National College of Ireland

Utilizing Machine Learning to Detect Diabetes Risk

MSc Research Project
AI for Business - MSCAIBUS

Tural Abdullayev
Student ID: X23140399

School of Computing
National College of Ireland

Supervisor: Dr. Muslim Jameel Syed

National College of Ireland

MSc Project Submission Sheet – 2023/2024

School of Computing

Student Name:	Tural Abdullayev		
Student ID:	X23140399		
Programme:	MSCAIBUS	**Year:**	2024
Module:	MSc Research Project		
Supervisor:	Dr. Muslim Jameel Syed		
Submission Due Date:	12/08/2024		
Project Title:	Utilizing Machine Learning to Detect Diabetes Risk		
Word Count:	33 **Page Count** 7610		

"I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project."

Signature: Tural Abdullayev

Date: 12/08/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

"Attach a completed copy of this sheet to each project (including multiple copies)"	□
"Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies)."	□
"You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer."	□

"Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office."

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Utilizing Machine Learning to Detect Diabetes Risk

Tural Abdullayev

X23140399

Abstract

Being in the 21st century, some of the problems facing individuals today should be bygones, and one of them is diabetes. With the current enhanced technologies, individuals should not be spending more money on treating and managing diabetes. This study will assess how Wearable Devices such as watches can be used to manage diabetes. Studies in the past have assessed the impact of diabetes globally. Scaling this down, the impact is based on different factors. This explains why some individuals are more prone to diabetes than others. Some factors include age, gender, type of food consumed, access to healthcare facilities, access to information, and also technological advancement. This study will focus on how technologies such as A.I. and ML can be used in the management of diabetes. This study will use the Pima Indians Dataset (Kaggle, n.d.) as a case study to assess diabetes and its impact on individuals. The Pima dataset has an accuracy of 76%. The methodology used in this study is Knowledge Discovery Databases (KDD). This is a comprehensive methodology that ensures accurate results. Its objective was to conduct intensive research on the implications of wearable devices such as smartwatches in the management of diabetes.

Key Words - Diabetes, ML, A.I., Smart Watch, Wearable Devices

1 Introduction

Diabetes or diabetes mellitus is a health condition resulting from an increment in blood clucose levels (IDF, 2019). In the 21st century, diabetes is considered one of the fastest-growing health concerns globally, considering that the number of people with diabetes over the past two decades has tripled. This is because of increased new cases and prevalence of diabetes. The International Diabetes Federation (IDF) indicates that the number of people infected with diabetes in 2019 was approximately 463 million. According to IDF, if the situation is not dealt with, approximately 580 million people will have diabetes by 2030, and this will increase to over 700 million by 2045 (IDF, 2019). Diabetes has various impacts on individuals' general health as it can lead to other complications such as damaging nerves, stroke, heart attack, leg amputation, and kidney failure and vision loss. It also has an impact on families and the economy since individuals with diabetes might be unable to work, failing to cater to their family's needs as well as support the economy through paying taxes.

According to IDF (2019), diabetes is one of the non-communicable diseases leading to disability and mortality. Various solutions have been implemented to help deal with the increased cases of diabetes, such as insulin, which helps manage glucose levels. Although this is the case, there has been a need to incorporate technology to help manage diabetes. According to the World Health Organization (WHO), no solution can be found to manage and monitor diabetes (WHO, 2017). Although this is the case, smart devices such as smartwatches have effectively predicted blood glucose levels, detected risk events early, and effectively adjusted insulin doses. This can also help increase the patient's quality of life. Due to the changing technology, there is room for improvement to ensure that smartwatches and other AI-generated solutions effectively manage glucose levels and diabetes.

1.1 Motivation and Background

Suppose diabetes is detected at an early stage. In that case, it can help effectively manage the disease and one of the ways of ensuring this is through monitoring blood glucose levels through a diabetes smartwatch. The more the condition is undetected, the worse the diagnosis outcome. Various devices today, such as smart bands and smart watches, are used to monitor diabetes. Although this is the case, more effective, affordable, and easily accessible devices should be invented to help manage diabetes and improve patients' quality of life in a non-invasive manner. Most of them also require assistance from a health expert; thus, it becomes challenging to control the condition in the comfort of their homes. This

study was selected to develop a smartwatch to effectively monitor and manage diabetes and reduce its implications on patients. Individuals without diabetes can also use it to determine their glucose levels and take necessary precautions early enough.

1.2 Research Objectives

Objectives	Description
Objective 1	Effects of wearable devices such as smartwatches on diabetes management
Objective 2	Impact of diabetes on today's society
Objective 3	Use of AI in managing diabetes
Objective 4	Accessing real-time diabetes information using AI

Table 1. Research Objectives

2 Related Work

2.1 Introduction

There are various studies conducted in the past regarding diabetes and strategies that have been used in the past to control diabetes. After conducting extensive research, various relevant studies that inform this study were determined. This section will provide a critical review of the available studies and assess both the studies' negative and positive sides. The subsections (health implications of diabetes, use of A.I. to manage diabetes, and cost of treating and managing diabetes and its complications) will be discussed in detail.

2.2 Health Implications of Diabetes

There are various studies assessing how diabetes impacts an individual's health, and one of them is a study by Latts (2018). This study was conducted on 300,000 patients with type 2 diabetes. The study conducted medical therapy among the selected patients for three months. The study indicated that 31% of patients discontinued their medication, which increased with time as by 6 months, the discontinuation rate was at 44% and 58% by 1 year. One of the main challenges being faced by persons with diabetes is the lack of real-time and crucial health data crucial to influence informed decisions linked to intensive therapy and detailed diabetes management. The study indicates that despite various technological advancements enhancing how diabetes is managed, there is still a long way to go to ensure that individuals do not have to pay much money to control or prevent diabetes.

The other challenge hindering the effective use of technology to manage diabetes is the rapid expansion of medical knowledge. As a result, the technologies need to be updated regularly to incorporate the new medical knowledge available. The other study is research by Manaf et al. (2020) that assesses the quality of health and life of individuals with diabetes. According to the study, it is crucial to assess the health-related quality of life among individuals with type 2 diabetes. The study indicated a moderate health-related quality of life among patients with type 2 diabetes. Thus, while managing diabetic patients, the patient's quality of life should be a priority. Various factors should be considered while enhancing health-related quality of life, such as gender and age, among other factors. This is related to another study by Pham et al. (2020), which assesses the health complications of type 2 diabetes.

According to the Pham et al., (2020), "type 2 diabetes has a negative impact on health-related quality of life. The study's findings indicated that there was significant reduction of physical functioning, role emotional, role physical, social functioning among patients with type 2 diabetes." The study indicated that the health-related quality of life among Vietnamese patients with diabetes was significantly low. This was more profound in mental and social health perspectives. The study recommends that there should be effective strategies that will help prevent diabetic complications and manage diabetes. Another study by Hill-Briggs et al. (2021) assesses the social determinants of health. According to the study, diabetes significantly impacts racial and ethnic minority and low-income adult populations in the United States. The study indicates that there is a shift in population health outcomes, and the social determinants of health are one of the essential strategies that can help achieve health equity.

2.3 Cost of Treating and Managing Diabetes and its Complications

There are various studies on the implications of diabetes; one of the major ones is cost implications. Different studies have assessed the lifetime cost of diabetes to families. Zhao et al. (2014) conducted a study assessing the cost implications of diabetes. The study used the U.S. National Population data representative to approximate the lifetime medical cost of people with diabetes. The data was compared to the lifetime medical cost of people without diabetes. The estimates were made based on various factors such as age of diagnosis, duration, and sex.

The study indicates that after providing a 3% discount on future spending, "the lifetime medical spending for people with diabetes was 124,600, 19200, 53,800, and 35900

when diagnosed with diabetes at ages 40, 50, 60, and 65 years respectively." The individuals found to have higher medical expenditure attributed to diabetes were young females. The study indicates that diabetes comes with a higher medical expenditure despite being linked to decreased life expectancy (Zhao et al., 2014; Yang et al., 2020).

Another study by Yang et al. (2020) indicates that diabetes is a costly affair, and this is because it leads to other complications, which require a significant amount of medical expenses to deal with. This study indicates that a significant percentage of costs associated with diabetes are because of managing the disease and dealing with various complications. This study researched a longitudinal data panel. The study was comprised of approximately 700,000 people with both type 1 and type 2 diabetes. The cost was estimated based on the number of years.

This is closely related to another study by Zhuo, Zhang & Hoerger (2013), which assesses the direct cost of treating type 2 diabetes and its complications. Unlike the other study that focuses on both types of diabetes, this only focuses on type 2 diabetes and also focuses only on new infections. The results indicated that the cost of treating and managing type 2 diabetes among men aged 25-44 and 45-54. 55-64 and over 65 years is $130,000, $110,400, $85,500 and $56,000. The study also indicated that 53% of the costs went to treating diabetes while approximately 57% went to treating diabetes complications. The above studies have indicated that diabetes is a costly health issue, and the majority of the cost goes to managing its complications. Thus, early diagnosis and intervention could help reduce the cost of treating this disease. As a result, accurate ways to detect the condition at an early stage or prevent the disease are required.

Another study by Mapa-Tassou et al. (2019) assesses the economic burden of diabetes in Africa. Africa is one of the top continents where there are a significant number of diabetes cases, and this is coupled with lack of adequate knowledge on how to prevent, detect and treat diabetes. Due to the lack of adequate statistics in different countries, it also becomes a hustle to assess the impact of the disease on particular populations. As a result, the study predicts that the greatest prevalence of diabetes will happen in Africa. The study found that just like in other continents, Africa also spends a significant amount of its expenditure on diabetes.

Although this is the case, there are inadequate studies regarding the same. Thus, there is a need of more studies to ensure that there is enough information and statistics about diabetes, its impact on societies and the amount being spent to control the conditions. The study indicates that the cost of treating diabetes differs from one country to the other. For

example, the national annual costs of diabetes in Nigeria were estimated to be over $1 billion. In Cameroon, individuals with diabetes spend over $148 monthly on the treatment of the same, while in Sudan, treating type 2 diabetes costs Sudanese approximately $175 annually (Mapa-Tassou et al., 2019).

2.4 Use of AI to Manage Diabetes

Artificial Intelligence, over the past few years, has been embraced in various sectors, and the healthcare sector has not been left behind. There are various studies available assessing the use of A.I. to monitor and prevent diabetes. According to Dankwa-Mullan et al. (2019), AI is a subsection of computer science whose aim is to generate methods or systems that assess data and allow the handling of sophistication in a variety of applications. According to this study, considering that individuals with diabetes globally are over 400 million and over 12% of global expenditure goes to treating and managing diabetes, the use of AI is an effective alternative that can help deal with the issue. The study conducted an online search using PubMed and included 450 articles related to the use of AI to manage diabetes.

One of the strengths of this study is that it acknowledges that A.I. is crucial in managing diabetes. According to the author, due to the increased advances in technologies such as A.I., there is hope of getting real-time unstructured and structured health data to cater to individuals with diabetes. One of the challenges today is the lack of adequate and real-time data; thus, it becomes a challenge to know what one should do if they have diabetes or how to administer medications such as insulin without having to seek medical assistance. If this information is available, then managing diabetes could be easy. A.I. can help ensure that there is real-time and accessible data. The other strength of the study is that it provides some of the current innovative approaches that are AI-powered and can help monitor and manage diabetes. Examples given include predictive modeling programs, retinal imaging programs, glucose sensors, smartphone applications, and insulin pumps. Some of the tasks that these AI-powered applications are that they can monitor blood glucose levels, decrease diabetes comorbidities and complications, and reduce hypoglycemic episodes. Thus, the study will help me gain more knowledge regarding what aspects best suit an effective smartwatch that can help monitor and manage diabetes.

Although this is the case, there are also limitations in this study, and one of them is that the study does not give details of how the indicated A.I. solutions can effectively manage diabetes. The study only mentions some of the applications that can help deal with diabetes

but do not give examples of ready AI solutions available in the market, their advantages and disadvantages. This information would be crucial, as it would help an individual to pick the best solution based on their budget what they want the application or gadget to offer and its durability. The other limitation is that since the study used secondary data, the new and more advanced AI solutions were not incorporated. Some of the studies included were conducted more than five years before the study was published. Thus, the chances of assessing even the outdate solutions were high.

The other one is a study by Ellahham (2020). This study indicates that AI is the future of diabetes care. Considering that almost 8% of the world's population by 2017 had diabetes, and the figures are expected to increase to 10% by 2045, there is a need to introduce effective measures that will help deal with the issues. One of the most effective strategies that has shown significant results in dealing with the issue is the use of AI Its use have been effective in reforming the diagnosis and management of chronic diseases such as diabetes. This has been made possible through the use of principles of machine learning (ML) where algorithms are being built to support predictive models that helps determine various factors leading to diabetes or its consequent complications. One of the key areas where AI is being used in the healthcare sector to deal with health conditions such as diabetes include “clinical decision support, automated retinal screening, patient self-management tools and predictive population risk stratification.” According to the study, digital therapeutics have been shown to lead to effective interventions for the management of diabetes, such as the provision of lifestyle therapy.

Another study by Ahmed et al. (2023) indicates that the use of wearable devices (W.D.s) such as watches is very significant when it comes to gathering, storing, transmitting and processing data. As a result, they can effectively be used to predict, manage, treat and assess diabetes. Although this is the case, W.D.s effectively work with the interconnection of other gadgets or services such as smartphones, Wi-Fi or Bluetooth, and cloud computing services. Integration of W.D.s with cloud services is that it helps facilitate patient monitoring by healthcare professionals without the need for hospitalization. W.D.s effectively use different sensors: accelerometer, galvanic skin response, photoplethysmography, electrocardiogram (ECG), and near-infrared sensors. They are effective in sensing skin temperature, heart rate, physiological signs among other crucial details that can help manage or prevent diabetes.

3 Research Methodology

3.1 Introduction

This study's purpose is to analyze and develop a smartwatch that will help monitor and manage diabetes. The research procedure selected is Knowledge Discovery in Databases. The features of KDD will be of great help while conducting this research. After understanding the data, accurate analysis and predictions will be conducted. Considering that the information is available, the process of knowledge discovery and data mining will not be sophisticated. The research will be guided by KDD-modified methodology, which contains five stages. They include "selection, pre-processing, transformation, data mining and evaluation or interpretation" (Fayyad, Piatetsky-Shapiro and Smyth, 1996). The stages will be assessed one at a time in the subsections below. The selected data set is a case study of Pima Indians (Kaggle, n.d.) and Diabetes with a sample size of 768 and 9 data attributes. The data attributes include Pregnancy, Glucose, BloodPressure, SkinThickness, Isukin, DiabetesPedegreeFunction, BMI, Outcome and Age.

3.2 Methodology

Knowledge Discovery Databases (KDD) will be used to conduct this study, and one of the main reasons why it was selected is that it will provide feature selection, which is vital in this study. According to Maimon and Rokach (2005), KDD is an exploratory and automatic method used to assess and model large data repositories. It helps to determine novel, valid, understandable and useful patterns from huge and sophisticated data sets. This is another reason why this approach was selected, as it will help determine unknown patterns from the available dataset. Data will be selected from the Pima Indians(Kaggle, n.d.) and Diabetes case study. Considering that the data is saved in a zipped file, the data will be extracted, saved as . CSV and uploaded into Jupyter Notebook. Jupyter Notebook will help access the dataset. After this, the features will be selected, feature engineered, balance the data set using SMOTE analysis, evaluate models and record their difference on balanced and unbalanced datasets and interpret the data with accuracy.

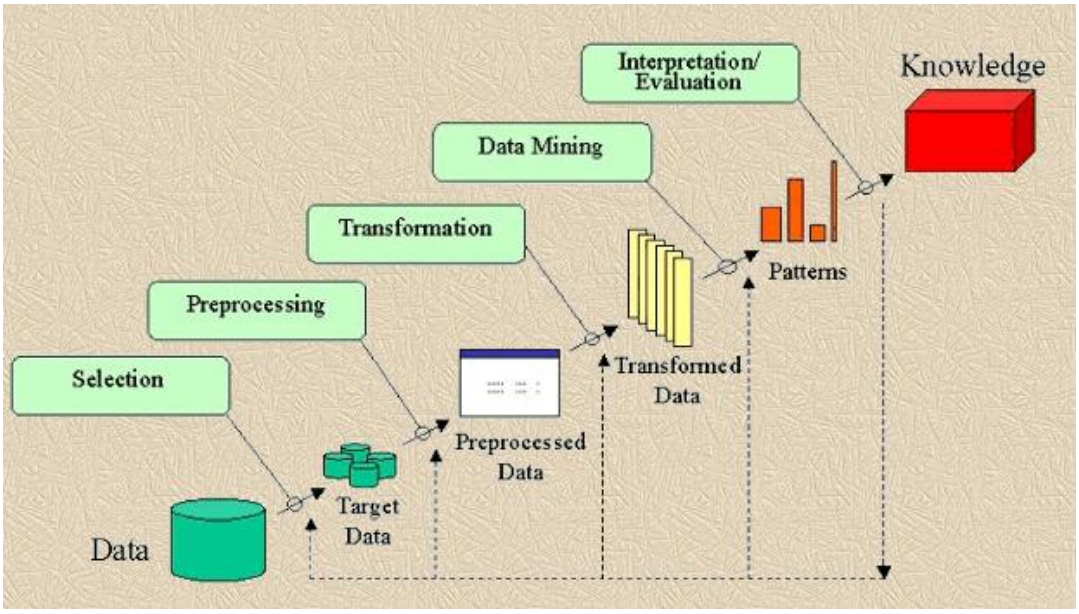


Figure 1. KDD Methodology

3.2.1 Selection of Target Data

The Pima Indians and Diabetes database (Kaggle, n.d.) was selected since it consists of the Pima Indians diabetic dataset (Kaggle, n.d.) and a data.csv file. The study settled on the Pima Indians dataset (Kaggle, n.d.) since it has various factors that make them more prone to diabetics. They live around the Gila and Salt Rivers of Arizona. It is heavily affected due to various factors such as genetics; they consume food rich in fibers and vitamins such as corn, beans, wheat, and pumpkins. Thus, the data will help predict whether a patient is prone to diabetes depending on multiple features. It was challenging to gain insight since the dataset was unbalanced.

3.2.2 Pre-processing

The dataset was gathered to provide an opportunity to assess the historical cause-effect on the Pima females and support it with numerical data. Considering that the data contained both diabetic and non-diabetic data, it need to undergo pre-processing to facilitate data mining. It accounts for over 50% of the duration spent on data. Pre-processing is crucial as it will help get consistent data that will significantly influence the data mining stage and lead to higher results (Duggal et al., 2016).

Numerical Features having too many unique values: It was challenging to gain any insight from the data while visualizing since there were too many numerical features. For the purpose of visualization, the numerical features were divided into two: Part 1 features

(Pregnancies, BloodPressure, SkinThickness, and Age) and Part 2 features (Glucose, Insulin, BMI, DiabetesPedigreeFunction).

Data Cleaning: Its target feature is in categorical format, whereas the input features are in numerical format. Categorical data is grouped into different categories but lacks numerical value or natural order. Although this is the case, one of the main challenges, more so when it comes to machine learning algorithms, is that they understand numbers, not categories. Thus, the categorical data was converted to numerical data to ensure that the data can be easily interpreted. Transforming the data into numerical data ensured that the data could be easily comprehended by machine learning models.

Class Imbalance: The quality of the results, as indicated in the mean values, shows that there is a difference between the cases of diabetes and non-diabetes, and this leads to imbalance. “Class imbalance is an uneven distribution between the majority and minority classes and this causes a bias in the majority class.” SMOTE analysis will be used to balance the dataset and record the difference between the model performances when trained on balanced and unbalanced datasets. According to Saez et al. (2015), SMOTE is an approach that considers the output to be the whole dataset and generates new instances from a minority class. The minority class is increased based on the total classes.

3.2.3 Transformation

The transformation was crucial in converting data with approaches such as feature selection, feature engineering, SMOTE analysis and normalization. Thus, it helps ensure that the data is more consistent.

Feature Division: The study found that some of the numerical features have too many unique values present. Thus, for the purpose of visualization the features were divided into two parts; Part 1 Features: Pregnancies, BloodPressure, SkinThickness, and Age and Part 2 Features: Glucose, Insulin, BMI, DiabetesPedigreeFunction. The analysis also found out that the data was unbalanced by a 2:1 ratio for Non-diabetes and Diabetes cases. The Non-diabetes cases were 500 representing 65.1%, while diabetes cases were 268 representing 34.9%, as indicated in Figure 2 below

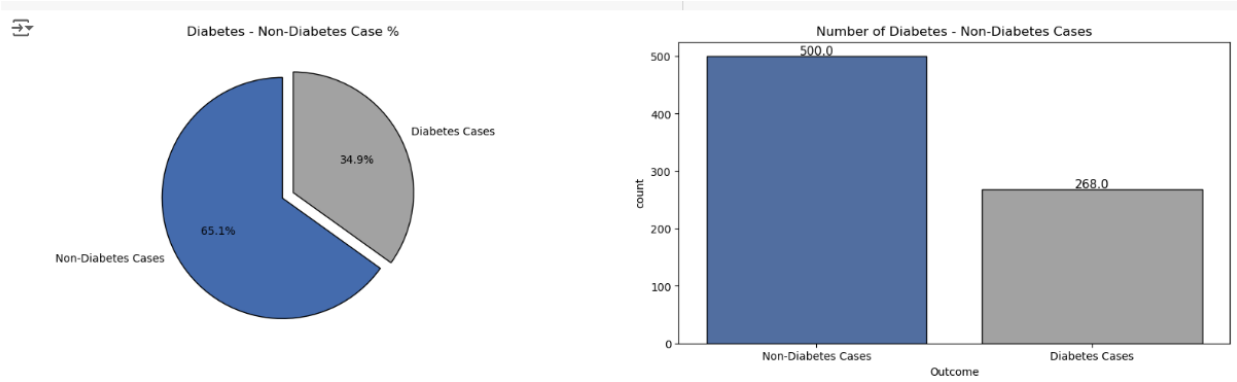


Figure 2. Diabetes-Non-Diabetes Cases (%)

SMOTE Analysis will be used to handle the dataset's class imbalance. The dataset has a 1 2 imbalance. The process of selecting the SMOTE percentage effectively lead to generation of same minority classes and increased the number of nearest neighbour by 1. The role of normalization is to scale numerical values in a dataset. Each feature was rescaled to [0,1] interval using Min-Max. It was conducted based on the algorithm. Normalization is conducted by splitting the features into two, part 1 features were visualized with respect to target variables (Outcome). The next step is visualizing part 2 features with respect to target variables (Outcome).

3.2.4 Data Mining

This section comprises of explanatory data analysis conducted to visualize data, descriptive analysis, linear correlation and modelling for problem classification. This stage was started by making use of Exploratory Data Analysis (EDA) strategy. According to Mueller and Tukey (1980), this technique was introduced by John Tukey in the 60s who focused on creating easy to draw arithmetic and pictures.

Descriptive Statistics: Table 2 below shows the "dataset's significant numbers such as standard deviation, mean, max and min for each numeric variable."

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.00	3.85	3.37	0.00	1.00	3.00	6.00	17.00
Glucose	768.00	120.89	31.97	0.00	99.00	117.00	140.25	199.00
BloodPressure	768.00	69.11	19.36	0.00	62.00	72.00	80.00	122.00
SkinThickness	768.00	20.54	15.95	0.00	0.00	23.00	32.00	99.00
Insulin	768.00	79.80	115.24	0.00	0.00	30.50	127.25	846.00
BMI	768.00	31.99	7.88	0.00	27.30	32.00	36.60	67.10
DiabetesPedigreeFunction	768.00	0.47	0.33	0.08	0.24	0.37	0.63	2.42
Age	768.00	33.24	11.76	21.00	24.00	29.00	41.00	81.00
Outcome	768.00	0.35	0.48	0.00	0.00	0.00	1.00	1.00

Table 2. Descriptive Statistics

3.2.5 Interpretation

This section interprets the Mean values of each feature for both non-diabetes and diabetes cases. The average number of Pregnancies determined for non-diabetes and diabetes cases were 3.30 & 4.87, respectively. The average Glucose values for diabetes cases is 109.98. This feature is perfect for forming an indicator to estimate the cases of diabetes. The results indicated that there was no significant in the average values of BloodPressure and SkinThickness among non-diabetes and diabetes cases. There was a higher average in Insulin level, as was the case with Glucose for diabetes compared to non-diabetes, which was at 68.79. There were also higher values for diabetes cases than non-diabetes cases for BMI, which were 35.14 and 30.30, respectively.

In DiabetesPedigreeFunction, the average was higher in diabetes cases than in non-diabetes at 0.55 and 0.43, respectively. The Mean values for diabetes and non-diabetes cases were 37.07 and 31.19, respectively. From the indicated mean values, it is evident that various features show precise differences between diabetes and non-diabetes cases. Although this is the case, more effort is needed to determine the link between some of the features and the target variable. Assessing the statistics is crucial in determining the data types that need modification. "During this analysis categorical data were transformed to numerical data. This led to classification of the data into two features."

Linear Correlation Matrix

Linear correlation was used to determine the relationship between variables. Plotting the data on correlation metrics will help visualize the relationship between variables. The

Pearson correlation scale ranges from -1 to 1 where 1 has a strong positive correlation and -1 has a strong negative correlation. 0 signifies there is no linear relationship between the variables. Pearson correlation is sensitive to outliers and performs best with clean, normally distributed numeric data.

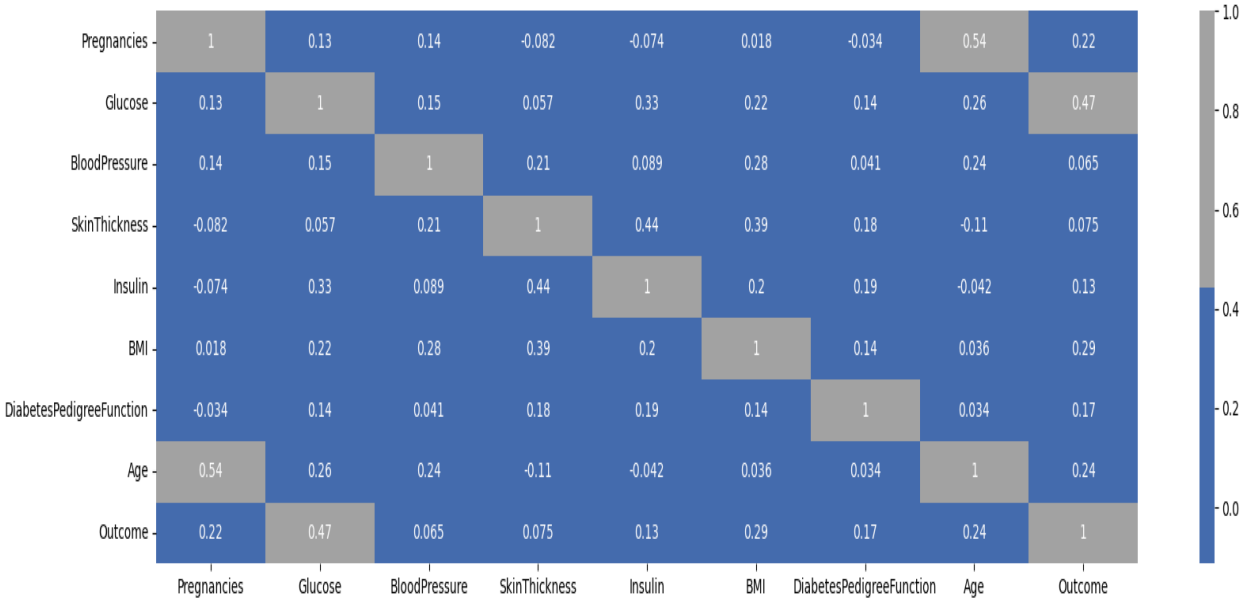


Figure 3. Linear Correlation Matrix

4 Implementation

4.1 Dataset Narrative

The dataset selected is a case study of Pima Indians and Diabetes(Kaggle, n.d.), and it contains numerous features. Data set formation was influenced by the selection of the main features of interest, as the objective of the research was to predict whether a patient is prone to diabetes depending on multiple features. The data set offers an opportunity to assess the historical cause-effect on the Pima Females and back it with numerical data. The necessary libraries were imported, and the overview of the main dataset attributes is shown in Figure below.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.60	0.63	50	1
1	1	85	66	29	0	26.60	0.35	31	0
2	8	183	64	0	0	23.30	0.67	32	1
3	1	89	66	23	94	28.10	0.17	21	0
4	0	137	40	35	168	43.10	2.29	33	1

Table 3. Dataset Narrative

A deep copy of the original dataset was created and label encoding the text data of the categorical features was conducted. The modifications made on the original dataset were not highlighted on the dep copy. Thus, the deep copy dataset which incorporated all the features converted into numerical values was used for visualization and modelling purposes.

4.2 Exploratory Data Analysis

This was comprised of conducting a preliminary analysis of the data and assessing the distribution and data normality, identifying outliers, and assessing the association between features visually. The process began by dividing features into numerical and categorical. Categorical features were defined if the attribute had less than 6 unique elements. If this was not the case, it was considered as a numerical feature. The distribution of categorical features was conducted and indicated in the Figure below

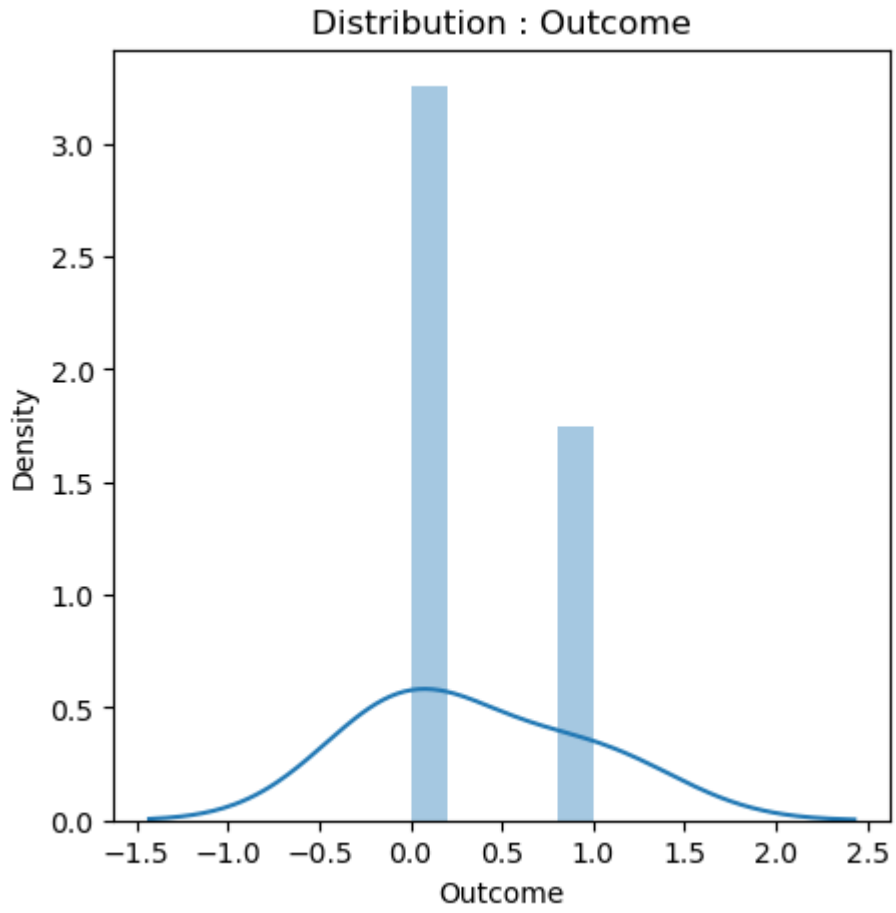
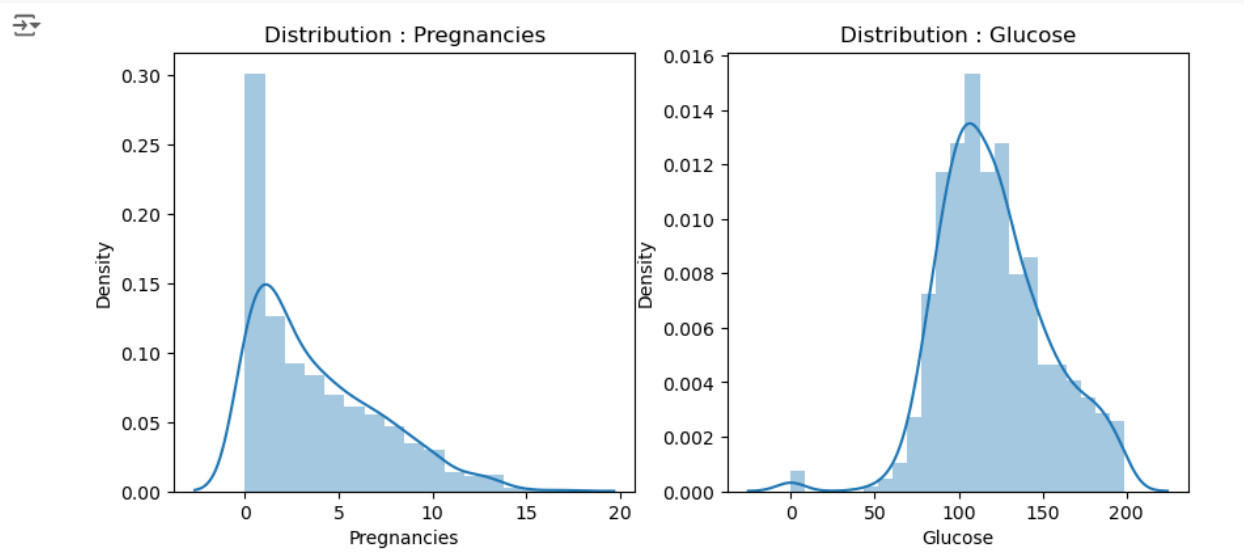
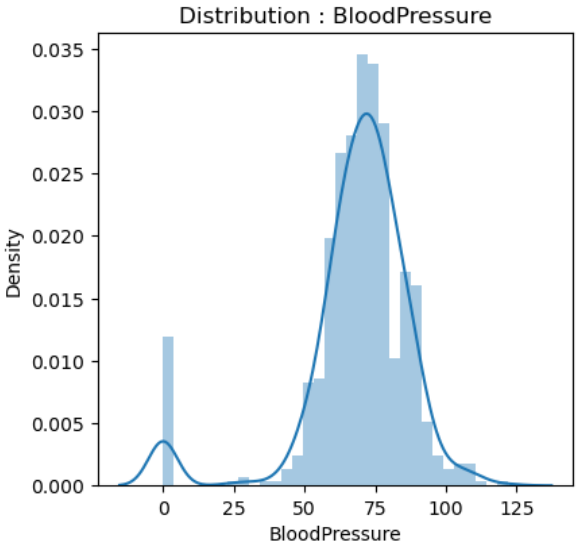
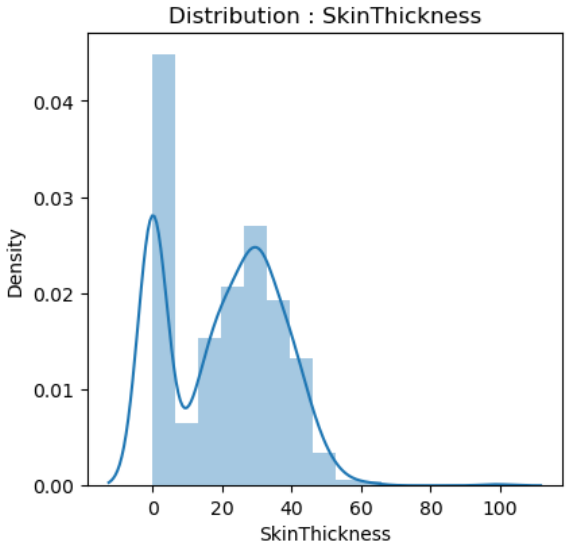


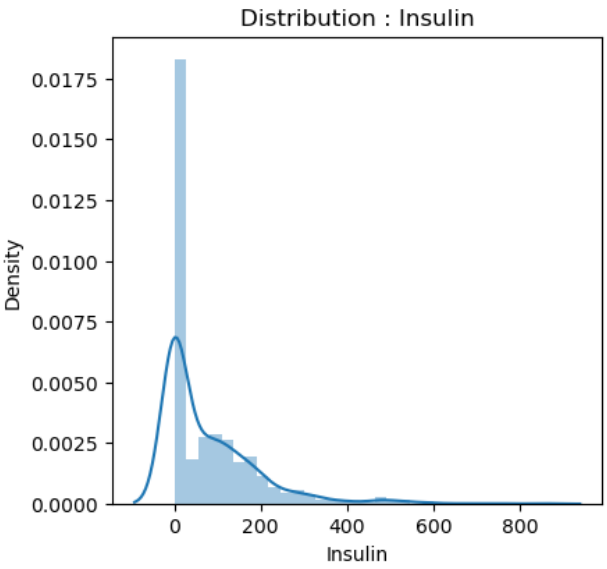
Figure 4. Distribution of Categorical Features

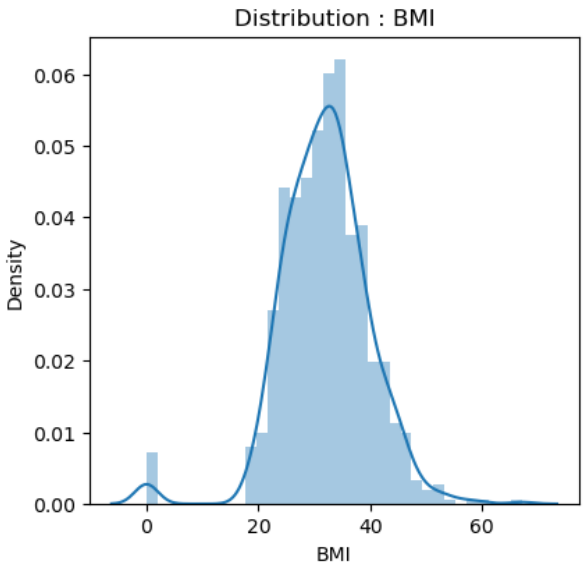
The outcome shows normally distributed data. It was followed by the distribution of numerical features

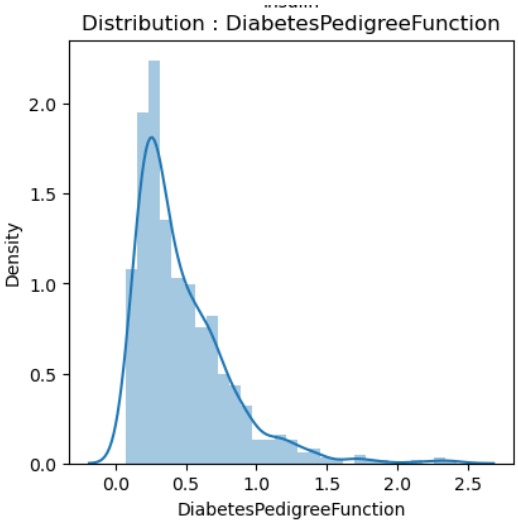



Distribution : BloodPressure
Density
BloodPressure


Distribution : SkinThickness
Density
SkinThickness


Distribution : Insulin
Density
Insulin


Distribution : BMI
Density
BMI


Distribution : DiabetesPedigreeFunction
Density
DiabetesPedigreeFunction

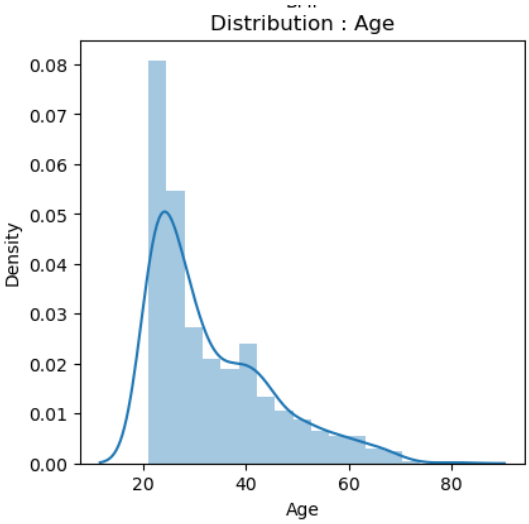

Distribution : Age
Density
Age

Figure 5. Distribution of Numerical Features

Pregnancies, Insulin, DiabetesPedigreeFunction and Age were found to have positively or rightly skewed data distribution. BloodPressure and Skin Thickness displayed a bidmodal data distribution. Data distributions of Glucose & BMI were a bit tricky. This is because they nearly highlighted a normal distribution or bimodal distribution. This is because of the small peak present at the value 0. The distributions were considered as bimodal.

4.2.1 Target Variable Visualization (Outcome)

The next step was visualizing the target variables. Considering that the dataset had an unbalanced ratio of 2:1 for Non-diabetes: Diabetes cases, it would be challenging to visualize the data. The other factor that was considered was the numerical features that had too many unique values. The features were divided into two for the purposes of visualization. Figure 6 below shows a visualization of part 1 features w.r.t. outcome.

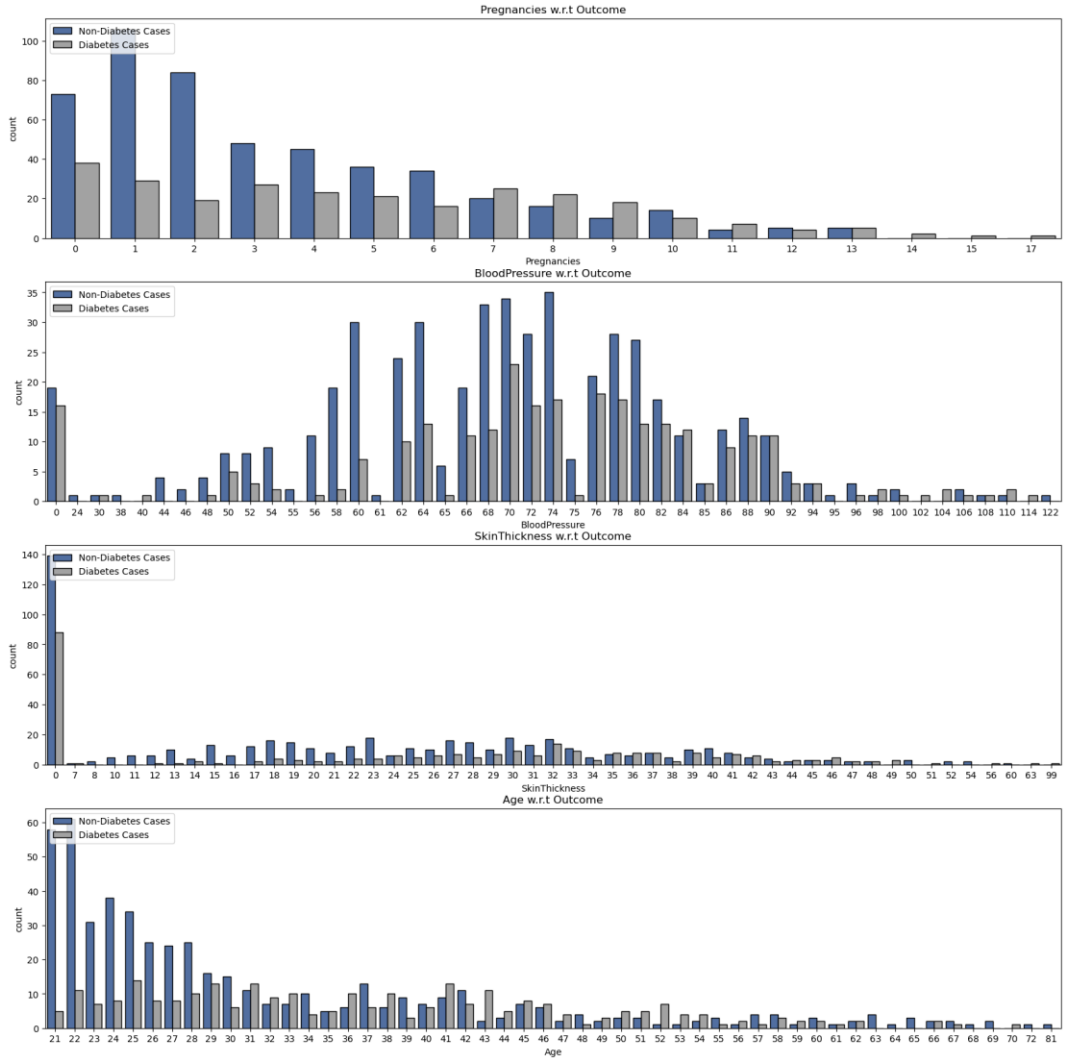


Figure 6. Visualizing Part 1 Features w.r.t Target Variables (Outcome)

For pregnancies, diabetes cases are present throughout the data. Thus, there is no particular range of values for which higher diabetes cases are found. Pregnancies range of values from 7-9 highlights more cases of diabetes than non-diabetes for the first, showing a pattern, but this case is rejected if values ahead are observed. BloodPressure values ranges between 60-90 mm/hg and shows a high number of diabetes patients. SkinThickness shows a low number of diabetes cases for all the values. Out of the values indicated, 24-42 has some prominent peaks of diabetes cases. When assessing the data based on Age, young women are more prone to diabetes than older women. The number of diabetes cases decrease with age. The data indicates that based on Age group 21-50 displays a higher probability of being diagnosed with diabetes.

The next step is visualizing part 2 features. Since the dataset has too many unique data points, they will be converted into categorical features for the purpose of visualization and

gaining insights. The values of the features were scaled down to attain constant value and varied data points. The data points of the numerical features were divided and “assigned the resulting value as the representative constant for that data point. The scaling constants of 5, 10 & 100 are decided by looking into the data & intuition.”

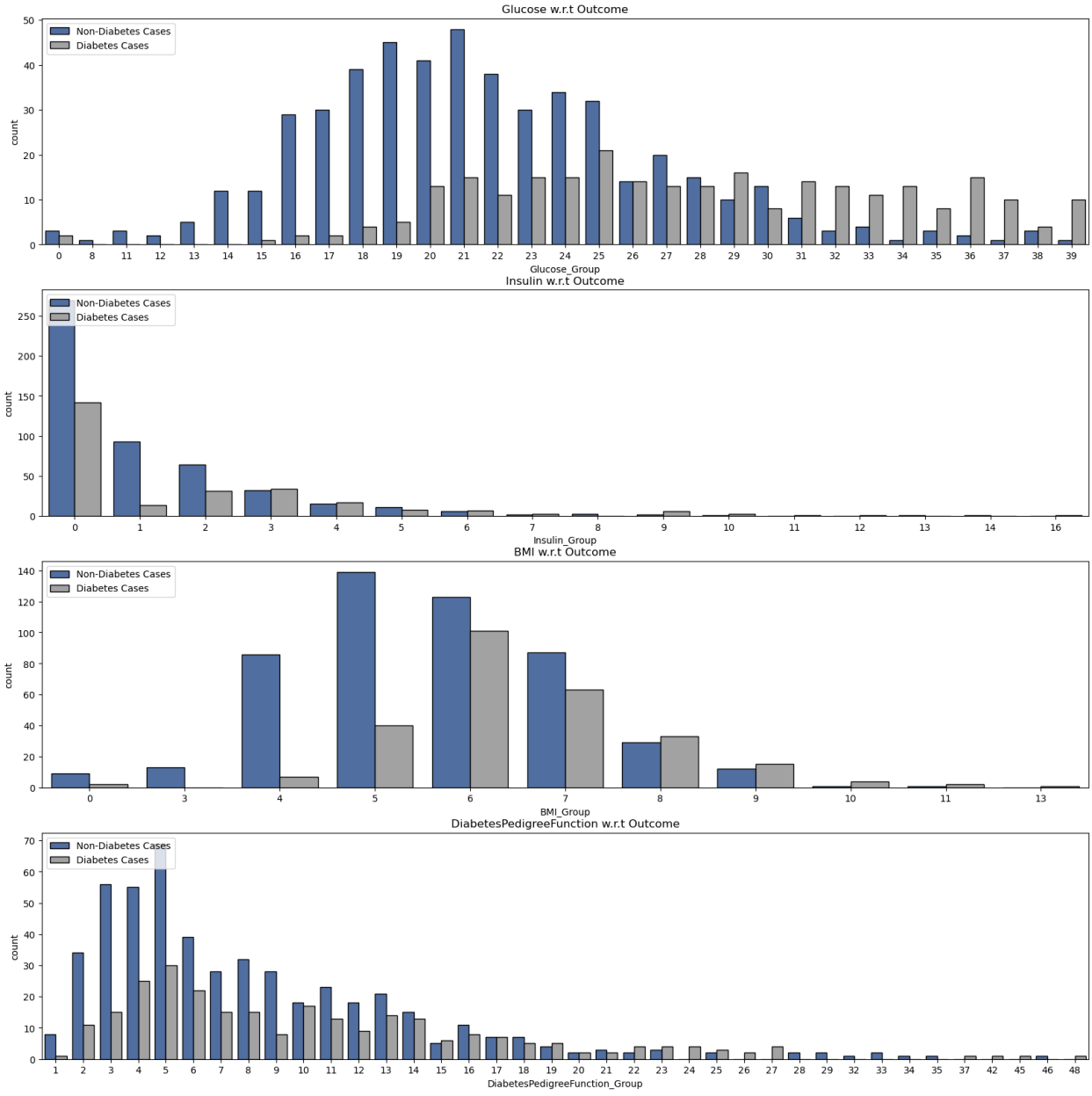
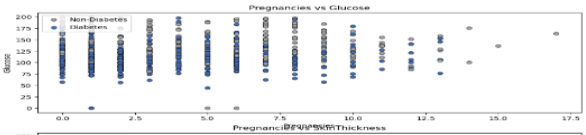
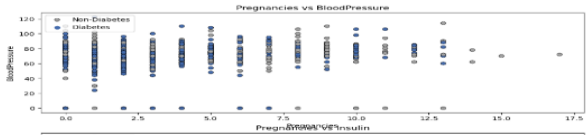


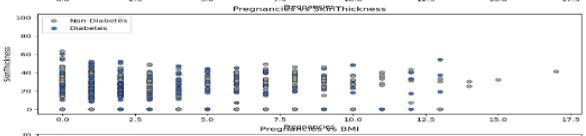
Figure 7. Visualizing Part 2 Features w.r.t Target Variables (Outcome)

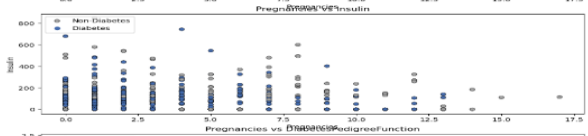
From the Glucose group data, early values display very low number of diabetes cases. Cases then increase and become constant starting from 100 (20x5) - 195 (39x5). Insulin levels between 0 (0x50) - 300 mu U/ml (6x10) are highly susceptible to diabetes. Insulin's data distribution displays a high number of diabetes cases for value 0. This is probably because of the no data recorded for those females. For the BMI readings, diabetes cases have a higher probability for the range of values from 20 (4x5) - 45 kg/m^2 (9x5).

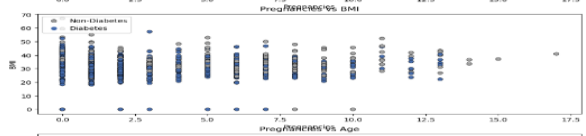
DiabetesPedigreeFunction values also display positive diabetes cases throughout. 0.1 (2x5/100) - 1.25 (25x5/100) range of values have detected higher number of diabetes cases.

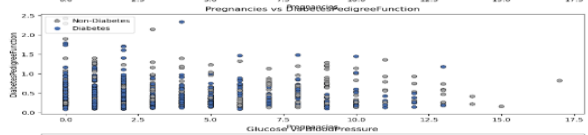


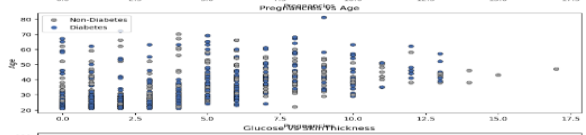


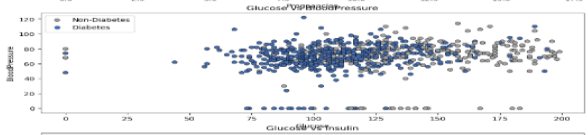


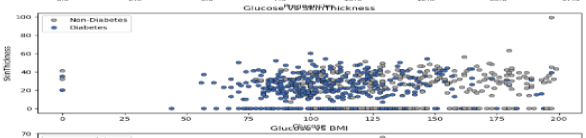


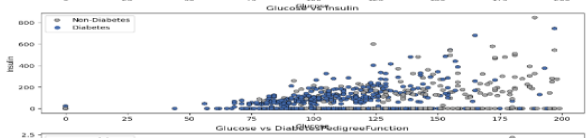


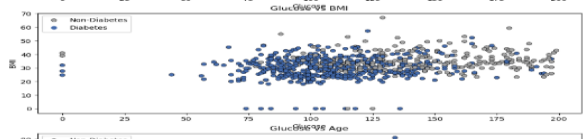


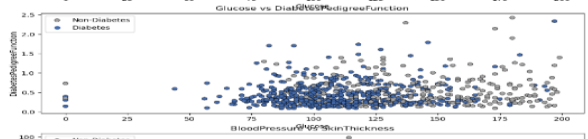


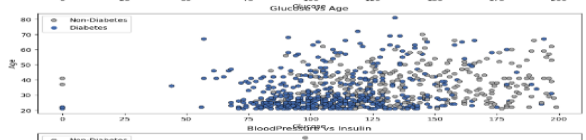


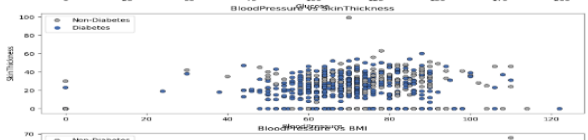


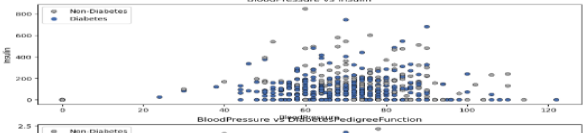


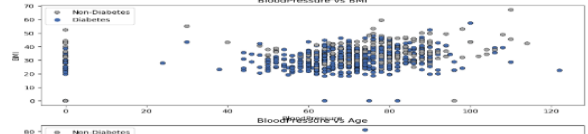


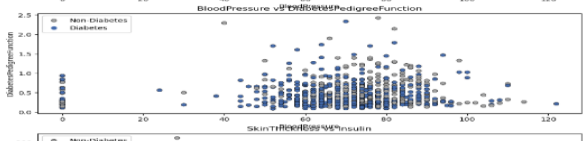


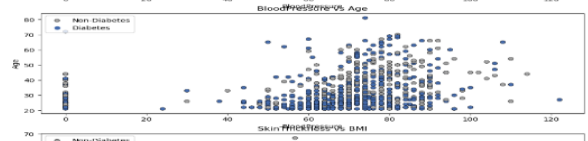


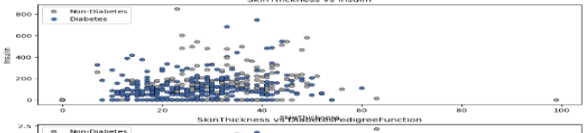


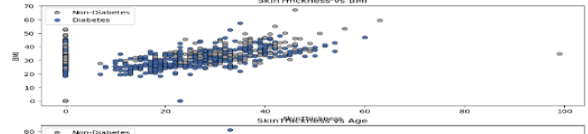


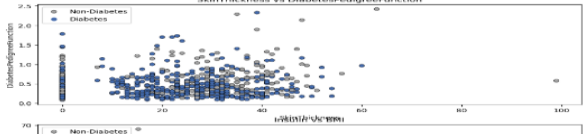


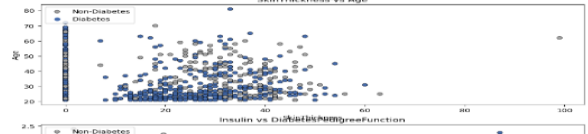


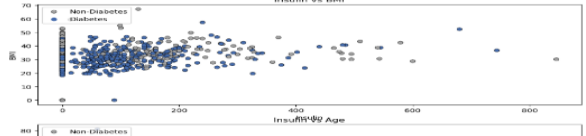


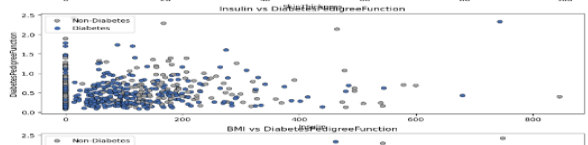


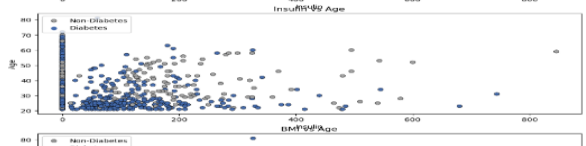


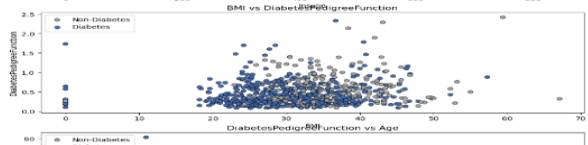


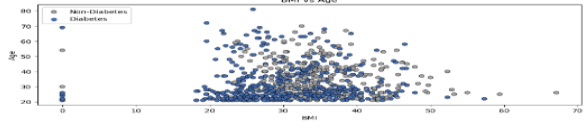












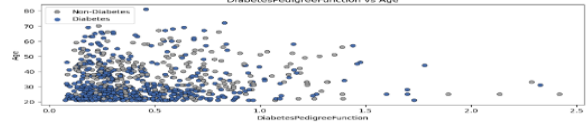


Figure 8. Numerical Features vs Numerical Features w.r.t Target Variable (Outcome)

Pregnancies with values between 7-10 have high chances of diabetes. This range does not display a complete dominance but it has some presence. Glucose values higher than 125 indicate very high chances of diabetes. BloodPressure values 60-100 highlight many cases of diabetes coupled with any feature. When both, BMI and SkinThickness, feature values are 20 - 50, probability of diabetes is very high. Insulin values between 0 - 300 increases the risk of diabetes. When above 400, the female has a sure shot chance of being diabetic. For Age group of 20 - 50 as well as DiabetesPedigreeFunction values ranging from 0 - 1.5 results in a diabetic condition.

4.2.2 Feature Engineering

The next step was feature selection and feature engineering was used. Considering that we have a gap between the raw data and algorithm, it is crucial to bridge the gap to ensure that the data can be led by machine learning. The features units of value are not understood by the ML model. The model treats the input as mare number but does not understand the actual meaning of its value. Thus, it is crucial to scale the data. Two different options were used to scale the data which included normalization and standardization. Considering that majority of algorithms assumes that is normally distributed, normalization was used to scale data. “It was done for features whose data does not display normal distribution and standardization is usually conducted for features that are normally distributed but their values are huge or very small compared to other features.”

Normalization: Pregnancies, Insulin, DiabetesPedigreeFunction and Age features are normalized as they displayed a right skewed data distribution. BloodPressure, Skin Thickness, Glucose & BMI highlight a bidmodal data distribution.

Standardization: None of the features are standardized for the above data. Correlation matrix was used to check their correlation with respect to outcome considering that the matrix was too huge with too many features as indicated in Figure 9 below

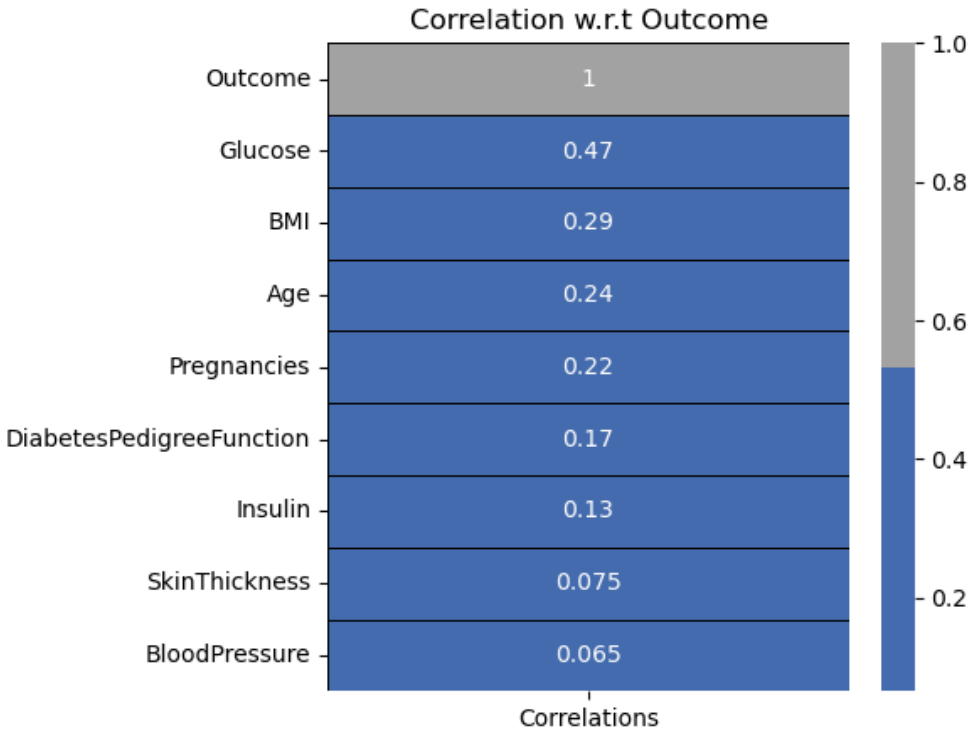


Figure 9. Correlation w.r.t Outcome

Some of the features that did not display any kind of correlation include SkinThickness and BloodPressure. Glucose displayed the highest positive correlation with respect to Outcome followed by BMI, Age, Pregnancies, DiabetesPedigreeFunction and Insulin.

4.2.3 Feature Selection for Numeric Features: ANOVA Test

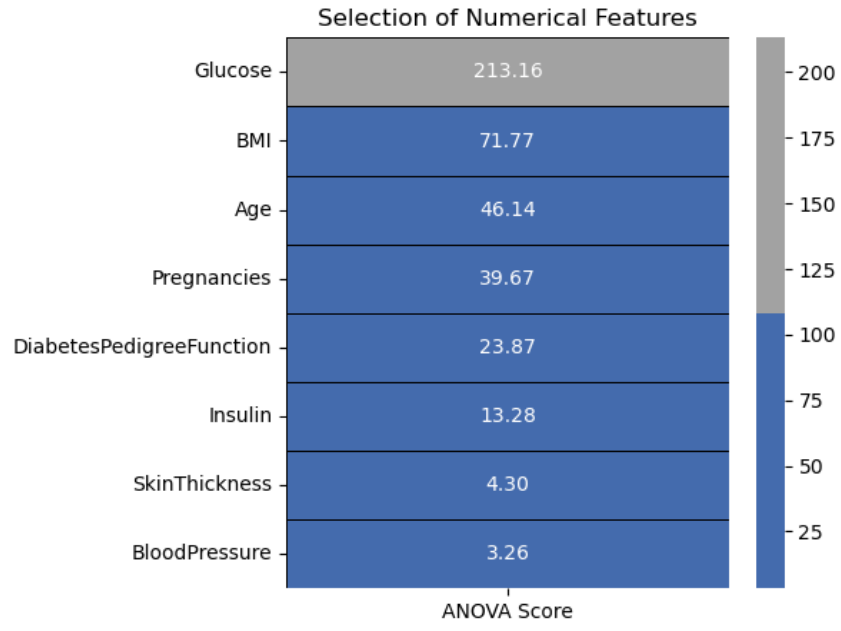


Figure 10. Selection of Numeric Features using ANOVA Test

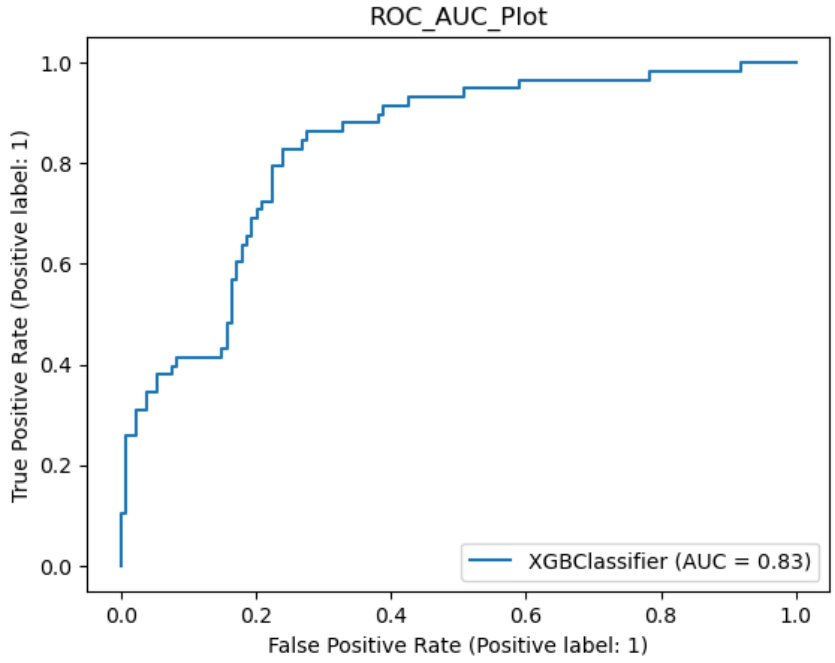
ANOVA test will help determine the importance of each feature based on the "ANOVA score. According to the ANOVA test, the higher the value of ANOVA score, the more is the importance of the feature. From the above results, we need to drop SkinThickness & BloodPressure. We will take the remaining features into consideration for modeling."

4.2.4 Data Balancing using SMOTE

The data need to be balanced and SMOTE technique will be used for data balancing. There are two options that can help balance unbalanced data. One of them is undersampling where majority of the samples of the target variable are trimmed down and the second one is oversampling where the minority samples of the target variable are increased to the majority samples. This study made use of oversampling the minority class. Imblearn was used for data balancing. This Pip statement was used to install imbalanced-learn "pip install imbalanced-learn." Due to the use of synthetic data, the models were not assessed using accuracy. Thus, the data was duplicated and using accuracy model would be misleading. ROC-AUC was used to provide the relation between True Positive and False Positive rate.

4.2.5 Modelling

Confusion Matrix and ROC-AUC graph were used for model evaluation. ROC-AUC provided the relation between True Positive and False Positive rate. Figures from above conducted tests were selected and the data split into 75-25 train – test groups.



	precision	recall	f1-score	support
0	0.80	0.84	0.82	134
1	0.58	0.52	0.55	58
accuracy			0.74	192
macro avg	0.69	0.68	0.68	192
weighted avg	0.73	0.74	0.74	192

Figure 11. Unbalanced Dataset's ROC_AUC Graph

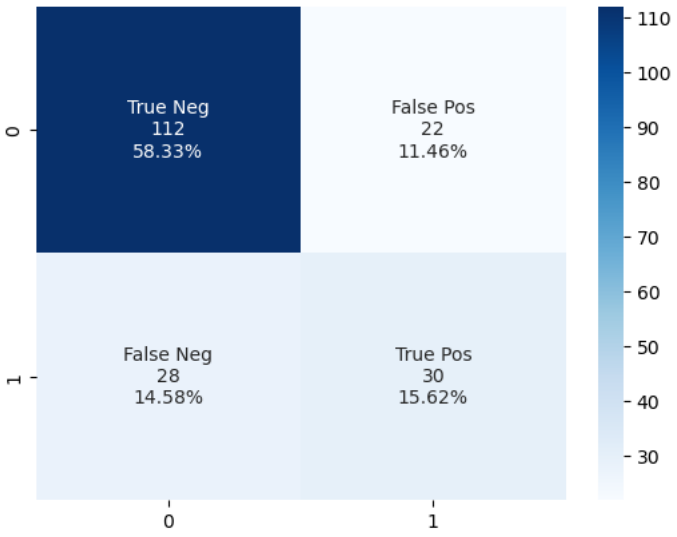


Figure 12. XGBClassifier Unbalanced Dataset (Confusion Matrix)

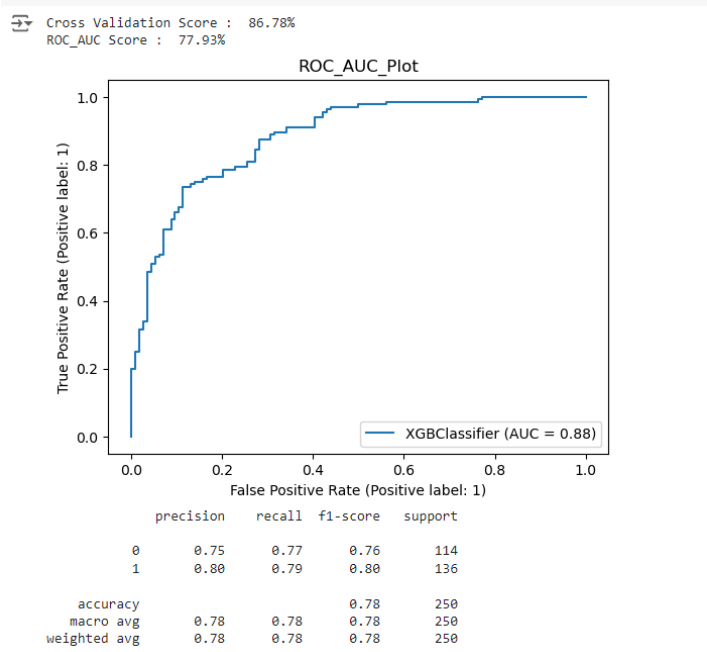


Figure 13. XGBClassifier Balanced Dataset (ROC-AUC Graph)

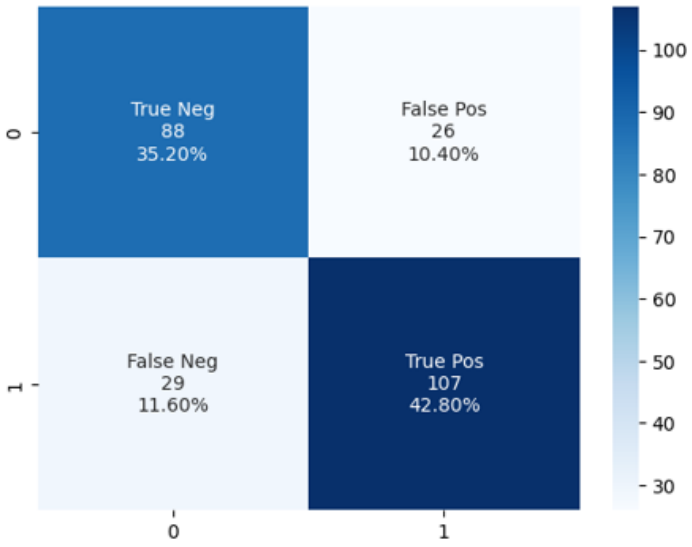


Figure 14. Stack of XGBClassifier and LightGBMClassifier

For the stacking of classifiers XGBClassifier and LightGBMClassifier were stack. It has an important hyperparameter known as final_estimator. It is the final classifier that makes the final prediction by using the predicted classes by the various classifier and predicts the final output.

Unbalanced Dataset :

Sr. No.	ML Algorithm	Cross Validation Score	ROC AUC Score
1	XGBClassifier	83.56%	67.65%
2	LightGBMClassifier	82.36%	64.20%
3	Stack of XGBClassifier & LightGBMClassifier	80.41%	70.38%

Balanced Dataset :

Sr. No.	ML Algorithm	Cross Validation Score	ROC AUC Score
1	XGBClassifier	86.78%	77.93%
2	LightGBMClassifier	86.96%	78.30%
3	Stack of XGBClassifier & LightGBMClassifier	84.51%	78.01%

Figure 15. Unbalanced and Balanced Dataset Algorithm

This is a great dataset that helps us to look into a very special case study and look through the lens of data. Its target feature is in categorical format, whereas the input features are in numerical format. Insights gained from the EDA section and the domain information match alot with a few mismatches. The domain information is generalized data, whereas the EDA insights gathered are based on a demographic area. SMOTE analysis is used for data balancing. Tree-based models are trained to highlight the performance difference when trained on unbalanced and balanced data. The performance of LGBM and XGB is neck and neck; however, the stack of the 2 classifiers did not outperform the other classifiers! Entirely, “a clear cut performance improvement can be observed when the models are trained on the balanced dataset.”

5 Evaluation

5.1 Introduction

This section will offer a comprehensive assessment of the research's outcomes and main findings, and its implications, both from an academic and practitioner perspective, will be presented. This section will also assess the PIMA Indian Database (Kaggle, n.d.) case study and provide a detailed discussion of the study.

5.2 Experiment / Case Study 1

A dataset from the UC-Irvine Machine Learning repository (Pima Indian Diabetes) dataset was used for this study's comparative analysis (Blake and Merz, 1998). The dataset contains

768 cases and 9 numeric attributes, including SkinThickness, blood pressure, Glucose, BMI, Age, Pregnancies, DiabetesPedigreeFunction, Outcome and Insulin. “The target variable, diabetes onset within 5 years, had a binary outcome (0,1). Class 0 show healthy patients while class 1 show patients with diabetes as indicated in Figure. Model statistics with 86% confidence interval and plot of influential predictors based on ROC measure are presented in Figure.” XGBClassifier and LightGBMClassifier were used for stacking classifiers, and final_estimator was used as the final classifier, which was used to make the final prediction by using the predicted classes.

The algorithms of the unbalanced and balanced datasets were different. For the unbalanced dataset, XGBClassifier had a cross-validation score of 83.56% and an ROC-AUC score of 67.65%. LightGBMClassifier had a cross-validation score of 82.36% and a ROC-AUC score of 64.20%. The stack of XGBClassifier and LightGBMClassifier had a cross-validation score of 80.41% and 70.31%. For the balanced dataset, the XGBClassifier had a cross-validation score of 86.78% and an ROC-AUC score of 77.93%. LightGBMClassifier had a cross-validation score of 86.83% and an ROC-AUC score of 78.30%. The stack of XGBClassifier and LightGBMClassifier had a cross-validation score of 84.51% and an ROC-AUC score of 78.01%. The details are shown in Figure.

5.3 Discussion

Diabetes is one of the chronic conditions that have a severe impact not only on individual and their families but also on communities and the whole world in general. Various strategies are beingVarioutodaytegies are b, being use,d daily to diagnose prevent and manage diabetes. Although this is the case, prevention and management of diabetes is still an issue today. Technologies such as A.I. and ML have been highly adapted to the management of diabetes. Although this is the case, there is need of introducing more advanced wearable devices that will help detect, prevent and manage diabetes. To have a vivid picture of the current issue PIMA Indians Diabetes Database (Kaggle, n.d.) was used to analyse the dataset. This the dataset provided an opportunity to assess the historical cause-effect on the PIMA females and it was backed with numerical data. The data set was effective in predicting whether a patient is prone to diabetes depending on the 9 dataset attributes.

Although this is the case, the dataset was collected a few decades ago; thus, the study would have utilized a more recent dataset that would accurately reflect the challenges individuals with diabetes face today. The study utilized a KDD methodology and its main advantage in this study is that it provides feature selection. This methodology was crucial in

data mining and heavily relied on the study's objective. It helped determine different forms of attributes.

The attributes will be crucial while developing a diabetes W.D. that will be effective in managing diabetes (Razavian et al., 2015). ANOVA test was used to determine the importance of the features. The test indicates that the higher the value of the ANOVA score, the more important it is as indicated in Figure. From the results the features with least importance were SkinThickness and BloodPressure which were dropped and the remaining features were put into consideration. The most important features were selected for the purpose of enhancing learning algorithm. Although this is the case, dropping the two was not an accurate move since they are important while developing W.D.s for managing diabetes. Thus, in future, studies should include the two to ensure that they effectively include features that are crucial while developing wearable devices used to manage diabetes.

For the purposes of visualization, the dataset was split into numerical and categorical features. This was for visualization and modelling. After data visualization, the data set had an imbalance of 2:1 ration for Non-diabetes: Diabetes cases. This could have led to a biased predictions towards Non-bias, making it challenging to gain insights. Data balancing was conducted using SMOTE analysis. There are two options one can choose from when data balancing and they include undersampling and oversampling. Considering that for this study the target variable is the minority sample, oversampling was selected. This involved increasing the minority samples of the target variable to the majority samples. Accuracy was not used to evaluate the models since synthetic data was used. For model evaluation confusion matrix, ROC-AUC graph score were used for determining the relation between True Positive and False Positive rate.

6 Conclusion and Future Work

One of the fastest growing chronic disease is diabetes. Although there are various efforts to enhance the diagnoses, prevention and treatment of the disease, there is need of more enhanced solutions to effectively manage diabetes. Machine learning and AI have been effectively adopted in the healthcare sector as there are various commercial solutions that have been introduced to diagnose, prevent, and assist in the treatment of various disease. Diabetes is not an exception as there are multiple solutions such as W.D.s and smarphone applications available. Although this is the case, there is need of more enhanced solutions that will keep with the ever-changing technologies and the new medical practices being introduced. There are various studies in the past that have assessed the same and based on the past studies, there are various recommendations.

One of them is that, researchers should analyse how real-time information can be used with the wearable devices for easy assessment and treatment of patients. Cloud could also be utilized to ensure that the information can be accessed with ease. Effective use of the wearable devices could help reduce the high mortality rates of diabetes and enhance quality of life. This can be attained if the M.D.s are effective and accurate in preventing and managing the condition.

Integration of W.D.s with cloud services is that it helps facilitate patient monitoring by healthcare professionals without the need of hospitalization. W.D.s effectively use deffirent sensors accelerometer, galvanic skin response, photoplethysmography, electrocardiogram (ECG), and near infrared sensors. They are effective in sensing skin temperature, heart rate, physiological signs among other crucial details that can help manage or prevent diabetes. Based on the assessed information, there study would provide an accurate and more enhanced watch that assesses the patient's history, access real-time information to assess and manage diabetes.

References

Ahmed, A., Aziz, S., Alzubaidi, M., Schneider, J., Irshaidat, S., Serhan, H. A., ... & Househ, M. (2023). Wearable devices for anxiety & depression: a scoping review. Computer Methods and Programs in Biomedicine Update, 3, 100095.

Dankwa-Mullan, I., Rivo, M., Sepulveda, M., Park, Y., Snowdon, J., & Rhee, K. (2019). Transforming diabetes care through artificial intelligence: the future is here. Population health management, 22(3), 229-242.

Duggal, A., Pinto, R., Rubenfeld, G., & Fowler, R. A. (2016). Global variability in reported mortality for critical illness during the 2009-10 influenza A (H1N1) pandemic: a systematic review and meta-regression to guide reporting of outcomes during disease outbreaks. PloS one, 11(5), e0155044.

Ellahham, S. (2020). Diabetes and its associated cardiovascular complications in the Arabian Gulf: challenges and opportunities. J Clin Exp Cardiol, 11, 1-5.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996, August). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In KDD (Vol. 96, pp. 82-88).

Hill-Briggs, F., Adler, N. E., Berkowitz, S. A., Chin, M. H., Gary-Webb, T. L., Navas-Acien, A., ... & Haire-Joshu, D. (2021). Social determinants of health and diabetes: a scientific review. Diabetes care, 44(1), 258.

IDF Diabetes Atlas. Key Figures from the IDF Diabetes Atlas 9th Edition. 2019. Available online: https://web.archive.org/web/20211208190021/https://diabetesatlas.org/ (accessed on 24 March 2021).

IDF Diabetes Atlas. Worldwide Toll of Diabetes. 2019. Available online: https://web.archive.org/web/20211118111050/https://www.diabetesatlas.org/en/sections/worldwide-toll-of-diabetes.html (accessed on 24 March 2021).

Latts, L. (2018). ADA/IBM Watson Health Study (N> 300,000) finds that nearly 60% of people with T2D discontinue therapy after one year. American Diabetes Association 78th Scientific Session.

Maimon, O., & Rokach, L. (2005). Introduction to knowledge discovery in databases. In Data mining and knowledge discovery handbook (pp. 1-17). Boston, MA: Springer US.

Mapa-Tassou, C., Katte, J. C., Mba Maadjhou, C., & Mbanya, J. C. (2019). Economic impact of diabetes in Africa. Current diabetes reports, 19, 1-8.

Manaf, H., Harvey, W. S., Armstrong, S. J., & Lawton, A. (2020). Differences in personality and the sharing of managerial tacit knowledge: an empirical analysis of public sector

managers in Malaysia. Journal of Knowledge Management, 24(5), 1177-1199.Mapa-Tassou, C., Katte, J. C., Mba Maadjhou, C., & Mbanya, J. C. (2019). Economic impact of diabetes in Africa. Current diabetes reports, 19, 1-8.

Pham, L. A. Huynh, G., Tran, T. T., Do, T. H. T., Truong, T. T. D., Ong, P. T., & Nguyen, T. N. H., (2021). Diabetes-related distress among people with type 2 diabetes in Ho Chi Minh City, Vietnam: prevalence and associated factors. Diabetes, Metabolic Syndrome and Obesity, 683-690.

Saez, A., Rigotti, M., Ostojic, S., Fusi, S., & Salzman, C. D. (2015). Abstract context representations in primate amygdala and prefrontal cortex. Neuron, 87(4), 869-881.

WHO (2017). 'WHO | Global report on diabetes', WHO. World Health Organization. Available at: http://www.who.int/diabetes/global-report/en/ (Accessed: 1 December 2018).

Yang, D., Cheng, B., Chen, J., Peng, A., Yang, C., ... & Huang, K. (2020). Clinical characteristics and outcomes of patients with diabetes and COVID-19 in association with glucose-lowering medication. Diabetes care, 43(7), 1399-1407.

Zhang, B., Kumar, R. B., Dai, H., & Feldman, B. J. (2014). A plasmonic chip for biomarker discovery and diagnosis of type 1 diabetes. Nature medicine, 20(8), 948-953.

Zhuo, X., Zhang, P., Barker, L., Albright, A., Thompson, T. J., & Gregg, E. (2014). The lifetime cost of diabetes and its implications for diabetes prevention. Diabetes care, 37(9), 2557-2564.

Zhuo, X., Zhang, P., & Hoerger, T. J. (2013). Lifetime direct medical costs of treating type 2 diabetes and diabetic complications. American journal of preventive medicine, 45(3), 253-261.

Kaggle. (n.d.). Pima Indians Diabetes Database. Retrieved from https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database