|  |
| --- |
| COVENTRY  UNIVERSITY |
| Faculty of Engineering, Environment and Computing |
| School of Computing, Electronics and Mathematics |
| Master of Science in Data Science and Computational Intelligence |
| 7151CEM  Computing Individual Research Project |
| A Comparative Study on Predicting Diabetes using Machine learning and Deep learning methods |
| Author: Rohith Chityala |
| SID: 10013258 |
| Supervisor: Dr. Rochelle Sassman |
| Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Data Science and Computational Intelligence |
| Academic Year: 2021/2022 |

Declaration of Originality

I declare that this project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for the purposes of plagiarism prevention and detection.

Statement of copyright

I acknowledge that the copyright of this project report, and any product developed as part of the project, belong to Coventry University. Support, including funding, is available to commercialise products and services developed by staff and students. Any revenue that is generated is split with the inventor/s of the product or service. For further information please see [www.coventry.ac.uk/ipr](http://www.coventry.ac.uk/ipr) or contact [ipr@coventry.ac.uk](mailto:ipr@coventry.ac.uk).

Statement of ethical engagement

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (https://ethics.coventry.ac.uk/) and that the application number is listed below (Note: Projects without an ethical application number will be rejected for marking)

Signed: Rohith Chityala Date: 19/04/2022

Please complete all fields.

|  |  |
| --- | --- |
| First Name: | Rohith |
| Last Name: | Chityala |
| Student ID number | 10013258 |
| Ethics Application Number | P133139 |
| 1st Supervisor Name | Dr. Rochelle Sassman |
| 2nd Supervisor Name | Dr. Michael Richards |

This form must be completed, scanned and included with your project submission to Turnitin. Failure to append these declarations may result in your project being rejected for marking.

# Abstract

Diabetes is a condition where a human body fails to generate enough insulin to reduce the glucose levels in blood. It has the capability to cause some serious long-term effects on human, if left untreated. Until now, no perfect cure or medicine has been found to treat diabetes. Research suggests that diabetes has a higher death rate among women compared to men. So, it is important to detect diabetes in its early stage and take necessary steps to avoid its serious effects. In this project, several machine learning and deep learning techniques are implemented to predict diabetes in women depending on their previous health conditions. The data used for this project is PIMA Indian diabetes dataset available on Kaggle that consists of 768 observations with 8 health factors. The target variable whose value is to be predicted is the Outcome variable that indicates if a person has diabetes or not. Firstly, preliminary data analysis has been carried out which includes exploring the data, visualising the data, and cleaning the data to implement the models. Cleaning the data includes filling missing values, treating outliers, and standardising the data. The data has been split into train and test sets in the ratio of 70:30. Machine learning models applied on the dataset are Decision tree, Logistic regression, Support vector machine, K Nearest neighbour, XGBoost, and Random Forest classifier. Different Deep learning algorithms used in this project are Artificial neural networks, Convolutional neural networks, and LSTM. Among the ML models used, Support vector machine has come out as the best performer with 79.5% test accuracy. With an accuracy score of 77.3%, Convolutional neural networks have shown the best performance among Deep learning algorithms. XGBoost and Random Forest algorithms being the ensemble models used, they are also used to analyse the importance of various features. Glucose has been found as the best feature followed by BMI and Insulin as the next important features.

Table of Contents

[Abstract 2](#_Toc101278666)

[Acknowledgements 7](#_Toc101278667)

[Chapter 1 Introduction 8](#_Toc101278668)

[1.1 Background to the Project 9](#_Toc101278669)

[1.2 Project Objectives 11](#_Toc101278670)

[1.2.1 Gather client and user requirements 11](#_Toc101278671)

[1.2.2 Analyse and model the requirements 11](#_Toc101278672)

[1.3 Overview of the Report 12](#_Toc101278673)

[Chapter 2 Literature Review 13](#_Toc101278674)

[2.1 Diabetes- Causes and Effects 13](#_Toc101278675)

[2.2 Applications of Machine Learning in Predicting Diseases: 14](#_Toc101278676)

[2.3 Predicting Diabetes using Deep Learning: 16](#_Toc101278677)

[Chapter 3 Methodology 19](#_Toc101278678)

[3.1 Exploratory Data Analysis: 20](#_Toc101278679)

[3.1.1 Tools Required for EDA 20](#_Toc101278680)

[3.1.2 Basic Statistical analysis 21](#_Toc101278681)

[3.1.3 Univariate Analysis 21](#_Toc101278682)

[3.1.4 Bivariate Analysis 21](#_Toc101278683)

[3.1.5 Correlation Analysis 24](#_Toc101278684)

[3.1.6 Missing or Null Values 24](#_Toc101278685)

[3.1.7 Outlier Analysis 25](#_Toc101278686)

[3.2 Machine Learning Algorithms: 25](#_Toc101278687)

[3.2.1 Decision Tree: 26](#_Toc101278688)

[3.2.2 Logistic Regression: 26](#_Toc101278689)

[3.2.3 Random Forest Classifier: 26](#_Toc101278690)

[3.2.4 Support Vector Machine: 27](#_Toc101278691)

[3.2.5 K Nearest Neighbour: 28](#_Toc101278692)

[3.2.6 XGBoost Algorithm: 28](#_Toc101278693)

[3.3 Deep Learning Algorithms: 29](#_Toc101278694)

[3.3.1 Artificial Neural Networks (ANN): 29](#_Toc101278695)

[3.3.2 Convolutional Neural Network (CNN): 30](#_Toc101278696)

[3.3.3 Long Short-Term Memory (LSTM): 31](#_Toc101278697)

[3.2 Evaluation Metrics: 31](#_Toc101278698)

[3.2.1 Confusion Matrix: 31](#_Toc101278699)

[3.2.2 Classification Report: 32](#_Toc101278700)

[3.2.3 AUC-ROC Score: 33](#_Toc101278701)

[Chapter 4 Requirements 35](#_Toc101278702)

[4.1 The Dataset: 35](#_Toc101278703)

[4.2 Tools and Libraries Required: 37](#_Toc101278704)

[4.2.1 Python: 37](#_Toc101278705)

[4.2.2 Jupyter Notebook: 37](#_Toc101278706)

[4.2.3 Pandas and NumPy: 37](#_Toc101278707)

[4.2.4 Seaborn and Matplotlib: 38](#_Toc101278708)

[4.2.5 Scikit-learn: 38](#_Toc101278709)

[4.2.6 TensorFlow and Keras: 38](#_Toc101278710)

[Chapter 5 Analysis 39](#_Toc101278711)

[5.1 Descriptive Statistics: 39](#_Toc101278712)

[5.2 Target Variable: 40](#_Toc101278713)

[5.2.1 Visualising Target variable: 40](#_Toc101278714)

[5.3 Univariate Analysis: 41](#_Toc101278715)

[5.4 Outlier Analysis: 44](#_Toc101278716)

[5.5 Pair Plot: 45](#_Toc101278717)

[5.6 Bi-variate analysis: 46](#_Toc101278718)

[5.7 Correlation analysis: 49](#_Toc101278719)

[5.8 Finding and Filling Missing values: 50](#_Toc101278720)

[Chapter 6 Implementation 51](#_Toc101278721)

[6.1 Train-Test Split: 51](#_Toc101278722)

[6.2 Model Building: 51](#_Toc101278723)

[6.2.1 Machine learning algorithms: 51](#_Toc101278724)

[6.2.2 Deep learning algorithms: 52](#_Toc101278725)

[Chapter 7 Experimental Results 54](#_Toc101278726)

[7.1 Confusion Matrix 54](#_Toc101278727)

[7.2 Classification Report: 58](#_Toc101278728)

[7.3 AUC-ROC Score & Curve 61](#_Toc101278729)

[7.4 Feature Importance: 65](#_Toc101278730)

[7.5 Comparison of models: 66](#_Toc101278731)

[Chapter 8 Project Management 67](#_Toc101278732)

[8.1 Project Schedule 67](#_Toc101278733)

[8.2 Quality Management 67](#_Toc101278734)

[8.3 Social, Legal, Ethical and Professional Considerations 68](#_Toc101278735)

[Chapter 9 Critical Appraisal 69](#_Toc101278736)

[Chapter 10 Conclusions 70](#_Toc101278737)

[10.1 Achievements 70](#_Toc101278738)

[10.2 Review of Objectives: 70](#_Toc101278739)

[10.3 Future Work 71](#_Toc101278740)

[Chapter 11 Student Reflections 72](#_Toc101278741)

[Bibliography and References 73](#_Toc101278742)

[Appendix A – Project Specification 76](#_Toc101278743)

[Appendix B – Project Presentation 1](#_Toc101278744)

[Appendix C – Certificate of Ethics Approval 1](#_Toc101278745)

**List of Figures**

[Figure 1 17](#_Toc101276832)

[Figure 2 20](#_Toc101276833)

[Figure 3 21](#_Toc101276834)

[Figure 4 25](#_Toc101276835)

[Figure 5 26](#_Toc101276836)

[Figure 6 38](#_Toc101276837)

[Figure 7 39](#_Toc101276838)

[Figure 8 39](#_Toc101276839)

[Figure 9 40](#_Toc101276840)

[Figure 10 40](#_Toc101276841)

[Figure 11 40](#_Toc101276842)

[Figure 12 41](#_Toc101276843)

[Figure 13 41](#_Toc101276844)

[Figure 14 41](#_Toc101276845)

[Figure 15 43](#_Toc101276846)

[Figure 16 44](#_Toc101276847)

[Figure 17 45](#_Toc101276848)

[Figure 18 45](#_Toc101276849)

[Figure 19 46](#_Toc101276850)

[Figure 20 46](#_Toc101276851)

[Figure 21 47](#_Toc101276852)

[Figure 22 47](#_Toc101276853)

[Figure 23 48](#_Toc101276854)

[Figure 24 50](#_Toc101276855)

[Figure 25 52](#_Toc101276856)

[Figure 26 52](#_Toc101276857)

[Figure 27 52](#_Toc101276858)

[Figure 28 54](#_Toc101276859)

[Figure 29 54](#_Toc101276860)

[Figure 30 55](#_Toc101276861)

[Figure 31 55](#_Toc101276862)

[Figure 32 56](#_Toc101276863)

[Figure 33 56](#_Toc101276864)

[Figure 34 56](#_Toc101276865)

[Figure 35 57](#_Toc101276866)

[Figure 36 57](#_Toc101276867)

[Figure 37 58](#_Toc101276868)

[Figure 38 58](#_Toc101276869)

[Figure 39 58](#_Toc101276870)

[Figure 40 59](#_Toc101276871)

[Figure 41 59](#_Toc101276872)

[Figure 42 60](#_Toc101276873)

[Figure 43 60](#_Toc101276874)

[Figure 44 60](#_Toc101276875)

[Figure 45 61](#_Toc101276876)

[Figure 46 61](#_Toc101276877)

[Figure 47 61](#_Toc101276878)

[Figure 48 62](#_Toc101276879)

[Figure 49 62](#_Toc101276880)

[Figure 50 63](#_Toc101276881)

[Figure 51 63](#_Toc101276882)

[Figure 52 64](#_Toc101276883)

[Figure 53 64](#_Toc101276884)

[Figure 54 65](#_Toc101276885)

[Figure 55 66](#_Toc101276886)

[Figure 56 66](#_Toc101276887)

[Figure 57 67](#_Toc101276888)

**List of Tables**

[Table 1 31](#_Toc101278087)

[Table 2 35](#_Toc101278088)

[Table 3 35](#_Toc101278089)

[Table 4 38](#_Toc101278090)

[Table 5 38](#_Toc101278091)

[Table 6 39](#_Toc101278092)

[Table 7 43](#_Toc101278093)

[Table 8 43](#_Toc101278094)

[Table 9 49](#_Toc101278095)

[Table 10 50](#_Toc101278096)

[Table 11 65](#_Toc101278097)

Acknowledgements

Firstly, I would like to express my deep gratitude to Dr. Rochelle Sassman, who is my supervisor. Although I got very short period to work under her guidance, I believe she supported me in the best way possible. I thank Dr. Rochelle Sassman for understanding my situation and mentoring me when I was struggling to meet the project deadline. I sincerely appreciate my previous supervisor Dr. Matthew Hocking for his best supervision throughout the project execution. I’m extremely thankful to my university’s CAW team in helping me how to structure and write a dissertation project professionally. I am very grateful to my parents and friends for taking care of my mental health and well-being right from starting of this project until the end.

# Chapter 1 Introduction

In recent years, development of various long-term diseases has become more due to the unhealthy lifestyle choices being made by humans. While many of these diseases have a mild effect on humans, some of them are becoming life threatening in long-term. One such disease which has a noticeable effect on human body is diabetes. As the saying goes, prevention is always better than cure, it is beneficial to detect diabetes in early stages of human’s life rather than curing it after diagnosis. A lot of research has been carried out and is in progress which is mainly focused on inventing ways to detect, prevent, and cure diabetes. With the availability of variety of data in the healthcare sector, researchers can perform multiple operations on the data to generate new insights. Technologies such as machine learning and data science have seen a remarkable growth in recent years. These techniques are being applied extensively in almost every sector to attain state-of-art results. By implementing these techniques on the data available in the health care sector, many diseases are being detected in advance by which the treatment can be done at the starting stages of such diseases. This project mainly focuses on predicting whether a person has diabetes or not by considering their health factors. In addition, this project also aims at developing, utilising, and comparing various machine learning and deep learning techniques by implementing them on the diabetes data. Not only machine learning algorithms are helping in predicting diabetes, but also several other diseases such as skin cancer, heart problems, brain tumours, etc. With the implementation of Deep Learning algorithms, it has become possible to analyse various image datasets. There are several health factors and conditions which play a vital role in developing diabetes in a person’s body. And, with the unhealthy lifestyle that is being practiced by humans, certain new factors are being added to the traditional list of factors that can cause diabetes. The data used for this project consists of different health factors of females who are aged above twenty-one. This project also aims at understanding how the chosen machine learning and deep learning models perform and how the results vary by varying the data. This report discusses various stages of how the project has been carried out. It gives a detailed explanation about the factors that motivated in doing this project. In addition, it demonstrates things such as how the research has been carried out in collecting the data required for the project, analysing the data, implementing the models, and the reasons that inspired in choosing specific machine learning models and deep learning models. In the end, the results obtained are thoroughly understood, and compared with works done by other researchers.

## 1.1 Background to the Project

According to the global statistics survey conducted by the International Diabetes Federation, it is estimated that one in eleven grown-ups have diabetes in which one in two grown-ups have undiagnosed diabetes. Not only in grown-ups, but also the number of diabetes cases in children is increasing drastically. Diabetes is one of the seriously affecting chronical illnesses that are present in today’s world. The death rate is quiet high in diabetes compared to many other long-term diseases. It is estimated by the International Diabetes Federation that there’s a death recorded in every six seconds due to diabetes. It also predicts that by the year 2040, one in eleven grown-ups in the world would be diagnosed with diabetes. Many countries have been spending a lot of amounts in funding research works related to diabetes. Getting treated for diabetes in low- and middle-income countries has become expensive to many people. People in countries such as Africa find it difficult to get diagnosed and treated with diabetes due to lack of awareness in health and low income. Previously, diabetes was only diagnosed in older people who tend to have less physical activity in their lives. But in recent years, researchers have been finding many uncommon causes that are leading to diagnosis of diabetes in young adults and children. Diabetes is known to affect a human’s body both directly and indirectly. Diabetes is known to be a stage where a human body cannot produce or struggles to produce required amounts of insulin in blood. This lack of insulin in blood results in build-up of the sugar levels, which is scientifically termed as glucose. The body needs to make sure that there is enough insulin in the body to break down the rising levels of glucose. There are different types of diabetes of which type 1, type 2, and gestational diabetes are mostly found these days. Type 1 diabetes can be found in people of all ages including children. This type of diabetes can cause serious health complications such as high blood pressure, risk of kidney infection, increased probability of heart stroke, skin infections, eye problems etc. While type 1 diabetes possess serious health issues, it can be controlled if required amount of insulin is taken on time. But as per health experts, it is always important to take required preventive measures beforehand rather than curing long lasting diseases like diabetes. Whereas type 2 diabetes is known to be the most common type of diabetes occurring in today’s world. It is mainly found in people who are aged 45 or above, who have a family history of diabetes, and who are obese. Like the type 1 diabetes, type 2 diabetes is also caused by the inability of the body to utilise available insulin that is used to reduce the excess glucose level in the blood. National Institute of Diabetes and Digestive and Kidney diseases states that type 2 diabetes is the most found type of diabetes in most of the cases. Despite having high chances of extreme health complications with type 2 diabetes, it can be controlled effectively if diagnosed in early stages. Some of the early signs to observe in terms of type 2 diabetes are increased frequency of urination, blurred vision, feeling thirsty and hungry more often than usual, sudden weight loss, itching infections on skin, slow recovery of wounds and cuts etc. The history of diabetes in family also called as genetic history plays crucial part in predicting how precisely a human’s body can be diagnosed with diabetes. To differentiate type 1 and type 2 diabetes, it can be concluded that type 1 diabetes is mainly detected in early stages of life or due to genetic disorder, whereas type 2 diabetes develops in later stages of life due to the unhealthy diet and lifestyle choices that are made by a person. Lack of physical exercise or activity, excess mental stress also results in type 2 diabetes. There is another type of diabetes condition known as prediabetes which is not usually considered as regular diabetes. It is a condition where the blood sugar levels or glucose levels are considered high, but not as high to be called as type 2 diabetes. Most human beings possess this type of diabetes condition without even knowing they have it. Prediabetes is found in almost every adult and sometimes in children too, which can often trigger chances of causing type 2 diabetes if there is no change in lifestyle like diet and physical activity. Many researchers have found ways to prevent diabetes by following a strict diet and exercise routine in premature stages of life. The third most well-known type of diabetes is Gestational diabetes. Gestational diabetes is a kind of diabetes that is confined to females only, which is caused because of decreased ability of body to produce essential insulin required to control blood sugar level. It is developed in some of the females’ body during the third of fourth terms of pregnancy. Having diagnosed with gestational diabetes during pregnancy doesn’t necessarily mean that the person has diabetes for quite a long time before pregnancy or will have it continued after giving birth. However, if not taken proper precautions or not managed properly, there is a high probability of risk that gestational diabetes can advance into type 2 diabetes. Such type of diabetes is usually diagnosed by doctors beforehand by thorough monitoring of both mother’s and child’s health. It is important for both doctor and patient to concentrate on the symptoms that indicate diabetes during pregnancy. Women living in countries with low income, countries with poor health hygiene, and countries where women with lack of knowledge about health are prone to gestational diabetes. The accurate or exact cause of gestational diabetes is still not found, but it is believed and observed by the health experts that gestational diabetes occurs by excess generation of hormones which possess the capability of resisting body to produce glucose. While in most of the cases, these hormones tend to go away with pregnancy, but they are still present after pregnancy in some cases. Few health conditions in women that are responsible to develop gestational diabetes are being obese or overweight, history of high blood pressure, history of gestational diabetes in previous pregnancy times etc. Research suggests that race and ethnicity also matter in being affected by diabetes. Women from Asian, Hispanic, and African women are more prone to diabetes even if all the necessary precautions are taken. Although, the data shows that the percentage of men with diabetes is more than women, but it shows that the death rate in women with diabetes is more than men. Women tend to develop critical health complications faster than men, which often lead to life and death situations. This study is meant to analyse the impact of diabetes specifically on women and the causes that can make women vulnerable to diabetes. The data that is referred to in this project, consists of different health factors of 768 women who are at least twenty-one years old. The main aim of this study is to help researchers understand what health factors contribute most to one of the chronical illnesses called diabetes. It also provides researchers or analysts a new perspective to look at the problem, which can pave a path to further developments to solve the problem.

## 1.2 Project Objectives

The main aim of this project is divided into two stages or objectives that are discussed below.

### Gather client and user requirements

The primary objective of this project is to research about the causes of the disease diabetes and its effects on humans by using machine learning and deep learning technologies. To implement these technologies, a suitable dataset must be chosen. So, the initial aim of this project is to explore various datasets available and choose the one that serves the purpose of this project. The dataset chosen to implement further project objectives is a publicly available dataset on a platform called Kaggle. Before deciding on the dataset, necessary research has to be done in understanding the root cause of the problem and an effective study has to be done on the solutions proposed in past research.

### Analyse and model the requirements

The secondary objective of this project is to analyse the data gathered and implement proposed methodology on the data to solve the problem. The aim of this project is to focus on performing necessary data analysis in python using several statistical and visualisation techniques. After analysing the data, the next aim is to apply machine learning and deep learning algorithms on the data. In addition to this, the comparison of both these technologies is to be done based on their accuracy scores.

## 1.3 Overview of the Report

This report is structured in such a way that it provides a clear description of every aspect of the project. The way this report is organised is as described below.

Chapter one discusses about the introduction, project objectives, and background of the project. Chapter two consists of literature review which examines different studies or papers previously proposed by some researchers that motivated in implementing this project.

Next, chapter three gives a clear description of the methods that have been adopted to achieve the necessary results.

Chapter four focuses on the requirements of the project such as the data collection and the platform used to perform this project.

Chapter five and six deal with analysis and implementation of the project, which are mainly concentrated on the practical analysis and model building of the problem.

Chapter seven is more into testing the proposed methodology and it includes results generated. Chapter eight consists of details about project management that helps the reader understand different stages of the project implementation.

Chapter nine is the critical appraisal section which is an unbiased discussion explaining both positives and negatives of the project.

Finally, in chapter 10, the report is ended with the conclusion section that describes the achievements and future work of the proposed study.

# Chapter 2 Literature Review

Over the last few years, numerous studies and works have been published by various researchers that are concentrated on the causes of diabetes, its effects on human, preventive measures, and methods to cure it. As there is a lot of data being produced in the healthcare sector every day, it has become possible for researchers and scholars to understand more about how various diseases can affect a human’s body. With the help of technology, the collection and storage of healthcare data is being done successfully. In this section of the report, some of the studies conducted by researchers that motivated to do this project are discussed. This section also discusses about how the research were conducted and how different this project is from previous works done.

## 2.1 Diabetes- Causes and Effects

1. A study proposed by (American Diabetes Association, 2019) reviews and explains how diabetes has developed over the years into a life-threatening disease. This report published by the American Diabetes Association mainly focuses on explaining the classification, diagnosis, and types of therapeutic methods for different types of diabetes. It states that 90-95% of all kinds of diabetes cases are classified under type 2 diabetes. People with type 2 diabetes may not need to take insulin in their lifetime which means that it can be controlled by maintaining balanced diet and with proper exercise. In terms of confirming the diagnosis of diabetes, the American Diabetes Association recommends having two sample tests or two results from a single sample test. These tests are conducted to examine the plasma glucose level present in the blood. It suggests that by measuring the level of plasma glucose, it can make it easy to confirm that the symptoms are due to diabetes. This report published by the American Diabetes Association states that age, BMI, and ethnicity are some of the most important determining factors of presence of diabetes. The data suggests that the BMI level for people from Asian, American backgrounds should be set low compared to the BMI level of other ethnic groups because they are more prone to diabetes. The ADA has published in this report that women with gestational diabetes most often tend to develop type 2 diabetes than type1 diabetes, so it is important to test pregnant women for diabetes in their prenatal visit. This report on diabetes published by the American Diabetes Association has provided a reasonable information to understand and explore more about one of the chronic illnesses called diabetes.
2. Another paper published by (Apicella et al., 2020) discusses the impact of a recent pandemic known as covid-19 on people suffering with diabetes. It analyses how the effects of covid-19 had worsen in people with diabetes when compared to people without diabetes. Initially, it was observed that people with type 2 diabetes had more health complications due to covid-19 than those with type 1 diabetes. But, in later examinations, people with any type of diabetes were at high risk of covid-19. This research also found that the probability of occurring death is more in people with diabetes who are affected by covid-19. Diabetes when combined with other health conditions such as hypertension, cardiovascular diseases etc, can put a person in life-threatening situations. People living in low-income countries with lack of education are known to have undiagnosed diabetes. This state of undiagnosed diabetes would put a person in danger when they are affected with any other viral infections such as covid-19. To conclude, this report suggest that it is essential to discover ways to diagnose and treat diabetes in its initial stages. This paper has been taken into reference for this project because it discusses required information that helps in interpreting the importance and purpose of this project.

## 2.2 Applications of Machine Learning in Predicting Diseases:

1. According to (Ibrahim & Abdulazeez, 2021), machine learning techniques have become some of the most useful and impactful technologies that are being used in the field of medicine and healthcare. This report illustrates how machine learning has changed the approach of treating diseases by healthcare professionals in recent years. Especially, in terms of predicting and diagnosing few of the critical diseases such as cancer, heart diseases, brain tumours, depression etc, machine learning and deep learning techniques are known to produce state-of-art results. Additionally, this paper analyses how the data generation methods are emerging in healthcare industry every day. Not only this report gives an overview of implementing various machine learning techniques in detecting diseases, but it also helps in understanding how the accuracies of these techniques vary by varying the size of the data. In total, this report has referred to nineteen papers published by various scholars and researchers, between 2018 and 2020, which deal with applications of machine learning algorithms on different types of standard datasets. In addition to this, a clear explanation of the datasets used, machine learning algorithms implemented, and how their accuracies are used to determine the best algorithms is given. This report concluded that for larger datasets, to achieve greater accuracy, advanced techniques such as Convolutional Neural Networks should be used. The researchers in this report also found that to gain better accuracy, it could be advantageous to use ensemble models (combining multiple models or datasets) rather than using general models. This paper has motivated the current project in comprehending how different machine learning and deep learning models can be used is diagnosing diseases like diabetes and the factors that need to be changed to gain better accuracy.
2. In (Mujumdar & Vaidehi, 2019), authors proposed a study which is focused on predicting diabetes using machine learning methods. This study has used two different datasets which are PIMA Indian diabetes dataset and diabetes dataset. Also, this study discusses how the accuracies vary when different machine learning algorithms are implemented on both the datasets. Initially, the data is cleaned, and K Means Clustering has been performed on the data, which resulted in different clusters and machine learning models are fitted on to the data. Different machine learning techniques used in this report are Logistic Regression, AdaBoost classifier, Linear Discriminant Analysis, Random Forest, Gradient Boost Classifier, Extra Trees Classifier, Gaussian NB, Support Vector Machine, Perceptron, and K nearest neighbour. The evaluation metrics used in this study to compare a model’s performance are classification accuracy, F1 Score, Recall, Confusion Matrix, and Precision. Among all the supervised and unsupervised algorithms used, Logistic Regression has shown the highest accuracy of 96%. This accuracy was achieved by using the Diabetes dataset, whereas in the case of PIMA Indian diabetes dataset, the accuracy of Logistic Regression was found to be 76%. By this, the authors concluded that by using different datasets, a model’s accuracy could be improved. Furthermore, this study has performed pipelining of machine learning algorithms which had highest accuracies before. After pipelining, AdaBoost Classifier model showed the highest accuracy score of 98.8%. With the application of pipelining techniques, this study has proved that the accuracy score can be boosted. This study has been referred for the current project to understand how the performance scores of different models can be varied using multiple datasets and by implementing further techniques such as pipelining. The current study is focused on using the same dataset but understanding how the accuracies can be improved by implementing Deep Learning techniques instead of performing pipelining on the data.
3. A paper published by (Kopitar et al., 2020) is aimed at predicting the plasma glucose level in blood, and the presence of type 2 diabetes mellitus. It also focuses on finding out if there is any improvement in accuracy scores when machine learning models are used compared to traditional regression techniques. The machine learning models implemented in this study are Glmnet, Light Gradient Boosting Machine, XGBOOST model, Linear Regression, and Random Forest model. This study suggests that performance of a machine learning model in practical, may or may not be the same when new data gets added. So, it is important to consider measuring and understanding the stability of selected variables with addition of new data. This helps in choosing the best model for prediction. The evaluation of all the machine learning models used in this study is performed by various metrics such as Area Under the Curve (AUC), Area Under the Precision-Recall curve (AUPRC), and Root Mean Square Error value. The methods used for calculation of stability of the variables or variable importance are different for different machine learning models used. Some of the variables that had more variable importance are hyperglycaemia, level of cholesterol, and patient’s age. This study concludes that with the availability of new data, the model’s performance could be improvised, and the variable importance of different features can be monitored.

## 2.3 Predicting Diabetes using Deep Learning:

1. Whereas (Diab et al., 2020) published a paper to predict and classify diabetes by implementing Artificial Neural Networks algorithms. The dataset used in this study is the most popular PIMA Indian diabetes dataset. The techniques used for pre-processing of the dataset in this study are eliminating the outliers and filling the missing values. After pre-processing of the dataset, the three artificial neural network algorithms were fitted on the training, testing, and validation sets. The three models utilised for the prediction and classification of diabetes are Pattern Network, Feed Forward Neural Network, and Cascade Forward Network. The difference between these three neural networks is the number of neurons present, several layers of neural networks present, and different training functions used. Of all the three models used, Cascade Forward Network has shown better results with highest accuracy of 91.1% compared to other models. Whereas, Feed Forward and Pattern Neural networks showed an accuracy of 87.5% and 86.7%. Other performance metrics used apart from accuracy in this study are sensitivity and specificity. This paper has laid a good foundation to understand how different Artificial Neural Network algorithms can be used to predict presence of diabetes. As per the authors, the accuracy scores of these algorithms can further be improved by performing the pre-processing stages of the dataset more effectively.
2. A report published by (Rahman et al., 2020) studies the application of some of the mostly used Artificial Neural Network algorithms, that are Convolutional Neural Networks (CNN), Traditional Long Short-Term Memory (LSTM), Conv-LSTM, and a combination of both Convolutional Neural Networks (CNN) and Long Short-Term Memory models (CNN-LSTM). The performances of these models were examined by fitting the PIMA Indian Diabetes dataset on these models. Apart from performing the prediction of diabetes, this study also used grid search algorithm to perform hyperparameter optimization on the data. This grid search algorithm resulted in identifying the optimal hyperparameters for each model and its accuracy. By doing this, it becomes possible to known which parameters play an important role in predicting the required outcomes. The performance metrics used to analyse the models applied are sensitivity, specificity, and accuracy scores. Of all the models used, Conv-LSTM model has shown highest accuracy of 91.38%. This accuracy score has been further improvised to 97.26% using cross-validation technique. This study not only demonstrates the application of various Artificial Neural Network algorithms to predict diabetes, but also proves that feature selection is an important factor to consider in gaining better accuracy scores.
3. Finally, a study published by (Refat et al., 2021) mainly focused on providing a comparison between machine learning and deep learning algorithms used in predicting diabetes. The data used in this study is a publicly available dataset on UCI machine learning repository website which consists of data related to 520 patients at a hospital in Bangladesh. This dataset contains 16 different features or attributes of which 300 are of positive results and 220 are of negative results. The authors in this study considered various health factors of a patient such as age, sex, sudden weight loss, polyuria, polyphagia, weakness, visual blurring, genital thrush, irritability, itching, partial paresis, delayed healing, alopecia, muscle stiffness, obesity, and class to predict presence of diabetes. The methodology used was to pre-process the dataset initially by converting nominal values to numeric values which means that yes is categorised and 1 and no is categorised as 0. They used both machine learning and deep learning algorithms in this study. The Machine Learning classifier algorithms used to predict diabetes are XGBoost Classifier, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, and K nearest neighbour. Similarly, three different deep learning algorithms were used which are, Artificial Neural Networks, Multilayer Perceptron, and Long Short-Term Memory. All the models used are compared based on their performances by calculating their accuracy, recall, F1-score, precision, and ROC curve. In this experiment, among all the models used, XGBoost classifier has shown state-of-the-art results with 100% accuracy followed by Random Forest and Decision Tree algorithms. According to the authors, the limitation of this study is the availability of small size of dataset. This study paves a path where a dataset with more values can be used, and the models’ performances can be evaluated more effectively.

# Chapter 3 Methodology

In this section of the report, the methodology that has been adopted to achieve desired goals is discussed. Along with providing a good theoretical knowledge and understanding of the models and methods used, this section also discusses about the reasons that motivated to choose the specific methodology. The below subsections in this section explore the steps or methods employed to produce required results in this project by giving a detailed explanation of why these methods have been chosen and implemented.

Generally, the process of solving a problem in machine learning is divided into three or more stages. The first stage consists of performing Exploratory data analysis (EDA) followed by implementing the Machine Learning models, and finally evaluating the performance of models using certain metrics. This helps data scientists and analysts to produce desired results or expected outputs in solving a problem. The figure 1 shown below represents a flow chart that illustrates the proposed methodology. The below mentioned sections give an in-detailed explanation of the stages that have been adopted to implement this project.

Figure 1

*Stages of Proposed Methodology*

Loading the Dataset

Performing Exploratory Data Analysis

Cleaning the Data

Model Building and Fitting on the Data

Deep Learning Models

Machine Learning Models

Model Evaluation

Comparing the Models based on Accuracy scores

## Exploratory Data Analysis:

Exploratory data analysis often termed as EDA is the initial stage when addressing a problem using data science and machine learning technology. To put it in simple terms, it is the process of analysing the problem by exploring the data available that is related to the problem. Nowadays, data is being generated from many sources and most of the time, it is raw data, which means it contains many unnecessary factors that does not contribute to the problem. In most of the cases, the data available today may not always be structured, pure, or in ready to use form. Hence, it is essential to transform the data into required form according to the application by understanding and filtering it. This is where the exploratory data analysis stage comes into picture. For any problem to be solved using machine learning techniques, it is necessary to perform exploratory data analysis to gain better and true results. There are many advantages of performing EDA on the data being used, starting from knowing what the data represents to be able to use according to our requirements. By performing exploratory data analysis on a dataset, it becomes possible to gather various form of information of a particular dataset such as the number of columns, rows present, the type of data present in each column etc. In addition to this, EDA allows data scientists and analysts to find out if there are any missing values in the data and how they can be removed or filled according to their requirement. The process of exploratory data analysis also includes generating insights from the data using visualisations, i.e., it allows analysts to demonstrate the data in pictorial form. Some hidden trends present in the data can be observed, understood, and utilised according to the problem requirements by performing EDA. Exploratory data analysis is not only designed to examine the trends of a specific variable, but also to interpret the relationship between different variables present in the dataset. The correlation analysis allows the user to know how the variables are related to each other and how important they are in deciding the outcome of the problem. Some of the steps performed in this project as a part of EDA and the reasons these methods have been chosen are discussed below.

### Tools Required for EDA

There are various programming and non-programming tools that provide required functions and libraries to perform exploratory data analysis. Over the recent years, these tools have been expanded and developed various forms as per the required applications. Some of the tools that are extensively used for implementing EDA are Python, R, Microsoft Excel, etc. R and Python are the two programming languages that provide various kits and libraries which allow users to apply EDA effectively on variety of datasets. In this project, Python language has been used to perform exploratory data analysis. Python allows users to utilise some of the most famous libraries such as pandas, numpy, matplotlib, and seaborn. With the help of matplotlib and seaborn, data scientists and analysts produce visualisations which interpret the data. Visualisations include building charts such as boxplots, bar charts, histograms, violin plots, pie charts, etc. Additionally, correlation analysis among different variables in the dataset is represented pictorially using a graph called heatmap.

### Basic Statistical analysis

Statistical concepts hold an important role in designing and applying machine learning algorithms. Some of the concepts that help in understanding and solving data related problems are calculation of mean, median, minimum value, maximum value, first quartile, and third quartile. These statistical values provide a good description of all the columns present in the dataset. This helps in understanding the hidden patterns and trends present in the data.

### Univariate Analysis

The name Univariate analysis suggests that it is the process of analysing a single variable. Analysis of a single variable allows user to identify the trends present in that specific variable. It gives analysts and data scientists an opportunity to deep dive into understanding the importance of a variable. Univariate analysis includes visualising the variable data using box plots, violin plots, histograms, etc.

### Bivariate Analysis

Bivariate analysis is a type of analysis that consists of analysing two variables and interpreting the relation between them. With the help of bivariate analysis, the correlation among different variables present in the dataset can be understood. This analysis can be done between any two kind of variables such as numerical vs numerical, categorical vs categorical, categorical vs numerical. A bivariate analysis is usually performed by generating scatter plots, correlation coefficients, and regression analysis. By calculating the correlation coefficients between variables, the dependence of one variable on other can be measured.

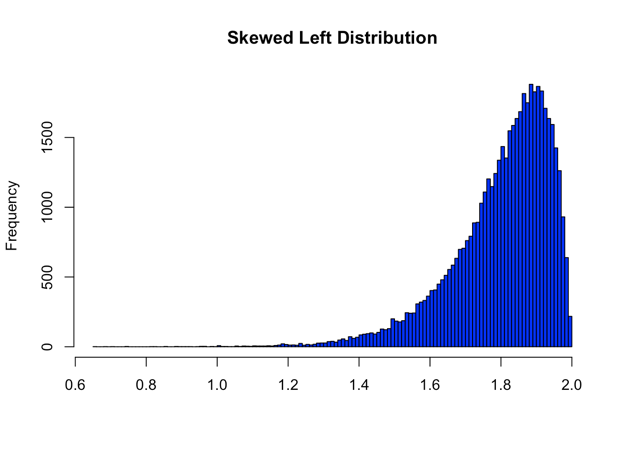
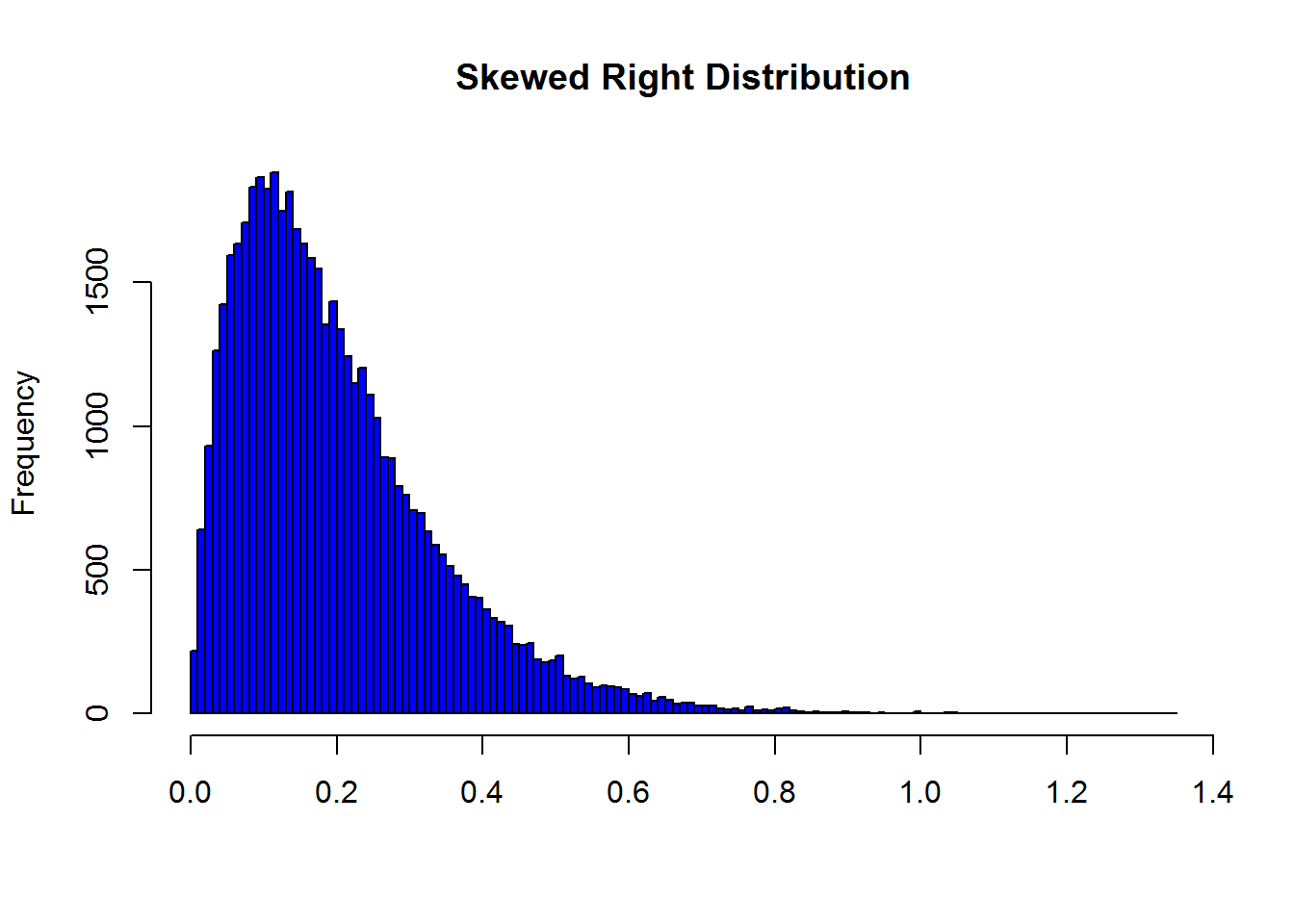
A brief description of the charts used in this project to perform exploratory data analysis are explained below.

* **Histogram:**

A histogram is a graphical representation of the distribution of a variable’s value in the dataset. It is basically a collection of bar graphs also called as bins. Histograms are widely used in most of the data representation applications. Histograms provide information about how skewed the data is i.e., if the data is left skewed, right skewed, or normally distributed. It also gives information about the presence of outliers and their distribution in the data. Histogram has a feature to choose the bin size when plotting the chart. It is essential to choose an ideal bin size when plotting a histogram because with the variance in bin size, the shape of the histogram changes which holds importance in representation of the data. Below shown figure 2 shows a graphical illustration of left, right and normally skewed histograms.

Figure 2

*Left, Right, and Normally Skewed Histograms*

Chart, histogram

Description automatically generated

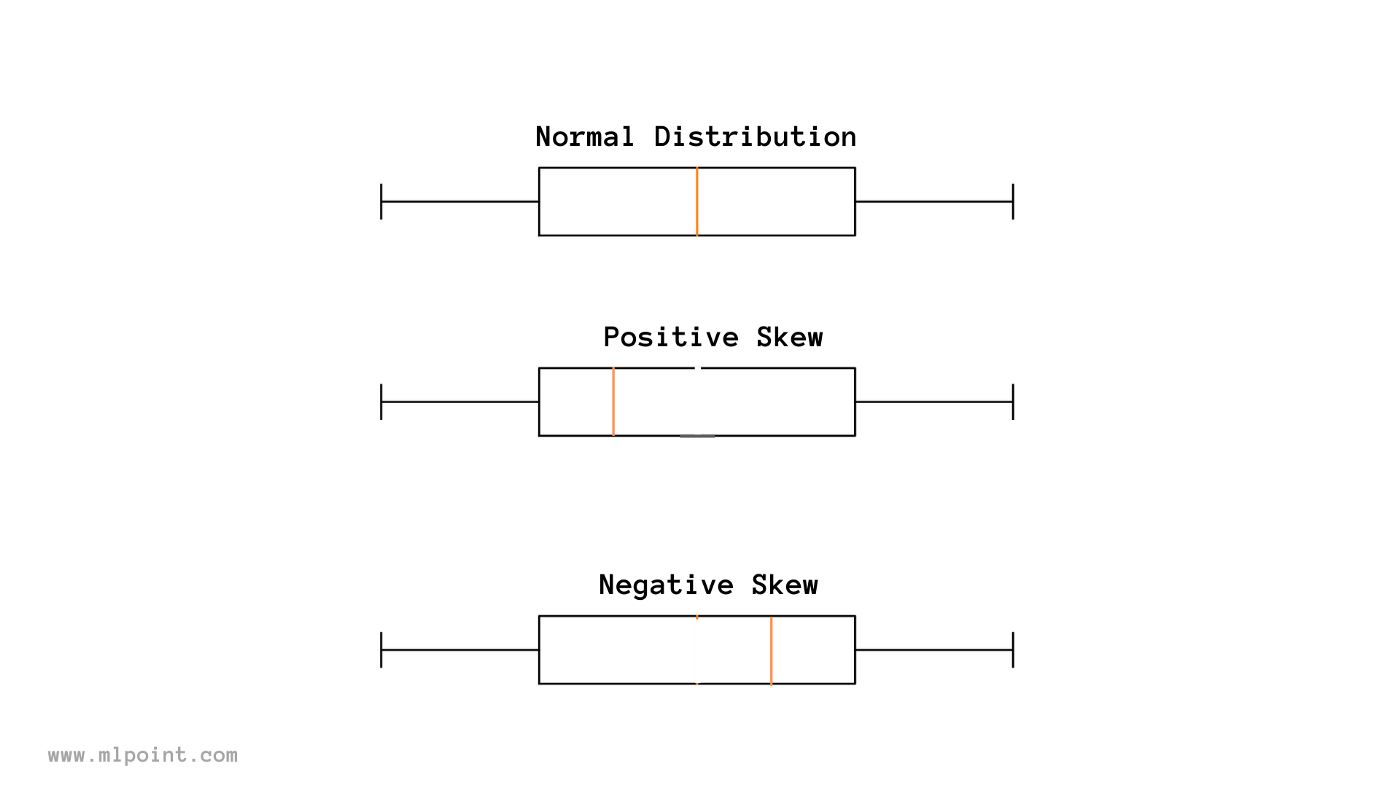
*Note: From Passion Driven Statistics* (<https://alanarnholt.github.io/PDS-Bookdown2/symmetric-distributions.html>)

* **Boxplot:**

As discussed in the above sections, statistics play a vital role in interpreting various kinds of data. One of the most used graphs that explains the required information about the data is a boxplot. A boxplot is an illustrative representation of the concept of five number summary in statistics. A five number summary is a method where five different factors of a variable could be defined. The five factors that define a variable’s data and its pattern in the dataset are minimum value, first quartile, median value, third quartile, and maximum value of the variable. These five factors broadly describe a variable in a dataset. Boxplots are designed to present a visual depiction of the theory of five number summary in a graphical format. Boxplots are also called as Whisker plots because the structure of a boxplot is more like a box that shows the interquartile range, and the minimum, maximum values are connected by lines called as whiskers. Like the histograms, the fundamental aim of the boxplots is to make user understand how the data is distributed. Boxplots not only provide a clear view of the skewness of the data, but also exhibit comprehensible outlook about the outliers present in the data. The below shown figure represents normally, positively, and negatively skewed boxplots.

Figure 3

*Normal, Positive, and Negatively skewed Boxplots*



*Note: From Boxplot: What does it tell you? by Medium* (<https://medium.com/mlpoint/box-plot-box-and-whiskers-plot-what-does-it-tell-you-99e827fac158>)

* **Pie Chart & Bar Graph:**

A pie chart is one of the basic visualisation charts that are used extensively in representing the data. The structure used in pie chart is basically a circle that is divided into different portions of areas called as pie. These areas represent the percentages of the data i.e., the size of the area is directly proportional to the percentages. More the percentage, more the area. Pie charts are used to demonstrate the quantity of a variable in a dataset.

Similar to a pie chart, bar graph is also a graphical illustration that is broadly used to represent the quantity of a variable in a data. But bar graph doesn’t use a circle to represent the data, instead it uses bars that describe the quantity of a variable. The higher the height of the bar, the higher the quantity. In this project pie chart and bar graphs are used to as a count plot that denotes the quantity or count of unique values present in a variable.

* **Pair Plot:**

Seaborn library in Python provides an excellent plotting option which makes it easy to perform univariate and bivariate plots with a single line of code usually known as Pair plot. A pair plot is basically a matrix of graphs that have been plotted between different variables and with the same variables. These graphs that represent the relationships among various combinations of the features can be histograms, KDE plots, bar graphs etc. In a pair plot, all the variables are considered on both x and y axes. The diagonal charts formed in a pair plot represent the univariate analysis as the same variable is taken on both x and y axes.

### Correlation Analysis

Correlation test or analysis is conducted on a dataset to learn the relation between any two numerical or categorical variables present in the dataset. Correlation between two variables defines the dependability of one variable on the other. By performing correlation analysis, the user can understand the trends of a variable and predict their future behaviour. With this, the importance of certain variables in predicting the outcome variables can be known. Correlation is generally calculated and measured using a feature called correlation coefficient. The correlation coefficient value ranges from -1 to 1, with value 0 being least correlated and value 1 being highly correlated. Based on the correlation coefficient value, the concept of correlation is divided into three types. Correlation coefficient value 0 denotes that no correlation exists between the two variables. Whereas values 1 and -1 indicate that the two variables are perfectly positive correlated and perfectly negative correlated. This correlation is often represented in the form a graph called as Heat map. A heat map basically represents the correlation coefficient value of all the variables by using a range of colours. The darker the colour, the more the variables are correlated and vice-versa. Seaborn library provides state-of-art tools to plot this heatmap with option to choose colours of our choice.

### Missing or Null Values

With the vast amount of data available, the probability of presence of missing values and outliers in a dataset is more. In many cases the data collected or gathered cannot be hundred percent accurate and complete. For example, in a data that has been gathered by conducting a questionnaire in a company, it is possible that some of the questions would not be answered by few employees based on their interests. In such cases, data engineers treat the unanswered questions i.e., the missing values are either replaced by zeroes or a term called Null value. The presence of these null values in a dataset can affect the data characteristics in different ways. So, when working on a dataset, it is essential to identify the missing values present in the data and clean the data, making it ready to use for machine learning models implementation. By following this procedure in the exploratory data analysis part, more effective results can be generated. There are different ways in treating the null values present in a dataset. Based on the type of data available, the process of filling null values is divided into three types- mean, median, mode. As discussed in the above sections, mean, median, and mode are some of the basic statistical concepts that play a vital role in solving a data-related problem. If the data available is in categorical form, then the missing values are filled by calculating mode of that column or variable. If the data is in numerical format, the null values can be treated by filling them with mean or median of that column. When the data consists of significant number of outliers, it is ideal to use the median value to fill the missing values. For data without outliers, both mean and median can be used to fill the null values. This way, the incomplete data could be made ready for applying the necessary models.

### Outlier Analysis

In general, outliers are the value that is considered to deviate from the normal distribution of the data. In every dataset, the way the data is distributed is examined to understand the trends it possesses. It is highly possible that a good number of outliers can be present in a dataset. Sometimes, these outliers can have a significant effect on the data distribution. So, it is important to identify these outliers and treat them accordingly. Various methods are available to identify the outliers present in a dataset. Outliers can be detected by few visualisation techniques such as boxplots, histograms, and violin plots. In few cases, scatter plots also provide a good knowledge of outliers present in a dataset. There are few statistical methods used to detect outliers such as the calculation of z-score and identifying the inter quartile range. After the outliers are detected, the dataset is treated by removing them, making it more effective for the implementation of machine learning algorithms.

## Machine Learning Algorithms:

In recent years, Machine learning and data science techniques have become some of the most reached out trends in technological applications. Machine Learning is a concept of artificial intelligence and computer science where a machine is designed to learn as a human by using the data and building specific algorithms. The main aim of machine learning is to improve the accuracy of its models by applying them on various kinds of data. Today, data is found everywhere around us, and it holds a lot of value if used appropriately. With variety of data being generated every day, more and more research is being conducted in developing machine learning techniques to solve many real-world problems. And these algorithms if implemented correctly, are producing state-of-the-art results. Machine learning is a combination of numerous concepts such as data analysis, statistics, probability, visualizations, etc. Over the past few years, data science and machine learning fields have seen a revolutionary growth in the way they are applied. Almost every sector today has applications of machine learning in solving different problems. Machine learning has been giving promising results to some of the complex issues being faced in the health sector. Many long-term diseases such as cancers and tumours are being detected and treated early with the help of machine learning algorithms. Basically, machine learning algorithms are divided into three types- Classification, Regression, and Clustering techniques. These techniques are further divided into many algorithms depending on the characteristics they possess. In this project, one of the most concerning long-term illnesses known as diabetes is predicted using machine learning techniques. The following sub-sections discuss various machine learning algorithms applied on the data to gain desired results in this project.

### Decision Tree:

Decision tree algorithm is known to be one of the most popular machine learning algorithms. It is a kind of supervised learning algorithm which uses previous data to make future decisions. Despite being a type of supervised algorithm, decision tree holds the ability to perform both classification and prediction problems. Depending on the type of target variable, decision tree learns to either classify or predict the output. The class of the dependent or the target variable is predicted or classified based on the independent variables available. For minute problems, decision tree can be used to visualise the decisions at every step. But, for real world data, it seems impossible to perform decision tree algorithm manually. So, machine learning provides a popular library called as scikit-learn which allows the implementation of several classification, regression, and clustering algorithms. To design a decision tree using scikit-learn library, there are few factors to be considered such as Criterion, Entropy, Gini gain, Gini index, Information gain, max\_depth, etc. These factors have been created to improve the accuracy results of decision tree. Different tree methods are available for application depending on the type of target variable, i.e., continuous, or categorical. The four types of decision tree methods are CART, CHAID, C4.5, and QUEST. The merits of using decision tree in this project are because of its ease of understanding and interpreting, its feature to simplify the relationship among dependent and independent variables. Decision tree algorithms are extensively used in the bank sector, health sector, customer analytics, and employee analytics.

### Logistic Regression:

Logistic Regression is one of the supervised learning algorithms in machine learning. Like Decision Tree algorithm, it works on the principle of classification technique. It uses the concept of probability to perform its operations by analysing the outcome when the target variable can be yes or no or in binary form (0&1). It has proved that Logistic Regression is one of the algorithms with most effective and accurate results. Unlike in Decision tree, the handling of missing values may not be possible in Logistic Regression.

### Random Forest Classifier:

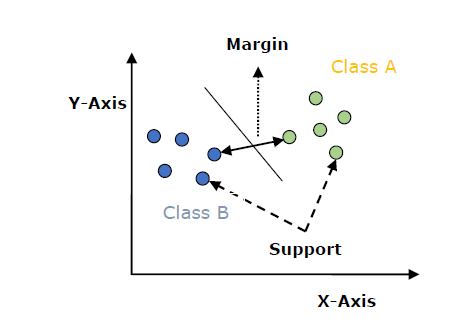
Random Forest is one of the widely used machine learning algorithms that has a proven history of generating productive results. It is a type of supervised learning algorithm that can be used as both classification and regression technique. Random Forest algorithm is a kind of ensemble models available which use multiple decision trees to form a single model. The name Random Forest suggests that it is a combination of random number of trees that forms into a forest. It outperforms the decision tree algorithm in things like over-fitting problem, and in improving accuracy. In Random Forest algorithm, the outcome is predicted by calculating the mean or average of all the decision trees used. For most of the machine learning algorithms, the accuracy could be increased by applying hyper parameter optimization technique, but Random Forest has the capability to produce best accuracy scores without using such technique. It is known for handling missing values better than decision tree algorithm.

### Support Vector Machine:

Support Vector Machine (SVM) is a most popular supervised machine learning algorithm. It is used for both classification and prediction problems. It can also be used as an unsupervised algorithm when the data is unlabelled. In case of unsupervised algorithm, support vector machine uses a clustering technique. Support Vector Machine uses a concept where the data is segregated into two different categories by drawing a line in between them. This line helps in deciding which category the new incoming data should be categorised as. Support vector machine is also applicable for image and text classification problems. For example, to identify whether a new image is a cat or dog, SVM can be trained on a dataset with many cat and dog images and by drawing a line between both these categories, the new image will be classified into category that shares more similarity.

Figure 4

*SVM dividing data into classes A & B using hyperplane*



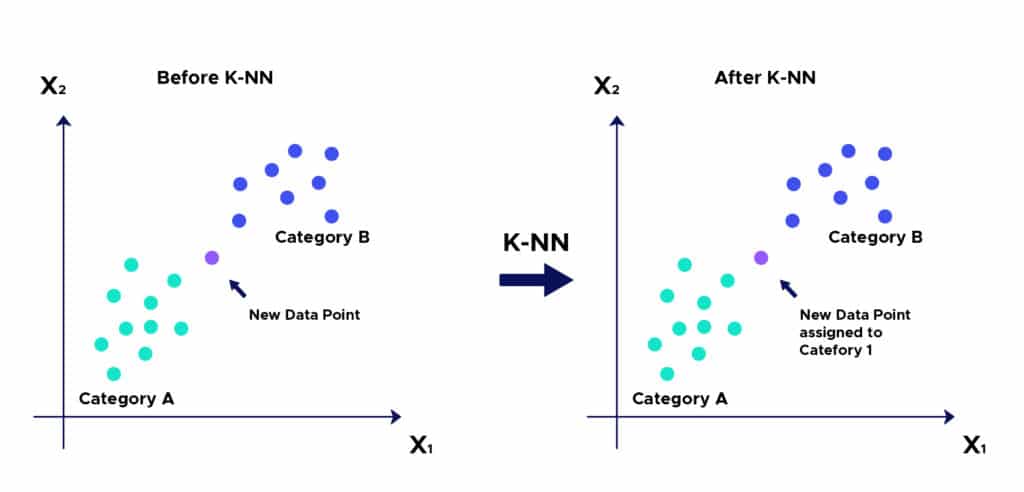
*Note: From Introduction to SVM by TutorialsPoint* (<https://www.tutorialspoint.com/machine_learning_with_python/classification_algorithms_support_vector_machine.htm>)

### K Nearest Neighbour:

K Nearest Neighbour usually referred as KNN is a type of supervised machine learning algorithm that is designed to solve both classification and regression problems. But in most of the cases, it is used as a classifier because of its features. It is known for its unique characteristic for being a non-parametric method in machine learning as it does not learn from the parameters in the data, but rather classifies the data based on the nearest point. It is also often referred to as a lazy learner among all the machine learning algorithms in view of the fact that it doesn’t train on the data when provided before. Instead, it stores the data when applied to it until any queries are performed on it. Assume the data is divided into two groups X and Y. Depending on the position of the new incoming data point, the distance is calculated to other nearby points. The point with the lowest Euclidian distance is selected as the target group (X or Y) of the new data point. The K in KNN denotes the number of points to be considered to calculate the distance. If K is 5, then distances between five new data points is calculated and the point with lowest distance is the chosen group or class. Choosing the K value can be experimental as there are no specific criteria followed in determining the K value. But it is recommended to choose the K value such that it’s not too low or too high. This can sometimes cause problems such as effect of outliers and over-fitting. The below shown figure 5 illustrates the working of KNN algorithm.

Figure 5

*Working of KNN algorithm*



*Note: From Introduction to KNN by Datascientist* (<https://datascientest.com/es/que-es-el-algoritmo-knn>)

### XGBoost Algorithm:

eXtreme Gradient Boost often indicated as XGBoost is one of the most widely used boosting methods in recent years. It is basically a development of Gradient Boosting algorithm and is designed to overcome the demerits of Gradient Boost algorithm. This algorithm has been showing promising and near-desired results in solving data problems with categorical, tabular, and structured data. Boosting methods or algorithms in machine learning are a way of generating new algorithms with multiple classification trees or algorithms to improvise the accuracy and models’ performance. It is known for being highly fast compared to other ensemble techniques. Both Random Forest and Boosting algorithms serve a similar purpose, but in different ways. In Random Forest, different decision trees are built parallelly and fitted onto different samples of the data. Whereas in boosting algorithms, different trees are combined sequentially by considering the output of previous tree. Additionally, XGBoost is designed in such a way that it consists various hyperparameters that can be changed according to the application. Although XGBoost is known for its way of handling missing values internally, but it is highly affected by presence of outliers.

## Deep Learning Algorithms:

Deep Learning is an advanced Machine learning concept that is designed in reference to the working of a human brain. It basically generates insights by trying to learn from the data like a human brain. The word deep suggests that the deep learning architecture consists of several layers. The input goes through these different layers and by performing some operations, output is generated. Every output of a layer is fed as an input to the next layers, this way the desired results will be produced. Deep learning has been showing state-of-the-art results in the fields of self-driven vehicles, automobiles, computer vision, image classification, natural language processing, text, and audio applications. Neural Networks is a concept of Deep Learning that is implemented on the phenomenon of working of human’s neural system. In this project, three most popular neural network models are used to achieve required output which are Artificial Neural Networks (ANN), Convolution Neural Network (CNN), and Long Short-Term Memory (LSTM). The below subsections discuss more in detail about these three models. Despite of their extensive applications in various fields, deep learning algorithms require a very high computational power to handle huge and various types of data.

### Artificial Neural Networks (ANN):

Artificial Neural Networks are developed to mimic the way a human brain thinks by using neural networks. It consists of multiple hidden layers through which the data is passed. At every stage of the layer, different feature extraction method takes place. For example, consider a face image is fed as input to ANN to recognise the gender of a person. So, in the initial layer, the neurons convert the image into a matrix of numbers, in the second layer, physical features like eyes, nose, mouth are extracted, and in the final layer, some more features are extracted to produce required output. The features extracted at every stage or layer are assigned certain weights that define the importance of that feature, by which the ANN generates outputs. The output of one layer is fed as an input to the next layers by the sum of all weights. The nodes or layers consist of an activation function which then processes the data and generates output. Some of the commonly used activation functions in Artificial neural networks are Identity function, Binary step function, Sigmoidal function, Bipolar step function, and Ramp function. Depending on the problem requirement, these functions are used. Some of limitations of ANN include consumption of high computational power, slow in operation, and no chance for the user to know what is happening inside the algorithm.

### Convolutional Neural Network (CNN):

Convolutional Neural Networks are one of the most sought out methods among Deep Learning models. CNN is a form of MLP that was first employed in image processing. They are designed to work effectively in feature extraction through the concept of backpropagation. It is basically a grid like structure consisting of different layers mainly convolution layer, pooling layer, and fully connected layer. The specified inputs are passed through a sequence of convolution layers with filters, pooling layer, fully connected (FC) layers, and finally the Softmax function to create the final prediction to train and test a deep learning CNN model. The convolution layer learns the relationship between the data to obtain the underlying data pattern.

Based on the relevance of each parameter, the pooling layer decreases the number of parameters. With the use of a Softmax function, the fully connected layer generates the likelihood function onto each class, which provides the ultimate prediction result. The input data was processed via a convolution layer, which learned multiple features to capture the data pattern. Then, using a dot product in between specified inputs and weight values, a feature map is created. To avoid vanishing gradient concerns, a bias data is assigned to each step.

The convolution layer employs a ReLU (Rectified Linear Unit) activation function, which brings non-linearity into our convolutional network. The corrected feature maps are then handed to the pooling layer, which down samples them. The max pooling operation, which extracts the heaviest features from the corrected feature maps, is used here. The down sampled feature maps are also processed through a combination of convolution and pooling layer, which serves the same purpose as the previous convolution and pooling layer. The feature map is then converted to a vector and supplied into a fully connected layer, similar to that of a neural network. The fully connected layer joins each of these vectors to form a model, and the Softmax function determines whether the patient is diabetic or not. The most major advantage of CNN is that it automatically extracts features. CNN takes less time to classify and has a high level of accuracy, especially in image processing. However, because of the high computational cost, it cannot perform well without a huge amount of training data.

### Long Short-Term Memory (LSTM):

The Long Short Term Memory Network (LSTM) is an enhanced Recurrent Neural Network (sequential network) that permits information to be stored indefinitely. It can deal with the diminishing gradient issue that RNN has. Apart from individual data points like photos, LSTM has feedback connections, which means it can process the complete sequence of data. This is useful in natural language processing, machine translation, and other areas. The LSTM is a type of RNN that performs exceptionally well on a wide range of issues. An LSTM model's central role is played by a memory cell called a 'cell state,' which maintains its state across time. In an LSTM, information can be provided to or withdrawn from the cell state, which is controlled by gates. These gates allow information to move into or out of the cell if desired. The method is aided by an element - wise multiplication operation and then a sigmoid neural net layer. The sigmoid layer outputs integers between 0 and 1, with 0 indicating that "nothing should be let through" and 1 indicating that "everything should be let through." LSTM neural networks can solve a variety of tasks that prior learning algorithms such as RNNs couldn't.

## Evaluation Metrics:

Developing machine learning models is based on the idea of constructive feedback. The main goal in implementing machine learning algorithms on different types of data is to understand how well a model performs on the data. Researchers found out various methods to calculate the performance of a machine learning model that are often called as Evaluation metrics. These metrics differ from model to model based on the problem to be solved. Evaluation metrics are not only used to compute a model’s performance, but they also allow the user to compare the model with other models. This way the user can decide which model performs best on the selected data. There are several most used evaluation or performance metrics such as Confusion matrix, Gini coefficient, F1 score, Gain and Lift charts, AUC – ROC, Kolmogorov Smirnov chart, Log Loss, Concordant – Discordant ratio, Cross-validation, and Root mean squared error. In this project, to analyse both machine learning and deep learning models performance, some of the standard metrics are used which are discussed below.

### Confusion Matrix:

To put in simple terms, a confusion matrix is a representation that generally consists of four different scores or values. These four values are True Negative (TN), True Positive (TP), False Negative (FN), and False Positive (FP). It basically represents the count of all these values. Confusion matrix is one of the most commonly used metrics to evaluate many classification algorithms.

Table 1

*Confusion Matrix*

|  |  |  |
| --- | --- | --- |
| Predicted Values | | |
| 0 | | 1 |
| Actual Values | 0 | TN | FP |
| 1 | FN | TP |

* **True Positive (TP):**

TP is the count of positive values that have been predicted correctly as positive. In this project, TP indicates the count of correctly predicted number of females with diabetes i.e., outcome 1 is predicted as 1.

* **True Negative (TN):**

TN is the count of negative values that have been predicted as negative by a model. In this project, TN denotes the number of correctly predicted females without diabetes i.e., outcome 0 is predicted as 0.

* **False Positive (FP):**

False positive is a number that denotes the count of negative values that are predicted as positive by a model. In our project, it represents how many females without diabetes (0) have been predicted as with diabetes (1). A False positive value sometimes is also termed as Type 1 error.

* **False Negative (FN):**

False negative is a value that indicates the count of positive values that are predicted as negative by a model. FN in our project indicates the number of times females with diabetes (1) predicted as females without diabetes (0). A False negative value sometimes is also termed as Type 2 error.

Based on the TP, TN, FP, FN values, various performance metrics can be calculated which are described below.

### Classification Report:

A Classification Report is kind of a tabular representation of various evaluation metrics available in machine learning. This can be implemented in python language using a famous library provided by Scikit-learn called as ‘classification\_report’. The classification report mainly consists of five metrics that are Accuracy, Precision, Recall, F1-Score, and Support. All these metrics are calculated based on the values provided by the Confusion matrix. A detailed explanation of the metrics displayed in a classification report is given below.

* **Accuracy:**

Accuracy of a model is calculated mathematically by dividing the total number of truly predicted values to the total predicted values. So, it is basically the ratio of the sum of True positive and True negative values to the sum of total values predicted.

Accuracy =

* **Precision:**

Precision, sometimes called as Positive predictive value is the ratio of actual positive values (TP) to the sum of total positive values (TP+FP) predicted by a model.

Precision =

* **Recall:**

Recall which is also known as Sensitivity is an evaluation metrics which is the ratio of correctly predicted positive values (TP) to the sum of actual positive values (TP+FN).

Recall =

* **F1-Score:**

F1-Score is considered as one of the important metrics in most of the cases due to its ability to use both Precision and Recall values. It is the weighted mean of both Precision and Recall values with its score ranging from 0 to 1, where 0 is the worst and value near to the 1 is considered as a good score. F1-Score is also called as F-measure.

F1-Score =

### AUC-ROC Score:

As discussed in the above section, confusion matrix offers several calculation parameters to analyse a model’s performance such as sensitivity, specificity, true positive, false positive. These parameters can be used in a different evaluation criterion known as AUC-ROC score which utilises visualisation as the method of representation. This score is mainly used in evaluating classification algorithms as it provides a clear understanding of a model’s ability to differentiate classes. Receiver Operating Characteristics curve (ROC) is a graphical representation metric that is formed by plotting a graph between True Positive rate (TPR) and the False Positive rate (FPR). Area Under Curve (AUC) is a metrics that is mostly used for binary classification problems. It computes the area under the ROC curve and this score determines the performance of the model in distinguishing both the classes. The higher the AUC score, the better the model’s performance.

# 

# Chapter 4 Requirements

In this section of the report, a clear explanation of the dataset and the system features that are required to generate the end results is provided. It not only also discusses about the availability of the data used for this project, but also gives information about how the data is collected. This section describes how the system features play an important role in executing this project.

## 4.1 The Dataset:

There are several datasets available publicly and privately about different health conditions of various groups of people. When considering the case of healthcare, lots of data is being generated every day by multiple health organisations all over the world. This data is being collected and stored by monitoring everyday health conditions of all kinds of patients who are reaching out to these organisations. With the help of different data storage models that are being developed in the field of machine learning and data science, it is becoming achievable to store this data on computers rather than storing them in traditional paper records. Some of this stored data is being made accessible to the public by the hospitals with the consent of the patients. The data collected this way is further being analysed, cleaned, and sorted in a way that makes it easier for researchers and scholars to experiment their ideas on the data. These datasets are then published online by few websites such as UCI Machine Learning repository and Kaggle. The dataset used in this project to predict diabetes is one of the famous publicly available datasets on Kaggle, which is the PIMA Indian Diabetes dataset (<https://www.kaggle.com/uciml/pima-indians-diabetes-database>). Kaggle is a platform that hosts hundreds of datasets which have been made available publicly by various researchers and analysts. This dataset has been provided to the public by the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset contains data related to women who are of PIMA Indian heritage and are of age at least 21 years old. It consists of 768 rows and 9 columns. The columns represent the attributes or features of the data, which are different health factors of a woman. In these nine attributes, eight of them are considered as predictor or independent variables and one variable is considered as target or dependent variable. Generally dependent variables are the ones whose value depends on rest of the independent variables present in the dataset. The target or dependent variable is to be predicted by the predictor or independent variables. The table 2 below shows the list of features or attributes present in this dataset.

Table 2

*List of Variables in the dataset.*

Dependent Variable (or) Target

Independent

Variables (or) Predictors

|  |  |
| --- | --- |
| **List of Attributes** | |
| 1 | Pregnancies |
| 2 | Glucose |
| 3 | BloodPressure |
| 4 | SkinThickness |
| 5 | Insulin |
| 6 | BMI |
| 7 | DiabetesPedigreeFunction |
| 8 | Age |
| 9 | Outcome |

The table 3 mentioned below provides a description of the columns or attributes present in the dataset.

Table 3

*Description of the Variables*

|  |  |  |
| --- | --- | --- |
| **Name of the Variable** | **Description of the Variable** | **Range** |
| Pregnancies | Number of times pregnant. | 0-17 |
| Glucose | Oral glucose test result. (In mg) | 0-199 |
| BloodPressure | Blood pressure value | 0-122 |
| SkinThickness | Triceps skin fold thickness value | 0-99 |
| Insulin | Level of Insulin in blood | 0-846 |
| BMI | Body mass index value | 0-67.1 |
| DiabetesPedigreeFunction | Family history of diabetes | 0.078-2.42 |
| Age | Age in years | 21-81 |
| Outcome | 1 – indicates a person has diabetes  0 – indicates a person doesn’t have diabetes | 0/1 |

## Tools and Libraries Required:

To apply the machine learning or deep learning techniques effectively, different libraries and tools are available. These tools come with their own advantages and disadvantages depending on the area of application. Some of the most used programming languages in machine learning include Python, R, SQL, etc. Whereas depending on the languages used, various tools or software are available in the market such as Jupyter Notebook, Matlab, Rstudio, etc. All these software tools come with a set of libraries that are applicable in different ways. The programming language, software, and libraries that have been used to implement this project are discussed below.

### Python:

Python is an open-source programming language that has become immensely popular in recent years for application in various fields. It is known for its ability to write programs in user-friendly language. Python has a wide range of applications such as in web development, Data Science, Internet of Things, Artificial Intelligence etc., Recently, python has been known for its usage as a primary language in implementing machine learning, and deep learning technologies. Starting from data collection to data analysis, and visualisations, python has become extremely popular for its ease of use. Python also allows users to perform all kinds of analysis on vast and variety of datasets. In this project, Python has been used as the main programming language to implement necessary machine learning tasks and achieve the expected results. In this project the version of python language used is 2.7.16. Python provides and allows users to analyse the data using various libraries which are discussed in the further sections below.

### Jupyter Notebook:

Jupyter Notebook is an open-source, integrated development environment that is available free both online and offline. Jupyter Notebook possess the quality of allowing to write, share, and execute already built codes. It allows users to write codes in different programming languages such as python, R, Scala, Julia, etc., While there are many available platforms to write and execute programs in python, Jupyter Notebook is considered one of the most widely used platform for implementing python due to its ease of availability, ability to perform user interface visualisations etc., In this project, python language along with Jupyter notebook has been used to accomplish the data analysis and machine learning tasks required to solve the problem.

### Pandas and NumPy:

Pandas and NumPy are some of the most successful and extensively used open-source libraries in data pre-processing stage of analysing a dataset. Pandas provide an excellent option to load data in a DataFrame format. It allows users to perform multiple actions on rows and columns in a DataFrame. Whereas NumPy is known for its ability to allow users to perform different numerical operations on data. These numerical functions include arithmetic operations, descriptive statistics etc.,

### Seaborn and Matplotlib:

Data visualisation is one of the popular methods of representing the data. It is a crucial component of data science that helps people comprehend and interpret complex data. Matplotlib and Seaborn are the foundations of data visualisation using Python. Matplotlib is mostly used to plot two dimensional graphs of arrays and create statical interferences such as histograms, box plots, violin plots, pie charts, bar charts, etc., Seaborn library is known as an extended version of matplotlib library. It is used to perform univariate and bivariate analysis by creating beautiful visualisations. It is also used in generating pair plots, heatmaps, time forecasting, etc.

### Scikit-learn:

Scikit-learn library is the most significant and core part of data science that is used to build and implement machine learning algorithms. It has a wide variety of features to choose ranging from ML algorithms to their metrics. All types of machine learning models are built in python by importing scikit-learn library. Additionally, scikit-learn library provides different tools to perform operations such as feature extraction and model evaluation.

### TensorFlow and Keras:

To solve some complex problems in machine learning, a free and open-source library is used known as Tensorflow. It is extensively used in the fields of machine learning and deep learning to train the models. Keras is a Google-developed high-level deep learning API for implementing neural networks. It is built in Python and is used to make neural network implementation simple.

It also allows for the computation of numerous neural networks in the backend.

# Chapter 5 Analysis

In this section of the report, various tasks performed as a part of data analysis are discussed. It includes visualising the data, univariate, and bi-variate analysis, correlation analysis and plotting heatmap. Some of the mostly used graphical representations in interpreting the data such as histograms, count plots, pair plots, box plots, violin plots are presented in this section.

## 5.1 Descriptive Statistics:

Table 4

*Statistical Parameters calculated*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
| Count | 768 | 768 | 768 | 768 | 768 | 768 | 768 | 768 | 768 |
| Mean | 3.84 | 120.89 | 69.1 | 20.53 | 79.79 | 31.99 | 0.47 | 33.24 | 0.34 |
| Std | 3.36 | 31.97 | 19.35 | 15.95 | 115.24 | 7.88 | 0.33 | 11 | 0.47 |
| Min | 0 | 0 | 0 | 0 | 0 | 0 | 0.07 | 21 | 0 |
| 25% | 1 | 99 | 62 | 0 | 0 | 27.3 | 0.24 | 24 | 0 |
| 50% | 3 | 117 | 72 | 23 | 30.5 | 32 | 0.37 | 29 | 0 |
| 75% | 6 | 140 | 80 | 32 | 127.25 | 36.6 | 0.62 | 41 | 1 |
| Max | 17 | 199 | 122 | 99 | 846 | 67 | 2.42 | 81 | 1 |

The table 4 shown above is the output obtained when the statistical description operation has been performed on the dataset. This is achieved by using a function in python called “.describe()”. As mentioned in the table 4, different statistical parameters such as count, mean, median, standard deviation, minimum, max values, 25th, and 75th quartile values of each variable are calculated. In the table 5 shown below, the type of data every variable consists is obtained by applying a function on the dataset called “.info()”.

Table 5

*Data Types of Variables*

|  |  |  |
| --- | --- | --- |
| **S.no** | **Variable** | **Data Type** |
| 1 | Pregnancies | Int64 |
| 2 | Glucose | Int64 |
| 3 | BloodPressure | Int64 |
| 4 | SkinThickness | Int64 |
| 5 | Insulin | Int64 |
| 6 | BMI | Float64 |
| 7 | DiabetesPedigreeFunction | Float64 |
| 8 | Age | Int64 |
| 9 | Outcome | Int64 |

## Target Variable:

In every problem, the variables present in a dataset can be divided into two types- independent and dependent variables. As mentioned in the section 4, the dependent variable considered in the present dataset is the Outcome variable. It is important to explore the target variable to understand more about what the algorithms need to classify or predict. So below mentioned is some of the analysis performed on the Output variable. The table 6 shown below demonstrates the properties of the Outcome variable. We can see from the table that the Outcome variable has two types of classes 0 and 1. Person with diabetes is represented with 1 and without diabetes with 0. The dataset has a higher percentage of people without diabetes compared to people with diabetes.

Table 6

*Classes in Output variable*

|  |  |  |
| --- | --- | --- |
| **Target Variable: Output** | | |
| Type of class | 0 (non-diabetic) | 1 (diabetic) |
| Number of values | 500 | 268 |
| Percentage of values | 65.1% | 35.9% |

### Visualising Target variable:

The figure 6 mentioned below are a representation of the pie chart and bar graph of Outcome variable. These visualisations are obtained by importing the matplotlib library.

Figure 6

*Pie and Bar charts of Output Variable*

Chart, pie chart

Description automatically generatedChart, bar chart

Description automatically generated

## 5.3 Univariate Analysis:

As a part of exploratory data analysis, univariate analysis has been performed on all the variables in the dataset. By performing univariate analysis, the trends, and variations a variable possess can be understood. In the figures mentioned below, both boxplot and histogram used to visualise the variables is shown.

Figure 7

*Distribution plot of Pregnancies*

Chart

Description automatically generated

Figure 8

*Distribution plot of Age*

Chart

Description automatically generated

Figure 9

*Distribution plot of DiabetesPedigreeFunction*

Chart, histogram

Description automatically generated

Figure 10

*Distribution plot of Insulin*

A picture containing table

Description automatically generated

Figure 11

*Distribution plot of BMI*

Chart, histogram

Description automatically generated

Figure 12

*Distribution plot of SkinThickness*

Chart

Description automatically generated

Figure 13

*Distribution plot of BloodPressure*

Chart, histogram

Description automatically generated

Figure 14

*Distribution plot of Glucose*

Chart, histogram

Description automatically generated

## 5.4 Outlier Analysis:

Every variable in the dataset consists of few outliers which have been observed by performing the univariate analysis in the above section. The table 7 shown below represents the range after which the outliers are present in each variable. The values above these ranges are considered as outliers and they have been removed successfully. After removing the outliers, the size of the dataset has been reduced from (768, 9) to (748, 9).

Table 7

*Range of Outliers in each variable*

|  |  |
| --- | --- |
| **Variables** | **Outliers value** |
| Pregnancies | >14 |
| Glucose | <50 |
| BloodPressure | >120 |
| SkinThickness | >80 |
| Insulin | >55 |
| BMI | >600 |
| DiabetesPedigreeFunction | >2 |
| Age | >70 |

Table 8

*Size of data before & after outlier treatment*

|  |  |
| --- | --- |
| **Outlier analysis** | **Size of the dataset** |
| Before Outlier removal | (768,9) |
| After Outlier removal | (748,9) |

## 5.5 Pair Plot:

A pair plot is a graphical representation that illustrates the pairwise relationship between all the variables present in a dataset. It is obtained in python with the help of seaborn library. A pair plot is generated by taking a set of independent variables on the x-axis as well as on the y-axes. In this dataset, all the independent variables have been considered in plotting a pair plot on basis of the dependent variable Outcome. In the below shown figure 15, the blue and orange colour indicates two different classes present in the Outcome variable- 0 and 1.

Figure 15

*Pair Plot of the dataset*

A picture containing bedclothes

Description automatically generated

## Bi-variate analysis:

Bi-variate analysis is the analysis of each variable’s behaviour with respect to the target variable. In this project, bi-variate analysis has been performed by visualising the data using box plots and violin plots. Below mentioned figures are the outputs obtained by importing seaborn library in python.

Figure 16

*Age vs Outcome*

Chart, box and whisker chart

Description automatically generated

Figure 17

*DiabetespedigreeFunction vs Outcome*

Chart, box and whisker chart

Description automatically generated

Figure 18

*Glucose vs Outcome*

Chart, box and whisker chart

Description automatically generated

Figure 19

*BloodPressure vs Outcome*

Chart, box and whisker chart

Description automatically generated

Figure 20

*SkinThickness vs Outcome*

Chart, box and whisker chart

Description automatically generated

Figure 21

*Insulin vs Outcome*

Chart, box and whisker chart

Description automatically generated

Figure 22

*BMI vs Outcome*

Chart, box and whisker chart

Description automatically generated

## Correlation analysis:

To understand more about how the variables are related to each other, correlation analysis has been done. This is achieved by visualising a heat map which represents all the variables on both x-axis and y-axis with their correlation coefficient values in each box of the heatmap. The below shown fig 23 is the correlation heatmap obtained. We can see from the figure that Glucose has the highest correlation coefficient value as 0.49 when compared to other variables. The next two variables with highest correlation coefficient value are Age and BMI with values 0.28 and 0.22.

Figure 23

*Correlation plot of the data*

Chart

Description automatically generated

## 5.8 Finding and Filling Missing values:

As discussed in the above section, it is advantageous to clean the data before implementing any machine learning algorithms. In this project, the data has been cleaned by filling the missing values with mean and median values of respective variables. The below shown table represents the number of missing values present in each variable.

Table 9

*Number of Missing values*

|  |  |
| --- | --- |
| **Variable** | **Number of Missing values** |
| Pregnancies | 0 |
| Glucose | 0 |
| BloodPressure | 34 |
| SkinThickness | 224 |
| Insulin | 366 |
| BMI | 11 |
| DiabetesPedigreeFunction | 0 |
| Age | 0 |
| Outcome | 0 |

The below shown figure 24 is a bar graph illustrating the number of missing values present in each variable.

Figure 24

*Bar Graph showing the number of missing values*

A picture containing text, clipart

Description automatically generated

# Chapter 6 Implementation

In this section, the process of implementing the proposed methods is discussed. The platform and programming language used to build and execute the models are explained in this section. Certain factors considered for building machine learning and deep learning models are mentioned and the reasons to choose those parameters are discussed.

## 6.1 Train-Test Split:

To make the data ready to be fit on the models, it is divided into training and testing in a ratio of 70:30. This means that 70% of the data is considered as training set which is used to train the models and 30% of the data is considered as testing set which is used to test the model’s performance. This is achieved by importing a function from scikit-learn library called “train\_test\_split”. Generally, when solving machine learning problems, the data used is split in the ratio of 70:30 because it produces optimal results. This ratio of dividing the data can be varied according to the application. The independent variables are grouped together and are referred to as X. While the dependent variables grouped together are represented by Y. Next, this X and Y sets are divided into training and testing sets of data. In the below shown table 10 the shape of training and testing data after splitting the data is represented.

Table 10

*Shape of Training & Testing data*

|  |  |
| --- | --- |
| Shape of X\_train | (523, 8) |
| Shape of X\_test | (225, 8) |
| Shape of y\_train | (523, 1) |
| Shape of y\_test | (225, 1) |

## 6.2 Model Building:

## 6.2.1 Machine learning algorithms:

Machine learning models used in this project are implemented by importing the scikit-learn library. The decision tree algorithm used is fitted on the training and testing sets of data. The random forest algorithm is fitted on the data by setting certain parameters which decide the performance of it. The n\_estimators value is set as 100 and the max\_features is set as sqrt. Similarly, logistic regression and support vector machine are fitted on the data without setting any specific parameters. The k nearest neighbour algorithm is supplied by the number of neighbours values being 5 which means it is supposed to calculate the distance between its five nearest neighbours. When implementing XGBoost algorithm, one of its hyper parameters called gamma is set as default value i.e., 0.

### 6.2.2 Deep learning algorithms:

By importing Keras and TensorFlow libraries in python, the deep learning algorithms implementation has been done. To implement Artificial neural network algorithm, three layers have been considered with activation functions used being ‘relu’ and ‘softmax’. The input dimensions value has been set as 8 which indicates the number of independent variables. The number of epochs is set as 100 combined using early stopping function. As shown in the figure 25 below, the loss function is set as ‘categorical\_crossentropy’ and optimiser is chosen as ‘adam’.

Figure 25

*Structure of Artificial Neural Network*

Table

Description automatically generated

Similarly, the Convolutional neural network algorithm is implemented by setting the loss function as ‘binary\_crossentropy’. The below shown figure 26 represents the different layers considered in implementing CNN algorithm. However, the activation functions and optimiser are set same as ANN algorithm.

Figure 26

*Structure of CNN model built*

Table

Description automatically generated

The below shown figure 27 depicts the LSTM model built for solving the problem in this project. Four layers have been used to implement LSTM model with specifying the input shape as (8,1).

Figure 27

*Structure of LSTM model built*

A screenshot of a computer

Description automatically generated with low confidence

# Chapter 7 Experimental Results

In this section of the report, the results obtained by implementing the proposed methodology are discussed. These results include various types of metrics applied to evaluate the performance of machine learning and deep learning algorithms. This section also includes comparison of accuracy scores of all the models applied, which explains the reason why a specific model has been chosen as the best one.

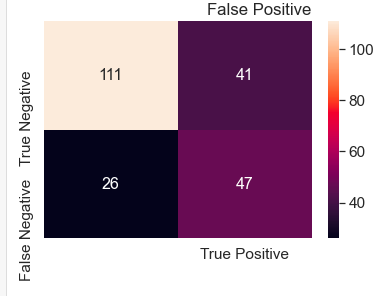
## 7.1 Confusion Matrix

1. **Decision Tree Algorithm:**

The figure 28 shown below is a 2\*2 confusion matrix of Decision Tree algorithm.

Figure 28

*Confusion matrix of Decision tree classifier*

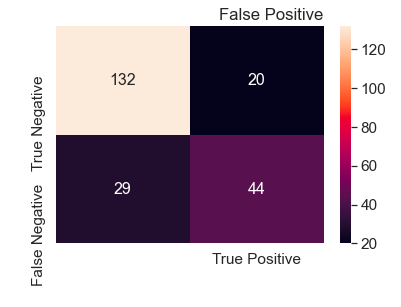


1. **Random Forest Algorithm:**

The figure 29 shown below is a 2\*2 confusion matrix of Random Forest Algorithm.

Figure 29

*Confusion matrix of Random Forest classifier*

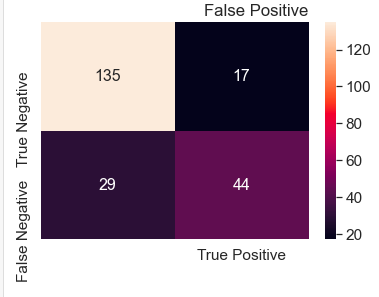


1. **Logistic Regression:**

The below shown figure 30 is a 2\*2 confusion matrix of Logistic Regression algorithm.

Figure 30

*Confusion matrix of Logistic Regression*

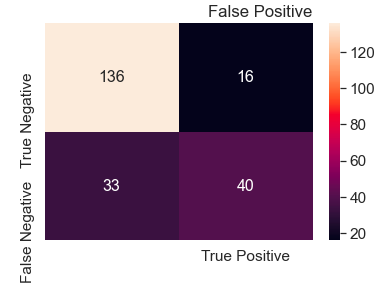


1. **Support Vector Classifier:**

The below shown figure 31 is a 2\*2 confusion matrix of Support Vector Classifier algorithm.

Figure 31

*Confusion matrix of Support Vector Classifier*

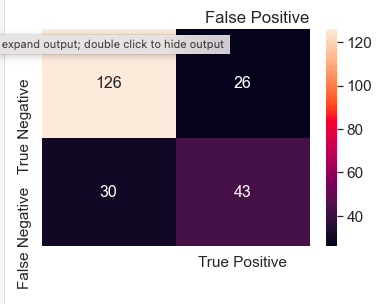


1. **K Nearest Neighbour Classifier:**

The below shown figure 32 is a 2\*2 confusion matrix of K Nearest Neighbour Classifier algorithm.

Figure 32

*Confusion matrix of KNN*

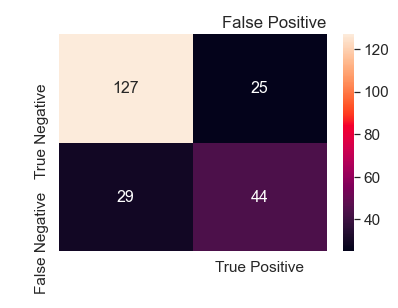


1. **XGBoost Classifier:**

The below shown figure 33 is a 2\*2 confusion matrix of XGBoost Classifier algorithm.

Figure 33

*Confusion matrix of XGBoost classifier*

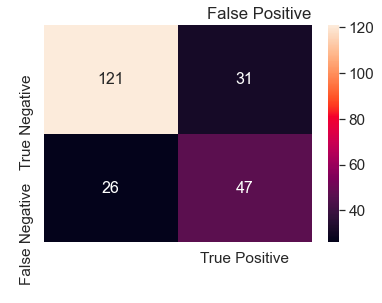


1. **Artificial Neural Network Algorithm:**

The below shown figure 34 is a 2\*2 confusion matrix of Artificial Neural Network algorithm.

Figure 34

*Confusion matrix of ANN algorithm*

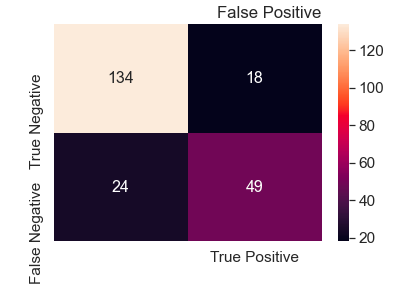


1. **Convolutional Neural Network Algorithm:**

The below shown figure 35 is a 2\*2 confusion matrix of Convolutional Neural Network algorithm.

Figure 35

*Confusion matrix of CNN algorithm*

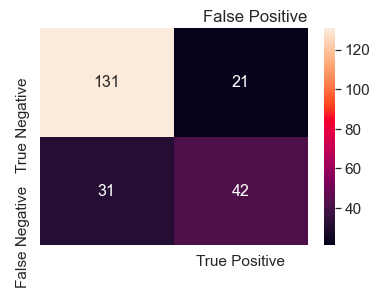


1. **Long Short-Term Memory Algorithm:**

The below shown figure 36 is a 2\*2 confusion matrix of Long Short-Term Memory algorithm.

Figure 36

*Confusion matrix of LSTM algorithm*



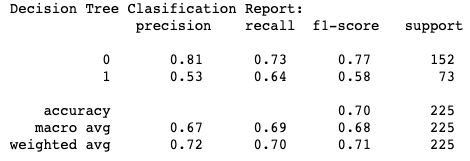
## 7.2 Classification Report:

1. **Decision Tree Classifier:**

The below shown figure 37 is the classification report obtained for Decision Tree algorithm.

Figure 37

*Classification report of Decision Tree*

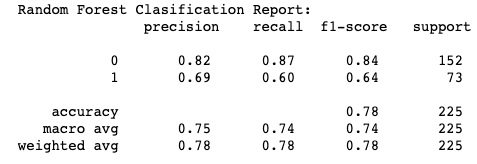


1. **Random Forest Classifier:**

The below shown figure 38 is the classification report obtained for Random Forest Classifier algorithm.

Figure 38

*Classification report of Random Forest*

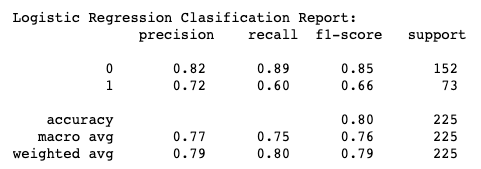


1. **Logistic Regression:**

The below shown figure 39 is the classification report obtained for Logistic Regression algorithm.

Figure 39

*Classification report of Logistic Regression*

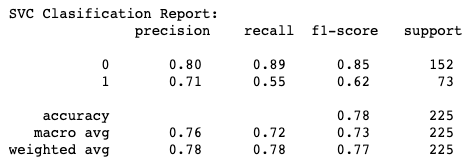


1. **Support Vector Classifier algorithm:**

The below shown figure 40 is the classification report obtained for Support Vector Classifier algorithm.

Figure 40

*Classification report of SVM*

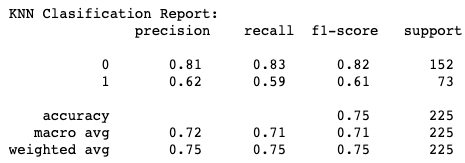


1. **K Nearest Neighbour algorithm:**

The below shown figure 41 is the classification report obtained for K Nearest Neighbour algorithm.

Figure 41

*Classification report of KNN*

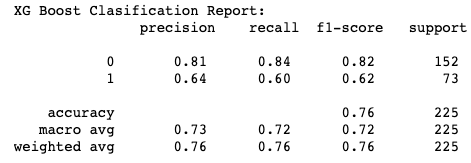


1. **XGBoost algorithm:**

The below shown figure 42 is the classification report obtained for XGBoost algorithm.

Figure 42

*Classification report of XGBoost*

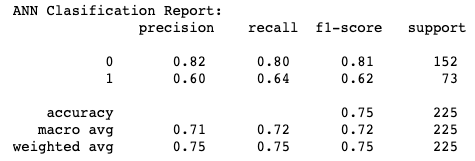


1. **Artificial Neural Network:**

The below shown figure 43 is the classification report obtained for Artificial Neural Network algorithm.

Figure 43

*Classification report of ANN*

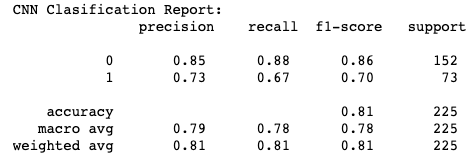


1. **Convolutional Neural Network:**

The below shown figure 44 is the classification report obtained for Convolutional Neural Network algorithm.

Figure 44

*Classification report of CNN*

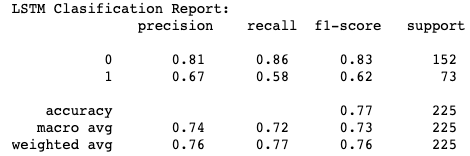


1. **Long Short-Term Memory algorithm:**

The below shown figure 45 is the classification report obtained for Long Short-Term algorithm.

Figure 45

*Classification report of LSTM*



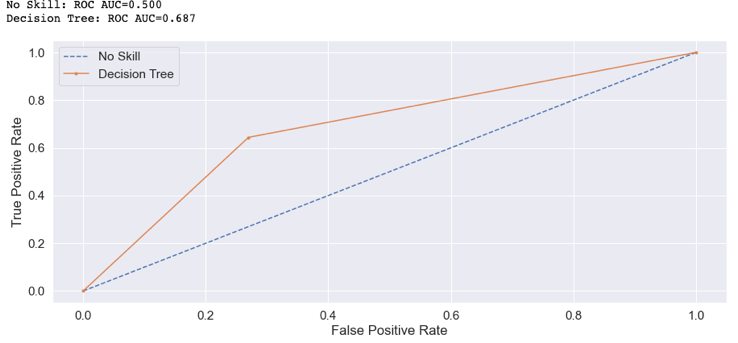
## 7.3 AUC-ROC Score & Curve

1. **Decision Tree Classifier:**

The below shown figure represents the AUC score and ROC curve of the Decision Tree Classifier algorithm.

Figure 46

*ROC Curve of Decision Tree*

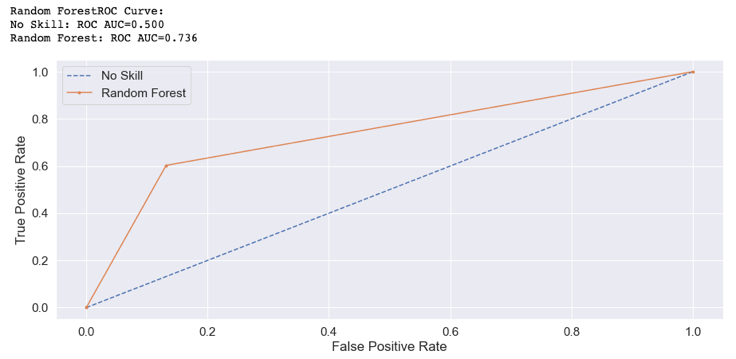


1. **Random Forest Classifier:**

The below shown figure 47 represents the AUC score and ROC curve of the Random Forest Classifier algorithm.

Figure 47

*ROC Curve of Random Forest classifier*

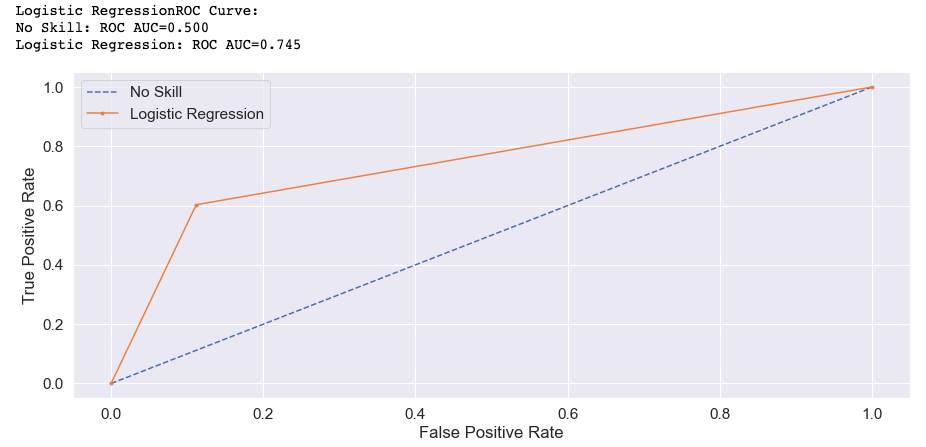


1. **Logistic Regression:**

The below shown figure 48 represents the AUC score and ROC curve of the Logistic Regression algorithm.

Figure 48

*ROC Curve of Logistic Regression*



1. **Support Vector Machine:**

The below shown figure 49 represents the AUC score and ROC curve of the Support Vector Machine algorithm.

Figure 49

*ROC Curve of SVM*

Chart, line chart

Description automatically generated

1. **K Nearest Neighbour algorithm:**

The below shown figure 50 represents the AUC score and ROC curve of the K Nearest Neighbour algorithm.

Figure 50

*ROC Curve of KNN*

Chart, line chart

Description automatically generated

1. **XGBoost algorithm:**

The below shown figure 51 represents the AUC score and ROC curve of the XGBoost algorithm.

Figure 51

*ROC Curve of XGBoost*

Chart, line chart

Description automatically generated

1. **Artificial Neural Network:**

The below shown figure 52 represents the AUC score and ROC curve of the Artificial Neural Network algorithm.

Figure 52

*ROC Curve of ANN*

Chart, line chart

Description automatically generated

1. **Convolutional Neural Network:**

The below shown figure 53 represents the AUC score and ROC curve of the Convolutional Neural Network algorithm.

Figure 53

*ROC Curve of CNN*

Chart, line chart

Description automatically generated

1. **Long Short-Term Memory:**

The below shown figure 54 represents the AUC score and ROC curve of the Long Short-Term Memory algorithm.

Figure 54

*ROC Curve of LSTM*

Chart, line chart

Description automatically generated

## 7.4 Feature Importance:

Feature importance in machine learning is a method of analysing the importance of different features or variables in a dataset. There are several techniques in machine learning that are used to perform feature analysis. One such method is using ensemble methods which include models like Random Forest, XGBoost etc. In this project, to analyse the importance of features, XGBoost algorithm has been used. The below shown figure 55 represents a bar graph showing the feature importance of all the variables present in the dataset. It can be seen that Glucose holds the most importance of all the variables followed by BMI and Insulin.

Figure 55

*Bar Graph showing Feature Importance*



## 7.5 Comparison of models:

All the machine learning and deep learning models implemented in this project are compared based on different accuracy metrics. The table () shown below represents the train and test accuracies obtained for all the models. Support Vector Machine has shown a highest accuracy among all other models used. The figure () mentioned below is a graph plotted between accuracies of all models.

Table 11

*Comparison of Models*

|  |  |  |  |
| --- | --- | --- | --- |
| **S.no** | **Model** | **Test accuracy** | **Train accuracy** |
| 1 | SVC | 79.55 | 76.2 |
| 2 | Random Forest | 78.22 | 100 |
| 3 | Logistic Regression | 78.22 | 81.64 |
| 4 | CNN | 77.3 | 74.7 |
| 5 | Gradient Boost or XGBoost | 76 | 100 |
| 6 | KNN | 75.11 | 81.64 |
| 7 | ANN | 74.2 | 80.6 |
| 8 | LSTM | 73.3 | 73.2 |
| 9 | Decision Tree | 71.11 | 100 |

Figure 56

*Bar Chart comparing Accuracy scores of Models*

Chart, bar chart

Description automatically generated

# Chapter 8 Project Management

## 8.1 Project Schedule

The project implementation is divided into various stages. It was designed in such a way that the first step was to explore my interests in different domains and finding a problem that has a good impact in real time. After finalising the topic, the next step was to find a suitable methodology. This was accompanied by writing an informative project proposal. After the project proposal was approved by my supervisor, I started working on the model building. There were several issues encountered when building the model, which were resolved later by using resources online. There was a delay in completing the report writing as my first supervisor had to leave the university. However, I got help from my new supervisor who guided in writing the report effectively. The below shown figure 57 is a Gantt chart illustrating the original schedule planned in designing and implementing the project.

Figure 57

*Gantt Chart*

## Quality Management

Quality has been given first preference throughout the period of planning, designing, and implementing the project. In the data analysis stage, methods such as filling missing values and outlier treatment has been carried out effectively. By doing this, the quality of data has been improved which made it suitable for the model implementation. Extensive research has been conducted on choosing various factors which can improve different machine learning and deep learning algorithms accuracy. Several theoretical and practical approaches of evaluating the algorithms have been explored before deciding the evaluation metrics. The evaluation metrics used are classification report, confusion matrix, and AUC-ROC score. The models’ performance was evaluated by comparing the accuracy score, AUC score, specificity, recall, and precision values.

## Social, Legal, Ethical and Professional Considerations

Health related data of a person is often considered very confidential. So, when handling such data, strict measures must be taken to avoid breaching of any laws. Many private, public, and governmental organisations working in health sector often collect and store the health data of their patients. Some of the organisations make this data available to the public for conduction research on it. The data considered in this project is available online for public usage. And all the necessary steps have been taken so that no personal identities of any person are revealed. The methodology adopted in this project is implemented by referring to various previous studies. All the references have provided appropriately in the bibliography section.

# Chapter 9 Critical Appraisal

Detecting diseases in an early stage has always proved to be an effective approach in treating those diseases. Based on this phenomenon, this project is focused on predicting diabetes of person considering their health condition. There can be different ways to predict diabetes, but in this project, an implementation of machine learning and deep learning techniques have been done to predict diabetes. The dataset used contains health data of females aged 21 or above with 768 observations in total. The implementation of chosen machine learning and deep learning techniques have been done successfully. Both the model’s accuracy has been compared based on the accuracy metrics. The limitations of this project include using a dataset with very less number of observations. Machine learning and deep learning algorithms have always shown a change in performance with different types of datasets. If there are more observations, a model learns more about the problem and outcome, which improves the model’s performance. The use of deep learning algorithms is found more in data related to images, but in this dataset majority of the data available is in numerical format. With variety kind of data, deep learning algorithms could have been improved. Different studies proposed have used various optimisation techniques such as gridsearchcv and randomsearchcv, which have shown an improvement in models accuracy scores. Research could have been done by exploring the usage of such techniques to improve models’ accuracy scores.

# Chapter 10 Conclusions

## Achievements

In conclusion, in this project, the causes and effects of diabetes on human body, in a long-term are explained. To avoid any serious health complications caused due to diabetes, it is advantageous to diagnose and treat diabetes at an early stage. In this project, two different approaches were adopted in predicting diabetes based on previous data which are by implementing machine learning and deep learning techniques. The data used to analyse and implement these models is PIMA Indian diabetes dataset which consists of data with various health conditions of females aged 21 or above. The dataset has been analysed by performing exploratory data analysis including visualising and cleaning the data. After making the data ready for model fitting, different machine learning and deep learning models are applied on the data. The implementation of this models is achieved by using python programming language in jupyter notebook. Various data and visualisation libraries have been used in this project. Machine learning algorithms have been implemented by importing sklearn library. Whereas deep learning algorithms have been employed by importing tensorflow and keras libraries. The models’ performance was measured using several evaluation metrics such as classification report, confusion matrix, AUC-ROC score, and curve. Apart from predicting diabetes, ensemble techniques such as random forest and XGboost algorithms have been used to analyse feature importance. With this, it is understood that among all the features, Glucose played an important role in predicting diabetes of a person. Support vector classifier algorithm has shown a highest accuracy of 79.5% and among deep learning algorithms Convolutional neural network algorithm has shown a good accuracy score of 77.3%.

## 10.2 Review of Objectives:

In this section, various objectives of the project are discussed. How and in what part of the project, these objectives are achieved is discussed. The below mentioned sections are the different objectives and aims considered before implementing the project.

**a) To understand Diabetes- Its causes and effects**:

This project is mainly focused on predicting diabetes. So, it was essential to understand the causes and its severe effects on diabetes, before implementing the proposed model. This was achieved by referring to various articles published online which explains diabetes in females. The detailed explanation of this objective is mentioned in the chapter one of the report.

**b) To implement Machine learning & Deep learning methods:**

The approach or methodology chosen for solving the problem in this project is discussed in chapters three, four, and five of the report. These methods were considered by referring to previously published papers which are focussed on predicting diabetes.

**c) To evaluate and compare the models’ performance:**

By implementing the machine learning and deep learning algorithms, diabetes was predicted in this project. These algorithms’ performance was measured using different metrics. In chapters six and seven, the results obtained in this project are discussed which include different evaluation metrics of the models.

## Future Work

There are diverse ways of tackling the diabetes prediction problem. The extensions to the current work can be done by finding wide variety of datasets which contain more observations and certain other variables. There is a scope of doing the work differently at every stage of implementing this project, starting from data collection, to evaluating the models. Apart from the current machine learning algorithms used in this project, other supervised algorithms can be used on the dataset and their accuracies can be compared. The model’s performance can be analysed using different metrics such as cross-validation score. The machine learning algorithms’ behaviour can be examined by varying the data, i.e., by adding new data or splitting the dataset in a different ratio. With the change in lifestyle of humans, new health factors are being added to the list of causes of diabetes. Research can be done in this area to understand what factors can be weighed most in predicting diabetes. The machine learning and deep learning algorithms’ performance can be improved by calculating the optimal parameters using hyperparameter optimisation techniques.

# Chapter 11 Student Reflections

I have discovered several challenges in different stages of implementing the project. Researching about various topics for project and choosing the best one was the initial challenge I faced as I had to choose a topic that fits into the domain, I’m interested in. I had to overcome this by exploring the way machine learning and deep learning are being applied in the field of healthcare. After shortlisting few topics, I had to take help from my supervisor in deciding about the final topic. The next stage was to find a suitable dataset related to the problem. I went through all the resources available online extensively and found a dataset that clearly explains the problem. To understand the problem and the data, I went through various journal articles and books which gave me a good theoretical knowledge about the problem. In analysing the data and implementing the models, I referred to different scholarly articles. This has thrown a spotlight on different approaches available to solve the problem and what could be done in overcoming the limitations of these papers. I was overcome by a situation where my supervisor was caught with covid, and I couldn’t ask for his guidance during that period. For me, this was the most difficult period as I was in the middle of writing project report. However, I tried my best to implement the project and design the report as per the guidelines available in the module. I was also benefitted by reaching out to the coursework assistance writing (CAW) team in adopting the best way to write my project report. I also found tremendous help from the library department in conducting a good literature review.

# Bibliography and References

Ahmad, H., Mukhtar, H., Alaqail, H., Seliaman, M., & Alhumam, A. (2021). Investigating

Health-Related Features and Their Impact on the Prediction of Diabetes Using Machine Learning. *Applied Sciences*, *11*(3), 1173. https://doi.org/10.3390/app11031173

Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A., & Alzakari, N. et al. (2021).

Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. *Applied Sciences*, *11*(2), 796. https://doi.org/10.3390/app11020796

*An introduction to machine learning with scikit-learn*. scikit-learn. (2022). Retrieved 18 April 2022, from https://scikit-learn.org/stable/tutorial/basic/tutorial.html.

Alex, S., Nayahi, J., Shine, H., & Gopirekha, V. (2021). Deep convolutional neural network for

diabetes mellitus prediction. *Neural Computing and Applications, 34(2),* 1319-1327.

https://doi.org/10.1007/s00521-021-06431-7

Apicella, M., Campopiano, M., Mantuano, M., Mazoni, L., Coppelli, A., & Del Prato, S. (2020). COVID-19 in people with diabetes: understanding the reasons for worse outcomes. *The Lancet Diabetes &Amp; Endocrinology*, *8*(9), 782-792. https://doi.org/10.1016/s2213-8587(20)30238-2

Chowdary, P., & Udaya, R. (2021). An Effective Approach for Detecting Diabetes using Deep

Learning Techniques based on Convolutional LSTM Networks*. International Journal Of*

*Advanced Computer Science And Applications, 12(4).*

https://doi.org/10.14569/ijacsa.2021.0120466

Das, R., Saha, I., Konai, D., & Medda, S. (2021). Effects of Insulin on Gestational Pima Indian

Heritage Females. *Endocrinology And Disorders, 5(7),* 01-04. https://doi.org/10.31579/2640-

1045/100

Denis, D. (2021). *Applied Univariate, Bivariate, and Multivariate Statistics Using Python*. John

Wiley & Sons, Incorporated.

Diab, M., Husain, S., & Jarndal, A. (2020). On Diabetes Classification and Prediction using Artificial Neural Networks. *2020 International Conference On Communications, Computing, Cybersecurity, And Informatics (CCCI)*. https://doi.org/10.1109/ccci49893.2020.9256621

Ibrahim, I., & Abdulazeez, A. (2021). The Role of Machine Learning Algorithms for Diagnosing Diseases. *Journal Of Applied Science And Technology Trends*, *2*(01), 10-19. https://doi.org/10.38094/jastt20179

Islam Ayon, S., & Milon Islam, M. (2019). Diabetes Prediction: A Deep Learning Approach. *International Journal Of Information Engineering And Electronic Business*, *11*(2), 21-27. https://doi.org/10.5815/ijieeb.2019.02.03

Kadowaki, T., Yamamoto, F., Taneda, Y., Naito, Y., Clark, D., & Lund, S. et al. (2021). Effects

of anti-diabetes medications on cardiovascular and kidney outcomes in Asian patients with

type 2 diabetes: a rapid evidence assessment and narrative synthesis. *Expert Opinion On*

*Drug Safety*, 1-14. https://doi.org/10.1080/14740338.2021.1898585

Kharroubi, A. (2015). Diabetes mellitus: The epidemic of the century. *World Journal Of Diabetes*, *6*(6), 850. https://doi.org/10.4239/wjd.v6.i6.850

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. (2021). 1D convolutional neural networks and applications: A survey. *Mechanical Systems And Signal Processing*, *151*, 107398. https://doi.org/10.1016/j.ymssp.2020.107398

Kiziloz, H. (2021). Classifier ensemble methods in feature selection. *Neurocomputing*, *419*, 97-107. https://doi.org/10.1016/j.neucom.2020.07.113

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-68771-z

Larabi-Marie-Sainte, Aburahmah, Almohaini, & Saba. (2019). Current Techniques for Diabetes Prediction: Review and Case Study. *Applied Sciences*, *9*(21), 4604. https://doi.org/10.3390/app9214604

Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*, *165*, 292-299. https://doi.org/10.1016/j.procs.2020.01.047

Nishat, M. (2021). Performance Assessment of Different Machine Learning Algorithms in

Predicting Diabetes Mellitus. *Bioscience Biotechnology Research Communications, 14(1),*

74-82. https://doi.org/10.21786/bbrc/14.1/10

Rahman, M., Islam, D., Mukti, R., & Saha, I. (2020). A deep learning approach based on convolutional LSTM for detecting diabetes. *Computational Biology And Chemistry*, *88*, 107329. https://doi.org/10.1016/j.compbiolchem.2020.107329

Refat, M., Amin, M., Kaushal, C., Yeasmin, M., & Islam, M. (2021). A Comparative Analysis of Early Stage Diabetes Prediction using Machine Learning and Deep Learning Approach. *2021 6Th International Conference On Signal Processing, Computing And Control (ISPCC)*. https://doi.org/10.1109/ispcc53510.2021.9609364

Singh, K., Elhoseny, M., Singh, A., & Elngar, A. *Machine learning and the internet of medical things in healthcare*.

Subrahmanya, S., Shetty, D., Patil, V., Hameed, B., Paul, R., & Smriti, K. et al. (2021). The role of data science in healthcare advancements: applications, benefits, and future prospects. *Irish Journal Of Medical Science (1971 -)*. https://doi.org/10.1007/s11845-021-02730-z

Team, K. (2022). *Keras documentation: The Sequential model*. Keras.io. Retrieved 17 April 2022, from https://keras.io/guides/sequential\_model/.

Xie, S., Yu, Z., & Lv, Z. (2021). Multi-Disease Prediction Based on Deep Learning: A

Survey*. Computer Modeling In Engineering &Amp; Sciences, 128(2)*, 489-522.

https://doi.org/10.32604/cmes.2021.016728

# Appendix A – Project Specification

The Dataset:

<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

The Code:

<https://colab.research.google.com/drive/1vcRfCt-orUIcbkaw2VT162TmXn9hyS5i?usp=sharing>

# Appendix B – Project Presentation

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing icon

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, application

Description automatically generated

Diagram

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Text

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

A picture containing text

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, diagram

Description automatically generated

Table

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated

Chart

Description automatically generated

Graphical user interface, application

Description automatically generated

A picture containing diagram

Description automatically generated

Graphical user interface, table

Description automatically generated with medium confidence

Chart

Description automatically generated

Graphical user interface, text

Description automatically generated with medium confidence

A picture containing table

Description automatically generated

Graphical user interface

Description automatically generated

Text

Description automatically generated

Graphical user interface

Description automatically generated

# Appendix C – Certificate of Ethics Approval

Graphical user interface, text, application, email

Description automatically generated