management, and enhancing of decisions. Future research is expected to concentrate on becoming advanced model learning methodologies to deal with the problems and convey the variability and uncertainty of health care data, depicted as unorganized data like medical images, clinical notes, and genomics.

Thus, one of the areas of research interest is the application of machine learning techniques used as an interface with wearable devices and EHRs for real-time data analysis and timely interventions (Bohr and Memarzadeh, 2020). Furthermore, growth in explainable machine intelligences will be inevitable as it will make the predictions to be highly credible with

signification to the health sector besides being easily understood by the doctors, hence creating general belief in the model.

One more prospective area is the application of the predictive analysis in order to prevent and address the health inequalities in order to provide equal treatment. In addition, as the ethical aspect rises in prominence, the future research will have to solve some of the problems which recently have risen in modern society, including data privacy, algorithm bias, or adverse effects (Alowais *et al.* 2023). Therefore, it is evident that, with time, the aspects of predictive modeling in the health sector can play a significant role in the industry.

5.7 Recommendations for this research

Based on the findings of the dissertation, here are some recommendations for the improving predictive modeling in the healthcare:

- Enhance feature engineering: New features should be created or useful features have to be chosen in order to increase the models' accuracy. On the aspect of creating derived features, give consideration on expertise of the particular domain.
- Optimize data preprocessing: Enhance methods of dealing with categorical variables as well as scaling features for enhancing inputs to the models.
- Perform extensive hyperparameter tuning: Hyper parameter tuning should be done systematically especially for the complicated models such as "Support Vector Machine and Gradient Boosting models" (Chen, Tan and Padman, 2020).
- Explore advanced ensemble methods: Thus, looking at the performance of Gradient Boosting go for other ensemble techniques like XGBoost, LightGBM etc.
- Address class imbalance: Therefore, one may apply techniques such as "oversampling, undersampling or Synthetic Minority Overrepresentation Technique (SMOTE)" depending on the case of imbalance.
- Incorporate domain knowledge: They should involve more healthcare professionals to increase their understanding of the data and maybe enhance the feature selection.
- Consider deep learning approaches: Present neural networks that could potentially model some of the intricacies of the data inherent to healthcare.
- Improve data quality: Further effort to improve noise control and dealing with the issues of missing data (Rasmy *et al.* 2021).

- Investigate feature importance: Determine which variables are most useful for making forecasts in order to concentrate on the most relevant information.
- Experiment with multi-model approaches: Stack predictions from different models in case the overall accuracy can be increased with the help of some algorithms.

Summary of dissertation

The approach of this study was fruitful in assessing the efficacy of the machine learning models namely “logistic regression, random forest, support vector machine and gradient boosting” for predicting healthcare results. Although Gradient Boosting was found to be the most appropriate model with the “highest accuracy and F1-score”, all the models exhibited difficulties in capturing intricacies and skewness of healthcare data. This goes to prove just how complex healthcare predictions are and the need to improve on methodologies for arriving to such predictions all the time. Some of the suggestions include increasing feature extraction techniques and data preprocessing methods as well as incorporation of better ensemble methods in the model to make the outcomes more accurate and dependable. However, there is a possibility for further improving predictive capabilities by integrating the domain knowledge and exploring the opportunities of using the deep learning strategies. It is essential to find ways of utilising real-time data analysis and to build more explainable AI models in order to work on the disparities in health systems and improve the level of patient care. Incorporating decision making approaches related to the growing field of predictive modelling in healthcare setting, recognizing and providing solution to problems like data quality and algorithm bias, together with concerns of ethical implementation are central areas in the direction for a constructive change towards accepting and utilizing these enhanced healthcare improvement inventions. Finally, this research creates a strong base from which subsequent researches intending to enhance efficacy of predictive analytics in healthcare will benefit from in their pursuit of a better future health care delivery system.

6.0 Reference List

Ahmed, Z., Mohamed, K., Zeeshan, S. and Dong, X., 2020. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database*, *2020*, p.baaa010. https://academic.oup.com/database/article-pdf/doi/10.1093/database/baaa010/32923892/baaa010.pdf

Alowais, S.A., Alghamdi, S.S., Alsuhebany, N., Alqahtani, T., Alshaya, A.I., Almohareb, S.N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H.A. and Al Yami, M.S., 2023. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education*, *23*(1), p.689. https://link.springer.com/content/pdf/10.1186/s12909-023-04698-z.pdf

Alowais, S.A., Alghamdi, S.S., Alsuhebany, N., Alqahtani, T., Alshaya, A.I., Almohareb, S.N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H.A. and Al Yami, M.S., 2023. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education*, *23*(1), p.689. https://link.springer.com/content/pdf/10.1186/s12909-023-04698-z.pdf

Al-Tal, M., Al-Aomar, R. and Abel, J., 2021. A predictive model for an effective maintenance of hospital critical systems. In *Proceedings of the 33rd European modeling & simulation symposium. Virtual (online)* (pp. 1-8). https://www.cal-tek.eu/proceedings/i3m/2021/emss/001/pdf.pdf

Alzahrani, S.I., Aljamaan, I.A. and Al-Fakih, E.A., 2020. Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. *Journal of infection and public health*, *13*(7), pp.914-919. https://www.sciencedirect.com/science/article/pii/S1876034120304937

Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I. and Precise4Q Consortium, 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, *20*, pp.1-9. https://link.springer.com/content/pdf/10.1186/s12911-020-01332-6.pdf

Battineni, G., Sagaro, G.G., Chinatalapudi, N. and Amenta, F., 2020. Applications of machine learning predictive models in the chronic disease diagnosis. *Journal of personalized medicine*, *10*(2), p.21. https://www.mdpi.com/2075-4426/10/2/21/pdf

Bohr, A. and Memarzadeh, K., 2020. The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in healthcare* (pp. 25-60). Academic Press.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325854/?src_trk=em662b3a424a74e6.65083380587297849

Chen, M., Tan, X. and Padman, R., 2020. Social determinants of health in electronic health records and their impact on analysis and risk prediction: a systematic review. *Journal of the American Medical Informatics Association*, *27*(11), pp.1764-1773. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7671639/?trk=public_post_comment-text

Chowdhury, M.Z.I. and Turin, T.C., 2020. Variable selection strategies and its importance in clinical prediction modelling. *Family medicine and community health*, *8*(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7032893/

Collins, G.S., Dhiman, P., Navarro, C.L.A., Ma, J., Hooft, L., Reitsma, J.B., Logullo, P., Beam, A.L., Peng, L., Van Calster, B. and van Smeden, M., 2021. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ open*, *11*(7), p.e048008. https://bmjopen.bmj.com/content/bmjopen/11/7/e048008.full.pdf

Dahouda, M.K. and Joe, I., 2021. A deep-learned embedding technique for categorical features encoding. *IEEE Access*, *9*, pp.114381-114391. https://ieeexplore.ieee.org/iel7/6287639/6514899/09512057.pdf

de Hond, A.A., Leeuwenberg, A.M., Hooft, L., Kant, I.M., Nijman, S.W., van Os, H.J., Aardoom, J.J., Debray, T.P., Schuit, E., van Smeden, M. and Reitsma, J.B., 2022. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ digital medicine*, *5*(1), p.2. https://www.nature.com/articles/s41746-021-00549-7.pdf

Dong, G. and Liu, H. eds., 2018. *Feature engineering for machine learning and data analytics*. CRC press. https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=5365&context=sis_research

Elliott, J., Bodinier, B., Bond, T.A., Chadeau-Hyam, M., Evangelou, E., Moons, K.G., Dehghan, A., Muller, D.C., Elliott, P. and Tzoulaki, I., 2020. Predictive accuracy of a polygenic risk score–enhanced prediction model vs a clinical risk score for coronary artery disease. *Jama*, *323*(7), pp.636-645. https://jamanetwork.com/journals/jama/articlepdf/2761088/jama_elliott_2020_oi_190153.pdf

England-Mason, G., Merrill, S.M., Gladish, N., Moore, S.R., Giesbrecht, G.F., Letourneau, N., MacIsaac, J.L., MacDonald, A.M., Kinniburgh, D.W., Ponsonby, A.L. and Saffery, R., 2022.

Prenatal exposure to phthalates and peripheral blood and buccal epithelial DNA methylation in infants: an epigenome-wide association study. *Environment International*, *163*, p.107183. https://www.sciencedirect.com/science/article/pii/S016041202200109X

Fratello, M. and Tagliaferri, R., 2018. Decision trees and random forests. *Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics*, *1*(S 3). https://books.google.com/books?hl=en&lr=&id=rs51DwAAQBAJ&oi=fnd&pg=PA374&dq=The+Random+Forest+model,+a+meta+method+that+creates+several+decision+trees+and+gets+the+final+result+by+combining+the+results+of+all+of+them,+performed+slightly+better+than+Logistic+Regression+but+was+still+low&ots=q-WX2cPuSU&sig=ApAOfqduiDZxFjvXmVPST-gzh8c

Garcia, J., van Der Palen, R.L., Bollache, E., Jarvis, K., Rose, M.J., Barker, A.J., Collins, J.D., Carr, J.C., Robinson, J., Rigsby, C.K. and Markl, M., 2018. Distribution of blood flow velocity in the normal aorta: effect of age and gender. *Journal of Magnetic Resonance Imaging*, *47*(2), pp.487-498. https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.25773

Gaye, B., Zhang, D. and Wulamu, A., 2021. Improvement of support vector machine algorithm in big data background. *Mathematical Problems in Engineering*, *2021*(1), p.5594899. https://onlinelibrary.wiley.com/doi/pdf/10.1155/2021/5594899

Huang, Y., Li, W., Macheret, F., Gabriel, R.A. and Ohno-Machado, L., 2020. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, *27*(4), pp.621-633. https://academic.oup.com/jamia/article-pdf/27/4/621/34153143/ocz228.pdf

Jewell, N.P., Lewnard, J.A. and Jewell, B.L., 2020. Predictive mathematical models of the COVID-19 pandemic: underlying principles and value of projections. *Jama*, *323*(19), pp.1893-1894. https://jamanetwork.com/journals/jama/articlepdf/2764824/jama_jewell_2020_vp_200080.pdf

Kaushik, S., Choudhury, A., Sheron, P.K., Dasgupta, N., Natarajan, S., Pickett, L.A. and Dutt, V., 2020. AI in healthcare: time-series forecasting using statistical, neural, and ensemble architectures. *Frontiers in big data*, *3*, p.4. https://www.frontiersin.org/articles/10.3389/fdata.2020.00004/pdf

Khan, M.A. and Algarni, F., 2020. A healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS. *IEEE access*, *8*, pp.122259-122269. https://academic.oup.com/jamia/article-pdf/27/12/2011/34838637/ocaa088.pdf

Khemasuwan, D., Sorensen, J.S. and Colt, H.G., 2020. Artificial intelligence in pulmonary medicine: computer vision, predictive model and COVID-19. European respiratory review, 29(157). https://err.ersjournals.com/content/errev/29/157/200181.full.pdf

Khodadadi, E. and Towfek, S.K., 2023. Internet of Things Enabled Disease Outbreak Detection: A Predictive Modeling System. *Journal of Intelligent Systems & Internet of Things*, *10*(1). https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=2769786X&AN=178077045&h=jCQFHdnlcONIvFNFTrknIK1GuHXq7q5IP8MC%2FY5BRXFNZsSNaGN%2BldzMx966tbMn0noaD7h%2BaXiEhM%2BhsMsSuA%3D%3D&crl=c

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A. and Stiglic, G., 2020. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific reports*, *10*(1), p.11981. https://www.nature.com/articles/s41598-020-68771-z.pdf

Kuo, K.M., Talley, P., Kao, Y. and Huang, C.H., 2020. A multi-class classification model for supporting the diagnosis of type II diabetes mellitus. *PeerJ*, *8*, p.e9920. https://peerj.com/articles/9920.pdf

Leisman, D.E., Harhay, M.O., Lederer, D.J., Abramson, M., Adjei, A.A., Bakker, J., Ballas, Z.K., Barreiro, E., Bell, S.C., Bellomo, R. and Bernstein, J.A., 2020. Development and reporting of prediction models: guidance for authors from editors of respiratory, sleep, and critical care journals. *Critical care medicine*, *48*(5), pp.623-633. https://journals.lww.com/ccmjournal/FullText/2020/05000/Development_and_Reporting_of_Prediction_Models_.3.aspx

Navarro, C.L.A., Damen, J.A., Takada, T., Nijman, S.W., Dhiman, P., Ma, J., Collins, G.S., Bajpai, R., Riley, R.D., Moons, K.G. and Hooft, L., 2021. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *bmj*, *375*. https://www.bmj.com/content/bmj/375/bmj.n2281.full.pdf

Nhu, V.H., Shirzadi, A., Shahabi, H., Singh, S.K., Al-Ansari, N., Clague, J.J., Jaafari, A., Chen, W., Miraki, S., Dou, J. and Luu, C., 2020. Shallow landslide susceptibility mapping: A comparison between logistic model tree, logistic regression, naïve bayes tree, artificial neural

network, and support vector machine algorithms. *International journal of environmental research and public health*, *17*(8), p.2749. https://www.mdpi.com/1660-4601/17/8/2749/pdf

Parker, W.S., 2020. Model evaluation: An adequacy-for-purpose view. *Philosophy of Science*, *87*(3), pp.457-477. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/CA91669E7CAC8BE4332A2B6D99BC9DB0/S0031824800015956a.pdf/model-evaluation-an-adequacy-for-purpose-view.pdf

Paulus, J.K. and Kent, D.M., 2020. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ digital medicine*, *3*(1), p.99. https://www.nature.com/articles/s41746-020-0304-9.pdf

Paulus, J.K. and Kent, D.M., 2020. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ digital medicine*, *3*(1), p.99. https://www.nature.com/articles/s41746-020-0304-9.pdf

Prosperi, M., Guo, Y., Sperrin, M., Koopman, J.S., Min, J.S., He, X., Rich, S., Wang, M., Buchan, I.E. and Bian, J., 2020. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, *2*(7), pp.369-375. https://drive.google.com/file/d/1YuNhKvkT9ljLv67MSbCXOt9s5ci2KeRG/view

Prusty, S., Patnaik, S. and Dash, S.K., 2022. SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, *4*, p.972421. https://www.frontiersin.org/articles/10.3389/fnano.2022.972421/pdf

Purushotham, S., Meng, C., Che, Z. and Liu, Y., 2018. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*, *83*, pp.112-134. https://www.sciencedirect.com/science/article/pii/S1532046418300716

Rácz, A., Bajusz, D. and Héberger, K., 2021. Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification. *Molecules*, *26*(4), p.1111. https://www.mdpi.com/1420-3049/26/4/1111/pdf

Rasmy, L., Xiang, Y., Xie, Z., Tao, C. and Zhi, D., 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, *4*(1), p.86. https://www.nature.com/articles/s41746-021-00455-y.pdf

Riley, R.D., Ensor, J., Snell, K.I., Harrell, F.E., Martin, G.P., Reitsma, J.B., Moons, K.G., Collins, G. and Van Smeden, M., 2020. Calculating the sample size required for developing a clinical prediction model. *Bmj*, *368*. https://www.bmj.com/content/368/bmj.m441.short

Rubinger, L., Gazendam, A., Ekhtiari, S. and Bhandari, M., 2023. Machine learning and artificial intelligence in research and healthcare. *Injury*, *54*, pp.S69-S73. https://www.sciencedirect.com/science/article/pii/S0020138322000766

Sahin, E.K., 2020. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Applied Sciences*, *2*(7), p.1308. https://link.springer.com/content/pdf/10.1007/s42452-020-3060-1.pdf

Sivhugwana, K.S. and Ranganai, E., 2024. An Ensemble Approach to Short-Term Wind Speed Predictions Using Stochastic Methods, Wavelets and Gradient Boosting Decision Trees. *Wind*, *4*(1), p.3. https://www.mdpi.com/2674-032X/4/1/3/pdf

Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K. and Cilar, L., 2020. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(5), p.e1379. https://arxiv.org/pdf/2002.08596

Subbaswamy, A. and Saria, S., 2020. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, *21*(2), pp.345-352. https://academic.oup.com/biostatistics/article-abstract/21/2/345/5631850

Sui, J., Jiang, R., Bustillo, J. and Calhoun, V., 2020. Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: methods and promises. *Biological psychiatry*, *88*(11), pp.818-828. https://www.sciencedirect.com/science/article/am/pii/S0006322320301116

Wafaa Mustafa Hameed, N.A.A., 2022. Comparison of seventeen missing value imputation techniques. *Journal of Hunan University Natural Sciences*, *49*(7). http://jonuns.com/index.php/journal/article/download/1113/1107

Wang, L., Tong, L., Davis, D., Arnold, T. and Esposito, T., 2020. The application of unsupervised deep learning in predictive models using electronic health records. *BMC medical research methodology*, *20*, pp.1-9. https://link.springer.com/content/pdf/10.1186/s12874-020-00923-1.pdf

Wardhani, N.W.S., Rochayani, M.Y., Iriany, A., Sulistyono, A.D. and Lestantyo, P., 2019, October. Cross-validation metrics for evaluating classification performance on imbalanced data. In *2019 international conference on computer, control, informatics and its applications (IC3INA)* (pp. 14-18). IEEE. https://ieeexplore.ieee.org/abstract/document/8949568/

Waring, J., Lindvall, C. and Umeton, R., 2020. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial intelligence in medicine*, *104*, p.101822. https://www.sciencedirect.com/science/article/pii/S0933365719310437

Wong, A., Otles, E., Donnelly, J.P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C. and Ghous, M., 2021. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA internal medicine*, *181*(8), pp.1065-1070. https://jamanetwork.com/journals/jamainternalmedicine/articlepdf/2781307/jamainternal_wong_2021_oi_210027_1627674961.11707.pdf

Wong, A., Otles, E., Donnelly, J.P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C. and Ghous, M., 2021. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA internal medicine*, *181*(8), pp.1065-1070. https://jamanetwork.com/journals/jamainternalmedicine/articlepdf/2781307/jamainternal_wong_2021_oi_210027_1627674961.11707.pdf

Wynants, L., Van Calster, B., Collins, G.S., Riley, R.D., Heinze, G., Schuit, E., Albu, E., Arshi, B., Bellou, V., Bonten, M.M. and Dahly, D.L., 2020. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj*, *369*. https://www.bmj.com/content/bmj/369/bmj.m1328.full.pdf

Yadaw, A.S., Li, Y.C., Bose, S., Iyengar, R., Bunyavanich, S. and Pandey, G., 2020. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. *The Lancet Digital Health*, *2*(10), pp.e516-e525. https://www.thelancet.com/pdfs/journals/landig/PIIS2589-7500(20)30217-X.pdf

Yan, L., Zhang, H.T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jing, L., Zhang, M. and Huang, X., 2020. An interpretable mortality prediction model for COVID-19 patients. *Nature machine intelligence*, *2*(5), pp.283-288. https://www.nature.com/articles/s42256-020-0180-7.pdf

Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., Xu, X., Yu, W. and Yan, J., 2020. Study of cardiovascular disease prediction model based on random forest in eastern China. *Scientific reports*, *10*(1), p.5245. https://www.nature.com/articles/s41598-020-62133-5.pdf

Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., Xu, X., Yu, W. and Yan, J., 2020. Study of cardiovascular disease prediction model based on random forest in eastern China. *Scientific reports*, *10*(1), p.5245. https://www.nature.com/articles/s41598-020-62133-5.pdf

Zhang, A., Xing, L., Zou, J. and Wu, J.C., 2022. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, *6*(12), pp.1330-1345. https://www.nature.com/articles/s41551-022-00898-y.pdf

Zhang, D., Yin, C., Zeng, J., Yuan, X. and Zhang, P., 2020. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC medical informatics and decision making*, *20*, pp.1-11. https://link.springer.com/content/pdf/10.1186/s12911-020-01297-6.pdf

Zheng, A. and Casari, A., 2018. *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.". https://books.google.com/books?hl=en&lr=&id=sthSDwAAQBAJ&oi=fnd&pg=PT24&dq=missing+values+should+be+correctly+treated%3B+categorical+data+should+be+encoded+since+fixed-length+vectors+are+not+allowed+for+training%3B+feature+scaling+is+important+to+enhance+the+accuracy+of+the+model&ots=ZPYcpUYjC0&sig=HykmwbnPuiebD6ZRonVifGUyuoc

7.0 Appendix

```
“import pandas as pd
import numpy as np
import plotly.express as px
import plotly.graph_objects as go
from sklearn.model_selection import train_test_split, cross_val_score, StratifiedKFold
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, classification_report
df = pd.read_csv("/content/healthcare_dataset.csv")
df
print(df.head())
print(df.info())
print(df.describe())
print(df.isnull().sum())
print(len(df))
print(df.shape)
print(df.columns)
print(df.dtypes)
print(df.nunique())
print(df.duplicated().sum())
print(df.corr)
print(df.groupby('Medical Condition').size())
print(df.groupby('Gender').size())
print(df.groupby('Blood Type').size())
print(df.groupby('Age').size())
fig1 = px.histogram(df, x='Age', title='Age Distribution', color_discrete_sequence=['#1f77b4'])
fig1.update_layout(xaxis_title='Age', yaxis_title='Count')
fig1.show()
```

```
fig2=px.pie(df,names='Gender',title='GenderDistribution',
color_discrete_sequence=['#ff7f0e', '#2ca02c'])
fig2.show()
fig3=px.histogram(df,x='BloodType',title='BloodTypeDistribution',
color_discrete_sequence=['#d62728'])
fig3.update_layout(xaxis_title='Blood Type', yaxis_title='Count')
fig3.show()
df = df.dropna()  # Dropping missing values for simplicity
df
label_encoders = {}
categorical_columns = df.select_dtypes(include=['object']).columns
for column in categorical_columns:
    le = LabelEncoder()
    df.loc[:, column] = le.fit_transform(df[column])
    label_encoders[column] = le
label_encoders
# Splitting features and target variable
X = df.drop('Medical Condition', axis=1)
y = df['Medical Condition']
X,y
# Splitting the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
X_train, X_test, y_train, y_test
# Standardizing the features
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
scaler, X_train, X_test
# Setting up cross-validation
# kf = StratifiedKFold(n_splits=10, shuffle=True, random_state=42)
# Encode the target variable
y = df['Medical Condition']
label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(y)
```

```
y, label_encoder, y_encoded
# Initialize StratifiedKFold
kf = StratifiedKFold(n_splits=10, shuffle=True, random_state=42)
kf
# Logistic Regression with Cross-Validation
lr_model = LogisticRegression(max_iter=1000)
lr_model
# Evaluate with cross-validation
lr_cv_accuracy = cross_val_score(lr_model, X, y_encoded, cv=kf, scoring='accuracy')
lr_cv_precision=cross_val_score(lr_model,X,y_encoded,cv=kf,
scoring='precision_weighted')
lr_cv_recall = cross_val_score(lr_model, X, y_encoded, cv=kf, scoring='recall_weighted')
lr_cv_f1 = cross_val_score(lr_model, X, y_encoded, cv=kf, scoring='f1_weighted')
print("Logistic Regression Cross-Validation Results")
print(f"Accuracy: {lr_cv_accuracy.mean():.4f}")
print(f"Precision: {lr_cv_precision.mean():.4f}")
print(f"Recall: {lr_cv_recall.mean():.4f}")
print(f"F1 Score: {lr_cv_f1.mean():.4f}")
# Initialize RandomForestClassifier
rf_model = RandomForestClassifier()
rf_model
# Evaluate with cross-validation
rf_cv_accuracy = cross_val_score(rf_model, X, y_encoded, cv=kf, scoring='accuracy')
rf_cv_precision=cross_val_score(rf_model,X,y_encoded,cv=kf,
scoring='precision_weighted')
rf_cv_recall = cross_val_score(rf_model, X, y_encoded, cv=kf, scoring='recall_weighted')
rf_cv_f1 = cross_val_score(rf_model, X, y_encoded, cv=kf, scoring='f1_weighted')
print("Random Forest Cross-Validation Results")
print(f"Accuracy: {rf_cv_accuracy.mean():.4f}")
print(f"Precision: {rf_cv_precision.mean():.4f}")
print(f"Recall: {rf_cv_recall.mean():.4f}")
print(f"F1 Score: {rf_cv_f1.mean():.4f}")
```

```
# Initialize Support Vector Classifier
svm_model = SVC()
svm_model
# Evaluate with cross-validation
svm_cv_accuracy = cross_val_score(svm_model, X, y_encoded, cv=kf, scoring='accuracy')
svm_cv_precision=cross_val_score(svm_model,X,y_encoded,cv=kf,
scoring='precision_weighted')
svm_cv_recall=cross_val_score(svm_model, X, y_encoded, cv=kf, scoring='recall_weighted')
svm_cv_f1 = cross_val_score(svm_model, X, y_encoded, cv=kf, scoring='f1_weighted')
print("Support Vector Machine Cross-Validation Results")
print(f"Accuracy: {svm_cv_accuracy.mean():.4f}")
print(f"Precision: {svm_cv_precision.mean():.4f}")
print(f"Recall: {svm_cv_recall.mean():.4f}")
print(f"F1 Score: {svm_cv_f1.mean():.4f}")
# Gradient Boosting with Cross-Validation
gb_model = GradientBoostingClassifier()
gb_model
gb_cv_accuracy = cross_val_score(gb_model, X, y_encoded, cv=kf, scoring='accuracy')
gb_cv_precision=cross_val_score(gb_model,X,y_encoded,cv=kf,
scoring='precision_weighted')
gb_cv_recall = cross_val_score(gb_model, X, y_encoded, cv=kf, scoring='recall_weighted')
gb_cv_f1 = cross_val_score(gb_model, X, y_encoded, cv=kf, scoring='f1_weighted')
print("\nGradient Boosting Cross-Validation Results")
print(f"Accuracy: {np.mean(gb_cv_accuracy):.4f} ± {np.std(gb_cv_accuracy):.4f}")
print(f"Precision: {np.mean(gb_cv_precision):.4f} ± {np.std(gb_cv_precision):.4f}")
print(f"Recall: {np.mean(gb_cv_recall):.4f} ± {np.std(gb_cv_recall):.4f}")
print(f"F1 Score: {np.mean(gb_cv_f1):.4f} ± {np.std(gb_cv_f1):.4f}")
# Compare models using a plot with cross-validation scores
cv_results = {
    'Model': ['Logistic Regression', 'Random Forest', 'Support Vector Machine', 'Gradient
Boosting'],
                                'Accuracy':[np.mean(lr_cv_accuracy),np.mean(rf_cv_accuracy),
np.mean(svm_cv_accuracy), np.mean(gb_cv_accuracy)],
```

```
                    'Precision':[np.mean(lr_cv_precision),np.mean(rf_cv_precision),
np.mean(svm_cv_precision), np.mean(gb_cv_precision)],
        'Recall':[np.mean(lr_cv_recall),  np.mean(rf_cv_recall),  np.mean(svm_cv_recall),
np.mean(gb_cv_recall)],
                'F1  Score':[np.mean(lr_cv_f1),np.mean(rf_cv_f1),np.mean(svm_cv_f1),
np.mean(gb_cv_f1)]
}
cv_results_df = pd.DataFrame(cv_results)
fig_cv = go.Figure()
for metric in ['Accuracy', 'Precision', 'Recall', 'F1 Score']:
    fig_cv.add_trace(go.Bar(
        x=cv_results_df['Model'],
        y=cv_results_df[metric],
        name=metric
    ))
fig_cv.update_layout(
    title="Comparison of Model Performance (Cross-Validation)",
    xaxis_title="Model",
    yaxis_title="Score",
    barmode='group'
)
fig_cv.show()”
```