



Department of Computer Science

MSc Data Science and Analytics

Academic Year 2020-2021

Health monitoring and threat detection through digital processing

TUSHAR TAYAL

2141441

A report submitted in partial fulfilment of the requirement for the degree of Master of Data Science
and Analytics

Brunel University

Department of Computer Science

Uxbridge, Middlesex UB8 3PH

United Kingdom

Tel: +44 (0) 1895 203397

Fax: +44 (0) 1895 251686

Abstract

The current study was conducted to develop an algorithm that could detect and identify health risks in patients through digital processing of patient data regarding health indicators. For contextualisation, the study was focused on the patient data obtained from the National Health Services, UK and Kaggle. The study employed a secondary quantitative methodology for data collection and analysis, which was concomitant with development of detection algorithm. For algorithm development and test of impacts, RStudio was utilised which also facilitated in examining the accuracy of the diagnostic tests as well as the significance of the models for prediction of each variable. The study found that whereas anemia could be considerably predicted by the suggested algorithms, creatinine phosphokinase could only be predicted insignificantly. The other blood pressure and diabetes indices are reasonably predictable. The suggested model's capacity to identify false negative outcomes is moderate in terms of diagnostic accuracy. Additionally, research shown that diagnostic costs and accuracy are negatively correlated, with rising precision leading to an exponential increase in related expenditures.

ACKNOWLEDGEMENT

It gives me immense pleasure in completing my dissertation, titled as “Health Monitoring and Threat Detection through Digital Processing”. This topic has helped me gain ample knowledge on concepts and importance of data science in health sector and research programs. I would like to acknowledge the assistance of all the people who have helped me complete this research program successfully. I would wish to show my gratitude towards my supervisor Mr. Fotios Spyridonis who had helped me extensively in this study and without whose support this would have been difficult. Mr Fotios Spyridonis has taken care of all the doubts and problems faced by me at all points of research program.

Thanking you all

I certify that the work presented in the dissertation is my own unless referenced

Signature... TUSHAR TAYAL

Date....5TH SEPT 2022..

TOTAL NUMBER OF WORDS: 11,080

Contents

Abstract	2
CHAPTER 1 – INTRODUCTION	8
1.1 Background of Study	8
1.2 Aim and Objectives	10
1.2.1 Aim	10
1.2.2 Objectives	10
1.3 Research Questions	10
1.4 Problem Statement	11
1.5 Significance of Study	11
1.6 Rationale of Study	11
1.7 Dissertation Format	12
Chapter Two: Literature Review	13
2.1 Introduction	13
2.2 Digital health and digital processing	13

2.3 Health Statistics and diseases that are led by Obesity	14
2.4 Use of AI in analysing the statistical data in health monitoring and threat detection	16
2.5 The efficiency of machine learning, deep learning and image processing algorithms in health monitoring and threat detection	17
2.6 Factors affecting health monitoring and threat detection through digital processing	18
2.6.1 Privacy	18
2.6.2 Smart devices	18
2.6.3 Physical and mental health.....	19
2.7 Theoretical framework.....	19
2.8 Literature Gap	20
2.9 Chapter Summary	20
CHAPTER THREE: METHODOLOGY	22
3.1 Introduction.....	22
3.2 Research design.....	22
3.3. Research approach.....	23
3.4 Data collection.....	24

3.5 Data analysis	25
3.6 Ethical considerations	25
3.7 Chapter summary	25
CHAPTER 4: FINDINGS AND DISCUSSION	26
4.1 Introduction.....	26
4.2 Findings.....	26
4.2.1 Descriptive Statistics.....	26
4.2.2 Regression Analysis.....	29
4.2.3 Logistic Regression Model	33
4.4.4 Interpreting the outcomes of a dataset through a logistic regression model.....	35
4.4.5 Assessing the predictive ability of the model	36
4.4.5 Support Vector Machine on Heart Monitoring Dataset.....	41
4.3 Discussion.....	47
CHAPTER FIVE: CONCLUSION.....	50
5.1 Summary of findings.....	50
Recommendations	51

Conclusion..... 52

Future implications..... 53

References..... 54

APPENDIX A: ETHICAL APPROVAL

APPENDIX B: RESULT CHUNKS AND CODE

CHAPTER 1 – INTRODUCTION

1.1 Background of Study

Health issues are increasing quickly day by day in the modern world. The number of people dying yearly is 55.3 million, and daily, it is 151,600. If the death rate is considered hourly, it is 6316 people dying in just an hour. World Health Organization (2020) stated in a report that the most common diseases that lead to death are heart, cancer, lung disease, stroke, Alzheimer's, diabetes, flu, kidney disease, HIV, and pneumonia. Furthermore, WHO (2020) elaborated that heart disease is responsible for 16 percent of the total deaths worldwide. From 2000 to 2019, deaths due to heart problems increased from 2 million to 8.9 million. The top health concerns that are common nowadays and can lead to chronic diseases are obesity, cancer, diabetes, flu, mental illness, AIDS, BMI, and blood pressure. Obesity in the world has increased three times since 1975, according to the report WHO (2021). In 2020, 39 million children who are under five years of age will be overweight. Another health issue that the world is facing is HIV and AIDS. Acquired Immunodeficiency Syndrome (AIDS) can result from an HIV infection (Human Immunodeficiency Virus). The immune system gradually deteriorates and falls due to AIDS, increasing cancer risk and life-threatening infections. In 2017, the rate of HIV/AIDS was 50 percent higher than the death rate of malaria (Roser and Ritchie, 2019).

Xu and Xu (2017) describe health monitoring as a technique or method to monitor activity during the operation or prevention of an organ failure through the integration of the structure and system. A health monitoring system was first developed in China in 1971 and used in the coal mine (Xu and Xu 2017). Health monitoring was integrated with computers in the late 1970s, which allowed the scanning process (Ovando et al., 2018). Ovando et al. (2018) further stated that health monitoring is essential for preventing and detecting different diseases in their early stage to reduce the patient's suffering and decrease the medical treatment cost. Detection of disease in the early stage does not cost, and the disease can be cured; however, the late detection of the disease might lead to surgery that costs a lot, and a patient may suffer pain. The diagnosis of the disease and rapid treatment can completely change the treatment process and procedures. This is true in the case of diseases like cardiovascular and diabetes, as the different sensors can determine the risks of disease by monitoring vital signs.

Information processing digitally has a long history of being created to make certain repetitive operations faster and simpler (Fieschi, 2018). Digital processing is considered independent that does not need assistance from anyone. Fieschi (2018) further explains that automation of several tasks makes it easy to solve complex and challenging functions that are impossible to solve manually or take much precious time. Digital processing captures the data and organizes it in a way that will make it easier to execute the operation. However, digital processing filters or algorithms use much computational power.

The internet of medical things (IoMT) is a crucial component of intelligent health monitoring (SHM) in the internet of things (IoT) technology. The internet of things is the idea of interconnecting electronic devices over a network to enable data exchange for a particular application domain (Sujith et al., 2022). The Internet of Medical Things is a network of interlinked electronic devices specifically designed for the medical and healthcare industry. Examples include remote and telemedicine treatment systems, disease, abnormalities, and patient monitoring and conditioning. IoT and SHM are additions to hospital medical systems that allow patients to be treated without carelessness in contrast to conventional approaches (Sujith et al., 2022). Over the past few decades, life safety and lowering inspection costs have been top priorities. Other emerging methods include SHM, General Smart Monitoring (GSM), and Artificial Intelligence, including machine learning, deep learning, transfer learning, and computer vision.

According to Yang et al. (2019), machine learning (ML) is divided into two categories: deep learning (DL) and transfer learning (TL), which consists of multiple layers that are utilised to acquire valuable data, which is then used in its applications to deal with big data that have been successfully confirmed on various platforms. Yang et al. (2019) further stated that by gathering confidential data with multiple layers of deep learning, DL gives valuable information like predicting the possible future disease in a person or a particular region. SHM and telemedicine are just two of the many study fields that use DL models. Deep learning and computer vision have significant benefits for monitoring health (Sujith et al., 2022). Computer vision and deep learning models provide accurate and efficient information that is useful for gathering large amounts of data and making future predictions (Yang et al., 2019). It reduces the time to report severe cases, make quick and straightforward diagnoses, and identify potential diseases that can develop due to existing ones that can save many lives.

Obesity has elevated to a serious global concern in the modern era. An excessive or abnormal amount of body fat is referred to as obesity. People are increasingly leading unhealthy lifestyles, indulging in excessive junk food consumption, sleeping late, and spending a lot of time sitting down. Adolescents in particular are impacted by their unconsciously held attitudes. It is a very complex sickness that is known as a medical issue. It encourages the spread of complicated disorders such as liver cancer, heart disease, and stroke. As deadly epidemics of diabetes, cardiovascular disease, cancer, osteoarthritis, persistent renal disease, stroke, hypertension, and other fatal diseases have shown, it can occasionally result in death. This makes it essential to predict the obesity and the upcoming diseases using the digital processing and digital health which can potentially decrease the chronic diseases and deaths caused by it.

1.2 Aim and Objectives

1.2.1 Aim

The study aims to monitor health statistics and predict possible threats of diseases caused by obesity on a large scale in a particular region to aware people and the government improve facilities.

1.2.2 Objectives

The study has the following objectives:

1. First, to make predictions of the upcoming diseases by monitoring health daily.
2. To analyse the health statistics data.
3. To determine the diseases that are caused by obesity and blood pressure
4. To recommend the best machine learning, deep learning, and image processing algorithm.

1.3 Research Questions

The study has the following research questions:

- What is the need to make predictions on health and its potential impact on the public?
- What role can artificial intelligence play in predicting upcoming diseases by analysing health statistics data?

- What are the possible diseases that are led by bold pressure and obesity?
- What are possible suggestions for machine learning, deep learning, and image processing algorithms?

1.4 Problem Statement

Poor health and the number of patients is increasing daily due to unhealthy lifestyles and eating habits. Covid 19 was a global pandemic that envisaged a considerable world population. The disease was viral, and special SOPs following were demanded to be safe from it, leading people to adopt a sedentary lifestyle that increases cholesterol, blood pressure, and obesity. The current study has raised the concern about monitoring health that can be useful for people to check their health and maintain their diet and health accordingly. Additionally, monitoring daily health can predict future diseases that can cause a problem. The modern world demands to predict medical conditions and diseases before they can cause a severe threat. The requirement of the modern world supports the monitoring of health and threat via digital processing.

1.5 Significance of Study

The study has discussed a vital concern to evaluate the problem of the modern world concerning health impacts. The modern age has to face modern pandemics as challenges. Covid 19 is an example of this context that has proved critical in the long run for threatening the planet (Lv et al., 2022). The research is necessary because it frames the importance of predicting the health issues that a large group of people might face. The use of machine learning, deep learning, and image processing algorithms are elaborated with the help of previous research. The study is essential for preparing for the upcoming waves of covid 19 and the issues caused by lifestyle, habits, and diet. The study recommends algorithms that can be used to track health and predict the possible diseases that result from previous diseases.

1.6 Rationale of Study

The study has a specific reason for the post-pandemic era and common health problems that could convert into chronic diseases. The study can contribute to predicting health issues. It reviews the health problems in the time of pre-pandemic and post-pandemic. It has focused on modern technology tools like machine learning, deep learning, image processing, and Building information models, which are essential in the long run to review health support (Costa and

Peixoto, 2020). The study has a specific reason not only for evaluating the past concern of the predicting diseases but also for recommendation support for the future to handle the covid 19 pandemic and health issues in a future scenario.

1.7 Dissertation Format

The study has been divided into the following chapters:

- Chapter 1: Introduction has covered the aim, objectives, research questions, and study background. It has discussed the significance, rationale, and problem statement.
- Chapter 2: The literature review has critically discussed the literature for the study's main objectives. It gave a theoretical framework and literature gap.
- Chapter 3: Methodology has framed the research mode by evaluating the philosophy, approach, and design. It evaluated the data collection, analysis, and display methods in addition to ethical considerations and limitations.
- Chapter 4: Findings and Discussion has summarized the main data compilation and displayed it in an understandable way.
- Chapter 5: Conclusion has summed up the prominent opinions and gives recommendations.

Chapter Two: Literature Review

2.1 Introduction

The cost of healthcare has been a major concern for many people recently. Numerous medical applications, such as early diagnosis and real-time monitoring, may be supported by wireless communications and the Internet of Things. The purpose of this chapter is to review the past studies and research on health monitoring and threat detection through digital processing. The review will be focusing on the possible threats of diseases caused by obesity.

2.2 Digital health and digital processing

Digital health and digital processing have started to gain attention in the modern era after technological development. In recent years, advances in digital information processing have been made to expedite and make routine processes easier (Fieschi, 2018). According to Fieschi (2018), automation of several processes makes it easy to handle challenging and complicated issues that would take a long time or be impossible to resolve manually. To carry out the procedure, data is gathered and digitally organised. However, algorithms and filters used in digital processing need a lot of computational power. According to Kalid et al. (2018), the general public has recognised how patients' increased use of health monitoring keeps them out of hospitals. Cyber-physical systems may smoothly combine the physical and digital worlds thanks to computer-based algorithms. A process is managed and controlled by a cyber-physical system. Physical and software-based structures are intricately connected, enabling them to function in a variety of spatial and temporal dimensions, exhibit a wide range of different behaviours, and interact in a variety of contexts. A wide, interdisciplinary phrase, "digital health," or "digital healthcare," refers to concepts that originate at the intersection of technology and healthcare. Digital health transforms the healthcare sector by combining software, hardware, and services (Mathews et al., 2019). Digital health includes mobile health (mHealth) apps, electronic health records (EHRs), electronic medical records (EMRs), wearable technologies, telehealth and telemedicine, as well as personalised treatment. The use of digital processing in digital health is helping to improve and advance human health everywhere.

Due to the substantial use of information and communication technology (ICT), healthcare has experienced major digital revolutions (Al-Shorbaji, 2021). However, due to ignorance of the availability of the service, and other potential causes, many patients do not have access to healthcare services (Vargo et al., 2021). Li et al. (2021) argued that the potential of telemedicine has been to offer safe, more affordable care without requiring the patient or individual to be in the exact location or room as the healthcare professional. This means the patient does not need to leave their current location to get healthcare services and reach the care site. Furthermore, Bhavnani, Narula, and Sengupta (2016) added that social and physical isolation during COVID-19 has led to a sharp rise in the use of telemedicine services across all nations.

Aman et al. (2021) stated that the Internet of Medical Things is a network of interlinked electronic devices specifically designed for the medical and healthcare industry. Examples include remote and telemedicine treatment systems, disease, abnormalities, and patient monitoring and conditioning. IoT and SHM are additions to hospital medical systems that allow patients to be treated without carelessness in contrast to conventional approaches (Sujith et al., 2022). Deep learning and computer vision have significant benefits for monitoring health (Sujith et al., 2022). Computer vision and deep learning models provide accurate and efficient information that is useful for gathering large amounts of data and making future predictions (Yang et al., 2019). The pandemic is showing high concern in this regard to impact the context of significant concerns. It reduces the time to report severe cases, make quick and straightforward diagnoses, and identify potential diseases that can develop due to existing ones that can save many lives. Technology plays a crucial role in improving healthcare, which can be observed in the time of covid 19 and the use of technology that helps individuals to maintain their health.

2.3 Health Statistics and diseases that are led by Obesity

Both developed and developing nations are experiencing levels of obesity. Obesity is becoming more common among people of all ages, but it is particularly prevalent in kids and teenagers (Bendor et al., 2020). Over the past five decades, the incidence of obesity has continuously climbed, and it may significantly reduce quality-adjusted life years. Additionally, obesity is strongly associated with an increased risk of mortality from any cause, including cancer and cardiovascular disease. Losing weight can greatly lessen the risk for the majority of these

comorbid conditions, despite the devastating implications of obesity. To raise awareness of potential negative effects, it is necessary to identify those comorbidities that are most directly related to obesity (Stavridou et al., 2021). The literature that was published between 1995 and 2008 and that included data from prospective longitudinal studies of obesity and concomitant medical disorders was found using a systematic search technique (Di Cesare et al., 2019). To give information helpful for the most effective patient care, this article analysed the evidence regarding conclusive relationships between obesity and comorbidities. People are increasingly leading unhealthy lifestyles, indulging in excessive junk food consumption, sleeping late, and spending a lot of time sitting down (Preston and Stokes, 2014). Adolescents in particular are impacted by their unconsciously held attitudes. It is a highly complex sickness that is known as a medical issue. It encourages the spread of complicated disorders such as liver cancer, heart disease, and stroke. Obesity leads to different types of minor and major diseases, as presented in the image below. It might also lead to minor diseases and then those minor complications can result in chronic disorders.

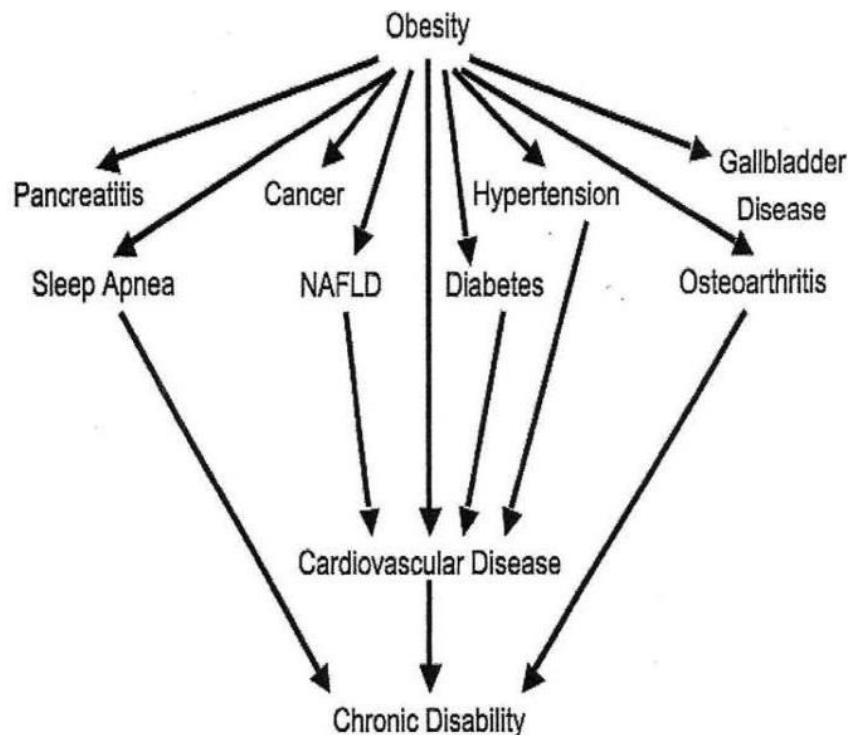


Figure 1. Major health conditions resulting from obesity

Body weight is an aspect affects almost every other component of human health. A healthy weight creates the conditions for the brain, muscles, heart, and other organs for functioning properly and effectively for a long time (Lavie et al., 2014). Obesity, in particular, has a detrimental effect on almost every aspect of health, including reproductive, pulmonary, and cognitive function. Diabetes, heart disease, and several cancers are among the deadly and incapacitating disorders that obesity increases the chance of developing. It does so by using a variety of mechanisms, some as straightforward as the mechanical strain of supporting extra weight and others requiring subtle changes to hormones and metabolism (O'Brien and Dixon, 2002).

Obesity is a contributing risk factor for cardiovascular disease (CVD) in the association to modify the effect, as the relationship between obesity and CVD and blood pressure has been debated (Gutierrez, Alloubani, Mari, and Alzaatreh, 2018). Because hypertension causes more markedly high blood pressure, an obese individual has a marginally increased chance of developing cardiovascular disease (CVD). Several researchers have recommended it (Liang et al., 2019). Becoming obese reduces life expectancy and quality of life while also raising personal, societal, and global healthcare expenses. The good news is that losing weight can reduce several hazards associated with obesity. Even if an obese person never reaches their "ideal" weight and even if they do not start losing weight until later in life, even a loss of 5 to 10 percent of body weight can have a significant positive impact on their health (Avolio et al., 2020). Numerous large, long-term epidemiological studies have demonstrated a clear correlation between obesity and a higher risk of death from all causes, cardiovascular disease, and cancer. Recently, researchers performed statistical analysis, or meta-analysis, of the data after conducting a systematic review of 89 studies on weight-related disorders. Diabetes was the disease with the highest risk among the 18 weight-related conditions they investigated (Guh et al., 2009). Numerous studies have shown a connection between coronary artery disease and being overweight. In comparison to diabetes and cardiovascular disease, the link between fat and cancer is not quite as obvious. This is partly because cancer is a group of distinct diseases rather than a single illness (Food, 2007).

2.4 Use of AI in analysing the statistical data in health monitoring and threat detection

Healthcare is also applying modern business and everyday life technologies with artificial intelligence (AI) to provide healthcare assistance to potential use artificial intelligence the

overcome faster challenges to improve the helping of patient care and administrative process. The healthcare field has a strong application, and also the healthcare technologies such as diagnosing disease, here the range of medical tasks which replaces human healthcare before artificial intelligence (AI). Technologies in the health sector that are driven by AI have the potential to address system-related issues, for instance, lowering the burden of health professionals while enhancing clinical decisions using AI-driven interventions (Guo, and Li, 2018). With the use of AI's novel discovery and conventional methods, the epidemic illness has been recognised (Lake,2019). The importance of these challenges has been acknowledged by more donor organisations, including many large global health communities. An overview of possible uses of artificial intelligence to control eating and nutritional habits, exercise routines, and weight reduction was the aim of the research. According to the study, weight loss with AI is still in its infancy. The researchers developed a framework for the application of AI in weight loss based on the findings of the current study, but we emphasised that it would depend on user participation and contextualization (Chew et al., 2021). utilising a proposed framework that integrates micro-interventions, real-time analytics, and machine perception, the study clarified the possible application of AI to enhance weight reduction. This, however, depends on additional contextual, environmental, and emotional elements that must be taken into consideration in the AI systems. To evaluate the resource efficiency of AI-assisted therapies, future research might compare the efficacy of existing behaviour modification programmes with AI-assisted self-regulation weight reduction programmes.

2.5 The efficiency of machine learning, deep learning and image processing algorithms in health monitoring and threat detection

It has been stated that machine learning had become a general-purpose technology in the sense that it is inevitable that it can be developed after a while and can eventually generate mutual development (Eckmanns et al. 2019). Making such advances will often lead to widespread financial disruption, with sectors corresponding and eliminated. Aiello et al. (2020) stated that the mechanised replacement of worker labour argues that computerization has a destructive effect when workers' labour is uprooted by machines in areas where machines are enjoyed by others separately. However, the power balance that increases interest in the work compensates for this difficult effect: efficiency, as tasks become more efficient and less expensive. This thus allows a

reserve to be accumulated for existing careless races and for the production of new non-automobile missions, some of which include computer enhancements.

Obesity has elevated to a serious global concern in the modern era. In a publication, the goal was to develop a machine-learning-based method for estimating the risk of obesity. In comparison to the other classifiers, the Logistic Regression Algorithm obtains the best accuracy of 97.09%. The gradient boosting approach also produced the lowest metric values and the worst accuracy, 64.08% (Ferdowsy et al., 2021). Another study presented the use of wearable sensors that have the potential to improve data collection's completeness and accuracy by reducing respondents' self-reporting burden, which has been linked to underreporting in up to 30% and 50% of individuals with normal and overweight body weights, respectively. This is frequently accomplished by the automated collecting of objective behavioural data, which removes typical obstacles to adherence such as a lack of desire, a lack of free time, and bad moods. The majority of the research did not examine the accuracy of food energy calculations, and none of the studies on machine perception investigated its impact on weight reduction or behaviour modification (Burgess et al., 2017).

2.6 Factors affecting health monitoring and threat detection through digital processing

Various factors affect health monitoring and threat detection through digital processing. These factors include:

2.6.1 Privacy

Majumder and Deen (2019) stated that one of the key factors influencing health monitoring and threat detection through digital processing, which is the use of machine learning, deep learning, and other algorithms, is privacy. In the modern world, everything is connected to the Internet. Devices used for health monitoring systems must be protected against attacks. A simple cyber-attack or breach can lead to a leak of personal data.

2.6.2 Smart devices

Smart watches and mobile phones are other devices that use medical internet of things technology, a breach of the application or the server where all the algorithms are working, and the

data is stored can cause serious harm. Catastrophic situation because the data in the database covers everyone using this technology. There is no proper government policy to ensure the safe use of these apps, which also affects health monitoring and prediction of future illnesses caused by previous apps. Through the use of digital treatments. Punj and Kumar (2019) stated that the contextual health monitoring and threat detection and life support systems for subjects such as elderly people monitor and assess the subject's health status and their ability to perform daily activities of the day.

2.6.3 Physical and mental health

Changes in perceptual ability, physical skills, and memory are other significant aspects that can affect human behaviour in addition to physical and mental health. Therefore, selecting proper approaches and strategies to effectively understand complex and shifting human behaviours is one of the key issues faced by smart health systems. Research has been done in this field to monitor and assess the participants' functional capacity as well as their psychophysiological and behavioural capabilities. This investigation looks at health monitoring systems' learning strategies and algorithms. Classification is advised as a first step in detecting activity in health monitoring and threat detection, two types of activities that are of major importance in health care and life support systems. The dynamic is determined by the sensors used and the processing methods selected to establish the topic context in light of the issue.

2.7 Theoretical framework

Weight loss improves patient outcomes, which has been seen in several epidemiological research. Numerous extensive, long-term research has looked at the impact of obesity on the risk and progression of CVD. Obesity is a contributor to and a risk factor for increased mortality and morbidity, but also from cancer, and other chronic and acute diseases such as kidney and liver disease, osteoarthritis, depression, and sleep apnea, according to large, high-quality longitudinal or prospective studies. The purpose of this study was to monitor health statistics and predict possible threats of diseases caused by obesity on a large scale in a particular region to aware people and the government to improve facilities. Reiblin et al. (2019) stated that the Systems Theory Paradigm marks a significant theoretical departure from approaches to understanding communication based on empirical laws and human rules. George Hegel (Kaufmann) introduced

systems thinking to the social and physical sciences in the 19th century, and Ludwig von Bertalanffy, a biologist, expanded on it in the 20th (Rakhmonov et al., 2020). Eckmanns et al. (2019) stated that applied systems theory helps to overcome problems caused by less technology being used in health monitoring and threat detection. For this document, the essential components required for health monitoring and threat detection have been described in detail. Aiello et al. (2020) stated that this theory would present a composite picture of the essential functions provided by an intelligent health system to monitor and detect the subject's behaviour, including concepts, approaches and treatment.

2.8 Literature Gap

This chapter has reviewed various articles and studies related to health monitoring and threat detection through digital processing. The focus was on the diseases led by obesity. Many studies have been conducted up till now on the diseases caused by obesity and threat detection through digital processing. Majumder and Deen (2019) stated that with the advent of computationally efficient smartphones, low-cost, high-resolution cameras, and robot sensors, a new era of intelligent next-generation surveillance systems for healthcare infrastructure has arrived. Liang et al., 2019 stated that the vibration-based state assessment has emerged as a great way to assess the state of large-scale infrastructure. The need to advance and develop alternatives to efficient sensor systems will lead to next-generation measurement techniques for structural health monitoring. Fieschi (2018) stated that by using an inexpensive camera, capture images and videos used to understand structural behaviour using advanced signal processing techniques. This paper provides a comprehensive overview of many next-generation smart sensing technologies developed in the context of structural conditions in healthcare monitoring in recent years. Future studies must predict possible threats of diseases caused by obesity on a large scale in various regions to aware people and the government to improve facilities.

2.9 Chapter Summary

A variety of approaches to monitoring and evaluating topics have been explored in this chapter. Health monitoring systems have emerged as a promising solution to the problem of ageing populations, providing contextual online health services. It has been claimed that rather than relying on medical knowledge, specialist knowledge, or gerontology, as well as a thorough

understanding of the subject's context, the strengths of health surveillance and threat detection concepts had been identified for older adults based on detection availability. In this research, the theoretical framework covers health monitoring and threat detection through digital processing. The characteristics of smart health systems and their impact on consumer interest has been discussed. Relationships between different topics according to research objectives have been found, and theoretical frameworks have been presented in this chapter.

CHAPTER THREE: METHODOLOGY

3.1 Introduction

Scientific research is a rigorous procedure to decipher the truth regarding the phenomenon. To attain procedural accuracy, a research methodology is established, which guides the researcher in navigating through various steps of the investigation. Research methodology is a means of enabling researchers to understand the procedures required to collect analytical data. Research technique is a systematic approach that aids in addressing research issues and providing answers to research inquiries. It contains all the pertinent details on the research's data, including where it was gathered from and how it would be analysed.

3.2 Research design

The research design mostly pertains to how the researcher carries out the investigation. The two main categories of research designs are quantitative and qualitative. While quantitative research design is focused on data that is done in numerical format, qualitative research is based on theoretical and conceptual facts (Bryman, 2017). This study is built on a quantitative design that is data-driven. This is because the study was conducted to learn more about obesity and other leading health indicators and examine how they could rise over the next few years. The goal of the research is to identify the disorders that are likely to develop due to obesity. The domain of health monitoring requires that the study be based on an investigation of recent real-time data obtained from credible sources about major health issues of the sample population (Akhtar and Khatak, 2016). A data-intensive research design in this current research facilitated the researcher in obtaining the maximum volume and depth of numeric data from the health institutions. Since the present research primarily relied on data collected and generated from credible healthcare institutions, this quantitative research design was also suitable to incorporate the findings from that data with the wider literature studied on this subject.

Choosing a suitable and workable research design is a crucial component of research to eliminate the considerable breadth of data and guarantee that the data information leads to a relevant conclusion (Ferdowsy, 2021). Unambiguity is removed and the core of the study subject is identified with the aid of an effective research design. Its primary goals include defining the genuine research topic, analysing published literature, highlighting hypotheses, emphasising the

data needed to reach meaningful conclusions, outlining the data collection approach, and selecting data analysis methods. From this perspective, the two main kinds of the study design are exploratory and conclusive. Exploratory research designs rely on freely available data sets and steer towards subject inquiry in an incrementally focusing manner. The data collecting and analysis is often qualitative, and the results provide new research opportunities. Whereas, a conclusive research design, in contrast, is used as a formal design that focuses on reaching a set of findings. As per the analysis of Vogt (2012), this method's strict and quantitative data collecting methodology results from its emphasis on conclusions.

The current study is centred on the extrapolation and interpretation of data from secondary sources; as a result, a conclusive research design is used throughout the investigation. The layout offers a comprehensive platform for using external data and analytical tools. Given the size of the sample and the number of sources, a conclusive design aided in focusing the research and producing trustworthy results.

3.3. Research approach

Research approach refers to the overall stance adopted by the researcher at the start regarding the kind of data that will be collected, and approach for its analysis. There is a myriad of research approaches available, including the major ones of inductive, deductive, and abductive approach (Osman et al., 2018). To get the required findings, a precise research strategy must be chosen. The acquired data cannot be correctly evaluated in the absence of an accurate and robust research strategy, which might hurt the goal of the entire research project. The deductive research method is appropriate when a scientific topic has to be investigated. It makes it easier to read previous ideas and locate ones that are relevant to the current study as well (Pandey, 2019). The research methodology chosen for this study is deductive approach. This is because the research aimed to examine the ongoing rise in health conditions among the populace, and devise a data-driven strategy to monitor the health outcomes. By learning from previous theories, the deductive method aids in evaluating assumptions and hypotheses (Jebb, Parrigon and Woo 2017). Therefore, this research strategy was very helpful in achieving the goals and objectives of this study. The goal of this study is to identify any diseases that may develop in the future as a result of rising obesity rates.

In investigations using the current deductive methodology, the scientist develops a set of hypotheses prior to the start of the investigation. Then, to determine if the hypotheses are correct or incorrect, significant exploratory approaches are chosen and used. When using a deductive approach, scientists start with a strong social hypothesis and then verify its predictions using data. Overall, the researcher follows the same steps as inductive analysis, but they will reverse the direction of the inquiry, going from broad to specific levels (Johnston, 2014). In the current case scenario, by using the deductive reasoning, the researcher started from the initial conception that patterns of health deterioration could be diagnosed and predicted through digital processing techniques. From this point, the searcher started gathering credible data about obesity related ailments faced by the target population. This data was used to guide the generation of digital processing method which could help the healthcare professionals in monitoring the health of patients, and larger population in general. Logical analysis is typically associated with insightful research methods (Woicshyn et al., 2018). The analyst focuses on what other people have done, reads current hypotheses about whatever peculiarity they are considering, and then evaluates ideas that emerge from those assumptions.

3.4 Data collection

The whole project was centred on addressing the fundamental issue of recognising possible health risks so that precautions may be taken. Since the researcher established the region of the United Kingdom for study contextualisation, hence, the two reliable sources of health-related data are Kaggle and National Health Service (NHS) During the study, the researcher examined crucial variables of health indication including blood pressure, body mass index (BMI), obesity, exercise routine, vitamin consumption, etc., and forecast probable risk diseases that the majority of individuals may experience. Information on illnesses caused by these variables was included in the project. Among a number of machine learning algorithms and other coding options available for digital processing of health indicators, the researcher selected the neural net, to forecast which causes are getting worse over time and which may eventually result in the diseases stated. Once the data had been collected and identified, it had to be checked for errors and subjected to exploratory data analysis. After fully comprehending the format and structure of the data, the researcher proceeded with the choose the machine learning method that worked best when applied to it. The project's final product is anticipated to assist improve the healthcare system for people.

The current healthcare system places more of an emphasis on disease treatment than disease prevention.

3.5 Data analysis

Since the current research focused on developing a predictive model through digital processing of health indicators, hence, data was analysed through R studio. This made it easier to determine how many patients have obesity. It also provides to be useful in assessing whether the number of patients with obesity rose or fell during the epidemic. Finding the effects on the health sector in the future was also much easier with the aid of this analysis technique because it automated processing of a huge volume of data collected from NHS and Kaggle. One of the leading benefits of using RStudio for data analysis is its inherent error-correcting functions, which made it easier for the researcher to omit repeated or contradictory data (Cirillo, 2016). Because of its versatility and data visualisation features, many professionals and relative beginners alike continue to transition from traditional statistical software like SPSS, SAS, and Stata to R.

3.6 Ethical considerations

While conducting research, ethical issues are crucial. The British Psychological Society's guidelines and the Helsinki Declaration's regulations are often followed by researchers (BPS) during medical science-oriented studies. It should be highlighted that even so, when responders are engaged, ethical issues are prioritised. However, the current research was completely data oriented in which no live participants were made part of the study at any stage. Hence, there was no need for a participant consent. Nonetheless, the researcher honoured the research ethics of information usage, and ensured proper referencing of all the data collected from secondary sources for data analysis.

3.7 Chapter summary

The current chapter elaborated the specificities of methods adopted to conduct this study. In order to maximise the generalisability of findings, the researcher adopted a secondary quantitative methodology. Meanwhile, data was collected regarding major health indicators of obesity, blood pressure, and vitamin consumption of the UK populace from credible sources of NHS and Kaggle. For data analysis, RStudio was used to arrive at statistically sound and meaningful results which guides the formulation of study findings.

CHAPTER 4: FINDINGS AND DISCUSSION

4.1 Introduction

This chapter elaborated on the results and thorough discussion on monitoring the health statistics and predicting possible threats of diseases caused by obesity on a large scale in a particular region to aware people and the government improve facilities. Notably, the research has utilised RStudio Software to compute statistics and graphics. This software has assisted in cleaning, evaluating, and graphing the data of significant variables of health indicators, including blood pressure, creatinine phosphokinase, diabetes, and other related variables, extracted from the national data of NHS and Kaggle. Besides, it forecasts probable risks of diseases that most individuals have experienced may experience. Additionally, an exhaustive discussion on the inclusive results of the research has been conducted. Finally, a concise summary of the chapter was presented.

4.2 Findings

4.2.1 Descriptive Statistics

Descriptive statistics have been employed to examine and perceive the obscured characteristics of the data of health indication variables such as blood pressure, body mass index (BMI), obesity, exercise routine, and vitamin consumption. In particular, the statistics assisted in elucidating the descriptions of health variables in an acceptable, measurable, comprehensible way (George and Mallery, 2018). Figures 1 and 2, presented below, provide the descriptive Statistics on the Dataset.

Health Monitoring and Threat Detection through Digital Processing

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time
1	75.000	0	582	0	20	1	265000	1.90	130	1	0	4
2	55.000	0	7861	0	38	0	263358	1.10	136	1	0	6
3	65.000	0	146	0	20	0	162000	1.30	129	1	1	7
4	50.000	1	111	0	20	0	210000	1.90	137	1	0	7
5	65.000	1	160	1	20	0	327000	2.70	116	0	0	8
6	90.000	1	47	0	40	1	204000	2.10	132	1	1	8
7	75.000	1	246	0	15	0	127000	1.20	137	1	0	10
8	60.000	1	315	1	60	0	454000	1.10	131	1	1	10
9	65.000	0	157	0	65	0	263358	1.50	138	0	0	10
10	80.000	1	123	0	35	1	388000	9.40	133	1	1	10
11	75.000	1	81	0	38	1	368000	4.00	131	1	1	10
12	62.000	0	231	0	25	1	253000	0.90	140	1	1	10
13	45.000	1	981	0	30	0	136000	1.10	137	1	0	11
14	50.000	1	168	0	38	1	276000	1.10	137	1	0	11
15	49.000	1	80	0	30	1	427000	1.00	138	0	0	12
16	82.000	1	379	0	50	0	47000	1.30	136	1	0	13
17	87.000	1	149	0	38	0	262000	0.90	140	1	0	14
18	45.000	0	582	0	14	0	166000	0.80	127	1	0	14
19	70.000	1	125	0	25	1	237000	1.00	140	0	0	15
20	48.000	1	582	1	55	0	87000	1.90	121	0	0	15
21	65.000	1	52	0	25	1	276000	1.30	137	0	0	16
22	65.000	1	128	1	30	1	297000	1.60	136	0	0	20

Figure 2: Dataset import in R Software

The figure displayed the illustration of the Dataset of health indication variables such as blood pressure, body mass index (BMI), obesity, exercise routine, and vitamin consumption imported from an excel sheet in the R Software. It could be witnessed that dummy variables have been created that took 0 or 1 value only to specify the existence or absence of a few categorical effects (Borenstein, 2022). Expressly, for anaemia, 0 signifies that the patient is anaemic, and 1 indicates that the patient is not anaemic. Moreover, for diabetes, 0 signifies that the patient has had diabetes, and 1 indicates that the patient did not have diabetes. Similarly, for high blood pressure, 0 signifies that the patient has high blood pressure, and 1 indicates that the patient does not have high blood pressure. Whereas for gender, 0 indicates that the patient was male, and 1 specified that the patient was female. For smoking, 0 indicates a smoker patient and 1 represented non-smoker patients.

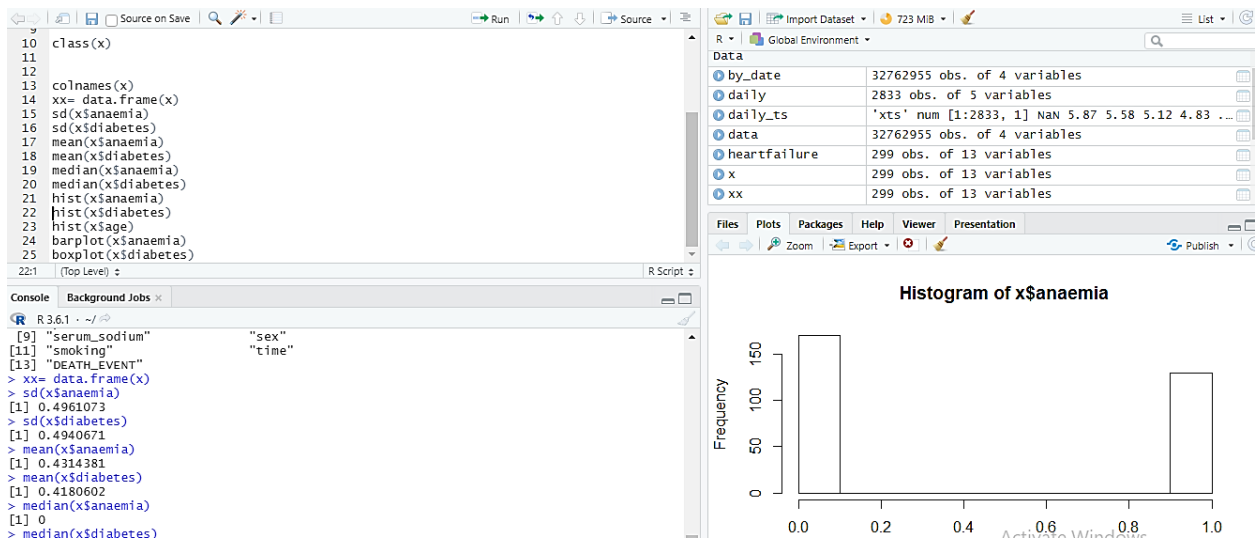


Figure 3: Descriptive Statistics on the Dataset

Figure 2 illustrates the descriptive statistics of the Dataset. It could be witnessed that the data of 299 patients for 13 variables are age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time, and death event have been presented. Furthermore, histograms have been constructed with the help of vector values of the variables, anaemia and age. Predominantly, the histogram characterised the frequencies of anaemia and age grouped into ranges. All the bars in the histogram signify the height of the frequency of variable values existing in that range (Ding and Xing, 2020).

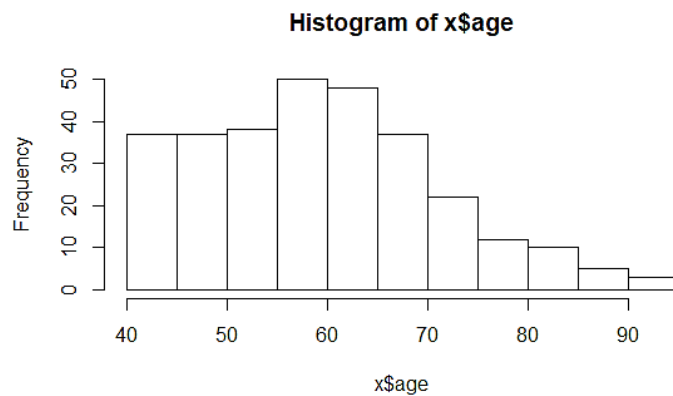


Figure 4: Histogram of Age variable

Figure 3 displayed the histogram for the variable "age", whereas the age group were represented on the x-axis and the frequency of occurrence on the y-axis. It could be seen that most of the patients were in the age group 40 to 55, and only a few patients had ages above 70.

4.2.2 Regression Analysis

Regression analysis is an analysis technique in statistics to investigate the extent to which the variables exert an impact on one another (Guryanova et al., 2020). This analysis assists in ascertaining those variables that induce a meaningful impact on the explanatory variable and the direction of the impact (Maulud and Abdulazeez, 2020). Figure 4, presented below, illustrates the regression between the variables, age and anaemia.

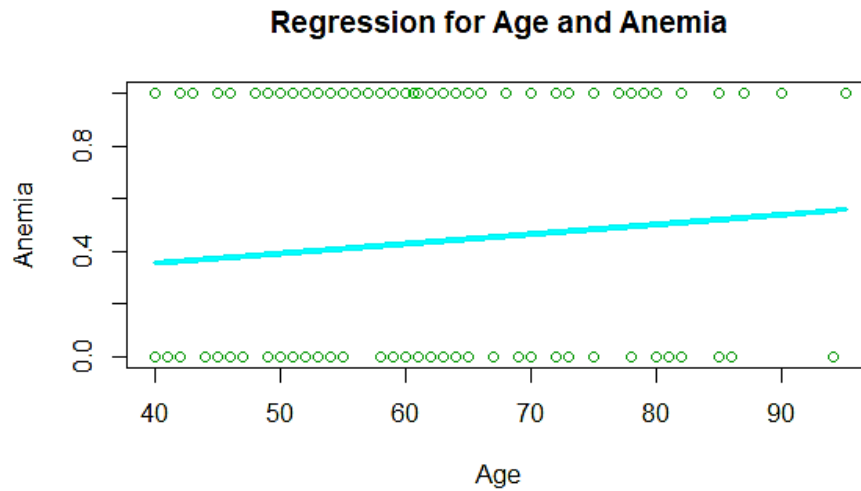


Figure 5: Regression Plot between Age and Anemia variables

Figure 4 displayed the regression plot between the variables, age and anaemia. From the figure, it is evident that there is a positive relationship between anaemia and age. Thus, with an increase in the patient's age, the patient would have more chance of getting anaemic. Moreover, the figure displayed that the chances of being anaemic have been lower at age 40 as the value is below 0.4. However, the chances gradually increase with an increase in age. According to Stauder et al. (2018), anaemia has been found to be most frequent in the elderly, who have age 65 years and above. Further, the regression plot delivers a seamlessly good explanation for the variables with higher variance. Khan et al. (2019) used a machine learning model to predict anaemia patient datasets and found that the predictive models delivered the most accurate predictions.

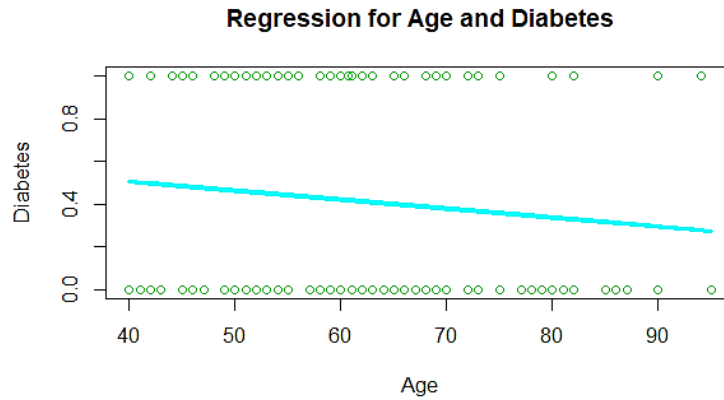


Figure 6: Regression between Age and Diabetes

Figure 5 displayed the regression plot between the variables, age and diabetes. From the figure, it has been observed that there exists a negative relationship between age and diabetes. Further, from the regression plot, it could be inferred that the model put forward an efficient explanation for the variables with higher variance. In particular, the model displayed the chances of having diabetes. The model is a good predictor of estimating diabetes from the age group of 40 as the value is above 0.4. However, the chances progressively decrease with an increase in age. Thus, the model is a good predictor for the early age patient. Although, with an increase in the patient's age, the explanatory power of the model has decreased because of the reduced regression coefficient. Sarwar et al. (2018) have found that the machine learning model tends to deliver higher and a more accurate predictions of diabetes.

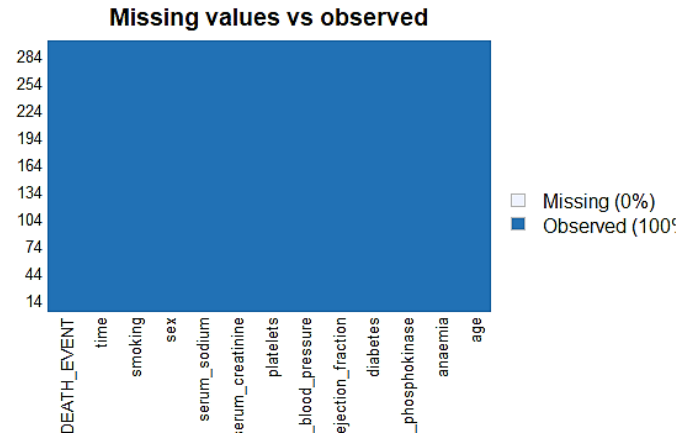


Figure 7: Missing Values Finding

Figure 6 illustrates the findings of missing values of age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time, and death event. Arumugam and Saranya (2018) have put forward that dealing with *missing values* has been substantially significant for conducting data analysis. In the R studio software, the missing values have been frequently characterised by “NA”. From the above figure, it could be seen that the model 100 percent predicts all the values, and none of the values of any variables has been missed.

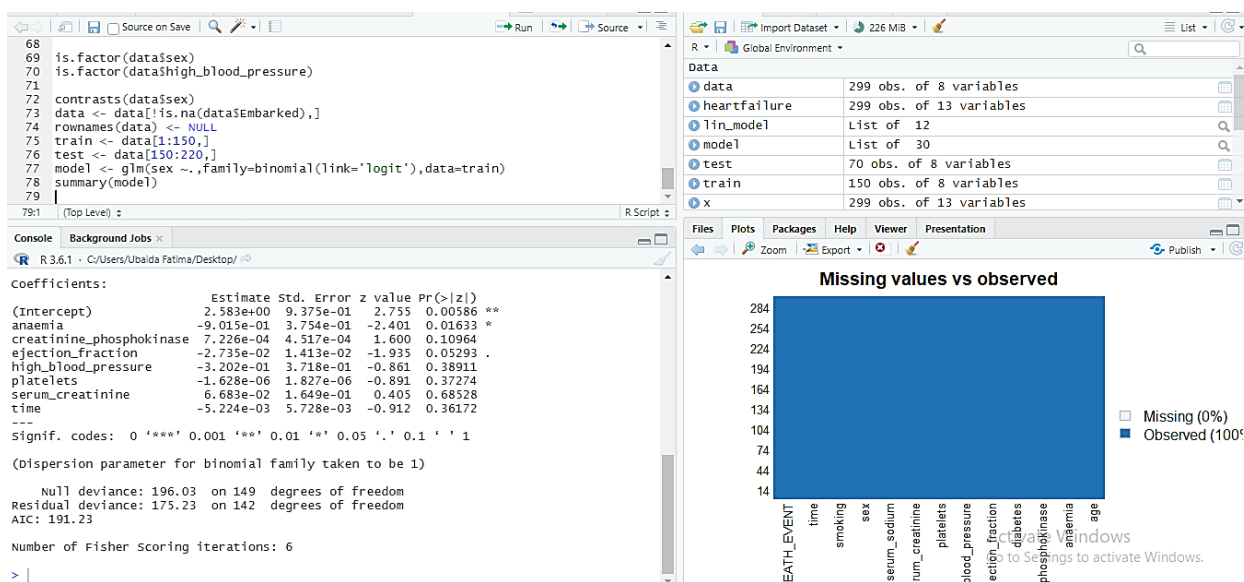


Figure 8: Training Dataset in R Software

Figure 9 displays the training dataset that has been included in the R Software. The major aim of the training dataset has likely to be the preliminary data employed to train the models of machine learning. In particular, such datasets have been nourished with machine learning algorithms with a view of making or performing a prediction of a targeted task (Kapur et al., 2021). Specifically, the data set included 150 observations of 8 variables, i.e., anaemia, creatinine phosphokinase, ejection fraction, high blood pressure, platelets, serum creatinine, time, and death event. It has been observed that none of the values of any variables was missed in the training data.

4.2.3 Logistic Regression Model

Logistic regression models have been employed in order to predict the category or class of entities founded upon one or more than one predictor variables. It is a model that includes the variables with binary outcomes, such as those variables that have a value of yes or no, 0 or 1, or diseased or non-diseased (Ray, 2019). In this study, seven variables tend to have a binary outcome that are anaemia, creatinine phosphokinase, ejection fraction, high blood pressure, platelets, serum creatinine, and time. Therefore, the logistic regression has been run on seven variables with binary outcomes to examine that either the model is efficient in making a good prediction of heart diseases or not. According to Nusinovici et al. (2020), logistic regression has been employed to precisely estimate the mean of the data; thus, the data variance is then estimated from the mean.

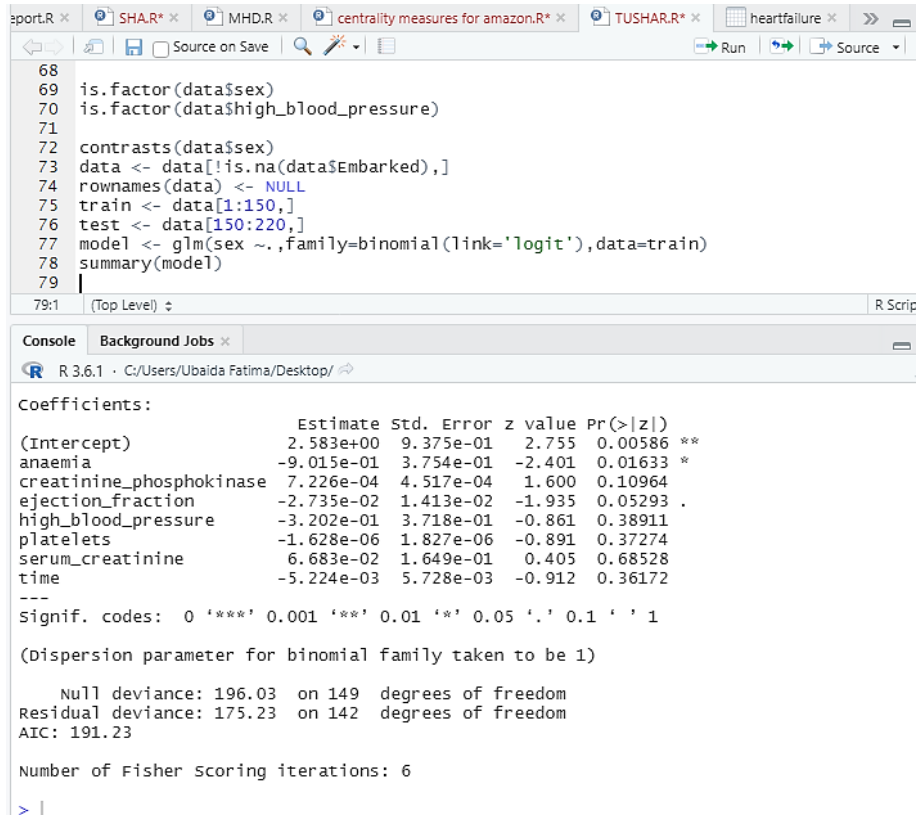
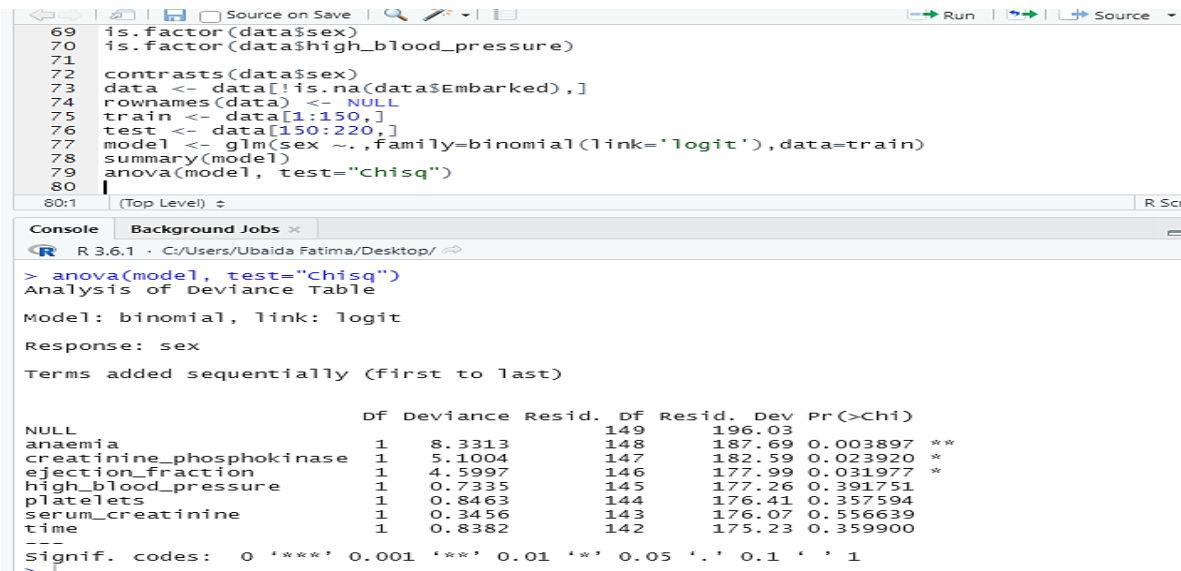


Figure 10: Logistic Regression Model Fitting on Dataset

The figure above displays the results of the logistic Regression Model Fitting on the Dataset. Particularly, this regression entailed the formation of null and residual deviance. However, the deviances have been employed for the computation of overall goodness of fit, i.e., R-square, as well as an inclusive p-value. Moreover, the Akaike information criterion (AIC) exhibited the adjustment of residual deviance from the existing number of parameters within the research model (Cavanaugh and Neath, 2019). The value of AIC displays that the model is a good fit for predicting the variables. Finally, the number of Fisher's scoring iterations demonstrated that the Generalised Linear Models (GLM) function quickly congregated with the estimates of the maximum likelihood of the coefficients (Dennis et al., 2019). The variable anaemic has been observed to be significantly predicted by the model as the p-value is $0.01633 < 0.05$. However, it could be seen that anaemia exerts a negative impact on heart diseases. The variable creatinine phosphokinase has been observed to be insignificantly predicted by the model as the p-value is $0.10964 > 0.05$. Besides, the variable ejection fraction has been perceived to be a significant predictor of heart diseases as the p-value is $0.05 = 0.05$ (level of significance). Nevertheless, the

variable ejection fraction has exerted a negative effect on heart diseases. This is a statistically significant finding as ejection fraction tends to demonstrate the strength of the heart (Chicco and Jurman, 2020). Particularly, it estimates the total oxygen-rich blood that is likely to be pumped out with each heartbeat. Thus, when ejection fraction is low, then it characteristically indicates failure of the heart.

4.4.4 Interpreting the outcomes of a dataset through a logistic regression model



```

69 is.factor(data$sex)
70 is.factor(data$high_blood_pressure)
71
72 contrasts(data$sex)
73 data <- data[!is.na(data$Embarked),]
74 rownames(data) <- NULL
75 train <- data[1:150,]
76 test <- data[150:220,]
77 model <- glm(sex ~.,family=binomial(link='logit'),data=train)
78 summary(model)
79 anova(model, test="Chisq")
80

```

Console

```

> anova(model, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit
Response: sex
Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                                149      196.03
anaemia                1      8.3313      148      187.69 0.003897 **
creatinine_phosphokinase 1      5.1004      147      182.59 0.023920 *
ejection_fraction      1      4.5997      146      177.99 0.031977 *
high_blood_pressure     1      0.7335      145      177.26 0.391751
platelets               1      0.8463      144      176.41 0.357594
serum_creatinine        1      0.3456      143      176.07 0.556639
time                   1      0.8382      142      175.23 0.359900
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

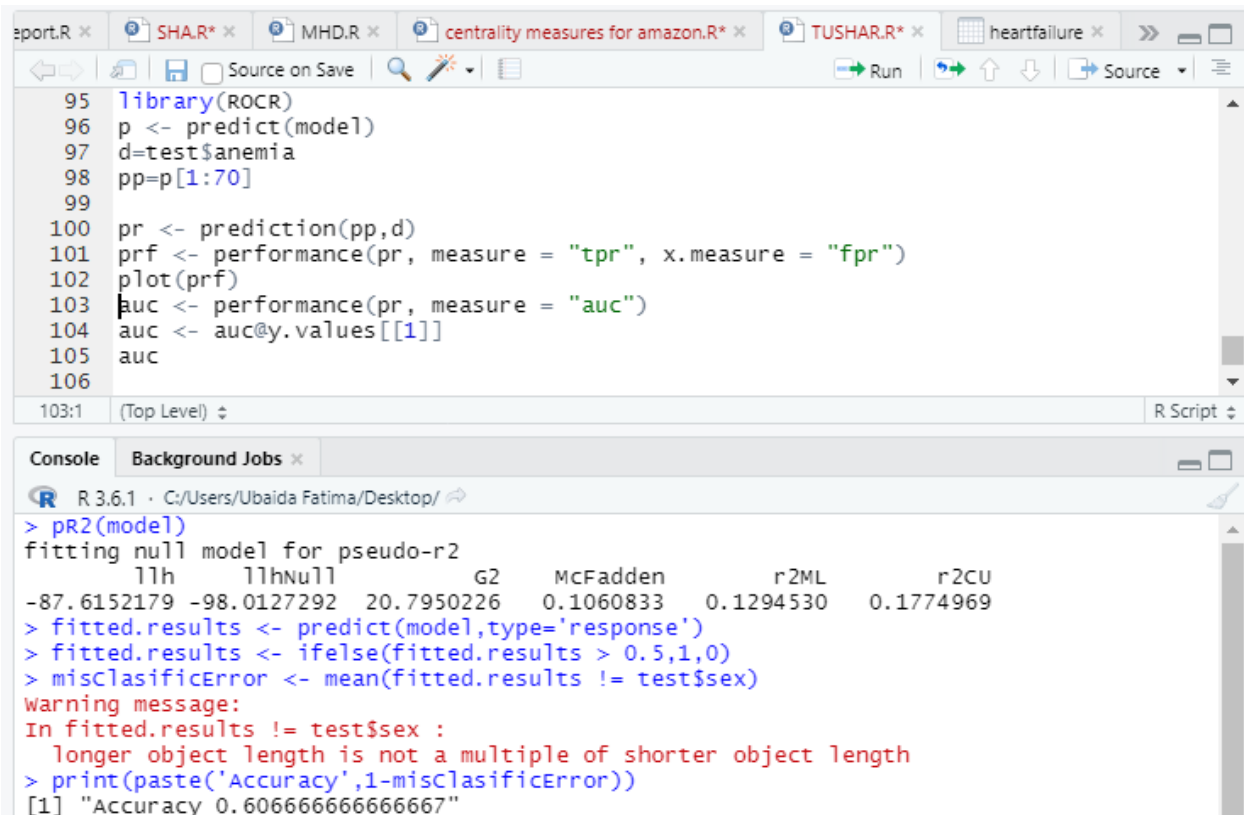
```

Figure 11: ANOVA on Dataset through R Software

Figure 9 presented the findings of the ANOVA table on the Dataset of anaemia, creatinine phosphokinase, ejection fraction, high blood pressure, platelets, serum creatinine, and time through R Software. It could be seen that the model has been able to make an accurate prediction of three variables, i.e., anaemia, creatinine phosphokinase, and ejection fraction. However, it failed to make a precise estimation of high blood pressure, platelets, serum creatinine, and time. It is because the p-values of the variables, i.e., 0.3917, 0.3575, 0.5566, and 0.35899, respectively, have been more than the significance value (0.05). Further, the table showed that the variance between the mean values of the variables is significantly high. Thus, this insinuates the presence of outliers in the model.

4.4.5 Assessing the predictive ability of the model

The models' predictive ability has been assessed by separating the composites into the training set. In other words, the main model is divided into two. The first is the set with which the researcher calculates the model, and the second is the training or the test set (Zhao et al., 2020).



```

95 library(ROCR)
96 p <- predict(model)
97 d=test$anemia
98 pp=p[1:70]
99
100 pr <- prediction(pp,d)
101 prf <- performance(pr, measure = "tpr", x.measure = "fpr")
102 plot(prf)
103 auc <- performance(pr, measure = "auc")
104 auc <- auc@y.values[[1]]
105 auc
106
103:1 (Top Level)
R Script

```

```

R 3.6.1 · C:/Users/Ubaida Fatima/Desktop/
> pr2(model)
fitting null model for pseudo-r2
      llh      llhNull      G2      McFadden      r2ML      r2CU
-87.6152179 -98.0127292 20.7950226 0.1060833 0.1294530 0.1774969
> fitted.results <- predict(model,type='response')
> fitted.results <- ifelse(fitted.results > 0.5,1,0)
> misclasificError <- mean(fitted.results != test$sex)
Warning message:
In fitted.results != test$sex :
  longer object length is not a multiple of shorter object length
> print(paste('Accuracy',1-misclasificError))
[1] "Accuracy 0.606666666666667"

```

Figure 12: Accuracy of Fitted Model

From figure 10, the accuracy of the fitted model has been found to be 60.67%. This implies that the model is a good fit for making a prediction of anaemia, creatinine phosphokinase, ejection fraction, high blood pressure, platelets, serum creatinine, and time.

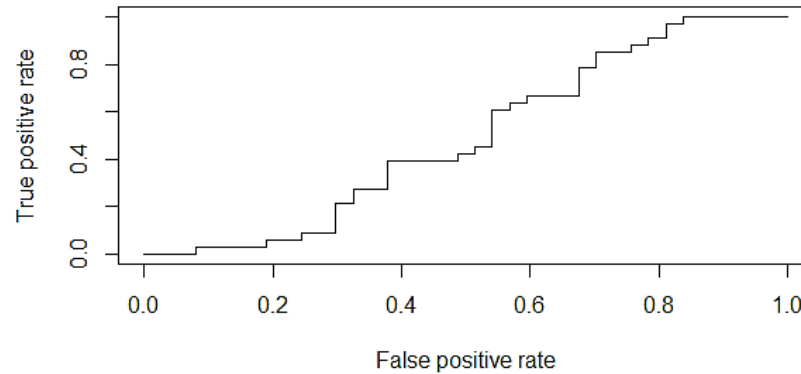
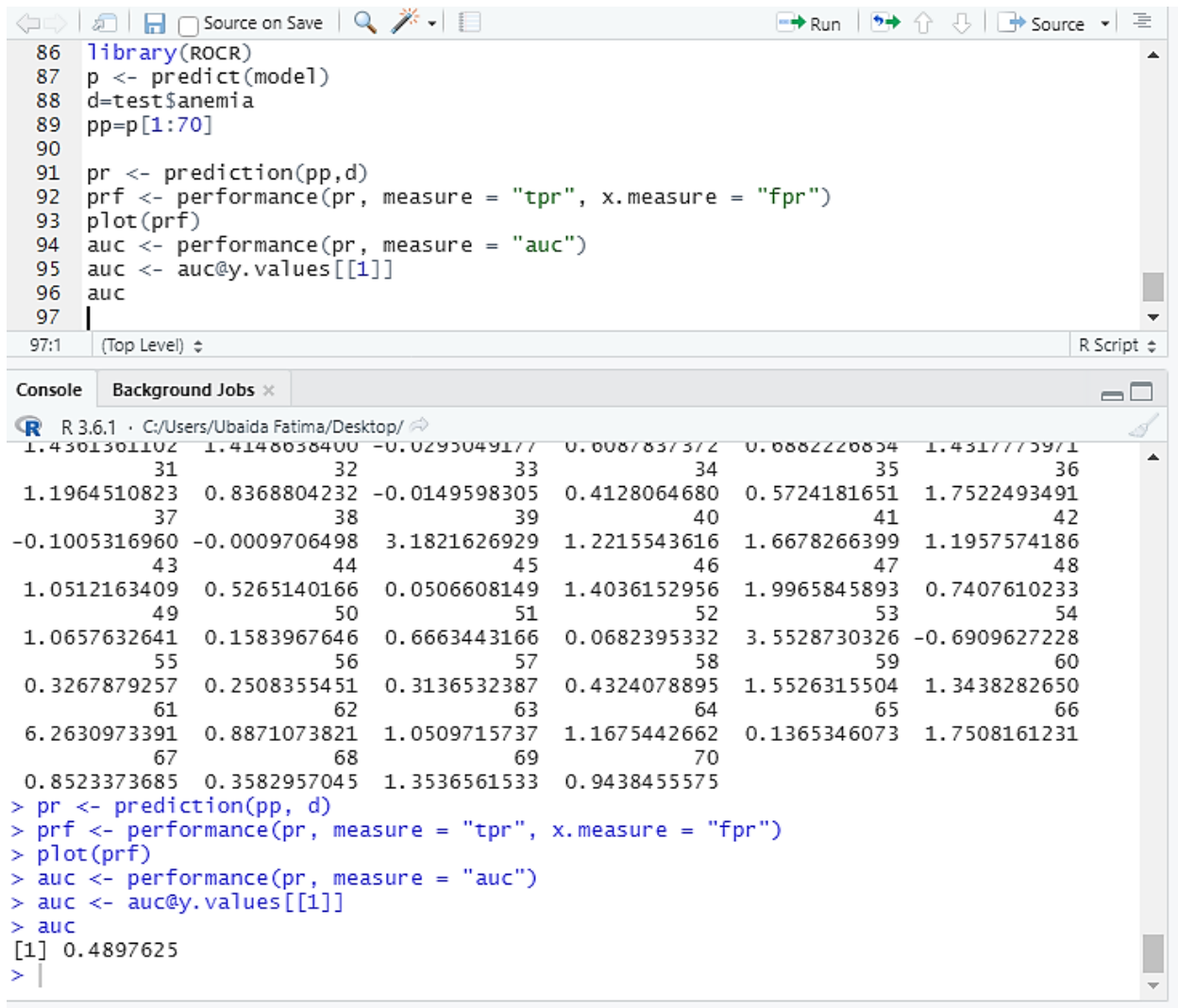


Figure 13: Prediction Plot for Anemia through Logistic Regression

Figure 11 displayed the prediction plot for anaemia exhibiting the actual values of anaemic patients from the Dataset in contrast with estimated or predicted ones that have been produced by the model. It shows two aspects of anaemia that are used in the diagnostic test that is either the test is truly positive or false positive. It could be inferred from the figure that the model failed to have the capability of differentiating the true and false positives. Thus, it means that the model contains errors and is unable to predict the true positives.



```

86 library(ROCR)
87 p <- predict(model)
88 d=test$anemia
89 pp=p[1:70]
90
91 pr <- prediction(pp,d)
92 prf <- performance(pr, measure = "tpr", x.measure = "fpr")
93 plot(prf)
94 auc <- performance(pr, measure = "auc")
95 auc <- auc@y.values[[1]]
96 auc
97
97:1 (Top Level)
R Script

```

Console Background Jobs

```

R 3.6.1 - C:/Users/Ubaida Fatima/Desktop/
1.4301301102 1.4148038400 -0.0293049177 0.6087837372 0.6882226834 1.4317773971
31 32 33 34 35 36
1.1964510823 0.8368804232 -0.0149598305 0.4128064680 0.5724181651 1.7522493491
37 38 39 40 41 42
-0.1005316960 -0.0009706498 3.1821626929 1.2215543616 1.6678266399 1.1957574186
43 44 45 46 47 48
1.0512163409 0.5265140166 0.0506608149 1.4036152956 1.9965845893 0.7407610233
49 50 51 52 53 54
1.0657632641 0.1583967646 0.6663443166 0.0682395332 3.5528730326 -0.6909627228
55 56 57 58 59 60
0.3267879257 0.2508355451 0.3136532387 0.4324078895 1.5526315504 1.3438282650
61 62 63 64 65 66
6.2630973391 0.8871073821 1.0509715737 1.1675442662 0.1365346073 1.7508161231
67 68 69 70
0.8523373685 0.3582957045 1.3536561533 0.9438455575
> pr <- prediction(pp, d)
> prf <- performance(pr, measure = "tpr", x.measure = "fpr")
> plot(prf)
> auc <- performance(pr, measure = "auc")
> auc <- auc@y.values[[1]]
> auc
[1] 0.4897625
>

```

Figure 14: Prediction Model and Accuracy of Prediction Model on Dataset through R Software

```

101 prf <- performance(pr, measure = "tpr", x.measure = "fpr")
102 plot(prf)
103 auc <- performance(pr, measure = "auc")
104 auc <- auc@y.values[[1]]
105 auc
106
107 install.packages("ISLR")
108 library(ISLR)
109 names(x)
110 summary(x)
111 head(x)
112
112:1 (Top Level)
R Script

```

```

> summary(x)
      age      anaemia  creatinine_phosphokinase  diabetes
Min.   :40.00   Min.   :0.0000   Min.    : 23.0         Min.   :0.0000
1st Qu.:51.00   1st Qu.:0.0000   1st Qu.: 116.5       1st Qu.:0.0000
Median :60.00   Median :0.0000   Median : 250.0       Median :0.0000
Mean   :60.83   Mean   :0.4314   Mean   : 581.8       Mean   :0.4181
3rd Qu.:70.00   3rd Qu.:1.0000   3rd Qu.: 582.0       3rd Qu.:1.0000
Max.   :95.00   Max.   :1.0000   Max.   :7861.0       Max.   :1.0000
ejection_fraction high_blood_pressure  platelets  serum_creatinine
Min.    :14.00   Min.    :0.0000   Min.    : 25100   Min.    :0.500
1st Qu. :30.00   1st Qu. :0.0000   1st Qu. :212500   1st Qu. :0.900
Median  :38.00   Median  :0.0000   Median  :262000   Median  :1.100
Mean    :38.08   Mean    :0.3512   Mean    :263358   Mean    :1.394
3rd Qu. :45.00   3rd Qu. :1.0000   3rd Qu. :303500   3rd Qu. :1.400
Max.    :80.00   Max.    :1.0000   Max.    :850000   Max.    :9.400
serum_sodium    sex      smoking      time  DEATH_EVENT
Min.    :113.0   Min.    :0.0000   Min.    :0.0000   Min.    : 4.0   Min.    :0.0000
1st Qu. :134.0   1st Qu. :0.0000   1st Qu. :0.0000   1st Qu. : 73.0   1st Qu. :0.0000
Median  :137.0   Median  :1.0000   Median  :0.0000   Median  :115.0   Median  :0.0000
Mean    :136.6   Mean    :0.6488   Mean    :0.3211   Mean    :130.3   Mean    :0.3211
3rd Qu. :140.0   3rd Qu. :1.0000   3rd Qu. :1.0000   3rd Qu. :203.0   3rd Qu. :1.0000
Max.    :148.0   Max.    :1.0000   Max.    :1.0000   Max.    :285.0   Max.    :1.0000
>

```

Figure 15: Summary of Heart Monitoring Dataset

In the figure, the mean value of anaemia and diabetes has been predicted to be 0.4314 and 0.4181, which implies that the predicted data is efficient and do not contain any outlier. However, the mean values of ejection fraction, high blood pressure, and serum creatinine has been more than 1, implying that the data set contains outliers.

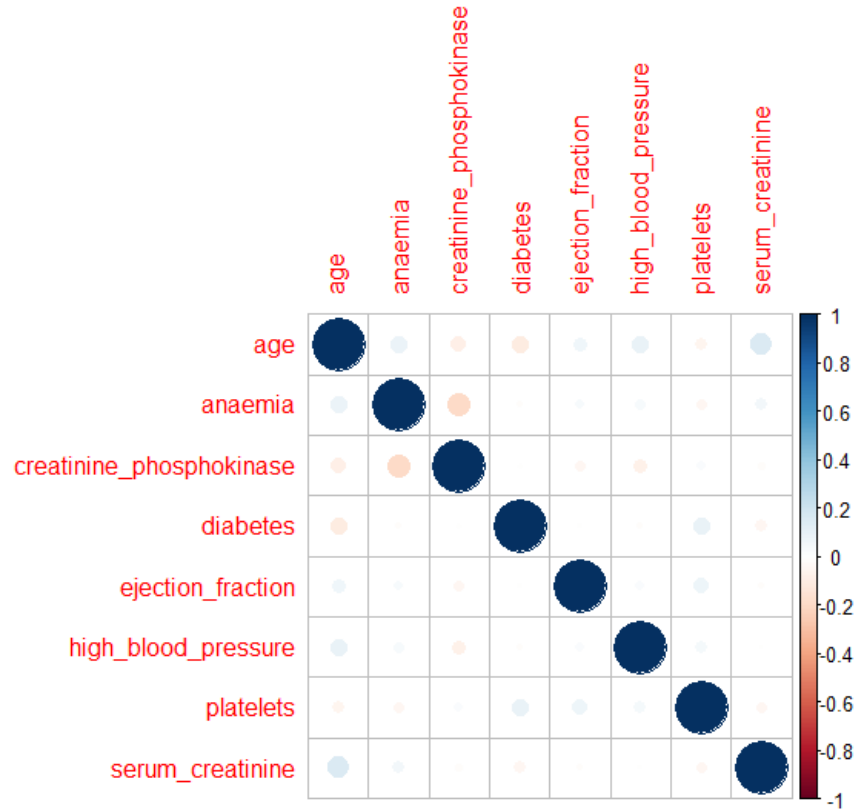


Figure 16: Correlation Plot for Dataset

In the figure above, the correlation amongst all the pairs of numeric variables has been calculated. In particular, within a correlation matrix plot, the pair-wise correlations have been plotted to deliver insights into the variables that have been changing together. The dot-representation has been employed; however, the blue dots signify the positive correlation, and the red dots implies the negative relation. Furthermore, the larger dot illustrates the correlation between variables is high. It has been evident that the matrix is symmetrical. Besides, the diagonal has been observed to be perfectly positively correlated since it demonstrates the correlation of all the variables by itself. Unfortunately, none of the variables has been found to be correlated with one another.

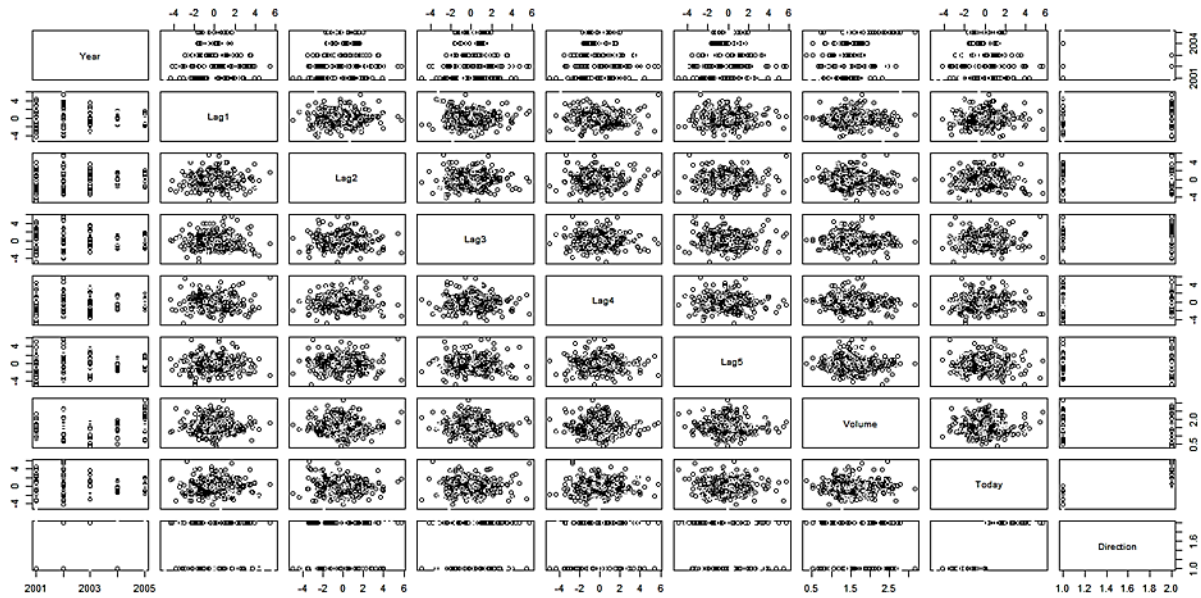


Figure 17: Density Distribution for every factor

Figure 18 exhibits the density distribution of anaemia, creatinine phosphokinase, ejection fraction, high blood pressure, platelets, serum creatinine, and time that have been broken down by the Sex factor of the data. This plot has assisted in identifying the parting of Up and Down.

4.4.5 Support Vector Machine on Heart Monitoring Dataset


```

133 anyNA(x)
134 summary(x)
135 training[["anaemia"]] = factor(training[["anaemia"]])
136 trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
137 install.packages("kernlab")
138 library(kernlab)
139 svmm <- train(anaemia ~., data = training, method = "svmLinear",
140               trControl=trctrl,
141               preProcess = c("center", "scale"),
142               tuneLength = 10)
143 svmm
144 |

```

144:1 (Top Level) ↕ R Scri

Console Background Jobs ×

R 3.6.1 · C:/Users/Ubaida Fatima/Desktop/ 

```

> svmm <- train(anaemia ~., data = training, method = "svmLinear",
+               trControl=trctrl,
+               preProcess = c("center", "scale"),
+               tuneLength = 10)
> svmm
Support Vector Machines with Linear Kernel

210 samples
 7 predictor
 2 classes: '0', '1'

Pre-processing: centered (7), scaled (7)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 189, 190, 189, 189, 188, 189, ...
Resampling results:

    Accuracy    Kappa
 0.5058081  -0.044373

Tuning parameter 'C' was held constant at a value of 1
> |

```

Figure 19: Support Vector Machine (SVM) results on R Software

One of the core purposes of the current research was to identify a machine learning mechanism which could be trained to identify potential health risks in the patients monitored. However, early in the research phase, the researcher realised that it is imperative to incorporate an iterative mechanism in the algorithm within this wider framework of the health risk identification system. This realisation came through the observation that health indicators among patients and even of a single patient vary considerably in a given period of time. This variation is a natural phenomenon originating from the biophysical microsystems of the body. Aspects such as temperature, blood pressure, heart rate, and even blood composition remain dynamic nature throughout the day. This is majorly influenced by the external environment. To overcome the issue, the researcher incorporated a learning algorithm into the program through the technique of Support-vector machines (SVMs) or Support-vector networks (SVNs). These are digitally supervised learning models which are integrally linked with the learning algorithms of R-studio.

```

133 anyNA(x)
134 summary(x)
135 training[["anaemia"]] = factor(training[["anaemia"]])
136 trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
137 install.packages("kernlab")
138 library(kernlab)
139 svmm <- train(anaemia ~., data = training, method = "svmLinear",
140              trControl=trctrl,
141              preProcess = c("center", "scale"),
142              tuneLength = 10)
143 svmm
144 |

```

144:1 (Top Level) ↕ R Scri

Console Background Jobs ×

R 3.6.1 · C:/Users/Ubaida Fatima/Desktop/ ↗

```

> svmm <- train(anaemia ~., data = training, method = "svmLinear",
+              trControl=trctrl,
+              preProcess = c("center", "scale"),
+              tuneLength = 10)
> svmm
Support Vector Machines with Linear Kernel

210 samples
 7 predictor
 2 classes: '0', '1'

Pre-processing: centered (7), scaled (7)
Resampling: Cross-validated (10 fold, repeated 3 times)
Summary of sample sizes: 189, 190, 189, 189, 188, 189, ...
Resampling results:

    Accuracy    Kappa
0.5058081 -0.044373

Tuning parameter 'C' was held constant at a value of 1
> |

```

Figure 20. Support Vector Machine (SVM) results on R Software

Proceeding further with the predictive value analysis, the researcher obtained information about the test cases of health predictors. This includes the evaluation of all the false positives, false negatives, true positives, and true negatives of a specific predictor variable. The current study analysed the predictors of anaemia, creatinine phosphokinase, diabetes, blood pressure, platelets, and serum creatinine. A confused matrix analysis technique was used to categorise these aspects of the developed machine learning algorithm. The tests conducted to check the significance of the predictive capacity is shown in the following figure.

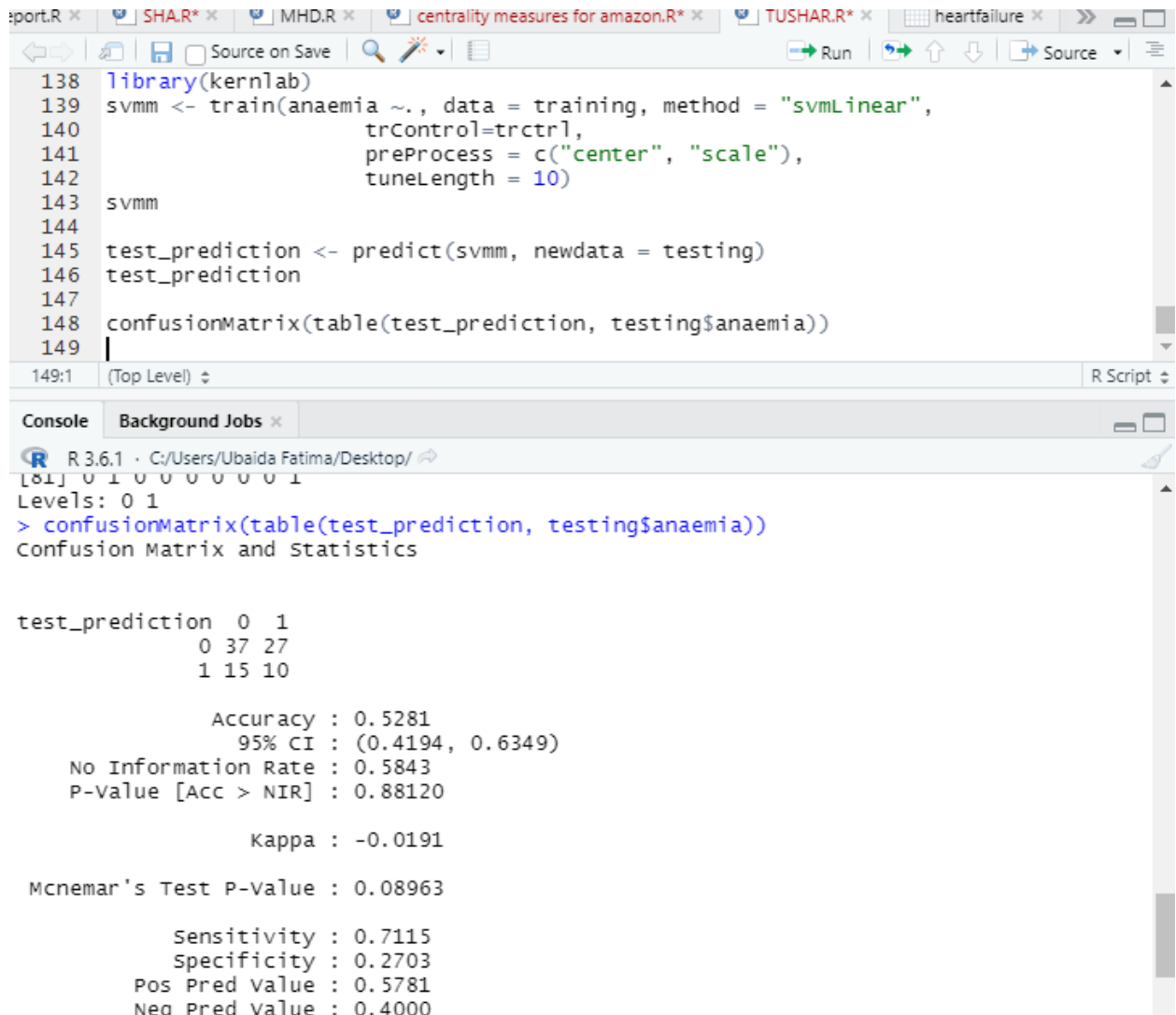


Figure 21. Test Prediction and Confusion Matrix Statistics

From the above figure, it can be observed that the predictive test had a coefficient of 0.5281, which indicates a moderate capacity to distinguish among the false positive, false negative, true positive, and true negative variations. Moreover, the p-value of the was also high, recorded at 0.88210 for the normal coefficient and McNemar's test P-value at 0.8963. Both of these values indicate that there is a high probability of the algorithm detecting false positives and false negative results. Although the model performs substantially well for the detection of true positives and true negatives, the high values for false positives and false negatives limit its practical application at this stage in clinical diagnostics and medical decision-making. This was further evidenced by the SVM accuracy test for the provided Dataset. The results are elaborated in the following figure.

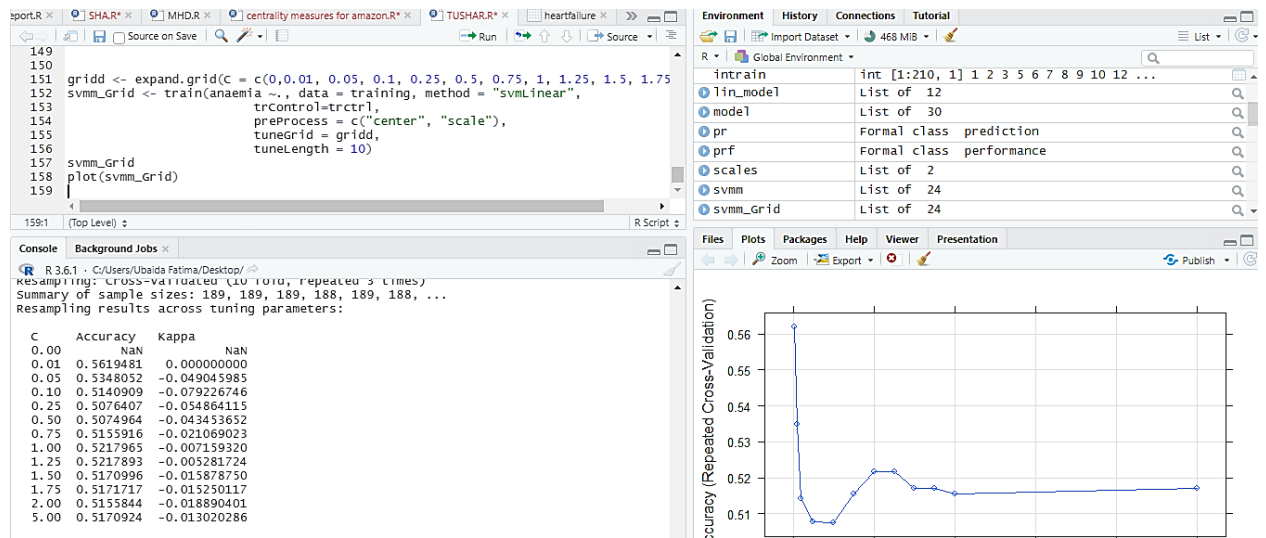


Figure 22. Support Vector Machine (SVM) Accuracy for Dataset

The predictability of the model presented by the Confusion matrix is shown in the following figure in the form of a plot. The plot depicts the accuracy obtained in the identification of the false positives, false negatives, true positives, and true negatives and the cost associated with the procedures to obtain that level of accuracy. As observable, the test prediction accuracy and cost are inversely proportional to each other. To attain a value of higher than 56% accuracy, the associated cost rise above the margins established. Whereas going down the trend, the associated cost reduces exponentially with a decline in accuracy. It is also to be noted that this inverse relation is not linear but exponential. This means that, firstly, there are limits to the statistical association between these two aspects, and they are not related to each other indefinitely. Secondly, going below the 51% accuracy mark does not lead to a further reduction in associated costs. This could be attributed to the cost structure of diagnostic tests of these indicators, which has certain essential criteria and procedures which could not be eliminated to further reduce costs. Moreover, from a practical point of view, it is unfeasible to go below such accuracy levels, as depicted in the figure, because medical diagnostics are generally preferred to be run at accuracy higher than 95%.

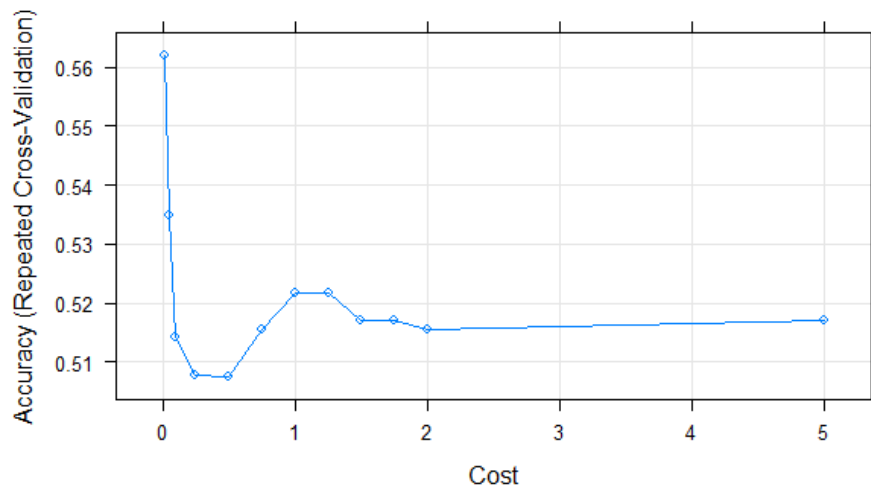


Figure 23. Support Vector Machine (SVM) plot for Dataset

The following figure shows the test for the significance of the model mentioned above.

Here, the accuracy is again moderate at 58.43%. However, the significance represented by the p-value is very high. This shows a satisfactory performance of the model applied for predicting the association between the aspects of the confusion matrix and the cost of the procedure.

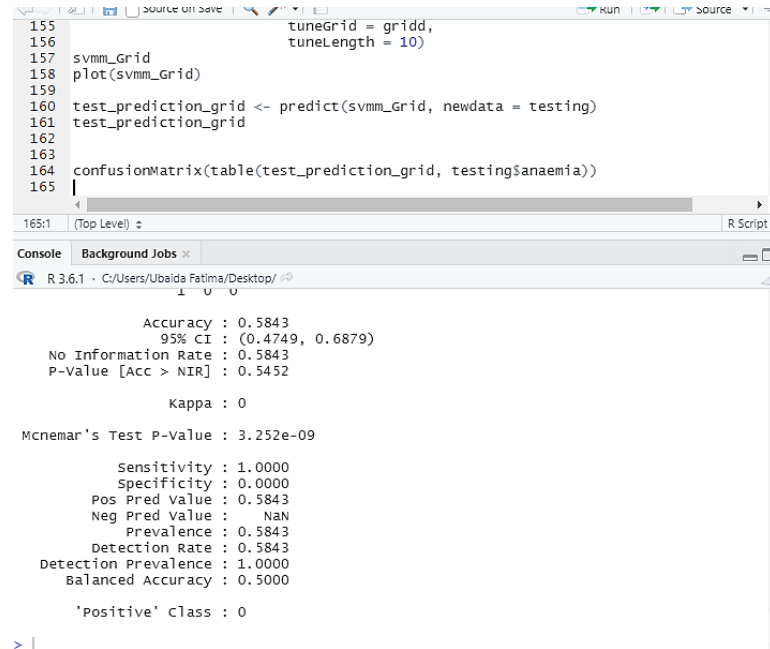


Figure 24. Confusion Matrix accuracy statistics for a test prediction model

4.3 Discussion

This research study has found that the practices of unhealthy lifestyles and eating habits have led to poor health outcomes and thus resulted in an increase in the number of patients. Moreover, it has been observed that machine learning, deep learning, and image processing algorithms have delivered advancements in the monitoring of health variables. Besides, the models have been found to be effective in making an accurate prediction of diseases and medical conditions before they exert a severe threat on individuals. This has been supported by Wang et al. (2021), which revealed the support of the requirement of monitoring health and subsequent risks with the help of digital processing. Further, Senbekov et al. (2020) have also corroborated the findings of this research by putting forward that modern technological tool such as machine learning and deep learning has remained to be efficient in delivering support to health care. Moreover, such digital tools have certified the prediction of threats and risks before imparting negative implications over the long run. According to Tucker et al. (2020), digital technologies and automation have made healthcare processes easier to be monitored by gathering and organising healthcare data digitally. Specifically, it has enabled the health care units to have access to all the pertinent information of their patients.

Furthermore, the study has determined the diseases that have been caused by obesity and blood pressure through R software. It has been observed that obesity has a prevalent concern among individuals of all ages. It has been observed that the prevalence of obesity has been more for the age group of 40 and less than for kids and teenagers. Brown et al. (2019) have observed similar findings that the rates of childhood obesity have substantially accelerated in England since the pandemic. There have been nearly two-fifths of children who lie in the age group of 10-11 have been facing the challenges of overweight and obesity. Thus, the occurrence of obesity has led to incessantly lessening the quality-adjusted life years of children. The study has observed that obesity has been significantly linked to the death rates of the patients together with a high risk of cardiovascular or heart diseases. The findings are supported by Kahathuduwa et al. (2019), which put forward that high levels of obesity tend to increase the risk of cancer, heart attacks, and thus mortality. The increase in obesity and blood pressure have been mainly the outcome of unhealthy lifestyles such as spending more time sitting down, sleeping late, and more consumption of junk food.

The study has also perceived that obesity is considerably associated with high blood pressure, anaemia, creatinine phosphokinase, ejection fraction, reduced platelets, and serum creatinine. This goes well with the findings of LeCroy et al. (2021) that obesity has been frequent comorbidity with the failure of the heart as well as is strongly linked with poor outcomes of health. Particularly, obesity or high blood pressure leads to heart failure as they have a tendency to develop an intricate collaboration of cytokine production, kidney disease, and deficiency of iron. According to Colmenarejo (2020), an obese individual carries more chance of developing cardiovascular disease. Gutierrez et al. (2018) have revealed that obesity has been contributing to a risk factor for heart disease and curtails the quality of life and life expectancy of an individual.

Furthermore, it has been put forward that if a model is able to make an accurate prediction of high blood pressure, anaemia, creatinine phosphokinase, ejection fraction, reduced platelets, and serum creatinine. Then it could result in lowering the risk and threats of heart diseases. In this regard, the R model has delivered substantial results. Berry et al. (2021) have employed the models of machine learning in order to predict the database of patients. The study has identified that machine learning models tend to deliver the most accurate predictions of their advanced technological features. Specifically, the variable anaemic has been observed to be significantly

predicted by the model. Jaiswal et al. (2019) have observed that machine learning models have been a good predictor of anaemia disease in patients. Moreover, anaemia has been common comorbidity of heart failure. In addition, it has also been linked with poor health outcomes (El-kenawy, 2019). Similar results have been put forward by Partridge and Redfern (2018), which found that the machine learning model tends to deliver higher and more accurate predictions of diseases by processing the huge patient database that is beyond the capability and magnitude of individuals. Moreover, it delivers more consistent and dependable data that could be easily converted for data analysis and gaining a medical understanding to deliver care and assist the clinician's plan. Furthermore, the study has analysed that the model has failed to deliver accurate predictions of a few health variables, such as creatinine phosphokinase, high blood pressure, platelets, and serum creatinine, as the data of patients have outliers. Rubbio, Rubbio et al. (2018) have proposed that, often, digital technologies have failed to deliver advantages to health care as the technologies have been linked with cyber-attacks and poor implementation. In all, the advancement of digital technologies has resulted in gaining more reliable and consistent outcomes for healthcare.

CHAPTER FIVE: CONCLUSION

5.1 Summary of findings

The current study was conducted with the aim of developing a rigorous learning algorithm to detect health predictors and indicators of patients from the available data. For this, the researcher utilised RStudio to formulate the algorithm upon which further statistical tests were conducted to evaluate the accuracy, predictability, and overall feasibility of the practical application of such an algorithm in real-time medical diagnostics. Although there is a myriad of health indicators used by medical practitioners to assess the health conditions of the public, the current research selected the indicators of anaemia, creatinine phosphokinase, diabetes, blood pressure, ejection fraction, platelets, and serum creatinine proportion. The idea behind the selection of these variables was to cover a wide range of health conditions which are generally assessable by these indicators. The results of regression analysis make it clear that anaemia and age have an interdependent association. As a result, the patient would have a higher risk of becoming anaemic as their age increased. Additionally, the figure showed that as long as the value is below 0.4, the likelihood of being anaemic at age 40 is low. The likelihood does, however, steadily rise with age. Showed the correlation between the variables diabetes and age in a regression plot. Therefore, based on the findings, it was possible to conclude that the model offered an effective explanation for the variables with a larger variation. The model specifically showed the likelihood of getting diabetes in higher age groups. Given that the value is over 0.4, the model is a good predictor of diabetes in those 40 years of age and older. However, the likelihood gradually decreases as the patient gets older. As a result, the model is a reliable prediction for patients who are young. With a p-value of 0.01633 0.05, it has been shown that the model strongly predicts the variable anaemic. However, it is clear that anaemia has a detrimental effect on cardiac problems. The model's prediction of the variable creatinine phosphokinase was found to be unreliable since its p-value was $0.10964 > 0.05$.

Finding a machine learning mechanism that might be trained to spot possible health hazards in the people being tracked was one of the main goals of the current research. However, the researcher realised that, within this larger framework of the health risk detection system, it is essential to add an iterative mechanism to the algorithm through the use of support vector machines (SVMs) to obtain this. These are digitally supervised learning models that are closely related to Rstudio's algorithms. The predictors of anaemia, creatinine phosphokinase, diabetes, blood

pressure, platelets, and serum creatinine were examined in the current study. These components of the created machine learning algorithm were divided into categories using a confused matrix analysis approach. A reasonable ability to discriminate between false positive, false negative, true positive, and true negative variants was shown by the predictive test's coefficient of 0.5281, which is a favourable sign. The p-value for the was likewise high, coming in at 0.88210 for the normal coefficient and 0.8963 for the McNemar test. Both of these numbers suggest that the algorithm has a good chance of identifying false positive and false negative outcomes. Although the model does a good job of detecting genuine positives and true negatives, its practical use in clinical diagnostics and medical decision-making is still constrained by the high values for false positives and false negatives. The test's cost and prediction accuracy have an inverse relationship. The related cost must exceed the stated margins in order to achieve a value of accuracy greater than 56%. On the other hand, when accuracy declines, the corresponding cost drops off tremendously. Additionally, it should be emphasised that this inverse relationship is exponential rather than linear. This indicates that, firstly, the statistical correlation between these two features has boundaries and that there is no constant relationship between them.

Recommendations

Since the current study focused on the detection and identification of health risks in patients of a myriad of conditions, it explored considerable volumes of data for a holistic digital processing of the real time patient information. Overall, the research provided evidence towards the initial assertion that digital processing techniques applied on patient data produce better diagnostic outcomes compared to conventional diagnostic mechanisms. This is mostly due to the fact that while computer-processed diagnostics are based on analysis of tremendously large volumes of data, conventional diagnostic mechanisms are limited in their ability to decipher hidden patterns in intricate data. In light of the above analysis, following recommendations are proposed to facilitate medical practitioners, healthcare institutions, and technology developers to better cope with the rising demand of digital algorithm-based patient care:

- Healthcare institutions should collaborate to initiate a framework development which could facilitate in creating a globally accessible and secure database of healthcare parameters of people from different cultures and regions. This is because for a fully functional self-learning algorithm to become viable, it is imperative that

it has access to a global databank of patients. This would also help in automating a number of steps in the diagnostic processes which are currently hindered due to requirement of human decision maker

- The current enthusiasm demonstrated in the field of medical science and machine learning needs to be harnessed in order to develop an applicable platform for practice. It is recommended that academia from healthcare and machine learning should initiate collaborative platforms on regional levels so that both could achieve optimum resource utilisation from each other's capabilities
- The current research found diagnostic costs with entailing financial burden as one of the leading limitations against implementation of digital diagnostics mechanisms. It is recommended diagnostics services are made open source and protected through the blockchain framework in order to ensure both high profitability and data security. Such a mechanism would also curb the possibility of emergence of data silos
- For medical institutions, it has become imperatively important to initiate programs on digital diagnostics for newly graduating medical students. This would help them in familiarisation with the major tools and techniques, which would, in turn, reduce the burden of learning later in their careers.

Conclusion

The current study aimed at developing a machine learning based algorithm for early detection and identification of health issues in patients. For specification, the study looked at the parameters of anaemia, blood pressure, diabetes, platelets, and creatinine hexokinase as the major indicators of health issues, and proceeded to categorise their data under the criteria segregation of age, sex, smoking, and time since test. Although the study focused on algorithm development, a number of tests were conducted to examine its viability for real time application. Through the descriptive analysis, the study revealed that the data obtained from model prediction had negligible variation in means, which is indicative of negligible occurrences of outliers. This in turn shows that the data would have high predictability in diagnostics. With respect regression analysis, the research revealed that while anaemia could be significantly predicted by the proposed algorithms, there is insignificance predictability for creatinine phosphokinase. The other indicators of diabetes and blood pressure are moderately predicted. With respect to diagnostic accuracy, the proposed

model has moderate power of detecting false negative results. Meanwhile, it also showed that diagnostic cost and accuracy are inversely proportional to each other, with increasing accuracy causing an exponential rise in associated costs.

Future implications

While the proposed algorithm is still in an infancy stage, it provides a productive ground for further iterations and developments. Results from the current study could be utilised as inputs for better developed models so that the need for data cleansing could be bypassed. Moreover, the current study was limited in terms of parameters selection. Future studies could apply the same mechanism for the examination of other health indicators. This could also be tested for patients from various demographics to check whether viability outcomes vary with respect to the population studied.

References

- Aiello, A.E., Renson, A. and Zivich, P., 2020. Social media-and internet-based disease surveillance for public health. *Annual review of public health*, 41, p.101.
- Alharahsheh, H.H. and Pius, A., 2020. A review of key paradigms: Positivism VS interpretivism. *Global Academic Journal of Humanities and Social Sciences*, 2(3), pp.39-43.
- Alkire, B.C., Peters, A.W., Shrimel, M.G. and Meara, J.G., 2018. The economic consequences of mortality amenable to high-quality health care in low-and middle-income countries. *Health Affairs*, 37(6), pp.988-996.
- Al-Shorbaji, N., 2021. Improving Healthcare Access through Digital Health: The Use of Information and Communication Technologies. *Healthcare Access*.
- Aman, A.H.M., Hassan, W.H., Sameen, S., Attarbashi, Z.S., Alizadeh, M. and Latiff, L.A., 2021. IoMT amid COVID-19 pandemic: Application, architecture, technology, and security. *Journal of Network and Computer Applications*, 174, p.102886.
- Arifin, S.R.M., 2018. Ethical considerations in qualitative study. *International Journal of Care Scholars*, 1(2), pp.30-33.
- Arumugam, P. and Saranya, R., 2018. Outlier detection and missing value in seasonal ARIMA model using rainfall data. *Materials Today: Proceedings*, 5(1), pp.1791-1799.
- Avolio, E., Gualtieri, P., Romano, L., Pecorella, C., Ferraro, S., Palma, G., Di Renzo, L. and De Lorenzo, A., 2020. Obesity and body composition in man and woman: associated diseases and the new role of gut microbiota. *Current medicinal chemistry*, 27(2), pp.216-229.
- Awuzie, B. and McDermott, P., 2017. An abductive approach to qualitative built environment research: A viable system methodological exposé. *Qualitative research journal*.

- Bendor, C.D., Bardugo, A., Pinhas-Hamiel, O., Afek, A. and Twig, G., 2020. Cardiovascular morbidity, diabetes and cancer risk among children and adolescents with severe obesity. *Cardiovascular diabetology*, 19(1), pp.1-14.
- Berry, R., Kassavou, A. and Sutton, S., 2021. Does self-monitoring diet and physical activity behaviors using digital technology support adults with obesity or overweight to lose weight? A systematic literature review with meta-analysis. *Obesity Reviews*, 22(10), p.e13306.
- Borenstein, M., 2022. Comprehensive meta-analysis software. *Systematic Reviews in Health Research: Meta-Analysis in Context*, pp.535-548.
- Brown, T., Moore, T.H., Hooper, L., Gao, Y., Zayegh, A., Ijaz, S., Elwenspoek, M., Foxen, S.C., Magee, L., O'Malley, C. and Waters, E., 2019. Interventions for preventing obesity in children. *Cochrane Database of Systematic Reviews*, (7).
- Burgess, E., Hassmén, P. and Pumpa, K.L., 2017. Determinants of adherence to lifestyle intervention in adults with obesity: a systematic review. *Clinical obesity*, 7(3), pp.123-135.
- Busnatu, S.S., Salmen, T., Pana, M.A., Rizzo, M., Stallone, T., Papanas, N., Popovic, D., Tanasescu, D., Serban, D. and Stoian, A.P., 2022. The Role of Fructose as a Cardiovascular Risk Factor: An Update. *Metabolites*, 12(1), p.67.
- Calle, E.E., Thun, M.J., Petrelli, J.M., Rodriguez, C. and Heath Jr, C.W., 1999. Body-mass index and mortality in a prospective cohort of US adults. *New England Journal of Medicine*, 341(15), pp.1097-1105.
- Cavanaugh, J.E. and Neath, A.A., 2019. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(3), p.e1460.
- Chen, J.M. and Luetz, J.M., 2020. Mono-/inter-/multi-/trans-/anti-disciplinarity in research. *Quality Education*, pp.562-577.

- Chew, H.S.J., Ang, W.H.D. and Lau, Y., 2021. The potential of artificial intelligence in enhancing adult weight loss: a scoping review. *Public health nutrition*, 24(8), pp.1993-2020.
- Chicco, D. and Jurman, G., 2020. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making*, 20(1), pp.1-16.
- Cirillo, A., 2016. *Rstudio for R statistical computing cookbook*. Packt Publishing Ltd.
- Cohen, A.L., Kang, N. and Leise, T.L., 2017. Multi-attribute, multi-alternative models of choice: Choice, reaction time, and process tracing. *Cognitive psychology*, 98, pp.45-72.
- Colmenarejo, G., 2020. Machine learning models to predict childhood and adolescent obesity: a review. *Nutrients*, 12(8), p.2466.
- Costa, D.G. and Peixoto, J.P.J., 2020. COVID-19 pandemic: A review of smart cities initiatives to face new outbreaks. *IET Smart Cities*, 2(2), pp.64-73.
- Dennis, B., Ponciano, J.M., Taper, M.L. and Lele, S.R., 2019. Errors in statistical inference under model misspecification: evidence, hypothesis testing, and AIC. *Frontiers in Ecology and Evolution*, 7, p.372.
- Di Cesare, M., Sorić, M., Bovet, P., Miranda, J.J., Bhutta, Z., Stevens, G.A., Laxmaiah, A., Kengne, A.P. and Bentham, J., 2019. The epidemiological burden of obesity in childhood: a worldwide epidemic requiring urgent action. *BMC medicine*, 17(1), pp.1-20.
- Ding, Z. and Xing, L., 2020. Improved software defect prediction using Pruned Histogram-based isolation forest. *Reliability Engineering & System Safety*, 204, p.107170.
- Eckmanns, T., Fuller, H. and Roberts, S.L., 2019. Digital epidemiology and global health security; an interdisciplinary conversation. *Life Sciences, Society and Policy*, 15(1), pp.1-13.

- El-kenawy, E.S.M.T., 2019. A Machine Learning Model for Hemoglobin Estimation and Anemia Classification. *International Journal of Computer Science and Information Security (IJCSIS)*, 17(2), pp.100-108.
- Ferdowsy, F., Rahi, K.S.A., Jabiullah, M.I. and Habib, M.T., 2021. A machine learning approach for obesity risk prediction. *Current Research in Behavioral Sciences*, 2, p.100053.
- Ferdowsy, F., Rahi, K.S.A., Jabiullah, M.I. and Habib, M.T., 2021. A machine learning approach for obesity risk prediction. *Current Research in Behavioral Sciences*, 2, p.100053.
- Fieschi, M., 2018. *Health data processing: Systemic approaches*. Elsevier.
- Food, N., 2007. Physical Activity, and the Prevention of Cancer Research. World Cancer Research Fund. American Institute for Cancer Research.
- Gao, W., Farahani, M.R., Aslam, A. and Hosamani, S., 2017. Distance learning techniques for ontology similarity measuring and ontology mapping. *Cluster Computing*, 20(2), pp.959-968.
- Gasser, U., Ienca, M., Scheibner, J., Sleight, J. and Vayena, E., 2020. Digital tools against COVID-19: taxonomy, ethical challenges, and navigation aid. *The Lancet Digital Health*, 2(8), pp.e425-e434.
- George, D. and Mallery, P., 2018. Descriptive statistics. In *IBM SPSS Statistics 25 Step by Step* (pp. 126-134). Routledge.
- Stauder, R., Valent, P. and Theurl, I., 2018. Anemia at older age: etiologies, clinical implications, and management. *Blood, The Journal of the American Society of Hematology*, 131(5), pp.505-514.
- Green, T.L., 2017. From positivism to critical theory: School-community relations toward community equity literacy. *International Journal of Qualitative Studies in Education*, 30(4), pp.370-387.

- Guh, D.P., Zhang, W., Bansback, N., Amarsi, Z., Birmingham, C.L. and Anis, A.H., 2009. The incidence of co-morbidities related to obesity and overweight: a systematic review and meta-analysis. *BMC public health*, 9(1), pp.1-20.
- Guo, J. and Li, B., 2018. The application of medical artificial intelligence technology in rural areas of developing countries. *Health equity*, 2(1), pp.174-181.
- Guryanova, L.S., Yatsenko, R., Dubrovina, N.A. and Babenko, V.O., 2020. Machine learning methods and models, predictive analytics and applications.
- Gutierrez, J., Alloubani, A., Mari, M. and Alzaatreh, M., 2018. Cardiovascular disease risk factors: hypertension, diabetes mellitus and obesity among Tabuk citizens in Saudi Arabia. *The open cardiovascular medicine journal*, 12, p.41.
- Hall, J.E., do Carmo, J.M., da Silva, A.A., Wang, Z. and Hall, M.E., 2019. Obesity, kidney dysfunction and hypertension: mechanistic links. *Nature reviews nephrology*, 15(6), pp.367-385.
- Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H. and Aerts, H.J., 2018. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8), pp.500-510.
- Jaiswal, M., Srivastava, A. and Siddiqui, T.J., 2019. Machine learning algorithms for anemia disease prediction. In *Recent trends in communication, computing, and electronics* (pp. 463-469). Springer, Singapore.
- Jebb, A.T., Parrigon, S. and Woo, S.E., 2017. Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, 27(2), pp.265-276.
- Jha, S. and Topol, E.J., 2016. Adapting to artificial intelligence: radiologists and pathologists as information specialists. *Jama*, 316(22), pp.2353-2354.
- Johnston, A., 2014. Rigour in research: theory in the research approach. *European Business Review*.

- Kahathuduwa, C.N., West, B.D., Blume, J., Dharavath, N., Moustaid-Moussa, N. and Mastergeorge, A., 2019. The risk of overweight and obesity in children with autism spectrum disorders: A systematic review and meta-analysis. *Obesity Reviews*, 20(12), pp.1667-1679.
- Kalid, N., Zaidan, A.A., Zaidan, B.B., Salman, O.H., Hashim, M. and Muzammil, H.J.J.O.M.S., 2018. Based real time remote health monitoring systems: A review on patients prioritization and related" big data" using body sensors information and communication technology. *Journal of medical systems*, 42(2), pp.1-30.
- Kapur, R., Kalra, S., Tiwari, K. and Arora, G., 2021. Training Software Engineers for Qualitative Evaluation of Software Architecture. *arXiv preprint arXiv:2105.09595*.
- Khan, J.R., Chowdhury, S., Islam, H. and Raheem, E., 2019. Machine learning algorithms to predict the childhood anemia in Bangladesh. *Journal of Data Science*, 17(1), pp.195-218.
- 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.
- Kleindienst, D., 2017. The data quality improvement plan: deciding on choice and sequence of data quality improvements. *Electronic Markets*, 27(4), pp.387-398.
- Kruk, M.E., Gage, A.D., Joseph, N.T., Danaei, G., García-Saisó, S. and Salomon, J.A., 2018. Mortality due to low-quality health systems in the universal health coverage era: a systematic analysis of amenable deaths in 137 countries. *The Lancet*, 392(10160), pp.2203-2212.
- Lake, I.R., Colon-Gonzalez, F.J., Barker, G.C., Morbey, R.A., Smith, G.E. and Elliot, A.J., 2019. Machine learning to refine decision making within a syndromic surveillance service. *BMC Public Health*, 19(1), pp.1-12.

- Lavie, C.J., McAuley, P.A., Church, T.S., Milani, R.V. and Blair, S.N., 2014. Obesity and cardiovascular diseases: implications regarding fitness, fatness, and severity in the obesity paradox. *Journal of the American College of Cardiology*, 63(14), pp.1345-1354.
- LeCroy, M.N., Kim, R.S., Stevens, J., Hanna, D.B. and Isasi, C.R., 2021. Identifying key determinants of childhood obesity: a narrative review of machine learning studies. *Childhood Obesity*, 17(3), pp.153-159.
- Li, J.P.O., Liu, H., Ting, D.S., Jeon, S., Chan, R.P., Kim, J.E., Sim, D.A., Thomas, P.B., Lin, H., Chen, Y. and Sakamoto, T., 2021. Digital technology, tele-medicine and artificial intelligence in ophthalmology: A global perspective. *Progress in retinal and eye research*, 82, p.100900.
- Liang, Y., Abbott, D., Howard, N., Lim, K., Ward, R. and Elgendi, M., 2019. How effective is pulse arrival time for evaluating blood pressure? Challenges and recommendations from a study using the MIMIC database. *Journal of clinical medicine*, 8(3), p.337.
- Lv, Z., Chen, D. and Lv, H., 2022. Smart City Construction and Management by Digital Twins and BIM Big Data in COVID-19 Scenario. *ACM Transactions on Multimedia Computing Communications and Applications*.
- Majumder, S. and Deen, M.J., 2019. Smartphone sensors for health monitoring and diagnosis. *Sensors*, 19(9), p.2164.
- Mathews, S.C., McShea, M.J., Hanley, C.L., Ravitz, A., Labrique, A.B. and Cohen, A.B., 2019. Digital health: a path to validation. *NPJ digital medicine*, 2(1), pp.1-9.
- Maulud, D. and Abdulazeez, A.M., 2020. A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, 1(4), pp.140-147.
- National Institutes of Health, 1998. National Heart, Lung and Blood Institute. Clinical Guidelines on the identification, evaluation and treatment of overweight and obesity in adults. Public Health Service.

- Nedungadi, P., Jayakumar, A. and Raman, R., 2018. Personalized health monitoring system for managing well-being in rural areas. *Journal of medical systems*, 42(1), pp.1-11.
- Nusinovici, S., Tham, Y.C., Yan, M.Y.C., Ting, D.S.W., Li, J., Sabanayagam, C., Wong, T.Y. and Cheng, C.Y., 2020. Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of clinical epidemiology*, 122, pp.56-69.
- O'Brien, P.E. and Dixon, J.B., 2002. The extent of the problem of obesity. *The American Journal of Surgery*, 184(6), pp.S4-S8.
- Osman, S., Mohammad, S., Abu, M.S., Mokhtar, M., Ahmad, J., Ismail, N. and Jambari, H., 2018. Inductive, Deductive and Abductive Approaches in Generating New Ideas: A Modified Grounded Theory Study. *Advanced Science Letters*, 24(4), pp.2378-2381.
- Ovando, M.A.W., Garcia, P.P., Escalante, F.D.A. and Nolasco, J.A.H. eds., 2018. *Intelligent data sensing and processing for health and well-being applications*. Academic Press.
- Pandey, J., 2019. Deductive approach to content analysis. In *Qualitative techniques for workplace data analysis* (pp. 145-169). IGI Global.
- Pandey, J., 2019. Deductive approach to content analysis. In *Qualitative techniques for workplace data analysis* (pp. 145-169). IGI Global.
- Parsons, V.L., 2014. Stratified sampling. *Wiley StatsRef: Statistics Reference Online*, pp.1-11.
- Partridge, S.R. and Redfern, J., 2018, June. Strategies to engage adolescents in digital health interventions for obesity prevention and management. In *Healthcare* (Vol. 6, No. 3, p. 70). MDPI.
- Preston, S.H. and Stokes, A., 2014. Obesity paradox: conditioning on disease enhances biases in estimating the mortality risks of obesity. *Epidemiology* (Cambridge, Mass.), 25(3), p.454.
- Punj, R. and Kumar, R., 2019. Technological aspects of WBANs for health monitoring: a comprehensive review. *Wireless Networks*, 25(3), pp.1125-1157.

- Ray, S., 2019, February. A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)* (pp. 35-39). IEEE.
- Reibling, N., Ariaans, M. and Wendt, C., 2019. Worlds of healthcare: a healthcare system typology of OECD countries. *Health Policy*, 123(7), pp.611-620. Rakhmonov, I.U., Nematov, L.A., Niyozov, N.N.,
- Reymov, K.M. and Yuldoshev, T.M., 2020, April. Power consumption management from the positions of the general system theory. In *Journal of Physics: Conference Series* (Vol. 1515, No. 2, p. 022054). IOP Publishing.
- Roser, M. and Ritchie, H. (2019). *HIV / AIDS*. [online] Our World in Data. Available at: <https://ourworldindata.org/hiv-aids> [Accessed 8 Jul. 2022].
- Rubbio, I., Bruccoleri, M., Pietrosi, A. and Ragonese, B., 2018. Digital health technology enhances resilient behaviour: evidence from the ward. *International Journal of Operations & Production Management*.
- Saltz, J.S. and Dewar, N., 2019. Data science ethical considerations: a systematic literature review and proposed project framework. *Ethics and Information Technology*, 21(3), pp.197-208.
- Sarwar, M.A., Kamal, N., Hamid, W. and Shah, M.A., 2018, September. Prediction of diabetes using machine learning algorithms in healthcare. In *2018 24th international conference on automation and computing (ICAC)* (pp. 1-6). IEEE.
- Senbekov, M., Saliev, T., Bukeyeva, Z., Almabayeva, A., Zhanaliyeva, M., Aitenova, N., Toishibekov, Y. and Fakhradiyev, I., 2020. The recent progress and applications of digital technologies in healthcare: a review. *International journal of telemedicine and applications*, 2020.
- Sony, S., Laventure, S. and Sadhu, A., 2019. A literature review of next-generation smart sensing technology in structural health monitoring. *Structural Control and Health Monitoring*, 26(3), p.e2321.

- Stavridou, A., Kapsali, E., Panagouli, E., Thirios, A., Polychronis, K., Bacopoulou, F., Psaltopoulou, T., Tsolia, M., Sergeantanis, T.N. and Tsitsika, A., 2021. Obesity in children and adolescents during COVID-19 pandemic. *Children*, 8(2), p.135.
- Sujith, A.V.L.N., Sajja, G.S., Mahalakshmi, V., Nuhmani, S. and Prasanalakshmi, B., 2022. A systematic review of smart health monitoring using deep learning and Artificial intelligence. *Neuroscience Informatics*, 2(3), p.100028.
- Swaroop, K.N., Chandu, K., Gorrepotu, R. and Deb, S., 2019. A health monitoring system for vital signs using IoT. *Internet of Things*, 5, pp.116-129.
- Taguchi, N., 2018. Description and explanation of pragmatic development: Quantitative, qualitative, and mixed methods research. *System*, 75, pp.23-32.
- Tucker, A., Wang, Z., Rotalinti, Y. and Myles, P., 2020. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine*, 3(1), pp.1-13.
- Wang, Q., Su, M., Zhang, M. and Li, R., 2021. Integrating digital technologies and public health to fight Covid-19 pandemic: key technologies, applications, challenges and outlook of digital healthcare. *International Journal of Environmental Research and Public Health*, 18(11), p.6053.
- Zhao, X., Yan, X., Yu, A. and Van Hentenryck, P., 2020. Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel behaviour and society*, 20, pp.22-35.
- Talal, M., Zaidan, A.A., Zaidan, B.B., Albahri, A.S., Alamoodi, A.H., Albahri, O.S., Alsalem, M.A., Lim, C.K., Tan, K.L., Shir, W.L. and Mohammed, K.I., 2019. Smart home-based IoT for real-time and secure remote health monitoring of triage and priority system using body sensors: Multi-driven systematic review. *Journal of medical systems*, 43(3), pp.1-34.

- Teubner, G., 2017. How the law thinks: toward a constructivist epistemology of law. In *Legal Theory and the Social Sciences* (pp. 205-235). Routledge.
- Vargo, D., Zhu, L., Benwell, B. and Yan, Z., 2021. Digital technology use during COVID-19 pandemic: A rapid review. *Human Behavior and Emerging Technologies*, 3(1), pp.13-24.
- Vogt, W.P., Gardner, D.C. and Haeffele, L.M., 2012. When to use what research design. Guilford Press.
- Woiceshyn, J. and Daellenbach, U., 2018. Evaluating inductive vs deductive research in management studies: Implications for authors, editors, and reviewers. *Qualitative research in organizations and management: An International Journal*.
- Woo, S.E., O'Boyle, E.H. and Spector, P.E., 2017. Best practices in developing, conducting, and evaluating inductive research. *Human Resource Management Review*, 27(2), pp.255-264.
- Woo, S.E., O'Boyle, E.H. and Spector, P.E., 2017. Best practices in developing, conducting, and evaluating inductive research. *Human Resource Management Review*, 27(2), pp.255-264.
- World Health Organization (2020). *The Top 10 Causes of Death*. [online] World Health Organization. Available at: <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>.
- World Health Organization (2021). *Obesity and Overweight*. [online] World Health Organisation. Available at: <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight> [Accessed 8 Jul. 2022].
- Xu, J. and Xu, L., 2017. *Integrated System Health Management: Perspectives on Systems Engineering Techniques*. Academic Press.
- Yang, L., Yao, T., Liu, G., Sun, L., Yang, N., Zhang, H., Zhang, S., Yang, Y., Pang, Y., Liu, X. and Hou, X., 2019. Monitoring and controlling medical air disinfection parameters of nosocomial infection system based on internet of things. *Journal of Medical Systems*, 43(5), pp.1-7.

APPENDIX A: ETHICAL APPROVAL



College of Engineering, Design and Physical Sciences Research Ethics Committee
Brunel University London
Kingston Lane
Uxbridge
UB8 3PH
United Kingdom
www.brunel.ac.uk

29 July 2022

LETTER OF CONFIRMATION

Applicant: Mr tushar tayal

Project Title: Health monitoring and threat detection through digital processing

Reference: 37603-NER-Jul/2022- 40470-1

Dear Mr tushar tayal

The Research Ethics Committee has considered the above application recently submitted by you.

The Chair, acting under delegated authority has confirmed that, according to the information provided in your application, your project does not require ethical review.

Please note that:

- You are not permitted to conduct research involving human participants, their tissue and/or their data. If you wish to conduct such research, you must contact the Research Ethics Committee to seek approval prior to engaging with any participants or working with data for which you do not have approval.
- The Research Ethics Committee reserves the right to sample and review documentation relevant to the study.
- If during the course of the study, you would like to carry out research activities that concern a human participant, their tissue and/or their data, you must inform the Committee by submitting an appropriate Research Ethics Application. Research activity includes the recruitment of participants, undertaking consent procedures and collection of data. Breach of this requirement constitutes research misconduct and is a disciplinary offence.

Health Monitoring and Threat Detection through Digital Processing

Good luck with your research!

Kind regards,

A handwritten signature in black ink, appearing to read 'Simon Taylor'. The signature is fluid and cursive, with a long horizontal stroke at the bottom.

Professor Simon Taylor

Chair of the College of Engineering, Design and Physical Sciences Research Ethics Committee

Brunel University London

APPENDIX : B

Health monitoring and threat detection through digital processing

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time
1	75.000	0	582	0	20	1	265000	1.90	130	1	0	4
2	55.000	0	7861	0	38	0	263358	1.10	136	1	0	6
3	65.000	0	146	0	20	0	162000	1.30	129	1	1	7
4	50.000	1	111	0	20	0	210000	1.90	137	1	0	7
5	65.000	1	160	1	20	0	327000	2.70	116	0	0	8
6	90.000	1	47	0	40	1	204000	2.10	132	1	1	8
7	75.000	1	246	0	15	0	127000	1.20	137	1	0	10
8	60.000	1	315	1	60	0	454000	1.10	131	1	1	10
9	65.000	0	157	0	65	0	263358	1.50	138	0	0	10
10	80.000	1	123	0	35	1	388000	9.40	133	1	1	10
11	75.000	1	81	0	38	1	368000	4.00	131	1	1	10
12	62.000	0	231	0	25	1	253000	0.90	140	1	1	10
13	45.000	1	981	0	30	0	136000	1.10	137	1	0	11
14	50.000	1	168	0	38	1	276000	1.10	137	1	0	11
15	49.000	1	80	0	30	1	427000	1.00	138	0	0	12
16	82.000	1	379	0	50	0	47000	1.30	136	1	0	13
17	87.000	1	149	0	38	0	262000	0.90	140	1	0	14
18	45.000	0	582	0	14	0	166000	0.80	127	1	0	14
19	70.000	1	125	0	25	1	237000	1.00	140	0	0	15
20	48.000	1	582	1	55	0	87000	1.90	121	0	0	15
21	65.000	1	52	0	25	1	276000	1.30	137	0	0	16
22	65.000	1	128	1	30	1	297000	1.60	136	0	0	20

Figure 25: Dataset import in R Software

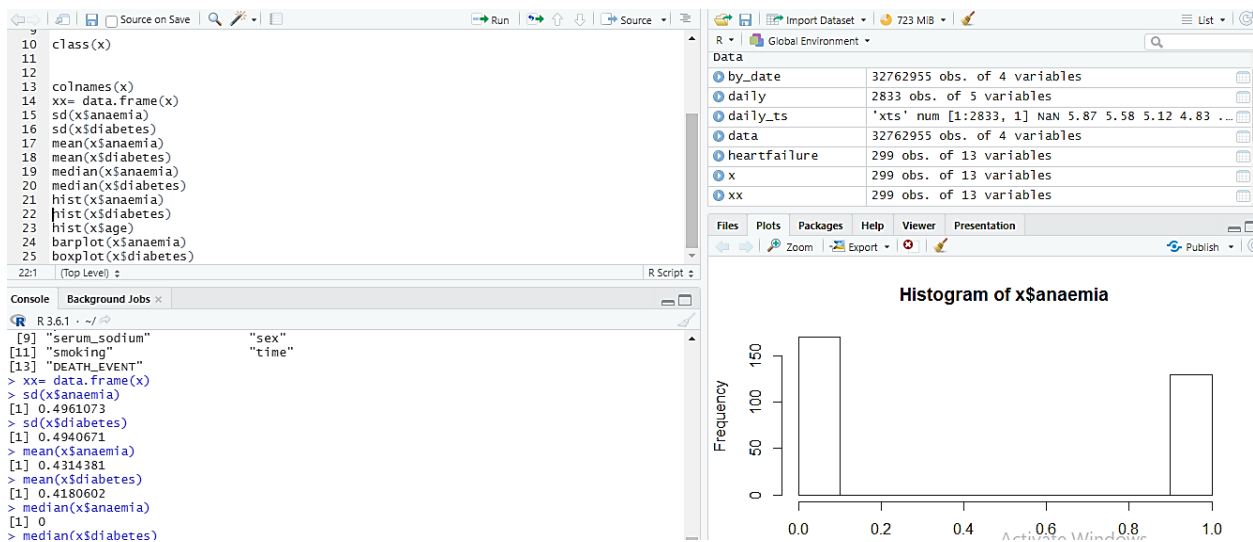


Figure 26: Descriptive Statistics on the dataset

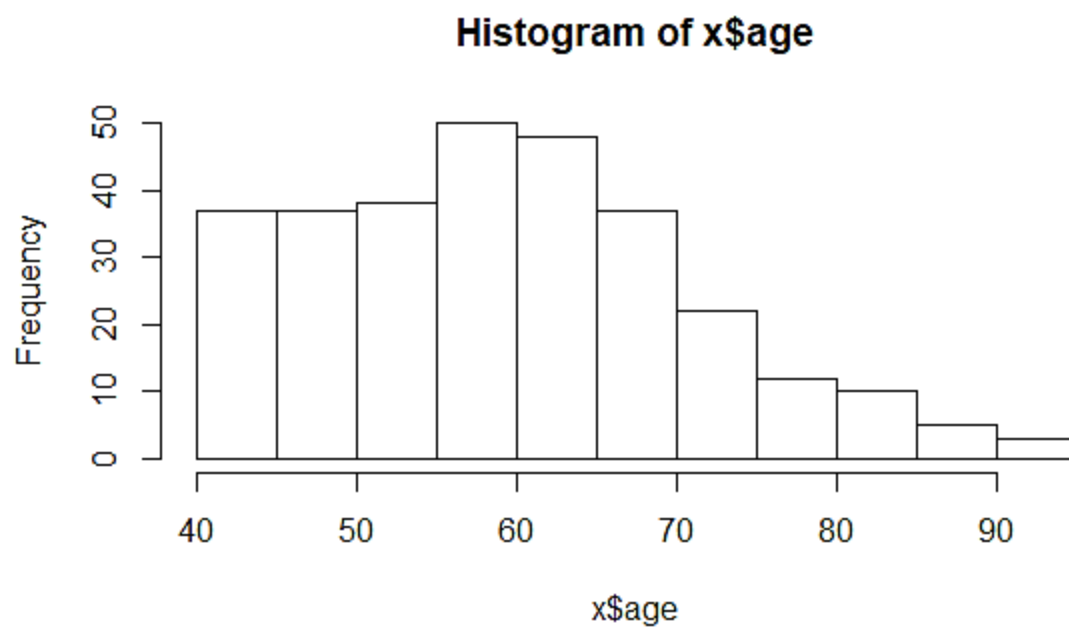


Figure 27: Histogram of Age variable

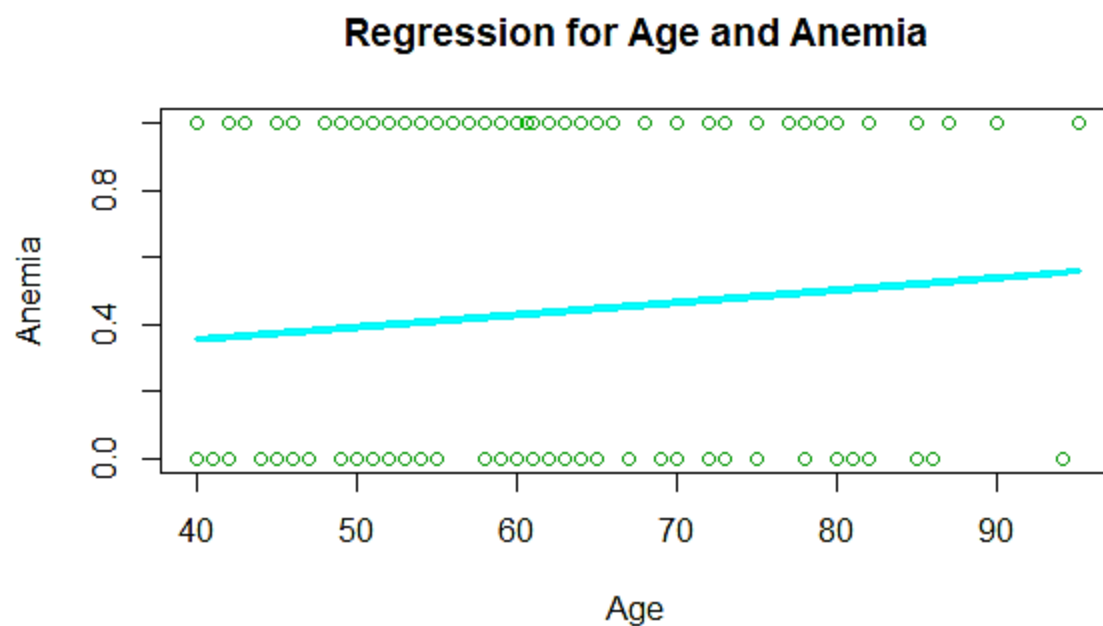


Figure 28: Regression Plot between Age and Anemia variables

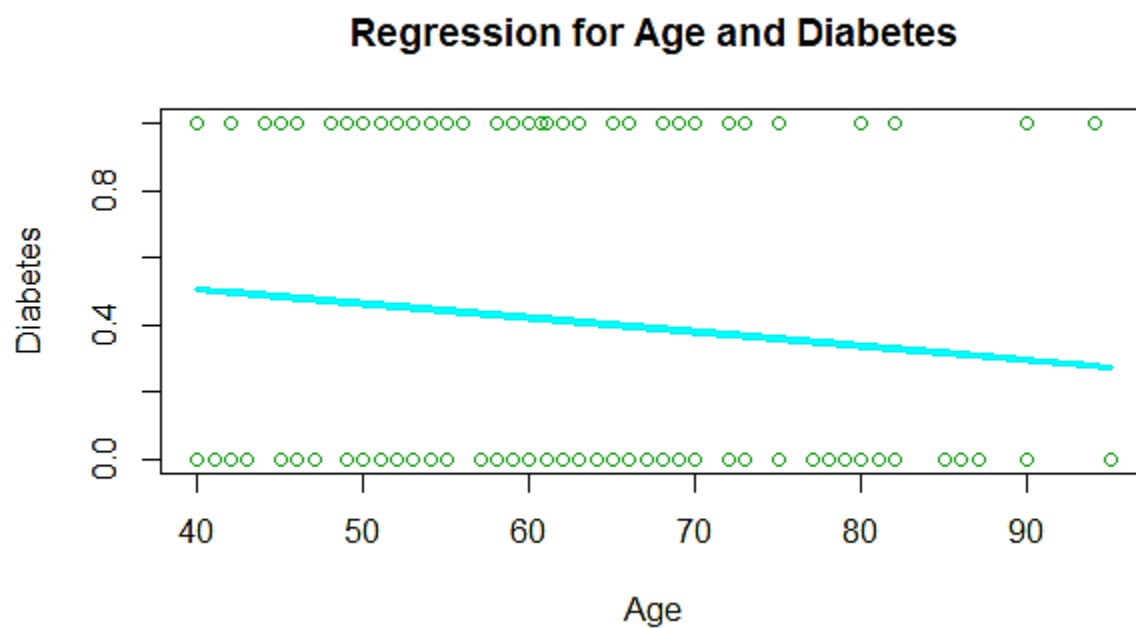


Figure 29: Regression between Age and Diabetes

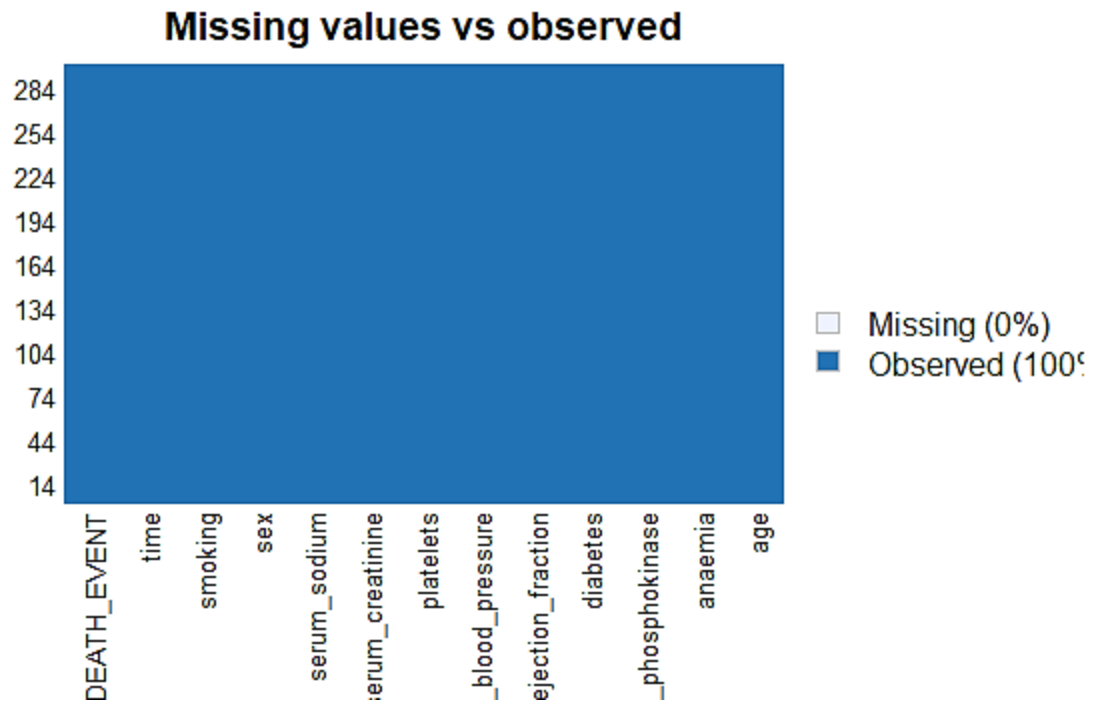


Figure 30: Missing Values Finding

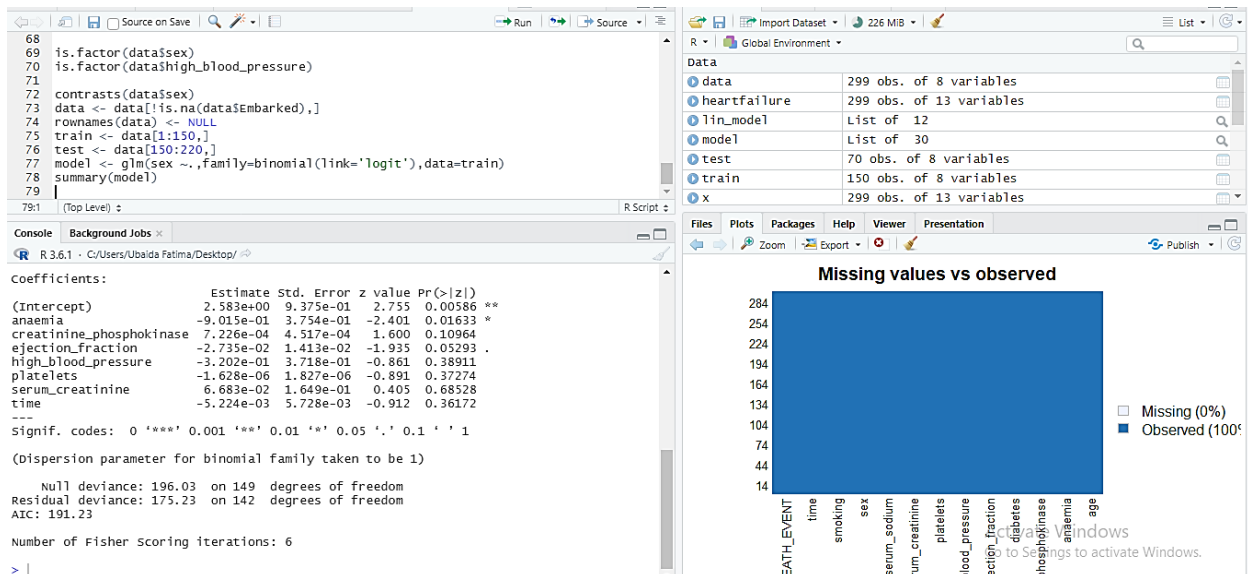


Figure 31: Training Dataset in R Software

```

68
69 is.factor(data$sex)
70 is.factor(data$high_blood_pressure)
71
72 contrasts(data$sex)
73 data <- data[!is.na(data$Embarked),]
74 rownames(data) <- NULL
75 train <- data[1:150,]
76 test <- data[150:220,]
77 model <- glm(sex ~.,family=binomial(link='logit'),data=train)
78 summary(model)
79 |

```

79:1 (Top Level) R Scrip

Console Background Jobs x

R 3.6.1 · C:/Users/Ubaida Fatima/Desktop/ ↗

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.583e+00	9.375e-01	2.755	0.00586 **
anaemia	-9.015e-01	3.754e-01	-2.401	0.01633 *
creatinine_phosphokinase	7.226e-04	4.517e-04	1.600	0.10964
ejection_fraction	-2.735e-02	1.413e-02	-1.935	0.05293 .
high_blood_pressure	-3.202e-01	3.718e-01	-0.861	0.38911
platelets	-1.628e-06	1.827e-06	-0.891	0.37274
serum_creatinine	6.683e-02	1.649e-01	0.405	0.68528
time	-5.224e-03	5.728e-03	-0.912	0.36172

 signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

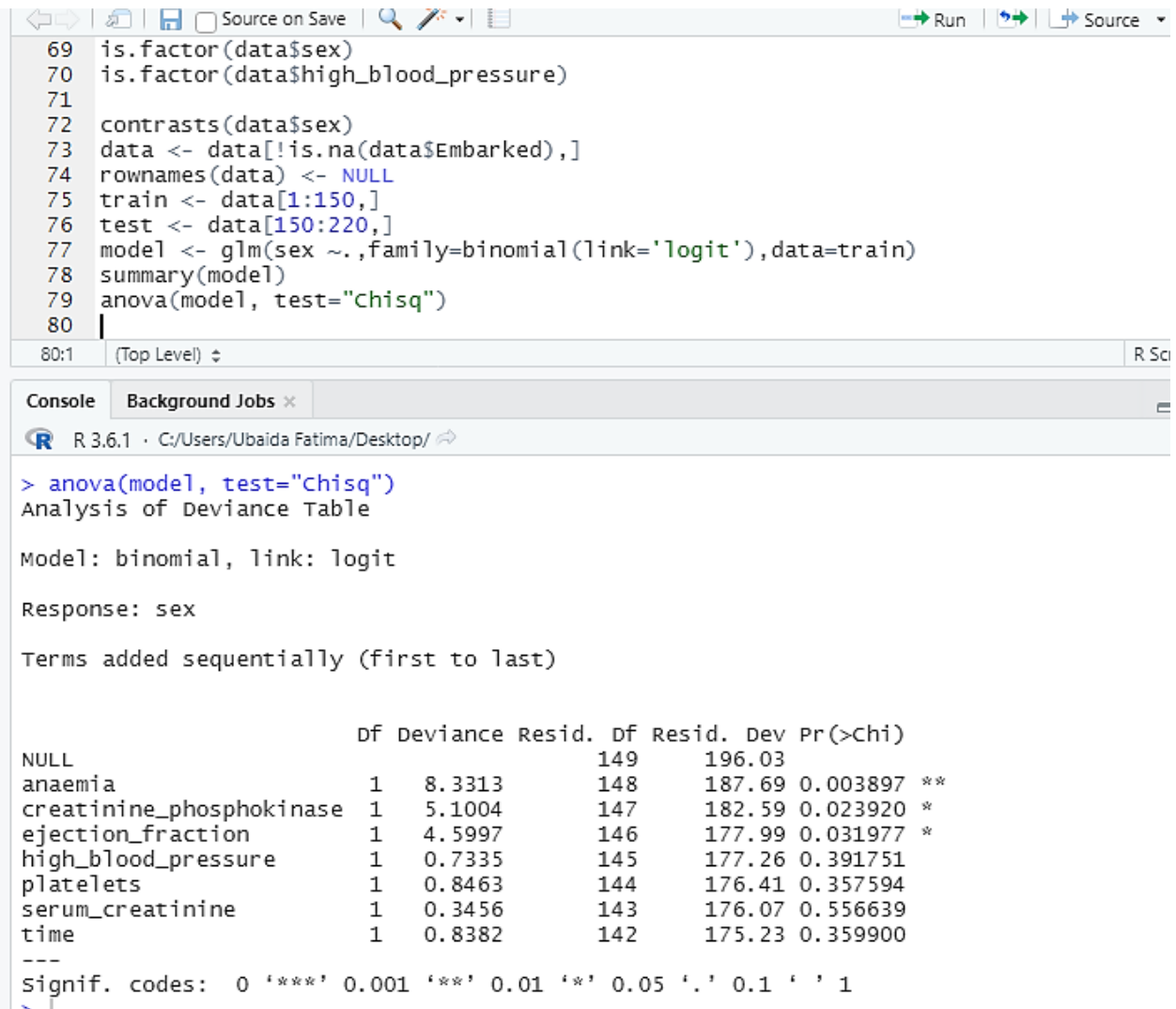
Null deviance: 196.03 on 149 degrees of freedom
 Residual deviance: 175.23 on 142 degrees of freedom
 AIC: 191.23

Number of Fisher Scoring iterations: 6

> |

Figure 32: Logistic Regression Model Fitting on Dataset

Interpreting the outcomes of dataset through logistic regression model:



```

69 is.factor(data$sex)
70 is.factor(data$high_blood_pressure)
71
72 contrasts(data$sex)
73 data <- data[!is.na(data$Embarked),]
74 rownames(data) <- NULL
75 train <- data[1:150,]
76 test <- data[150:220,]
77 model <- glm(sex ~.,family=binomial(link='logit'),data=train)
78 summary(model)
79 anova(model, test="Chisq")
80
80:1 (Top Level)
R Sci

```

Console Background Jobs x

R 3.6.1 · C:/Users/Ubaida Fatima/Desktop/

```

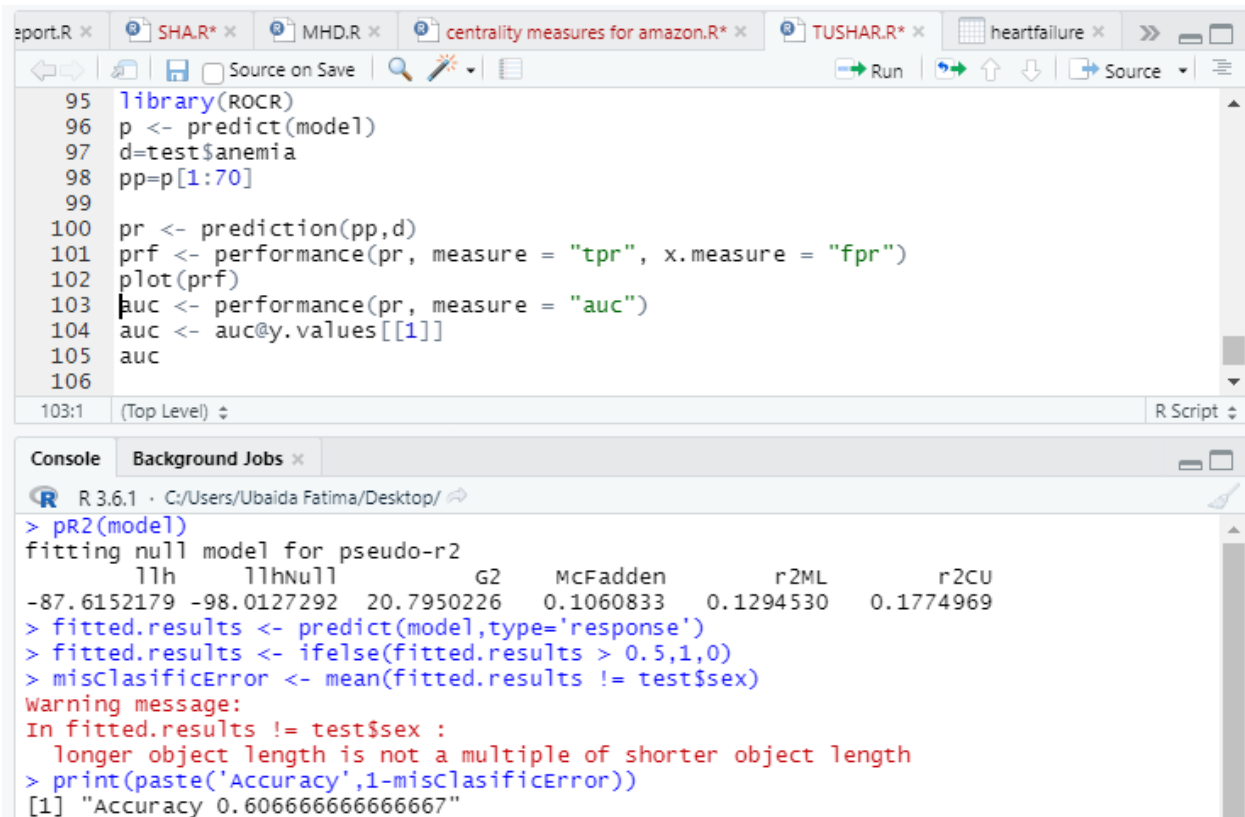
> anova(model, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit
Response: sex
Terms added sequentially (first to last)

              Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                      149      196.03
anaemia                   1    8.3313      148      187.69 0.003897 **
creatinine_phosphokinase  1    5.1004      147      182.59 0.023920 *
ejection_fraction         1    4.5997      146      177.99 0.031977 *
high_blood_pressure        1    0.7335      145      177.26 0.391751
platelets                  1    0.8463      144      176.41 0.357594
serum_creatinine           1    0.3456      143      176.07 0.556639
time                      1    0.8382      142      175.23 0.359900
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 33: ANOVA on dataset through R Software

Assessing the predictive ability of the model:


The screenshot displays the R Studio environment. The top pane shows a script with R code for model evaluation. The bottom pane shows the console output of the executed code.

R Script Code:

```

95 library(ROCR)
96 p <- predict(model)
97 d=test$anemia
98 pp=p[1:70]
99
100 pr <- prediction(pp,d)
101 prf <- performance(pr, measure = "tpr", x.measure = "fpr")
102 plot(prf)
103 auc <- performance(pr, measure = "auc")
104 auc <- auc@y.values[[1]]
105 auc
106

```

Console Output:

```

R 3.6.1 · C:/Users/Ubaida Fatima/Desktop/
> pr2(model)
fitting null model for pseudo-r2
      llh      llhNull      G2      McFadden      r2ML      r2CU
-87.6152179 -98.0127292  20.7950226  0.1060833  0.1294530  0.1774969
> fitted.results <- predict(model,type='response')
> fitted.results <- ifelse(fitted.results > 0.5,1,0)
> misClasificError <- mean(fitted.results != test$sex)
warning message:
In fitted.results != test$sex :
  longer object length is not a multiple of shorter object length
> print(paste('Accuracy',1-misClasificError))
[1] "Accuracy 0.6066666666666667"

```

Figure 34: Accuracy of Fitted Model

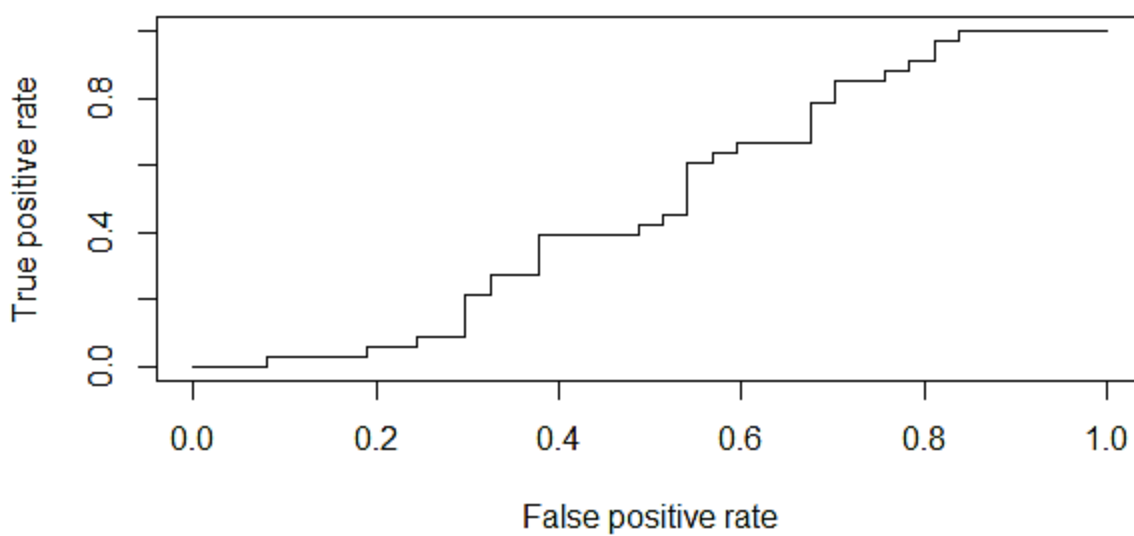


Figure 35: Prediction Plot for Anemia through Logistic Regression

```

86 library(ROCR)
87 p <- predict(model)
88 d=test$anemia
89 pp=p[1:70]
90
91 pr <- prediction(pp,d)
92 prf <- performance(pr, measure = "tpr", x.measure = "fpr")
93 plot(prf)
94 auc <- performance(pr, measure = "auc")
95 auc <- auc@y.values[[1]]
96 auc
97
97:1 (Top Level)
R Script

```

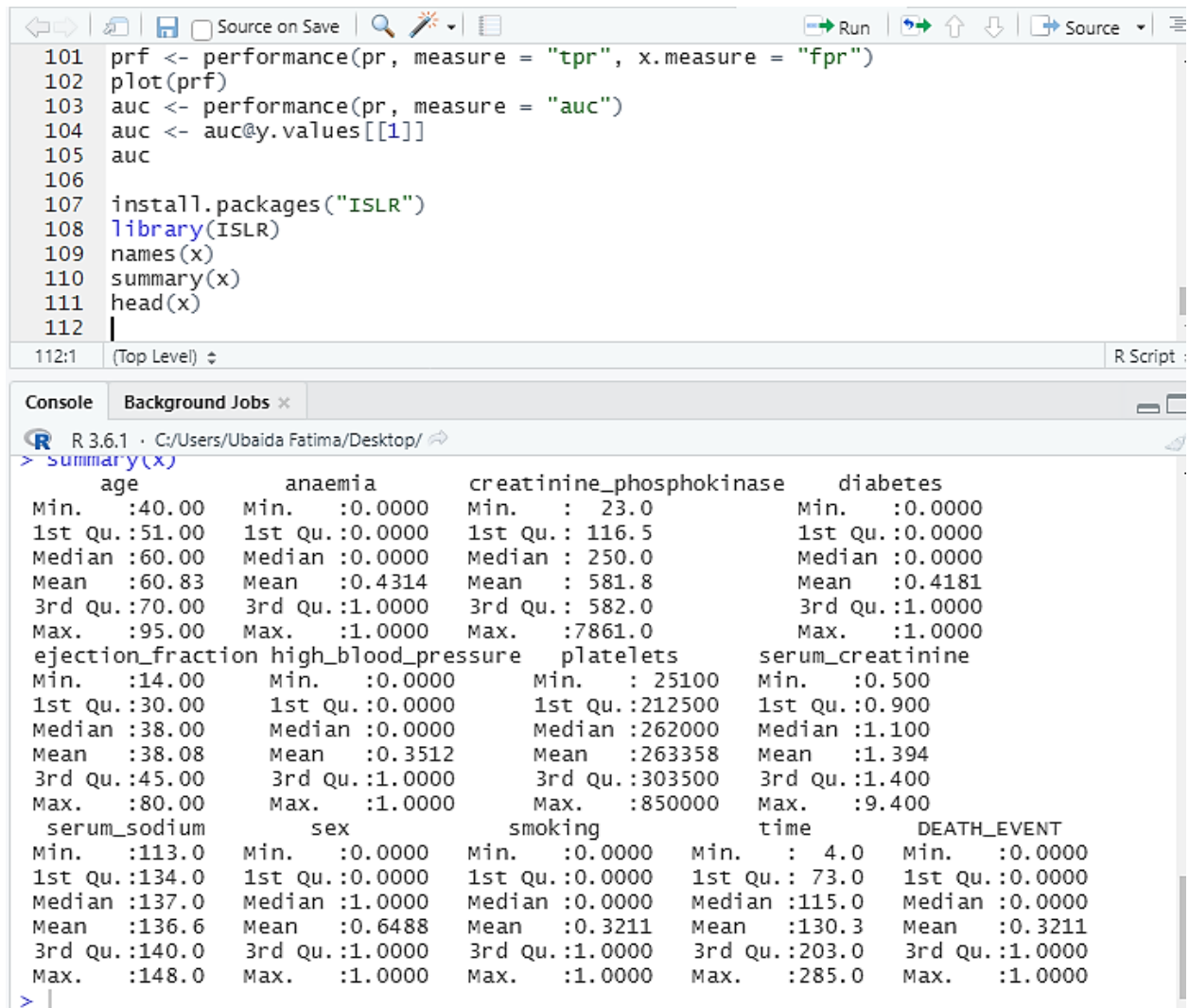
Console Background Jobs x

```

R 3.6.1 C:/Users/Ubaida Fatima/Desktop/
1.4361361102 1.4148638400 -0.0293049177 0.6087837372 0.6882226834 1.4317773971
31 32 33 34 35 36
1.1964510823 0.8368804232 -0.0149598305 0.4128064680 0.5724181651 1.7522493491
37 38 39 40 41 42
-0.1005316960 -0.0009706498 3.1821626929 1.2215543616 1.6678266399 1.1957574186
43 44 45 46 47 48
1.0512163409 0.5265140166 0.0506608149 1.4036152956 1.9965845893 0.7407610233
49 50 51 52 53 54
1.0657632641 0.1583967646 0.6663443166 0.0682395332 3.5528730326 -0.6909627228
55 56 57 58 59 60
0.3267879257 0.2508355451 0.3136532387 0.4324078895 1.5526315504 1.3438282650
61 62 63 64 65 66
6.2630973391 0.8871073821 1.0509715737 1.1675442662 0.1365346073 1.7508161231
67 68 69 70
0.8523373685 0.3582957045 1.3536561533 0.9438455575
> pr <- prediction(pp, d)
> prf <- performance(pr, measure = "tpr", x.measure = "fpr")
> plot(prf)
> auc <- performance(pr, measure = "auc")
> auc <- auc@y.values[[1]]
> auc
[1] 0.4897625
>

```

Figure 36: Prediction Model and Accuracy of Prediction Model on dataset through R Software



```

101 prf <- performance(pr, measure = "tpr", x.measure = "fpr")
102 plot(prf)
103 auc <- performance(pr, measure = "auc")
104 auc <- auc@y.values[[1]]
105 auc
106
107 install.packages("ISLR")
108 library(ISLR)
109 names(x)
110 summary(x)
111 head(x)
112
112:1 (Top Level)
R Script

```

```

> summary(x)
      age      anaemia  creatinine_phosphokinase  diabetes
Min.   :40.00   Min.   :0.0000   Min.    : 23.0         Min.   :0.0000
1st Qu.:51.00   1st Qu.:0.0000   1st Qu.: 116.5       1st Qu.:0.0000
Median :60.00   Median :0.0000   Median : 250.0       Median :0.0000
Mean   :60.83   Mean   :0.4314   Mean   : 581.8       Mean   :0.4181
3rd Qu.:70.00   3rd Qu.:1.0000   3rd Qu.: 582.0       3rd Qu.:1.0000
Max.   :95.00   Max.   :1.0000   Max.   :7861.0       Max.   :1.0000
ejection_fraction high_blood_pressure  platelets  serum_creatinine
Min.    :14.00   Min.    :0.0000   Min.    : 25100   Min.    :0.500
1st Qu. :30.00   1st Qu. :0.0000   1st Qu. :212500   1st Qu. :0.900
Median  :38.00   Median  :0.0000   Median  :262000   Median  :1.100
Mean    :38.08   Mean    :0.3512   Mean    :263358   Mean    :1.394
3rd Qu. :45.00   3rd Qu. :1.0000   3rd Qu. :303500   3rd Qu. :1.400
Max.    :80.00   Max.    :1.0000   Max.    :850000   Max.    :9.400
serum_sodium      sex      smoking      time  DEATH_EVENT
Min.    :113.0   Min.    :0.0000   Min.    :0.0000   Min.    : 4.0   Min.    :0.0000
1st Qu. :134.0   1st Qu. :0.0000   1st Qu. :0.0000   1st Qu. : 73.0   1st Qu. :0.0000
Median  :137.0   Median  :1.0000   Median  :0.0000   Median  :115.0   Median  :0.0000
Mean    :136.6   Mean    :0.6488   Mean    :0.3211   Mean    :130.3   Mean    :0.3211
3rd Qu. :140.0   3rd Qu. :1.0000   3rd Qu. :1.0000   3rd Qu. :203.0   3rd Qu. :1.0000
Max.    :148.0   Max.    :1.0000   Max.    :1.0000   Max.    :285.0   Max.    :1.0000
>

```

Figure 37: Summary of Heart Monitoring Dataset

Calculating the correlation between each pair of numeric variables. These pair-wise correlations can be plotted in a correlation matrix plot to give an idea of which variables change together.

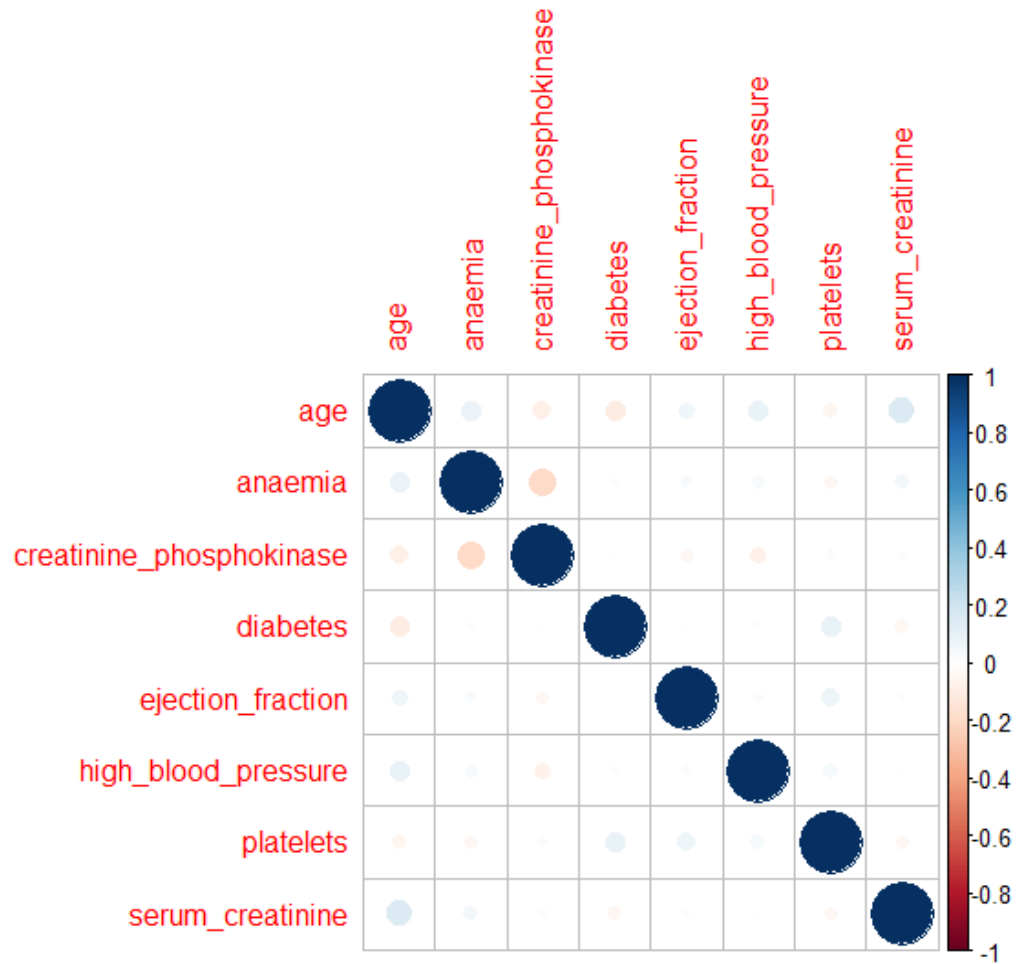


Figure 38: Correlation Plot for Dataset

A dot-representation was used where blue represents positive correlation and red negative. The larger the dot the larger the correlation. You can see that the matrix is symmetrical and that the diagonal are perfectly positively correlated because it shows the correlation of each variable with itself. Unfortunately, none of the variables are correlated with one another.

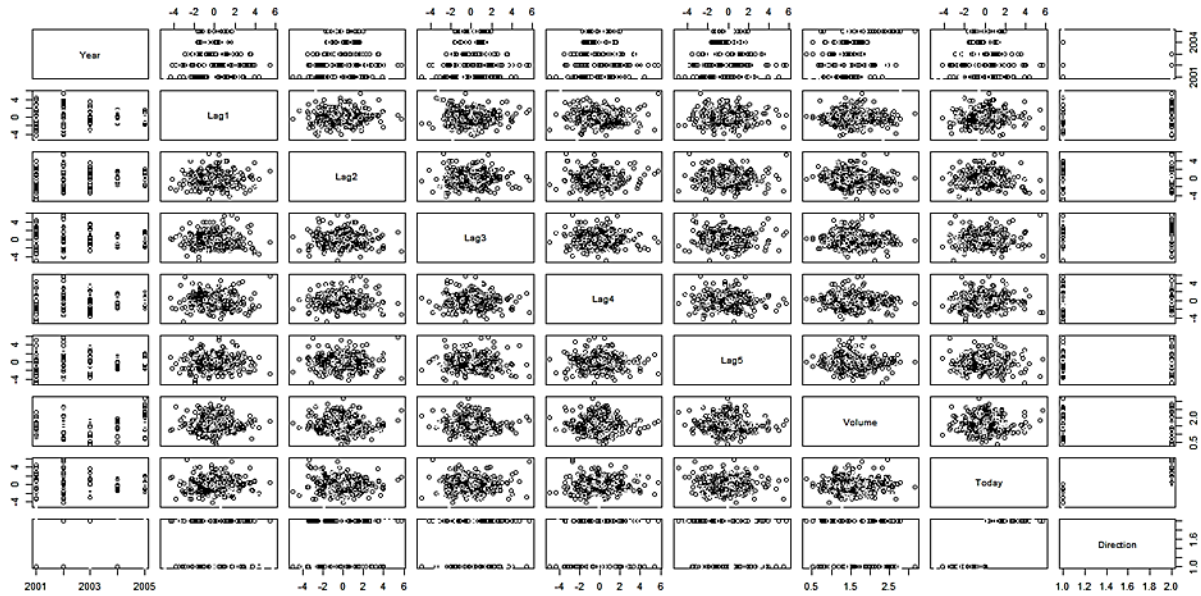


Figure 39: Density Distribution for every factor

The density distribution of each variable broken down by Sex factor in the data. Like the scatterplot matrix above, the density plot by Sex factor can help see the separation of Up and Down.

Support Vector Machine on Heart Monitoring Dataset


```

133 anyNA(x)
134 summary(x)
135 training[["anaemia"]] = factor(training[["anaemia"]])
136 trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
137 install.packages("kernlab")
138 library(kernlab)
139 svmm <- train(anaemia ~., data = training, method = "svmLinear",
140               trControl=trctrl,
141               preProcess = c("center", "scale"),
142               tuneLength = 10)
143 svmm
144 |

```

144:1 (Top Level) ↕ R Scri

Console Background Jobs ×

R 3.6.1 · C:/Users/Ubaida Fatima/Desktop/ ↗
atpna

```

> svmm <- train(anaemia ~., data = training, method = "svmLinear",
+               trControl=trctrl,
+               preProcess = c("center", "scale"),
+               tuneLength = 10)
> svmm
Support Vector Machines with Linear Kernel

210 samples
 7 predictor
 2 classes: '0', '1'

Pre-processing: centered (7), scaled (7)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 189, 190, 189, 189, 188, 189, ...
Resampling results:

    Accuracy   Kappa
 0.5058081 -0.044373

tuning parameter 'C' was held constant at a value of 1
> |

```

Figure 40: Support Vector Machine (SVM) results on R Software

```

138 library(kernlab)
139 svmm <- train(anaemia ~., data = training, method = "svmLinear",
140               trControl=trctrl,
141               preProcess = c("center", "scale"),
142               tuneLength = 10)
143 svmm
144
145 test_prediction <- predict(svmm, newdata = testing)
146 test_prediction
147
148 confusionMatrix(table(test_prediction, testing$anaemia))
149

```

149:1 (Top Level) R Script

Console Background Jobs

```

R 3.6.1 - C:/Users/Ubaida Fatima/Desktop/
[81] 0 1 0 0 0 0 0 0 1
Levels: 0 1
> confusionMatrix(table(test_prediction, testing$anaemia))
Confusion Matrix and Statistics

test_prediction  0  1
                0 37 27
                1 15 10

      Accuracy : 0.5281
      95% CI   : (0.4194, 0.6349)
No Information Rate : 0.5843
P-Value [Acc > NIR] : 0.88120

      kappa : -0.0191

McNemar's Test P-value : 0.08963

      Sensitivity : 0.7115
      Specificity : 0.2703
Pos Pred value : 0.5781
Neg Pred value : 0.4000

```

Figure 41: Test Prediction and Confusion Matrix Statistics

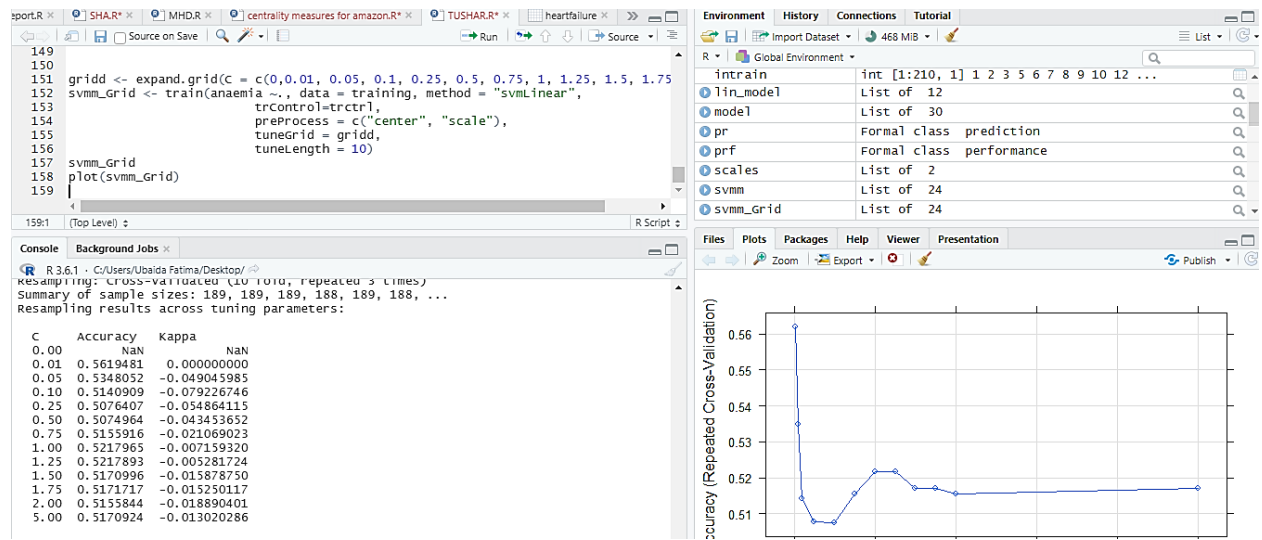


Figure 42: Support Vector Machine (SVM) Accuracy for dataset

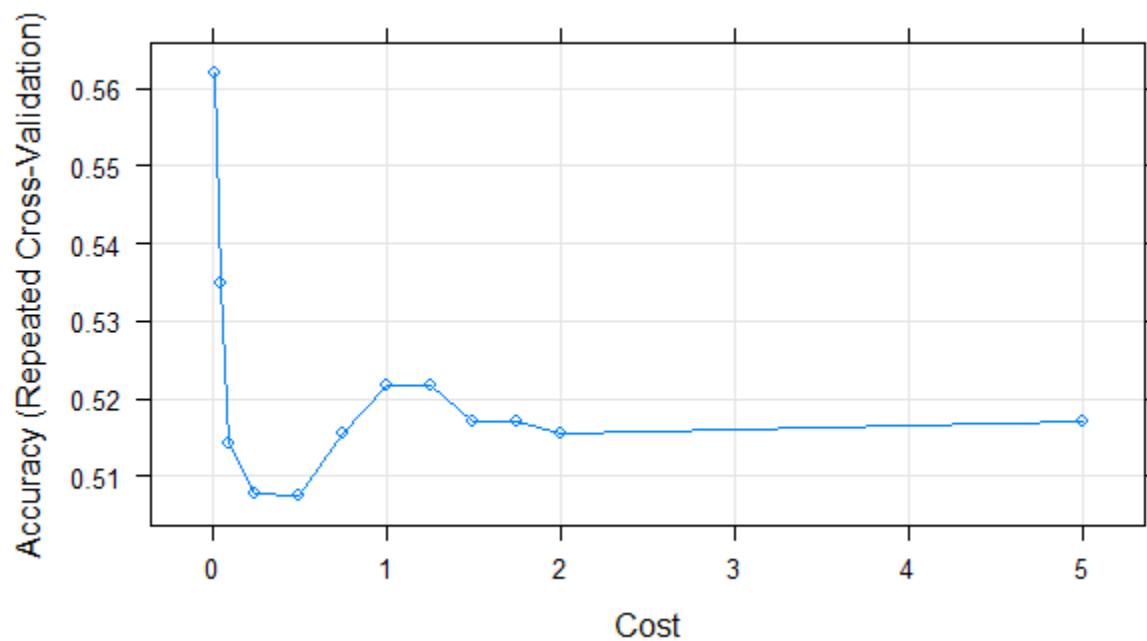


Figure 43: Support Vector Machine (SVM) plot for dataset

```

155 tuneGrid = grid(),
156 tuneLength = 10)
157 svmGrid
158 plot(svmGrid)
159
160 test_prediction_grid <- predict(svmGrid, newdata = testing)
161 test_prediction_grid
162
163
164 confusionMatrix(table(test_prediction_grid, testing$anaemia))
165 |

```

165:1 (Top Level) R Script

Console Background Jobs

R 3.6.1 · C:/Users/Ubaida Fatima/Desktop/

```

Accuracy : 0.5843
95% CI : (0.4749, 0.6879)
No Information Rate : 0.5843
P-value [Acc > NIR] : 0.5452

Kappa : 0

McNemar's Test P-Value : 3.252e-09

Sensitivity : 1.0000
Specificity : 0.0000
Pos Pred Value : 0.5843
Neg Pred Value : NaN
Prevalence : 0.5843
Detection Rate : 0.5843
Detection Prevalence : 1.0000
Balanced Accuracy : 0.5000

'Positive' Class : 0
> |

```

Figure 44: Confusion Matrix accuracy statistics for test prediction model

Code on R software:

```

install.packages("bnlearn")

install.packages("snow")

library('snow')

library(bnlearn)

library(readxl)

library(rlang)

```

```
x=heartfailure #Import data set heart failure
```

```
class(x)
```

```
colnames(x)
```

```
xx= data.frame(x)
```

```
sd(x$anaemia)
```

```
sd(x$diabetes)
```

```
mean(x$anaemia)
```

```
mean(x$diabetes)
```

```
median(x$anaemia)
```

```
median(x$diabetes)
```

```
hist(x$anaemia)
```

```
hist(x$diabetes)
```

```
hist(x$age)
```

```
barplot(x$anaemia)
```

```
boxplot(x$diabetes)
```

```
X=x$age
```

```
Y=x$anaemia
```

```
lin_model <- lm(Y~X)
```

```
coef(lin_model)

c <- 0.208143546

m <- 0.003670562

y_cap <- c + (m*X)

length(y_cap)


plot(Y~X,

     main='Regression for Age and Anemia',

     xlab='Age',ylab='Anemia', col=3)


lines(X,y_cap , col = 5 , lwd = 3)

Z=x$diabetes

lin_model <- lm(Z~X)

coef(lin_model)

c <- 0.673300159

m <- -0.004195687

y_cap <- c + (m*X)

length(y_cap)


plot(Z~X,

     main='Regression for Age and Diabetes',

     xlab='Age',ylab='Diabetes', col=3)


lines(X,y_cap , col = 5 , lwd = 3)
```

```
install.packages("Amelia")

library(Amelia)

missmap(x, main = "Missing values vs observed")

data <- subset(x,select=c(2,3,5,6,7,8,10,12))

zz=data$anaemia

zz[is.na(zz)] <- mean(zz,na.rm=T)


is.factor(data$sex)

is.factor(data$high_blood_pressure)


contrasts(data$sex)

data <- data[!is.na(data$Embarked),]

rownames(data) <- NULL

train <- data[1:150,]

test <- data[150:220,]

model <- glm(sex ~.,family=binomial(link='logit'),data=train)

summary(model)

anova(model, test="Chisq")

t= data$anaemia

install.packages("pscl")

library(pscl)

pR2(model)

fitted.results <- predict(model,type='response')

fitted.results <- ifelse(fitted.results > 0.5,1,0)
```

```
misClasificError <- mean(fitted.results != test$sex)

print(paste('Accuracy',1-misClasificError))

install.packages(ROCR)

library(ROCR)

p <- predict(model)

d=test$anemia

pp=p[1:70]

pr <- prediction(pp,d)

prf <- performance(pr, measure = "tpr", x.measure = "fpr")

plot(prf)

auc <- performance(pr, measure = "auc")

auc <- auc@y.values[[1]]

auc

install.packages("ISLR")

library(ISLR)

names(x)

summary(x)

head(x)

install.packages("corrplot")

library(corrplot)

correlations <- cor(x[,1:8])

corrplot(correlations, method="circle")

pairs(x, col=x$sex)

install.packages("caret")

library(caret)
```



```
x <- x[,1:8]

y <- x[,9]

scales <- list(x=list(relation="free"), y=list(relation="free"))

featurePlot(x=x, y=y, plot="density", scales=scales)

str(x)

head(x)

intrain <- createDataPartition(y = x$anaemia, p= 0.7, list = FALSE)

training <- x[intrain,]

testing <- x[-intrain,]

dim(training);

dim(testing);

anyNA(x)

summary(x)

training[["anaemia"]] = factor(training[["anaemia"]])

trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)

install.packages("kernlab")

library(kernlab)

svmm <- train(anaemia ~., data = training, method = "svmLinear",

             trControl=trctrl,

             preProcess = c("center", "scale"),

             tuneLength = 10)

svmm

test_prediction <- predict(svmm, newdata = testing)

test_prediction
```

```
confusionMatrix(table(test_prediction, testing$anaemia))

gridd <- expand.grid(C = c(0,0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2,5))

svmm_Grid <- train(anaemia ~., data = training, method = "svmLinear",
                  trControl=trctrl,
                  preProcess = c("center", "scale"),
                  tuneGrid = gridd,
                  tuneLength = 10)

svmm_Grid

plot(svmm_Grid)

test_prediction_grid <- predict(svmm_Grid, newdata = testing)

test_prediction_grid

confusionMatrix(table(test_prediction_grid, testing$anaemia))
```