Predicting Diabetes Using Machine Learning: A Comparative Study of Supervised Classification Models

By

Mahzebin Pushpo
19216002

A thesis submitted to the Department of Mathematics and Natural Sciences in partial fulfillment of the requirements for the degree of
B.Sc. in Mathematics

Department of Mathematics and Natural Sciences
Brac University
Summer 2023

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. I have acknowledged all main sources of help.

Student's Full Name & Signature:

Mahzebin Pushpo
19216002

Approval

The thesis/project titled "Diabetes Classification Using Machine Learning: A Comparative Study of Supervised Classification Models" submitted by

Mahzebin Pushpo (19216002)

of Summer, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Mathematics on August 29, 2023.

Examining Committee:

Supervisor:
(Member)

_____
Dr. Mohammad Rafiqul Islam
Professor
Department of Mathematics and Natural Sciences
Brac University

Program Coordinator:
(Member)

_____
Dr. Syed Hasibul Hassan Chowdhury
Professor
Department of Mathematics and Natural Sciences
Brac University

Departmental Head:
(Chair)

_____
Dr. A. F. M. Yusuf Haider
Professor
Department of Mathematics and Natural Sciences
Brac University

Abstract

Diabetes is a primary worldwide health concern that can develop at any age and has serious consequences. It results from imbalanced glucose levels in the body. As well as being a long-term disease, it has other associated risks, from life-threatening problems to financial loss. So, it is essential to correctly detect this condition as soon as possible to mitigate further complications. Due to developments in medical technology, many tools are available today for diagnosing diseases. To ensure faster predictions and diagnosis of patients, one such tool known as **machine learning (ML)** algorithms is used. It is a section of Artificial Intelligence (AI) that replicates a human's learning process to train a system. In this study, the algorithms used to predict diabetes patients are supervised classification ML algorithms like **Logistic Regression, K-Nearest Neighbor, Naïve Bayes, Decision Tree, and Random Forest**. The data used is primary data, which is collected from Bangladeshi adults from different age groups. It consists of all the demographic data, medical history, and family information necessary for the study. The dataset is collected and cleaned for repetition and errors. From these data, diabetes status is taken as the dependent variable, and the associated risk factors are the independent variable. Then, the model is deployed using the RapidMiner tool. The confusion matrices for each model are also produced, and a comparative analysis is carried out. After evaluating their performances, the highest accuracy achieved was 94.62% and 94.23%. From these findings, the best model can be determined. This selection of the ideal model is useful because it will help in the proper and timely identification of patients in the future in the healthcare sector so that treatment can be done to curb the disease.

Dedication

I would like to dedicate this thesis to my late grandmother, my late grandfather, and my parents.

Acknowledgment

First, I want to express my utmost gratitude to the Almighty Allah for allowing me to give my best to complete my thesis successfully.

Next, I would like to thank my supervisor, **Dr. Mohammad Rafiqul Islam**, Professor, Department of Mathematics and Natural Sciences, Brac University, who provided assistance and guidance throughout the study. His contribution to working with the datasets to help develop the model and produce the corresponding error matrices played an important role in the completion of this thesis.

I am extremely grateful to **Dr. Quamrun Nahar**, Principal Research Officer, Department of Clinical Pharmacology, BIRDEM General hospital. She helped me collect the data of diabetes patients by sharing my questionnaire with her patients and students. I also want to thank others who aided me in data collection, especially my family and friends.

Lastly, I want to show my deepest appreciation to my parents for being patient with me, and constantly supporting and motivating me to keep working.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Background

Diabetes is a long-term metabolic disorder, as defined by The World Health Organization (WHO), that changes how the body uses glucose (a type of sugar) from food and stores it. It develops when the body has trouble responding to insulin or the pancreas fails to produce enough of this hormone (*Diabetes*, n.d.-a).

The pancreas secretes the hormone insulin, which plays a key role in maintaining stable blood sugar levels. When we eat, our body breaks down carbohydrates into glucose, which is taken into the bloodstream. This glucose is transported into the cells to be stored as energy with the help of insulin.

Diabetes results from insulin deficiency or insulin resistance that can raise blood sugar, a condition known as hyperglycemia, creating health issues if untreated.

There are two main types of diabetes:

1. **Type 1 diabetes:** Type 1 diabetes typically contracts at younger ages but can also develop at any stage in life. It is known by names such as **juvenile-onset diabetes** or **insulin-dependent diabetes mellitus (IDDM).** In this type of diabetes, the pancreatic cells that make insulin are attacked and killed by the immune system (*What Is Diabetes?*, 2023). Thus to control their blood sugar, type 1 diabetes patients must inject or use an insulin pump.

2. **Type 2 diabetes:** Type 2 diabetes is the most prevalent form of the disease, making up about 90% of all cases (*What Is Diabetes?*, 2023). Typically diagnosed in adults, it is increasingly being diagnosed in children and adolescents. In this type of diabetes, the body becomes insulin-resistant or produces insufficient insulin to maintain normal blood sugar levels. The primary step in treating type 2 diabetes includes lifestyle changes, such as diet and exercise, taking oral medicines, or insulin therapy.

Further, there is **gestational diabetes**, which happens during pregnancy (*Diabetes*, n.d.-a); **monogenic diabetes**, which results from mutations in a single gene; and **diabetes brought on by taking certain drugs**.

Diabetes may be predicted using **machine learning (ML)** classification algorithms by studying a dataset of patient characteristics for patterns that are diagnostic of the condition. Some popular classification ML systems for diabetes prediction include **Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor (K-NN),** and **Naïve Bayes**. The selection of an algorithm is contingent on the nature of the problem and the available data. Testing numerous algorithms to find the best one for a task is typical.

1.2 Motivation

Nearly 422 million individuals around the world have diabetes, and the number of deaths per year due to the disease is 1.5 million (*Diabetes*, n.d.-b). The International Diabetes Federation (IDF) reports that the number of adults diagnosed with diabetes in Bangladesh in 2021 amounted to approximately 13.1 million individuals. This figure is expected to rise to 16.8 million by 2030 and nearly double to 22.3 million by 2045. In addition, it had been projected that in 2021, 43.5% of the country's population will have undiagnosed diabetes (*Bangladesh Diabetes Report 2000 — 2045*, n.d.). The reason for conducting this thesis is motivated by this rising prevalence of diabetes and its consequential effects on individuals and healthcare systems globally. This prominent and enduring medical condition, if not identified or adequately controlled, can give rise to significant adverse outcomes like blindness, kidney failure, heart attacks, stroke, and lower limb amputation (*Diabetes*, n.d.-a). The timely and precise detection of diabetes is of utmost importance to minimize subsequent complications and facilitate quick intervention, personalized treatments, and efficient management tactics. Usually, doctors prescribe blood tests to diagnose diabetes in a patient, i.e., they rely on three parameters: blood sugar level while fasting, glucose tolerance, and regular blood sugar levels at any point in time. (Cox & Edelman, 2009; *Diabetes Testing*, 2023). However, sometimes this practice may not always be prompt. Machine Learning algorithms have demonstrated the potential to expedite the prediction and diagnosis of diabetes. Additionally, by early detection, it can function as a valuable point of reference for medical professionals (Zou et al., 2018). This research endeavors to use ML algorithms to make a valuable contribution to advancing an effective and dependable system for the classification of diabetes. Such a system can potentially assist healthcare professionals in immediate interventions and improved patient management.

1.3 Objective

This thesis aims to construct a resilient and reliable predictive model for diabetes through supervised machine learning classification algorithms. To be more specific, other goals are to accomplish the following:

1. Compile and format an extensive dataset with essential features and variables linked to diabetes, and apply suitable machine learning algorithms to them.
2. Acquire an in-depth knowledge of the machine learning algorithms and investigate their performance.

3. Conduct a comparative analysis of the outcomes and select the one with the best performance.

Attaining these goals will allow this thesis to significantly contribute to healthcare analytics by delivering a robust model for diabetes prognosis. Selecting the ideal model can also aid medical experts, policymakers, and others in proper decision-making.

1.4 Thesis Orientation

In this thesis, there are a total of five chapters. **Chapter 1** discusses the thesis topic's background, motivation, and objective. **Chapter 2** introduces machine learning and gives ideas about the algorithms used and their theory. **Chapter 3** explains the methodology. This chapter explores the details of the research design, datasets used, the performance metrics, and tool used to model building and model evaluation. **Chapter 4** presents the results and its comparative analysis. Finally, **Chapter 5** concludes the thesis by summarizing the entire study.

Chapter 2 Algorithms and Theory

2.1 Machine Learning

First, let us begin by giving a brief overview of machine learning and its background. Artificial Intelligence (AI) is becoming more mainstream in today's technologically advanced society. AI uses rules-based algorithms to boost machine intelligence. Machine learning is a fast-expanding area of AI with many practical applications (Babcock University et al., 2017) in finance, healthcare, retail, data security, autonomous vehicles, image processing, computer vision, and more (Choudhary & Gianey, 2017). Presently, these machine learning algorithms are used in virtually every aspect of the digital world. Data has multiplied throughout time, making it critical to track it to make important choices efficiently. For this purpose, machine learning techniques are indispensable. The usage of machine learning algorithms is common in data mining, which refers to analyzing large amounts of data to uncover hidden links and patterns from big commercial databases containing lists of important records (Mitchell, 1997). Finding this pattern helps one to anticipate or focus on critical information to solve an issue (Berry et al., 2020). Hence, machine learning has a hype in the rapidly expanding discipline of data science.

Alan Turing, an English computer scientist, and mathematician, devised the "Turing Test" to evaluate if a computer is intelligent (Turing, 1950). To succeed, a computer must fool a human into thinking it is human as well. Arthur Samuel, a prominent figure in the field of machine learning, coined the phrase "machine learning" (Samuel, 1959). As an employee at "IBM" in 1959, he created a software that became proficient in defeating him in the game of checkers. Tom Mitchell, another machine learning specialist, provided a thorough formulation in 1998 with a Well-posed Learning Problem: Learning by a computer program occurs when their ability to perform a task T, as defined by the symbol P, improves over time due to accumulated experience E (Mitchell, 1997). Over the years, several breakthroughs have used artificial intelligence. As time has shown, these innovations outperform humans.

Now we can finally explain what machine learning is. The learning methodologies encompass the retention of new logical information, the attainment of proficiency in motor and cognitive abilities by training or execution, the transformation of fresh data into meaningful generalizations, and the empirical acquisition of new information (Carbonell et al., 1983). Machine learning is a field that seeks to analyze the many forms of learning processes, with the goal of formulating mathematical and computational models to represent these processes (Carbonell et al., 1983). It is the branch of AI that allows machines to emulate human learning and improvement over time (Edeh et al., 2022) [7]. Basically, it is the science of training computers to respond by continuously providing them with data and allowing them to pick up a few techniques from these past data without being explicitly programmed (Choudhary & Gianey, 2017). In turn, they gain experience and outperform humans in every aspect. The system is trained using either new or existing data to improve the

model's parameters during the learning process. However, the model may be parametrized to a certain extent. While the descriptive models are implemented to obtain insight into past occurrences or current states, the predictive ones are applied for forecasting upcoming events or outcomes. Since inference from samples is the primary goal of machine learning, the theory of statistics is used to develop mathematical models. Building an algorithm on an accurate prediction rule is difficult. Learning has two main functions: First, we need large-scale storage and fast processing methods to handle the vast amounts of data typically available during training and solve the optimized issue. Second, its representation and inference procedure must be effective after the model is learned. Predictive accuracy and the learning or inference algorithm's efficiency (in terms of space and time complexity) may be equally essential in some scenarios.

Some key steps in machine learning include:

- Importing relevant data and then cleaning it
- Splitting this adjusted data into training and testing sets
- Creating a model, i.e., selecting an appropriate algorithm to analyze the data
- Training the model and making predictions

Machine learning algorithms are grouped into four groups. **Supervised and unsupervised learnings** (Sarker, 2021) are the most important types. While **semi-supervised** and **reinforcement learning** (Sarker, 2021) are the other types, the main focus of this study is on supervised learning algorithms.

2.2 Supervised Learning:

Supervised learning involves feeding machines labelled data, where the inputs are independent features, and the goal output is dependent (Foster, 2021; Sarker, 2021). Humans tell the machine the input (typically vectors) and the expected result. In supervised learning, a "teacher" provides a training set of (X, Y) pairs. So, there is a labelled training data, using which a function is derived. A collection of training examples constitutes the training data. A supervised learner attempts to forecast the desired output of a function using input items. This prediction is based on the provided training examples.

Two main types of problems can be solved using supervised learning: **classification** (prediction of object's class) and **regression** (having a continuous output). Again, for each of these problems, further division of algorithms exist, which are linear regression, logistic regression (DeMaris, 1995), k-nearest neighbor (Silverman & Jones, 1989), support vector machine (SVM), decision tree (Rokach & Maimon, 2005) and many more. The diagram (Software, 2019) illustrates the mechanism behind supervised learning.

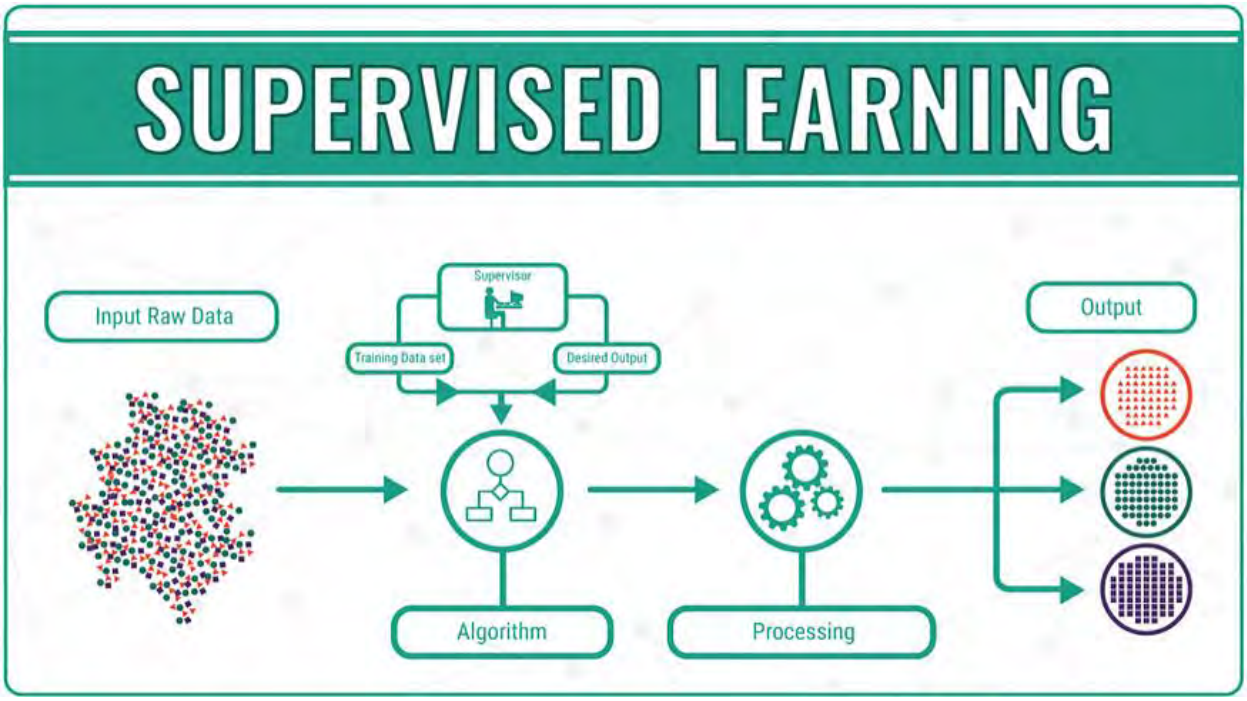


Figure 2.1: Supervised Learning Method

Classification provides a method for organizing data by separating it into distinct categories based on observable similarities. In mathematical terms, the process consists of a mapping function (*f*) that converts initial data (*X*) into the results (*Y*), which may be the groups or classes (Sarker, 2021).

2.3 Unsupervised Learning:

To teach a model in unsupervised machine learning, it does not rely on labels or a predefined output variable. Instead, the program must identify data patterns and relationships (Foster, 2021). In an unsupervised learning setting, knowledge acquisition may be achieved by focusing on the *Xs,* representing the input data, and an overall performance assessment function. The dataset consists of an experimental collection of vectors with no associated functional values. Due to the lack of labeling in the given cases, the learner cannot check the correctness of the structure produced by the corresponding algorithm. According to Hofmann, supervised learning tasks require labels in data, which can be established with unsupervised learning methods (Berry et al., 2020). Again, under unsupervised learning, there are two categories of problems: **association** and **clustering**. Unsupervised learning algorithms also have several categories, some of which are k-means clustering and hierarchical clustering. The mechanism behind unsupervised learning is described below in the diagram (Software, 2019).

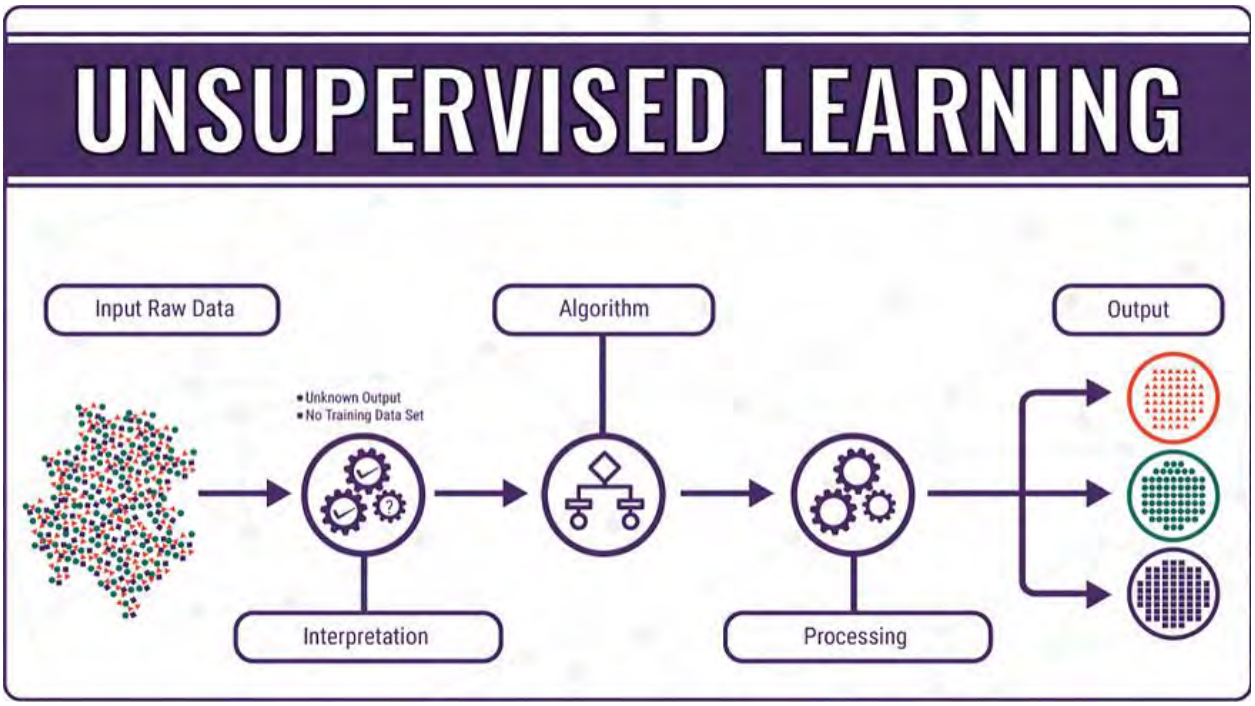


Figure 2.2: Unsupervised Learning Method

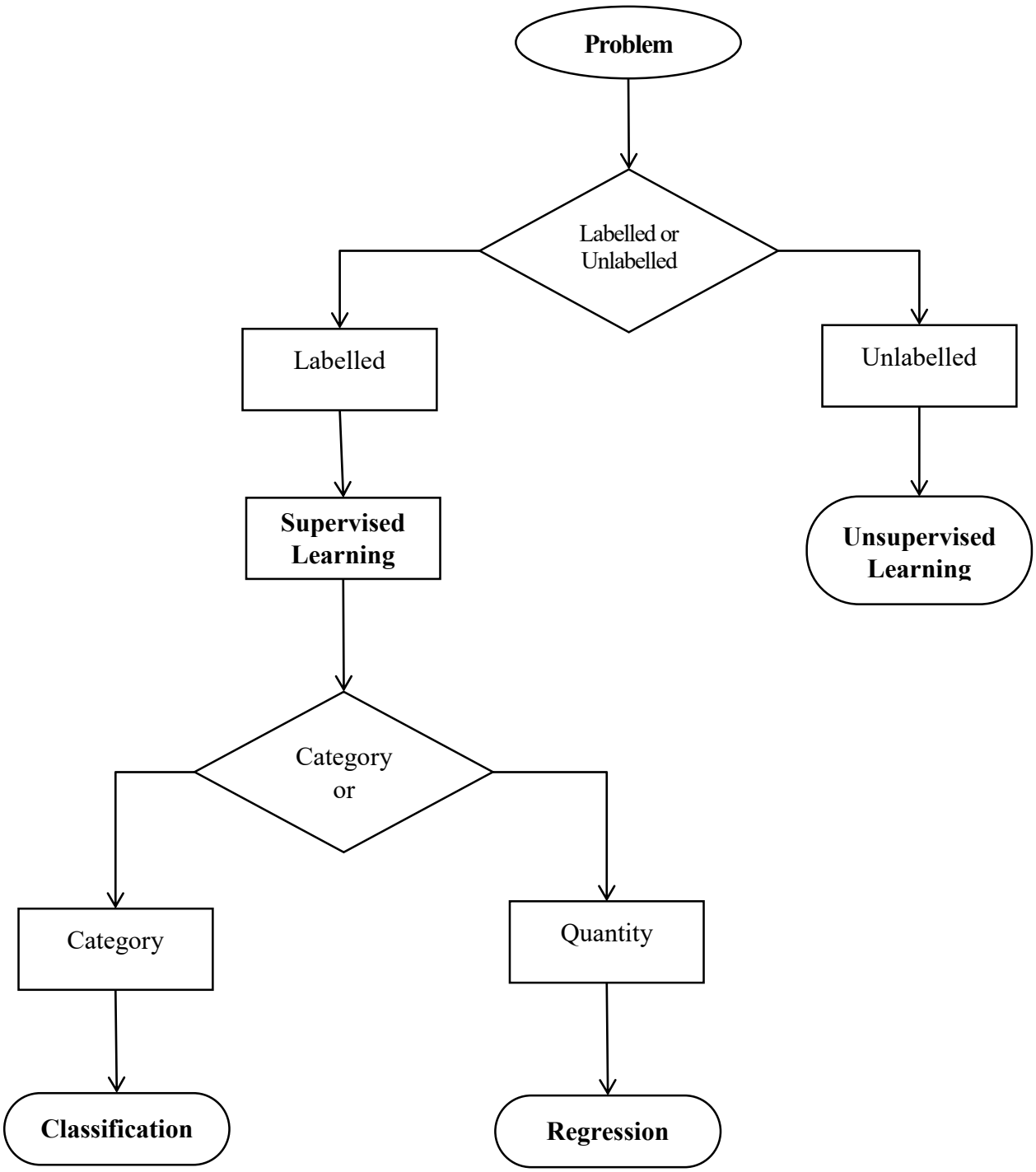


Figure 2.3: Flowchart for addressing a problem with supervised and unsupervised learning

2.4 Logistic Regression

The idea behind logistic regression, first put forth by Berkson in 1944, is to have linear regression generate probabilities. Analytically, logit models are used to generate a linear combination of explanatory factors *X* and the response variable *Y* via a **logit transformation** (Kotsilieris et al., n.d.). In applied statistics and the study of discrete data, logistic regression remains a widely employed technique.

In recent times, logistic regression has surpassed other analytic techniques for multivariate modelling of qualitative dependent variables (DeMaris, 1995). It is a supervised learning classification algorithm that enables one to express the degree to which the existence of a risk factor raises the probability of an outcome. So the type of value logistic regression estimate is **discrete**, and the model is of a **deterministic statistical** type. By considering a collection of independent variables, it calculates the likelihood of an event occurring that can only take on one of two possible values, such as yes, true, success, etc. (indicated by 1) or no, false, failure, etc. (indicated by 0). Therefore, the primary objective of logistic regression is to identify the best model that is capable of explaining how the dependent variable is effected by a collection of independent variables.

Logistic regression is used to build a classification function, and the typical method specifies the location of the class boundary, with class probabilities varying with the distance from the boundary (Babcock University et al., 2017). When the data collection is larger, the value tends to skew more quickly toward the extremes 0 and 1 (Babcock University et al., 2017). In addition, because the outcome is a probability, the dependent variable is also binary or binomial, with values stretching from 0 to 1. Thus, such a regression model is sometimes called the binary logistic model.

In a linear regression model, the hypothesis is a linear combination of feature variables given as follows:

$$z = b_0 x + b_1$$

where $z>1$ and $z<0$

This function needs to be transformed so that the output is characterized into two categories. As the output must be a probability, the value cannot exceed 1 or go below 0. For transformation, Logistic Regression uses an **activation function** to map any real value between 0 and 1. Overfitting occurs easily in high-dimensional datasets, but it excels when the data can be neatly divided along linear dimensions. The **logistic function**, or the **sigmoid function**, estimates the

probabilities. It uses an S-shaped curve, which is a logical choice for modelling. The equation for this can be stated as follows:

$$P(Y = 1) = h(x) = f(z) = \frac{1}{1 + \ exp^{-z}} = \frac{1}{1 + \ exp^{-(}b_0 x + b_1)}$$

P(Y=1) represents the probability of the occurrence of *Y*; then *P(Y=0)* is the complementary. This can be denoted using an odds ratio, which is given below:

$$odds = \frac{P(Y = 1)}{P(Y = 0)} = \frac{P(Y = 1)}{1 - P(Y = 1)} = \frac{f(z)}{1 - f(z)}$$

This **odds ratio** represents the probability of success by the probability of failure. It takes on positive values in the range *(0, ∞)*. As it can be seen, the range is constrained. Limiting the range of a variable makes it more challenging to model. Thus, the natural logarithm is applied to the odds, which is a mathematical operation that lies within the range *(-∞, +∞)*. Then it becomes known as the **natural logarithm of the odds** and is shown below:

$$\ln(odds) = ln\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right)$$

The right-hand side of this equation represents the **logit** of *Y*. Then we can write,

$$f(z) = \frac{1}{1 + \ exp^{-(b_0 x + b_1)}}$$

$$f(x) = Logit(b_0 x + b_1)$$

$$f(z) = ln\left(\frac{f(z)}{1 - f(z)}\right) = b_0 x + b_1 = Logit(z)$$

Thus, we have the sigmoid function *f(z)* which is our logit function and hypothesis for a Logistic Regression model, where *f(z)* is between 0 and 1, i.e., *(0,1)*. The function is employed to estimate

label probabilities. This S-shaped sigmoid curve has a roughly linear section in the center and a curvature when *X* becomes extremely small or very high values (DeMaris, 1995).

The diagram below shows how a 1-dimensional classifier can be trained using logistic regression. All the green points at the top are the positive examples, while those at the bottom are the negative examples. The decision boundary divides data into two groups (shown in red).

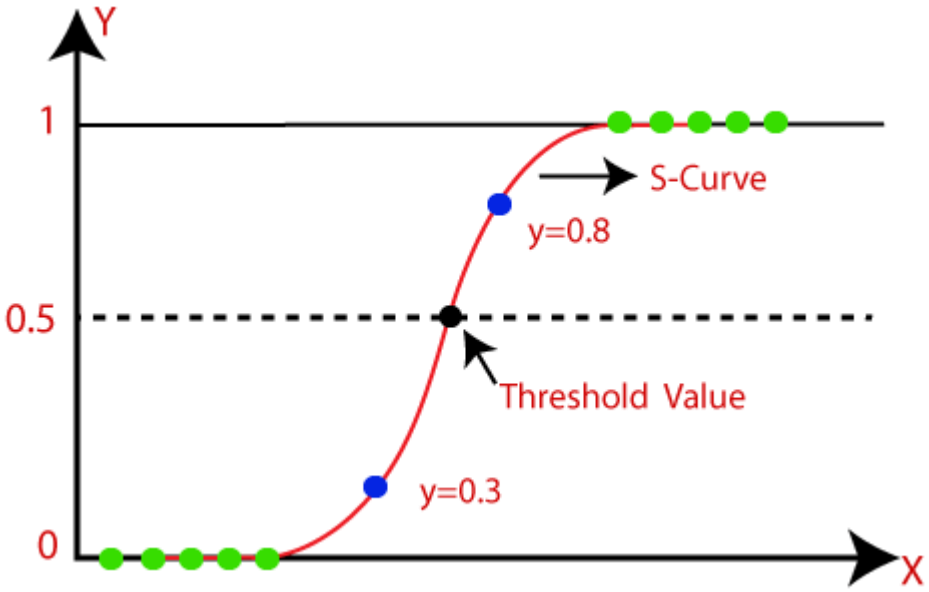


Figure 2.4: Logistic Regression

The sigmoid function's result is between 0 and 1 and never more than 1. At the origin, *X=0*, the value is precisely 0.5. The cutoff probability for establishing the different categories is half. Class-1 (*Y=1*) is assigned if the likelihood is more than 0.5, and Class-0 (*Y=0*) is set otherwise.

2.5 K-Nearest Neighbor (K-NN)

In 1951, the K-Nearest Neighbor algorithm was first proposed by Evelyn Fix and J.L. Hodges Jr in a study (Silverman & Jones, 1989). This report, while unpublished, presented a **non-parametric** approach to identifying patterns (Imandoust & Bolandraftar, 2013). Further modifications were made to the technique's fundamental characteristics in 1967 by T. Cover and P. Hart (Imandoust & Bolandraftar, 2013; Prasath et al., 2019). The term non-parametric refers to the idea that it is not necessary for individuals to possess any previous understanding or familiarity with the structure and characteristics of a frequency distribution. The determination of parameters would be contingent upon the size of the information used to train the sample (Prasath et al., 2019).

K-NN assigns a class label to an uncategorized test sample by identifying the samples that exhibit the greatest resemblance to its K nearest neighbors. Using a specialized distance function, it evaluates the test sample's gap from the samples in the training data (Prasath et al., 2019). K denotes a numerical value representing the amount of adjacent data points in the sample. Selecting the right value for K and the suitable distance function are crucial factors for the effectiveness of a K-NN classification model (Imandoust & Bolandraftar, 2013). Changing the value of K results in a corresponding alteration in the conditional class probabilities since the neighborhood is defined by the separation between the test sample and its Kth closest neighbor (Imandoust & Bolandraftar, 2013). K-NN neighborhood size selection is ambiguous. Hence weighted voting techniques have been developed to combat the ambiguity inherent in K-NN neighborhood size selection (Imandoust & Bolandraftar, 2013). In general, the test sample gets grouped in the closest neighbor class when K=1 (Prasath et al., 2019), and it is categorized by majority voting when K takes on greater values.

Due to its inability to master a discriminative function during the training process, the K-NN algorithm is often called a "lazy learner" or instance-based classifier (Imandoust & Bolandraftar, 2013). Nevertheless, it may still provide a rapid understanding of the dataset's structure by using a small sample. Additionally, the training phase is quicker since there is no time allotted for learning.

The K-NN method is characterized by its low complexity, that has shown its use in addressing a wide range of practical problems related to classification and is also most dependable for pattern recognition and regression models (Prasath et al., 2019). However, its speed decreases when using all features in distance calculations due to the impact of outliers on accuracy. Moreover, since most data in the real world do not conform to the theoretical assumptions made in linear regression algorithms, K-NN is a good choice for classification studies where the data distribution is unknown (Prasath et al., 2019). Yet, it should be mentioned that K-NN is readily comprehensible and user-friendly, exhibits reliance to training data that contains noise and excels in scenarios where one instance may possess several class labels (Jadhav & Channe, 2016). Data that has been distorted in some manner, resulting in changes to some of the values, is said to be noisy (Prasath et al., 2019).

In the K-NN algorithm, the **Euclidean distance** is the most popular distance metric. The choice of K, along with the sparsity and noise of data points, impacts the algorithm's performance. If K is too low, the result might be skewed, and if K is too high, processing time increases.

Euclidean distance function can be given using the following equation (Prasath et al., 2019):

$$Euclidean: D(x, y) = \sqrt{\sum_{i=1}^{k} |x_i - y_i|^2}$$

The quantity k is determined by squaring the aggregate data points denoted by n. When distinguishing the different categories of data, it is advisable to take a value that is odd for k.

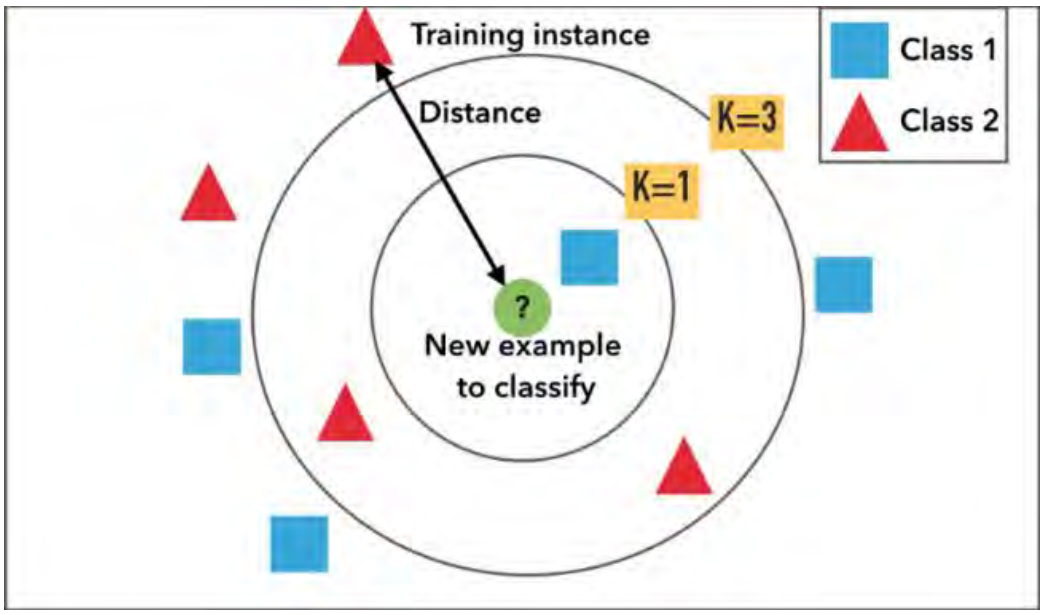


Figure 2.5: K-Nearest Neighbor

In the provided figure, the piece of data under test fits into either one of class 1 or class 2. The feature similarity tells us that when K=1, since there is only a square inside the circle, the unknown sample will be classified as class 1. Similarly, for K=3, the sample will be identified under class 2, as there are more triangles than squares inside the circle. However, when K=5, the sample will be labelled as class 1 since there are more squares than triangles outside the outer circle.

2.6 Naive Bayes

In 1988, J. Pearl introduced the **Bayesian network,** which is a **statistical classifier** that makes use of the **Bayes Theorem** and employs **conditional independence**. The model constitutes a **joint probability** distribution spanning several variables and is implemented in teaching the model (Taheri & Mammadov, 2013). Naïve Bayes may be seen as a specialized version of a broader

Bayesian network framework characterized by the imposition of certain constraints. These constraints include the absence of parent nodes for the class variable and the lack of links connecting the feature variable nodes. The use of this restricted foundation offers improved efficiency in comparison to generic Bayesian networks (Lowd & Domingos, 2005). When an arc connects two nodes, the node from which the arc originates is termed the **parent** of the node to which the arc terminates, which is the **child** node (Taheri & Mammadov, 2013).

The Naïve Bayes family of algorithms uses **conditional probability** to make predictions. Since it is a form of Bayesian network, Naïve Bayes is naturally based on the **Bayes Theorem**. Additionally, it is a classification technique that assumes the independence of the predictors (Mahesh, 2018). This idea that any feature from a class has no dependence on the presence of any other features, which is considered the "naïve" assumption (Lowd & Domingos, 2005), ignores the influence of other characteristics when determining the probability of whether or not an event will occur. This means that no single factor can increase or decrease the odds of any given event. Due to this assumption, Naïve Bayes is also named “Idiot’s Bayes.”

Bayes Theorem, which Thomas Bayes introduced in the 18th century, is sometimes referred to as **Bayes law** or the **Bayes rule,** and it presents a framework for calculating the probability of an event by incorporating already existing information potentially linked to the incident. This probability is a conditional probability, which by definition is the chance (probability) of an event occurring, considering the occurrence of another event. In mathematical terms, the Bayes law is given as the following (Kumari, n.d.):

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Given an uncategorized input data sample *X* and an arbitrary hypothesis H that assumes *X* belongs to a class *C*. The main objective is to calculate the **posterior probability**, denoted by *P(H|X)*, which is the probability that *H* is true, provided that the input data is *X*. *P(H|X)* is also called the **likelihood** (Jadhav & Channe, 2016) or the probability training data (Edeh et al., 2022). **Likelihood** is a quantitative measure that determines the degree of conformity between a certain hypothesis and the empirical evidence that has been observed. Here, *X* can also be considered as **evidence** (the evidence is an attribute of an unknown value). The **prior probability** of H, symbolized as *P(H)*, pertains to the initial probability assigned to *H* prior to the incorporation of any extra information, thus also called **priori** of *H* or **class prior probability** (Mahesh, 2018). Similarly, the prior probability of *X,* symbolized as *P(X)*, refers to the initial probability of *X*, which stays constant and is also called **predictor prior probability** (Mahesh, 2018).

Expressing this equation in words can be as follows (Kumari, n.d.):

$$Posterior = \frac{likelihood \times prior}{evidence}$$

Now, suppose *C* is a stochastic class variable that is being predicted and takes on values corresponding to different classes (Lewis, 1998); *c* is a class that belongs to *C;* the feature value *x* is given by the random variable *X* in a vector form with *n* components or features, i.e., $x = (x_1, x_2, \ldots, x_n)$. Using these variables in the Bayes Theorem stated above will give the following (Lewis, 1998):

$$P(C = c|X = x) = \frac{P(X = x|C = c)P(C = c)}{P(X = x)}$$

where *P(x)* can be calculated using the following formula (Lewis, 1998):

$$P(x) = \sum_{j=1}^{k} P(x|c_j)P(c_j)$$

Usually *P(x)* remains the same for all class so it is not necessary to calculate it. As each feature is conditionally independent, we can write the value of the feature data as (Lewis, 1998):

$$P(x|c) = \prod_{i=1}^{n} P(x_i|c)$$

Using frequency counts within arrays is a common practice in the training process, where values are derived from a singular iteration of the data utilized for learning. This approach aids in estimating *P(c)* and *P(x|c)* in the case of qualitative features(Webb, 2016). In the context of quantitative features, several other measures are taken, like data discretization or probability density estimation (Webb, 2016). Hence, we can adjust the equation in the Bayes law to calculate the unknown posterior probability. Finally, after calculating the probability for every class, the new data is put into the category of the class which has the highest probability. This is called the

Maximum a posteriori rule. Using this rule, the naïve Bayes classifier can be represented as (Al-Aidaroos et al., 2010):

$$\hat{y} = arg \max_{c \epsilon C} P(c) \prod_{i=1}^{n} P(x_i|c)$$

Naïve Bayes has little training time, and it eliminates redundant features to boost the effectiveness of classification and enable the system to function with high efficiency (Jadhav & Channe, 2016). However, to achieve a satisfactory performance, the classifier needs a substantial volume of data records, and it tends to be more biased than competing classifiers in some instances (Jadhav & Channe, 2016). Even so, due to its ease of construction along with little reliance on complex parameter estimation, the naïve Bayes approach is considered advantageous when dealing with huge amounts of information. The classifier has many applications in spam filtering, text classification, and a lot more (Sarker, 2021)

2.7 Decision Tree

The decision tree algorithm was first established in the 1960s (Song & Lu, 2015) and can solve both classification and regression problems. The use of this particular supervised learning algorithm is prevalent for data extraction, where it is employed to build predictive models pertaining to a certain variable of interest. Additionally, it is also used to develop data categorization systems that rely on a substantial number of factors. This method is preferred because of its ability to process vast amounts of data while being effective, easy to use, and devoid of any uncertainty, even when some values are missing. Decision trees may be constructed without prior knowledge of the target domain or specific parameter values (Gupta et al., 2017).

Decision trees have a central point from which several options sprout, with an increasing number of **branches**, decisions, and conditions. The entire population or sample is represented by the topmost node, which is the **root node** or the **decision node** (Song & Lu, 2015). From there, it is divided into two or more groups based on shared characteristics. Each value of an attribute in a decision node is associated with a set of possible tests. The **internal nodes** or the **chance nodes** that denote dataset attributes are the alternative outcomes to be chosen from at a certain juncture within the tree structure (Song & Lu, 2015). The branches are the links representing the decisions originating from the internal nodes (Song & Lu, 2015), and they indicate the respective values of the attributes (Rokach & Maimon, 2005). With further division, the tree terminates at the **leaf** or **terminal node** that carries a class label. The leaf node is the point where prediction is made and denotes a classification or decision which can be either categorical or a numerical value. By moving from the tree's root to a leaf, a specific choice is made, and depending on these choices, the final decision is reached following the interconnected paths. Basically, the intuition is that the

model asks a question and splits the data. The focus is to build a tree for all the data and retrieve results at each leaf by reducing errors.

The instance space is divided into multiple portions by an internal node of a decision tree, utilizing a discrete function of the input attributes (which can take on both continuous or discrete values (Song & Lu, 2015)). As only one attribute is considered, this division is done based on the weight of the attribute. If the attribute in question is quantitative, the splitting will be based on a specified range (Rokach & Maimon, 2005). The selection of possible input variables is determined by the features associated with the purity level in the child nodes that arise from the decision processes. This purity is measured by the number of nodes exhibiting the desired condition. Until a specific stopping condition or a required degree of similarity is reached, the splitting process is iteratively performed (Song & Lu, 2015). It is crucial to set a rule for discontinuing this splitting process so that the decision tree does not get overly intricate. In general, the reliability of a model diminishes as its complexity increases, affecting the predictive accuracy for future observations. Standard measures of tree complexity include the number of nodes, leaves, depth, and features used (Song & Lu, 2015).

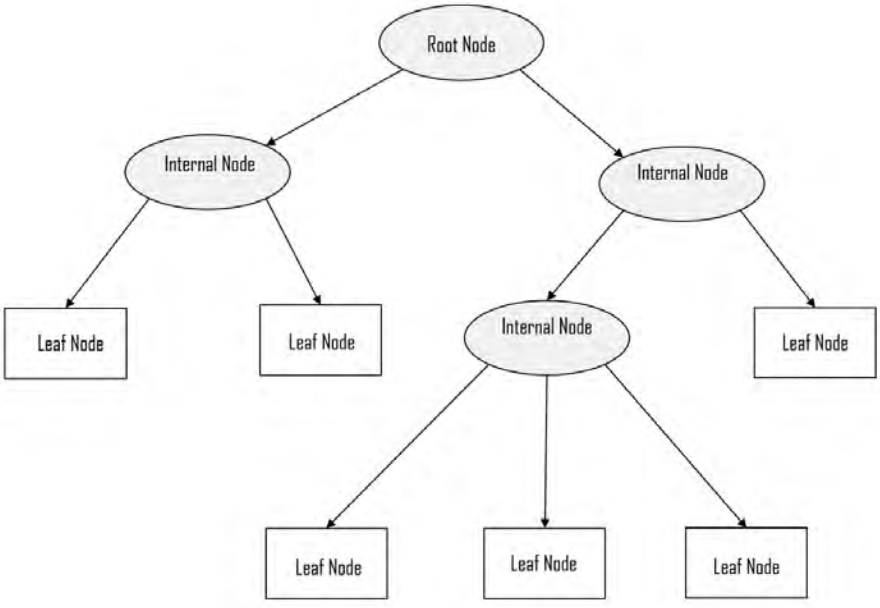


Figure 2.6 Decision Tree with labels (skeleton)

There are other concepts imperative in the decision tree framework that have to do with the splitting procedure, which will be discussed now.

2.7.1 Gini Index

Gini Impurity serves as a cost function to choose the splitting in the dataset. Quantitatively, the impurity level, denoted by D, can be evaluated by the Gini Index. This Gini for the Gini impurity can be mathematically expressed as (Gupta et al., 2017):

$$Gini(D) = 1 - \sum_{i=1}^{c} p_i^2$$

2.7.2 Entropy

Based on information theory, entropy is another impureness measurement metric (Foster, 2021). In other words, entropy is used to determine a sample's homogeneity. If the sample is totally homogenous, the entropy is 0; if it is evenly split, it is 1 (Foster, 2021). The equation for entropy can be mathematically stated as the following (Sarker, 2021):

$$Entropy: H(x) = -\sum_{i=1}^{n} p(x_i) log_2 p((x_i)$$

2.8 Random Forest

As a classification technique, the random forest algorithm, built on **ensemble modelling**, provides better efficiency and accuracy than decision trees (Sarker, 2021). This method combines a set of independent decision trees to lower variation and provides more precise data predictions. It can be used for both classification and regression and can work well for continuous as well as categorical data. Random forests can process a wide range of characteristics, even when many values are absent. Their highly parallelizable nature also enables smooth execution on big datasets (Foster, 2021). Even though understanding the trained model is challenging due to its substantial size and complexity, it still serves as a valuable means to assess the significance of characteristics, offering insights into the key factors that contributed to the construction of the classifier (Foster, 2021).

Random Forest has evolved over time through the works of different researchers. In 1995, Tin Kam Ho used the random subspace approach to put forward the random decision forest algorithm (Tin Kam Ho, 1995), while in 1996, the bagging sampling technique was put forward by Leo Breiman (Breiman, 1996). Then, in 1997, Amit and Geman (Amit & Geman, 1997) suggested a method to identify the tree's structure using the combined induction of shape features and tree classifiers (Fawagreh et al., 2014). In 1998, Tin Ho (Tin Kam Ho, 1998) offered a solution to tackle the conflict between overfitting and obtaining maximum accuracy (Fawagreh et al., 2014), while Dietterich (Dietterich, n.d.) pioneered the concept of randomized node optimization in the

same year, in which the choice at each node is made using a stochastic approach rather than a deterministic algorithm (Hastie et al., 2001). Out of all the previous research on random forests, Amit and Geman's work inspired Leo Breiman, and he created fresh training sets by randomizing the previous training set's outputs (Y. Liu et al., 2012). Consequently, in 2001, he (Breiman, 2001) and Adele Cutler introduced the random forest algorithm in a study.

All of these methodologies have one interesting thing in common: for any particular tree from the tree series, a multivariate random variable is generated for each of them. Despite following the same distribution, this vector variable is not influenced by the preceding random vectors, and through the utilization of the training set along with the vector in question, a tree is constructed, ultimately transforming into a classification tool. If k is the number of the tree sequence, and x is the input for the random vector denoted by Θ_k, then $h(x, \Theta_k)$ represents the classifier (Breiman, 2001).

According to Leo Breiman, Random Forest's definition can be put forward as: "A random forest is a classifier composed of a collection of tree-structured classifiers $\{h(x, \Theta_k), k = 1,...\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x" (Breiman, 2001). After k runs, a series of classifiers is obtained, which is then employed to form a system using several classification models, the ultimate result of which is established through a standard majority vote (Y. Liu et al., 2012).

So, in the initial stage of training, several decision trees are constructed. To create a set of decision trees with controlled variance, the technique combines Breiman's concept of "bagging," also known as "feature bagging", and the randomly chosen feature selection. The premise upon which the bagging process is based is that integration of many learned models yields better results. The final outcome does improve when we combine the learnings from various models. A single decision tree would otherwise result in an overfit model if the dataset were huge. In order to generate a decision tree for each sub-sample, we first partition our training dataset into k sub-samples. In the sample, instances are referred to as **in-bag instances**, while the remainder is referred to as out-**of-bag instances** (Fawagreh et al., 2014). Note that a dataset with n features (number of variables) needs to be chosen in such a way that it has a low correlation but high predictive power. According to Breiman, $\sqrt{n}$ features (number of overall variables) are typically used in each split or partition (Hastie et al., 2001), where $\sqrt{n}$ denotes the randomization when selecting the best node to split (Fawagreh et al., 2014). Then, we count the votes out of every

choice made by all the decision trees, and lastly, we combine the votes to obtain the random forest decision.

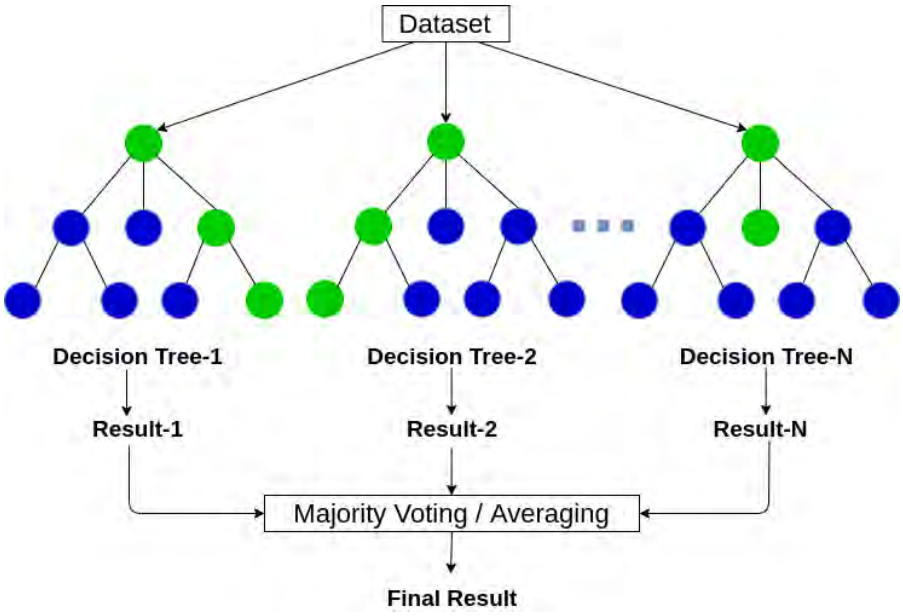


Figure 2.7 Random Forest

Chapter 3 Methodology

3.1 Research Design

The study employs a mixed method research design, so both nominal and numerical values are included in the dataset. Also, it takes a cross-sectional approach to its data collection. This means the data is collected at a single time point from participants of varying ages.

3.2 Data Description and Collection

In this study, all the information is gathered from a sample of Bangladeshi adults, encompassing individuals from various age groups. Initially, a questionnaire containing relevant questions was created, which included **demographic data, lifestyle data, clinical and biophysical data, and medical history of family related to diabetes**. The dataset consisted of both diabetes and non-diabetes participants. The records of the diabetes patients were collected from the **BIRDEM General Hospital**, and normal people with no diabetes also filled out the form. Ethical considerations are made to ensure privacy and confidentiality when dealing with participants' data. The data was recorded in an Excel sheet. The relevant features were also selected.

Features can be defined as discrete attributes or qualities of observable aspects of an event that can be measured. For an algorithm to be successful, selecting a distinct and instructive feature is essential. In the dataset, the relevant features that were included are shown in the table. Among all the features in the dataset, diabetes status is the dependent variable, whereas the other features act as the independent variable.

Table 1: List of features used in the dataset

No.	**Features**	**Unit**
1	gender	-
2	height	feet (converted to meters)
3	weight	kilogram (kg)
4	age	years
5	BMI	kg/m^2
6	blood pressure	mmHg
7	sleep duration	hours
8	living condition	-
9	profession	-
10	smoking habit	-
11	heart disease	-
12	kidney disease	-
13	other diseases	-
14	diabetes status	-
15	diabetes record in the family	-
16	family income	taka (Tk.)

The output of the data is the **class** or category. The data labels indicate the corresponding class. **Label** is the variable that is required to be estimated. Labelling means providing explanatory tags to every unlabelled data to enhance its meaning. In our case, the presence or absence of diabetes is the label.

3.3 Data Pre-processing

After collecting data, the next step is to preprocess it. Some irrelevant and incomplete values may not address the problem, and these need to be handled. Preprocessing allows cleaning data and fixing missing or duplicate values and outliers. Also, scaling features may be necessary because the magnitude of input variables affects machine learning algorithms, i.e., they are scale-sensitive. So, the numerical features are converted to a comparable scale. Basically, in this step, the data needs to be checked for any errors, inconsistencies, or ambiguity and transformed and formatted accordingly to allow learning. The discrepancies must be rectified to ensure the dataset is suitable and reliable for analysis.

3.4 Algorithms used

Through the application of various machine learning techniques, this study centers on predicting diabetes in the adult population across Bangladesh. In the context of diabetes prediction, supervised learning refers to the training of models using labelled data, wherein the output (diabetes status) is known for each input (participant's data). The algorithms used are:

- Logistic Regression
- K-Nearest Neighbor
- Naïve Bayes
- Decision Tree and
- Random Forest

3.5 Model Deployment

Once the machine learning algorithms have been chosen, the subsequent step involves their implementation using appropriate tools. In this case, the RapidMiner tool is selected for its user-friendly interface and capacity to manage a wide range of machine learning algorithms effectively.

3.6 Model Evaluation

Following the deployment of the models, the evaluation process is conducted to assess their performance using specialized performance indicators. In addition, confusion matrices are generated for each model. The performance indicators used are:

- Accuracy
- Precision
- Predicted value negative
- Specificity

3.6.1 Confusion Matrix

The class distribution of the dataset must be addressed when assessing a classifier's performance. In situations where there is a major class imbalance, conventional accuracy tests can wrongly classify certain categories over others. So sometimes, there are instances when a person is infected with the disease, but the laboratory test gives out negative test results, or when a person who is actually fine, tests positive for the disease. Hence, it is important to make sure the test is as much valid as possible. Also, the majority of error metrics quantify the model's overall error but not individual errors (*What Is a Confusion Matrix in Machine Learning?*, n.d.). This is where a **confusion matrix** comes into the picture.

A confusion matrix or **error matrix** (*Confusion Matrix in Machine Learning - Javatpoint*, n.d.; Sharma, 2021) is in the form of a contingency table that gives an overview of the classification algorithms' performance. The matrix provides a tabular representation of predicted and actual classes of a classification problem (*Confusion Matrix in Machine Learning - Javatpoint*, n.d.). In a problem involving binary classification, the matrix is a 2*2 consisting of four main components, which are **true positives (TP), true negatives (TN), false positives (FP)**, and **false negatives (FN)**. Now, we will define these terms in the context of this research.

Table 2: Confusion Matrix where TP, TN, FP, and FN denotes true positive, true negative, false positive, and false negative, respectively

		ACTUAL VALUES		
	Test Result	Disease Present (Positive)	Disease Absent (Negative)	Total
PREDICTED VALUES	Positive	TP	FP	Total Test Positive
	Negative	FN	TN	Total Test Negative
	Total	Total Diseased	Total Normal	Total Population

True Positive refers to the cases when the person who actually has the disease also tested positive by the tests performed. Thus, the model is said to have correctly identified the disease.

True Negative refers to the case when the person with no disease is correctly tested negative by tests carried out.

False Positive, also known as **Type-I error** (*Confusion Matrix in Machine Learning - Javatpoint*, n.d.), indicates the instances when the model detects the disease showing positive test results when, in reality, it is absent. Thus, the model is said to have wrongly identified the disease.

False Negative, also known as **Type-II error** (*Confusion Matrix in Machine Learning - Javatpoint*, n.d.), is when the model gives out negative test results and misclassifies a person with the disease as healthy.

Summarizing all the terms, we can make the following inferences. The **actual positives** consist of people who have diabetes (**TP and FN**), while the **actual negatives** include diabetes-free people.

The **predicted positives** are those who are expected to have the disease (**TP and FP**), while the **predicted negatives** comprise individuals who are predicted healthy.

Various reasons make the confusion matrix an important tool for data analysis and machine learning. It identifies not only the specific types of mistakes we made but also the classifiers' errors (*Confusion Matrix in Machine Learning - Javatpoint*, n.d.). This breakdown is what allows us to get around the barrier of using only classification accuracy. It allows an assessment of the model's performance. Using metrics like **accuracy, precision, recall,** and **specificity**, the efficacy of a model in predicting outcomes can be evaluated (*Confusion Matrix in Machine Learning - Javatpoint*, n.d.). Also, by analyzing the distributions of the TPs, TNs, FPs, and FNs, the models' strengths and weaknesses can be taken into account. Through error analysis, patterns and possibilities contributing to misclassifications can be recognized. These findings make it possible to make modifications, thus reducing mistakes and augmenting the model's success. Furthermore, a comparison of numerous models can be made using the matrix (*Confusion Matrix in Machine Learning - Javatpoint*, n.d.). After computing each of their performance metrics, it is possible to select an ideal model by examining which one is more effective. This makes decision-making easier, allowing businesses to pick the best model that matches their requirements. In conclusion, confusion matrix allows management of class imbalance, model validation, optimization, quality control and risk assessment.

There has been extensive application of confusion matrix derivatives. One of the most important applications of the confusion matrix is in medical diagnostics, which is our main topic of discussion. It is used to assess the illness detection tests and classification algorithms implemented for disease identification. In order to ensure proper treatment, a precise diagnosis is needed. Hence, it is important to evaluate the model's ability to correctly identify true positives and true negatives, which is done with the help of a confusion matrix. In addition, there are other applications in fraud or cyber-attack detection (Sharma, 2021), spam detection, stock market predictions (Dhingra, 2021), etc.

Other terms associated with the confusion matrix and performance evaluation used in this study are defined next.

3.6.2 Accuracy

Accuracy is a metric used to evaluate how well the model has rightly classified the values (*Confusion Matrix in Machine Learning - Javatpoint*, n.d.). When the data's desired variable classes are evenly distributed, accuracy is effective. However, if the dataset's target variable class is predominately a single class, the accuracy metric must be avoided (Sunasra, 2019). The equation for accuracy is given as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

3.6.3 Sensitivity

Sensitivity (Recall or **True Positive Rate)** is a performance indicator to measure a binary classification model's accuracy. It calculates the model's accuracy in identifying true positive samples or correct positive values, i.e., it shows how well the model detects the condition or target class (Dāsa, 2016). The equation for sensitivity is given below:

$$Sensitivity = \frac{P(positive\ test\ result \cap disease\ is\ present)}{P(disease\ is\ present)} = \frac{TP}{TP + FN}$$

3.6.4 Specificity

Specificity (or **True Negative Rate**) is also another performance metric similar to sensitivity. However, specificity evaluates how many true negative samples the model accurately identifies, i.e., the accuracy of finding people who do not have the disease (Dāsa, 2016). The equation for specificity is given below:

$$Specificity = \frac{P(negative\ test\ result \cap disease\ is\ absent)}{P(disease\ is\ absent)} = \frac{TN}{TN + FP}$$

A high sensitivity value means the model can recognize positive samples, while a high specificity value suggests the model can accurately classify negative samples.

3.6.5 Precision

Precision (or **Predicted Value Positive)** measures how many positive samples are successfully anticipated (*Confusion Matrix in Machine Learning - Javatpoint*, n.d.). It can be given by the conditional probability of disease present given a positive test result. A higher precision effectively detects positive samples as they exhibit a low rate of false positives. The equation for precision is given below:

$$Precision = \frac{P(disease\ present \cap positive\ test\ result)}{P(positive\ test\ result)} = \frac{TP}{TP + FP}$$

It is important to know when to focus on precision and when on recall. If the objective is to reduce the false negatives, it is desirable to maximize the recall while maintaining a sufficient amount of precision. Contrariwise, in case of false positives, the precision needs to be optimized as much as possible (Sunasra, 2019).

3.6.6 Predicted Value Negative

Predicted Value Negative (PVN), in contrast, measures how many negative samples are successfully anticipated. When this value is higher, the model can properly recognize negative instances and has a low rate of false negatives. The equation for the predicted value negative is given below:

$$PVN = \frac{P(disease\ absent \cap negative\ test\ result)}{P(negative\ test\ result)} = \frac{TN}{TN + FN}$$

3.7 Comparative Analysis

All these algorithms exhibit different accuracies for the classification task. So, a comparative analysis of the performance of each machine learning model is conducted to assess its strengths and weaknesses. Through comparison of accuracy and other metrics, the algorithm that yields the optimal results for diabetes prognosis is identified. However, the main emphasis is on the accuracy metric for selecting the best model, i.e., finding which model has the highest accuracy score in detecting diabetes.

Chapter 4 Result Discussion and Model Analysis

4.1 Results Obtained

This section shows the results from the collected data and provides their frequency distribution, bar chart, and pie chart representations.

4.1.1 Demographic data

Table 3: Frequency distribution of gender

Gender	Frequency	Percentage (%)
Male	**85**	**32.8**
Female	**174**	**67.2**

The study was conducted on a total of 259 subjects or individuals, of which the majority comprised females. With 174 female patients making up 67.2% of the population, the other 32.8% made up the 85 male patients in the study.

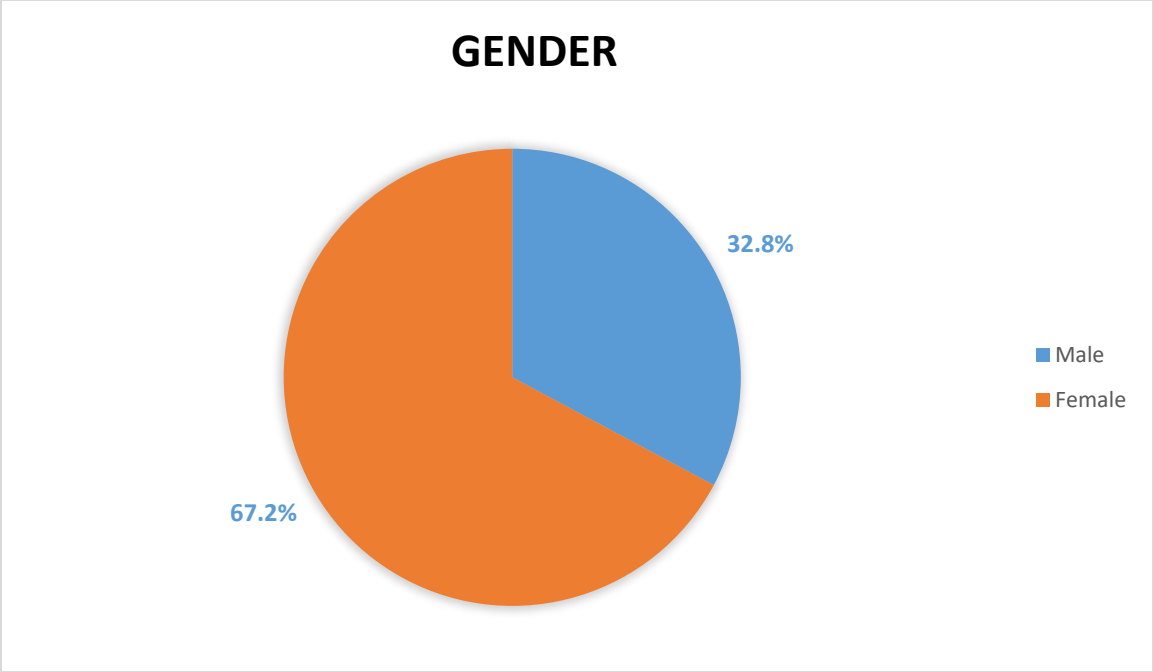


Figure 4.1: A pie chart demonstrating the percentage of males and females in the study

Table 4: Frequency distribution of age group

Age (years)	Frequency	Percentage (%)
17-25	**63**	**24.3**
26-35	**29**	**11.2**
36-45	**62**	**23.9**
46-55	**70**	**27.0**
56-65	**26**	**10.0**
66-75	**9**	**3.5**

According to the table above, out of the 259 subjects, the highest percentage of people falls within the age range of 46-55 years, comprising 27% of the population. This demography is followed by 24.3% and 23.9% being in the age range of 17-25 and 36-45, respectively. Hence, the study focused on this age category of people more to find out about the diagnosis of diabetes.

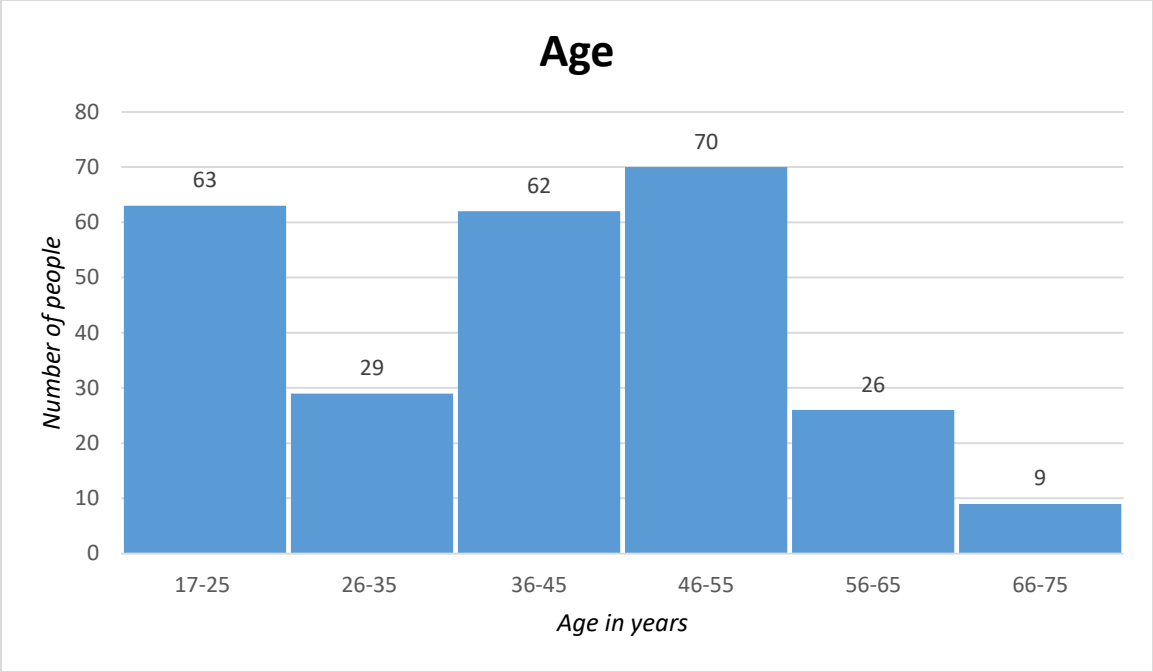


Figure 4.2: A histogram showing the age ranges of the total population in the study

Table 5: Frequency distribution of living conditions

Living Condition	**Frequency**	**Percentage (%)**
Urban	**208**	**80.3**
Rural	**51**	**19.7**

It can be seen from the table above that a large portion of the people in this study lives in the city. Only 19.7% of these people inhabit the rural area, while a huge 80.3% reside in urban regions.

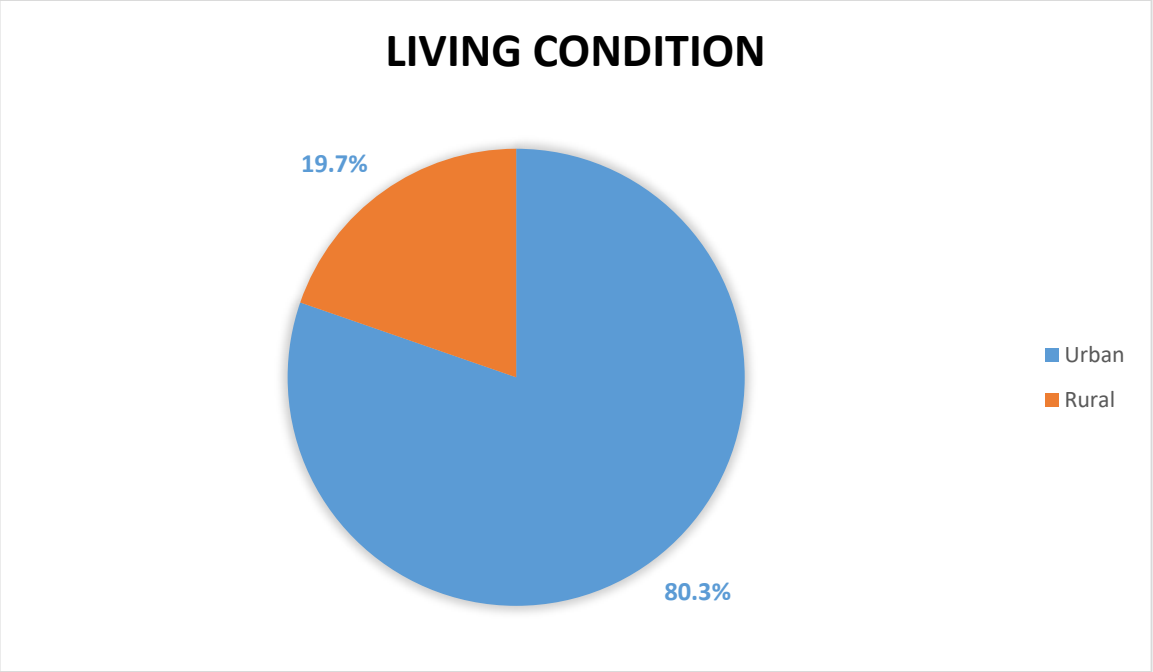


Figure 4.3: A pie chart demonstrating the living condition of the total population in the study

Table 6: Frequency distribution of occupations of profession

Profession	Frequency	Percentage (%)
Business	**31**	**11.9**
Housewife	**99**	**38.2**
Retired	**9**	**3.5**
Service	**60**	**23.2**
Student	**60**	**23.2**

As shown in the above table of professions, the highest percentage of people who took part in the study are housewives, comprising 38.2% of the population. As women make up 67.2% of the study population, this proportion is in line with the overall data. Furthermore, students and people who do service jobs equally consist of 23.2% of the total people. The remaining category of people is 11.9% of businessmen who either run their own business or work for an organization, and only 3.5% of individuals are retired.

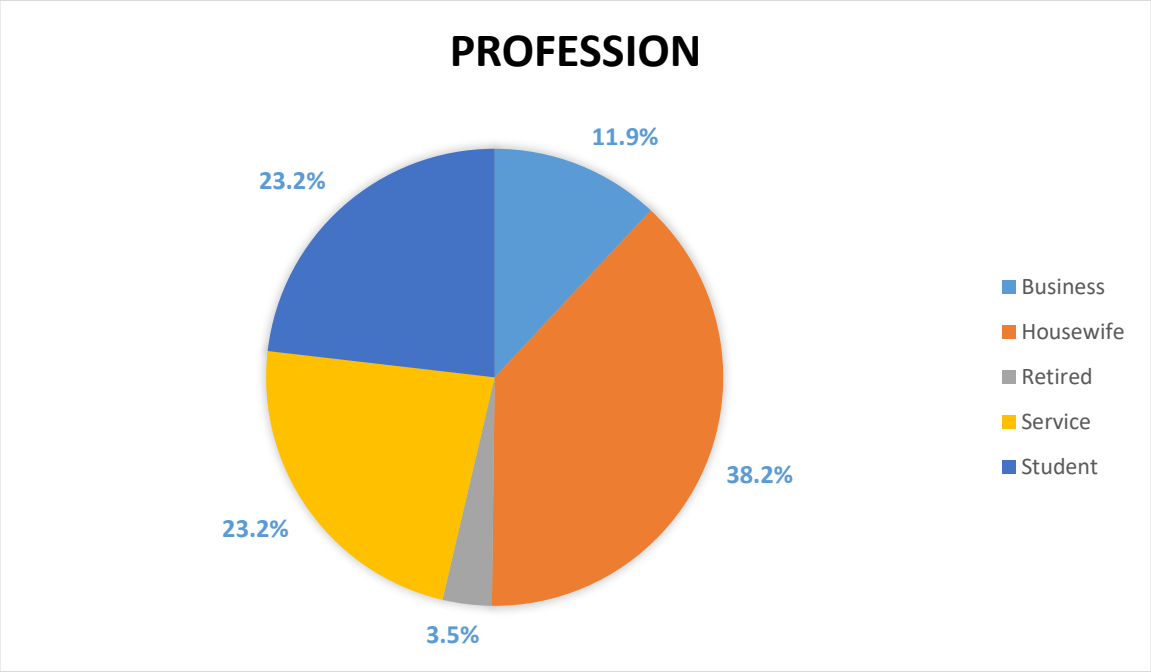


Figure 4.4: A pie chart representing the professional backgrounds of the total population in the study

Table 7: Frequency distribution of family/own income

Family/Own Income (Tk.)	**Frequency**	**Percentage (%)**
2000-5000	**24**	**9.3**
5001-10000	**28**	**10.8**
10001-20000	**47**	**18.1**
20001-30000	**39**	**15.0**
30001-40000	**31**	**12.0**
40001-50000	**38**	**14.7**
50001-100000	**47**	**18.1**
100000 above	**5**	**2.0**

From the data presented in the above table, it can be seen that the highest number of people had a salary between 10001-20000 Tk. and 50001-100000 Tk. both with a percentage of 18.1%. This salary is either the person's own income or the income of his or her family if the person does not work. The table illustrates that a small number of people have an income of 2000-5000 Tk. (9.3%) and 5001-10000 Tk. (10.8%), while only very few are retired (2%). There is a significant amount of individuals in the 20001-30000 Tk. (15%) and 40001-50000 Tk. (14.7%) range, followed by 12% of people in the 30001-40000 Tk. range.

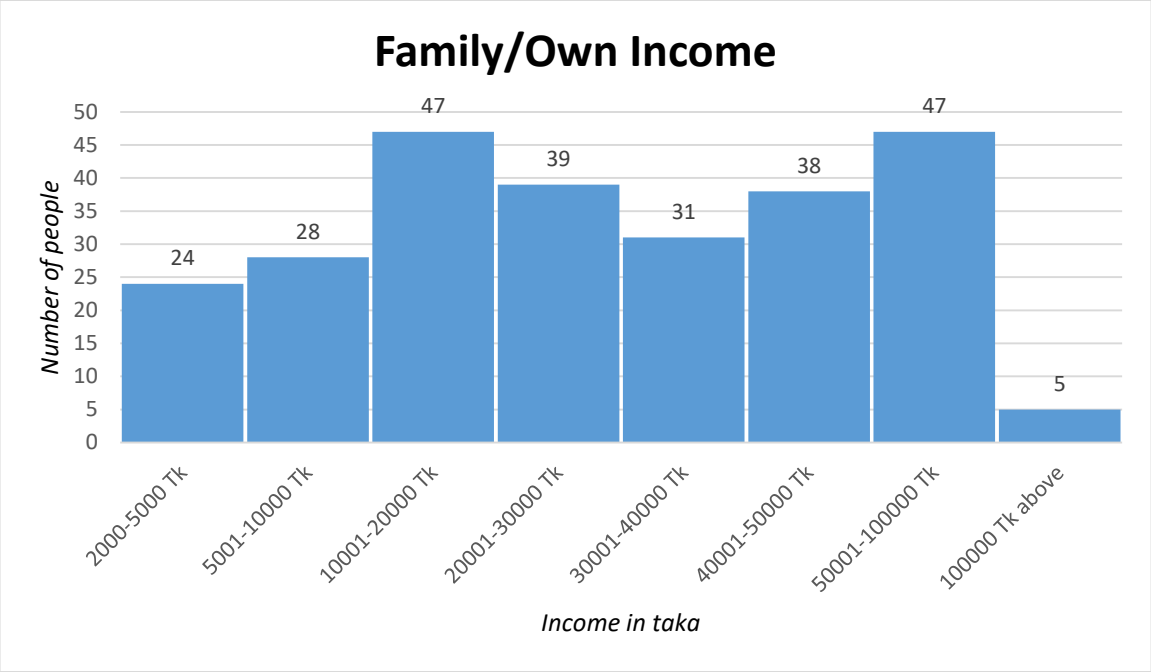


Figure 4.5: A histogram showing the income ranges of the total population in the study

4.1.2 Lifestyle Data

Table 8: Frequency distribution of sleep duration

Sleep Duration	Frequency	Percentage (%)
4 hours	**1**	**0.4**
5 hours	**17**	**6.6**
6 hours	**57**	**22.0**
7 hours	**81**	**31.3**
8 hours	**78**	**30.1**
9 hours	**19**	**7.3**
10 hours	**4**	**1.5**
12 hours	**2**	**0.8**

The sleep duration table shows the majority of the people sleep 7-8 hours with a cumulative percentage of 61.4%. 6 hours sleep duration is covered by 22% of the people. Some people sleep

9 hours, which is 7.3%. Another 6.6% are found to sleep just 5 hours of the day, while fewer sleep for 10 hours, comprising 1.5% and 12 hours, making up 0.8% of the population. Shockingly, 0.4% of the population gets by on only 4 hours of sleep each night.

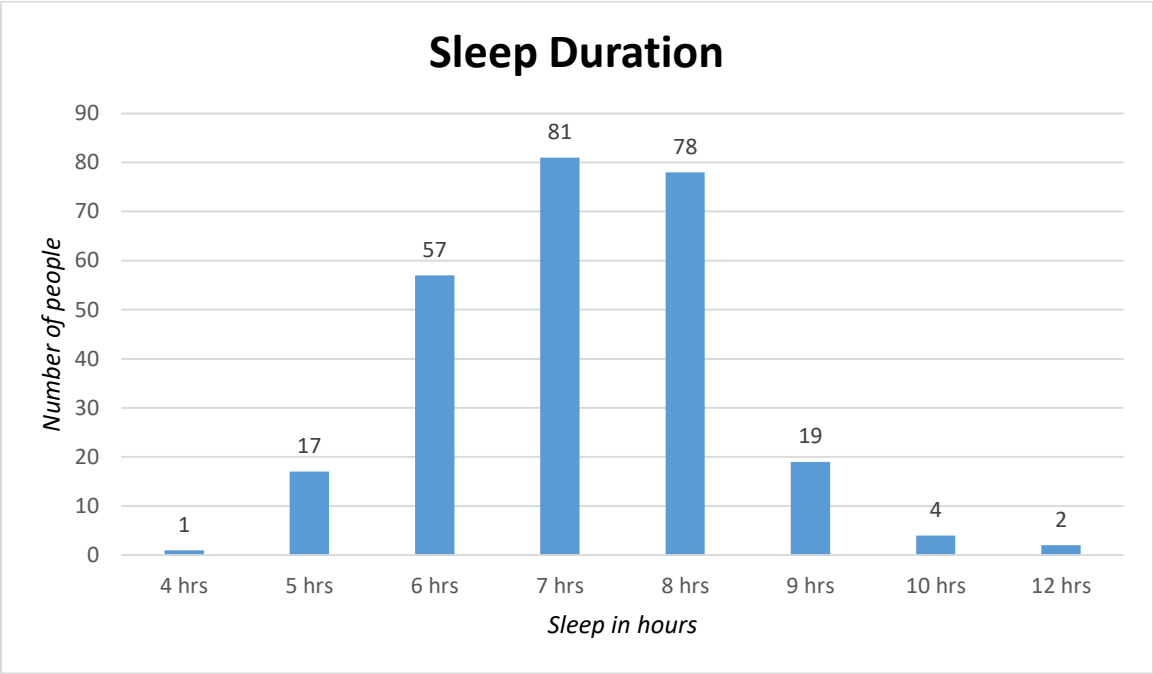


Figure 4.6: A bar graph representation of the sleep duration of the total population in the study

Table 9: Frequency distribution of smoking habits

Smoking Habit	**Frequency**	**Percentage (%)**
Yes	**28**	**10.8**
No	**231**	**89.2**

From the study, it can be seen from the table that a vast majority of the population was non-smokers, with a percentage of 89.2%, while only 10.8% were smokers. This can mainly be due to the fact that there are more females in the study than males, and most males are non-smokers. Also, none of the female smokes, which indicates that a majority of women are married housewives.

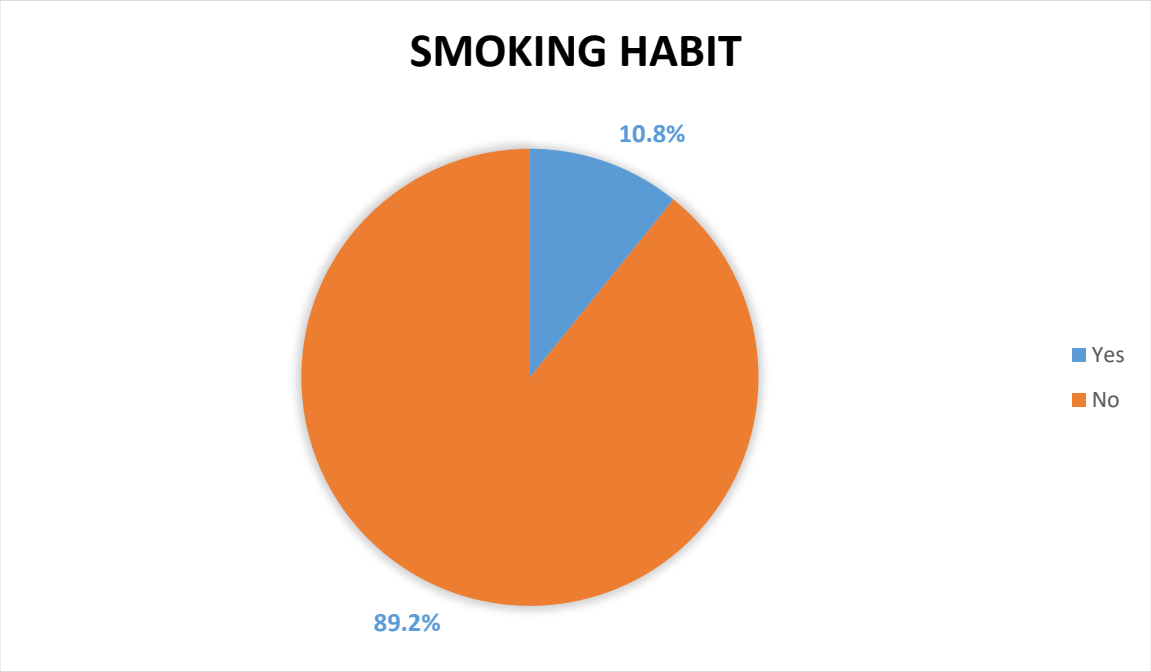


Figure 4.7: A pie chart representing the smoking habits of the total population in the study

4.1.3 Clinical and Biophysical Data

Table 10: Frequency distribution of Body Mass Index (BMI)

Body Mass Index (BMI) (kg/m^2)	**Frequency**	**Percentage (%)**
Underweight (under 18.5)	**6**	**2.3**
Healthy (18.5 to 24.9)	**61**	**23.6**
Overweight (25.0 to 29.9)	**34**	**13.1**
Obese (30.0 or higher)	**158**	**61.0**

The body mass index (BMI) was calculated using the height and weight variables, which is weight in kilograms divided by height in meters squared, i.e., Weight/ (Height*Height). It is an indicator that groups people into different categories according to their weight and height. The table illustrates that a BMI under 18.5kg/m^2 is considered underweight, comprising only 2.3% of the population. A good amount of 23.6% of the people fall under the healthy category, which is BMI range from 18.5

to 24.9kg/m^2. Next, the overweight category with a BMI of 18.5 to 24.9kg/m^2 is made up of 13.1%, while the remaining 61% consists of obese people.

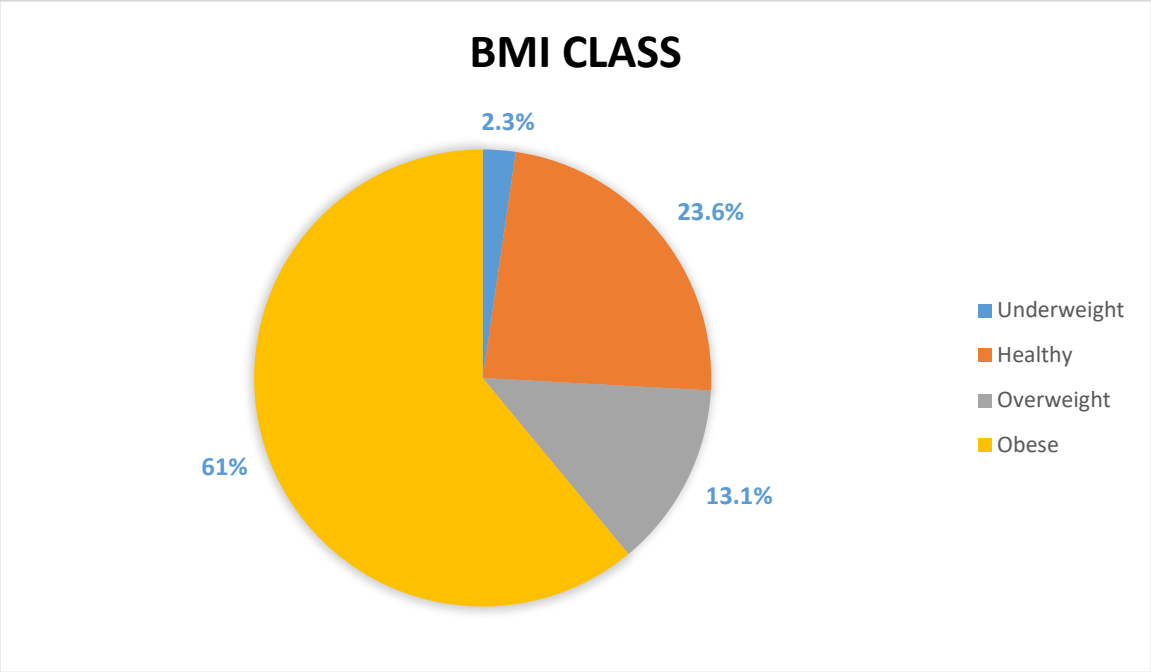


Figure 4.8: A pie chart representing the percentage of the Body Mass Index in the total population

Table 11: Frequency distribution of hypertension

Blood Pressure	**Frequency**	**Percentage (%)**
Yes	**110**	**42.5**
No	**149**	**57.5**

Using the systolic and diastolic blood pressure values, it was determined if the person had blood pressure. It can be seen that only 42.5% have blood pressure while the remaining 57.5% have normal blood pressure and hence no blood pressure.

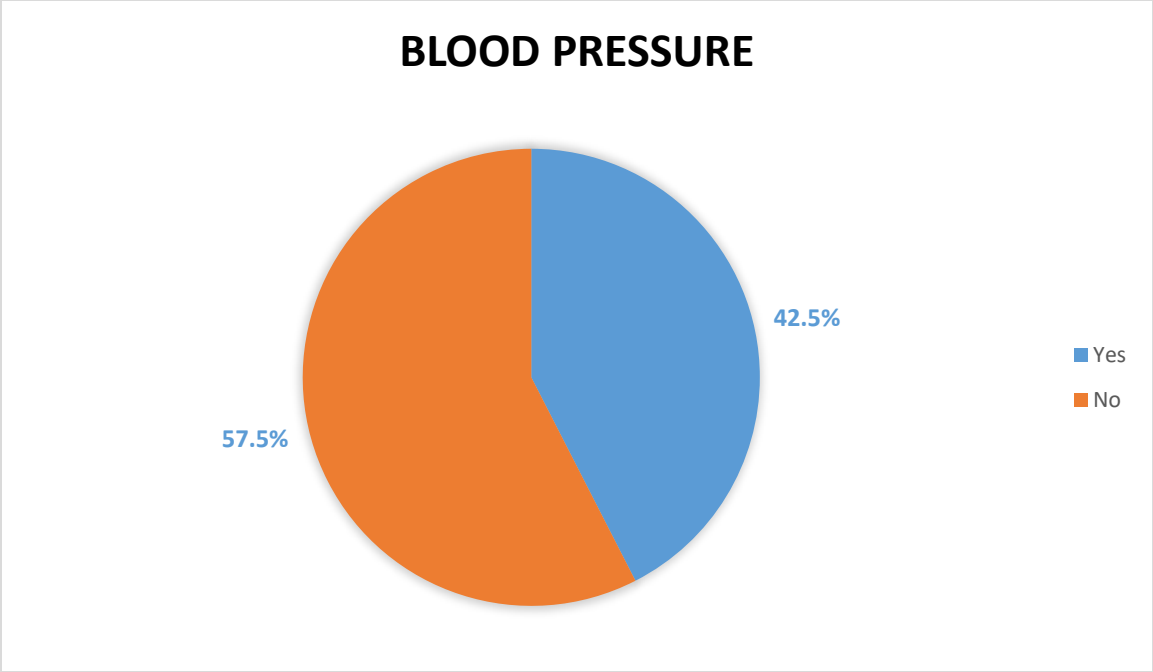


Figure 4.9: A pie chart indicating the percentage of prevalence of blood pressure in the study

Table 12: Frequency distribution of heart disease

Heart Disease	**Frequency**	**Percentage (%)**
Yes	**61**	**23.6**
No	**198**	**76.4**

From the above table, only 10.8% of the individuals suffer from heart disease. Most other people, with a massive percentage of 89.2%, do not have any heart diseases.

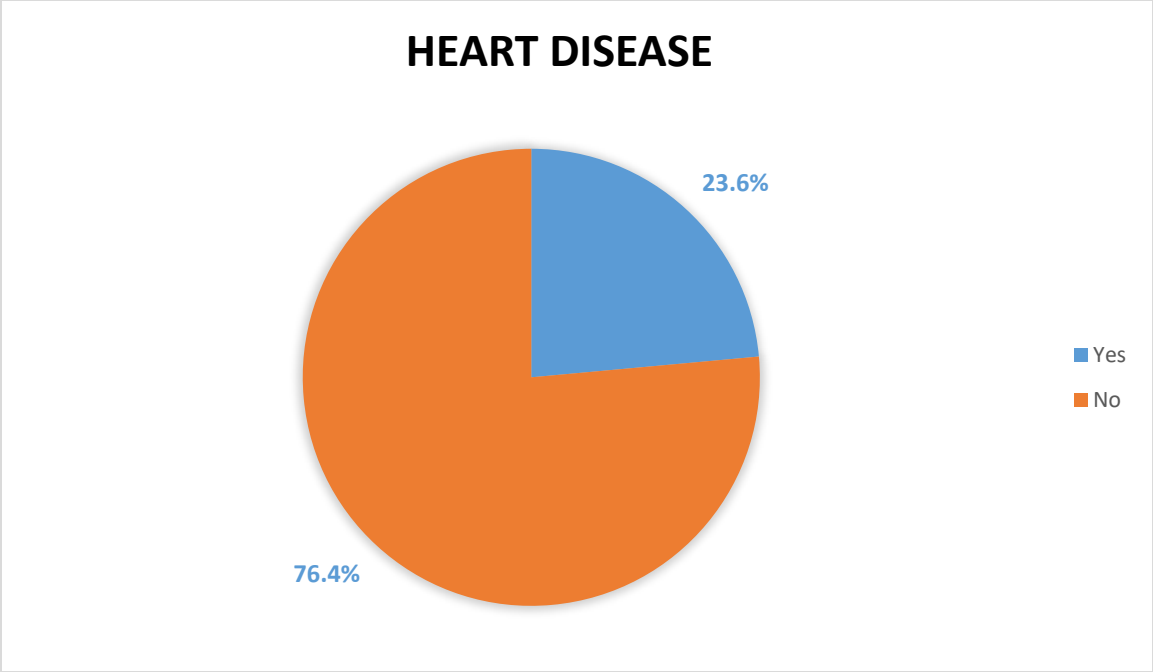


Figure 4.10: A pie chart indicating the percentage of prevalence of heart diseases in the study

Table 13: Frequency distribution of kidney disease

Kidney Disease	**Frequency**	**Percentage (%)**
Yes	**15**	**5.8**
No	**244**	**94.2**

According to this table, very few people have kidney diseases, only 5.8% of the total population. Again a significantly large amount of people are not afflicted with any kidney diseases.

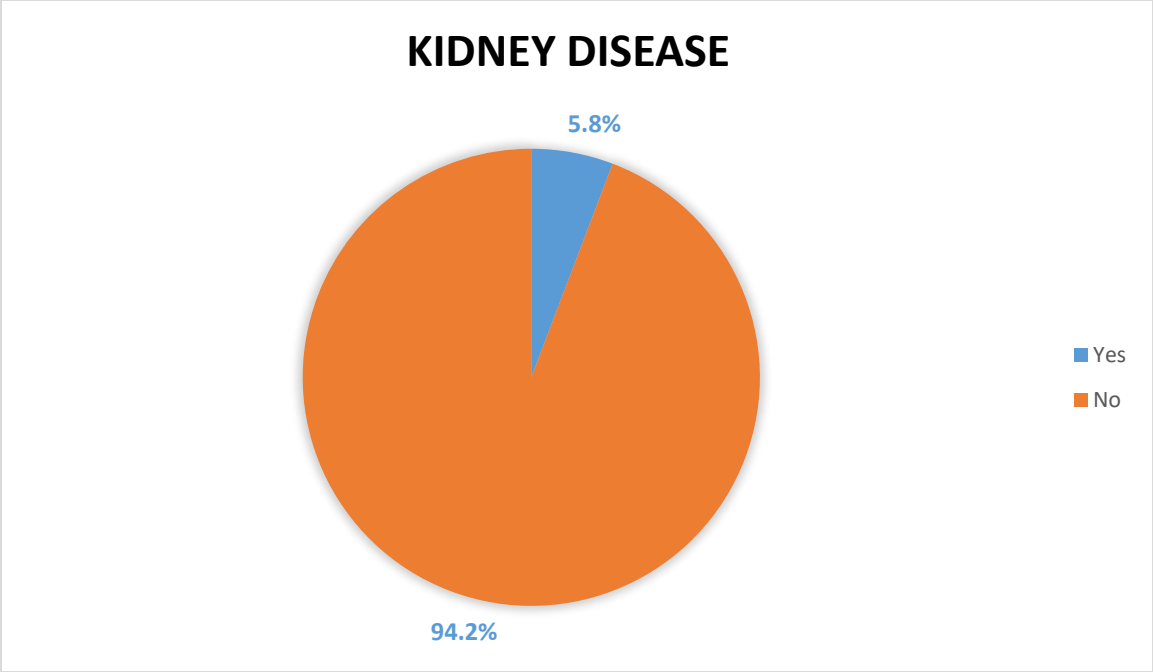


Figure 4.11: A pie chart indicating the percentage of prevalence of kidney diseases in the study

Table 14: Frequency distribution of other diseases

Other Disease	**Frequency**	**Percentage (%)**
Yes	**102**	**39.4**
No	**157**	**60.6**

The table of other diseases shows that only 39.4% suffer from any other illness besides those mentioned already. However, 60.6% do not have any other diseases.

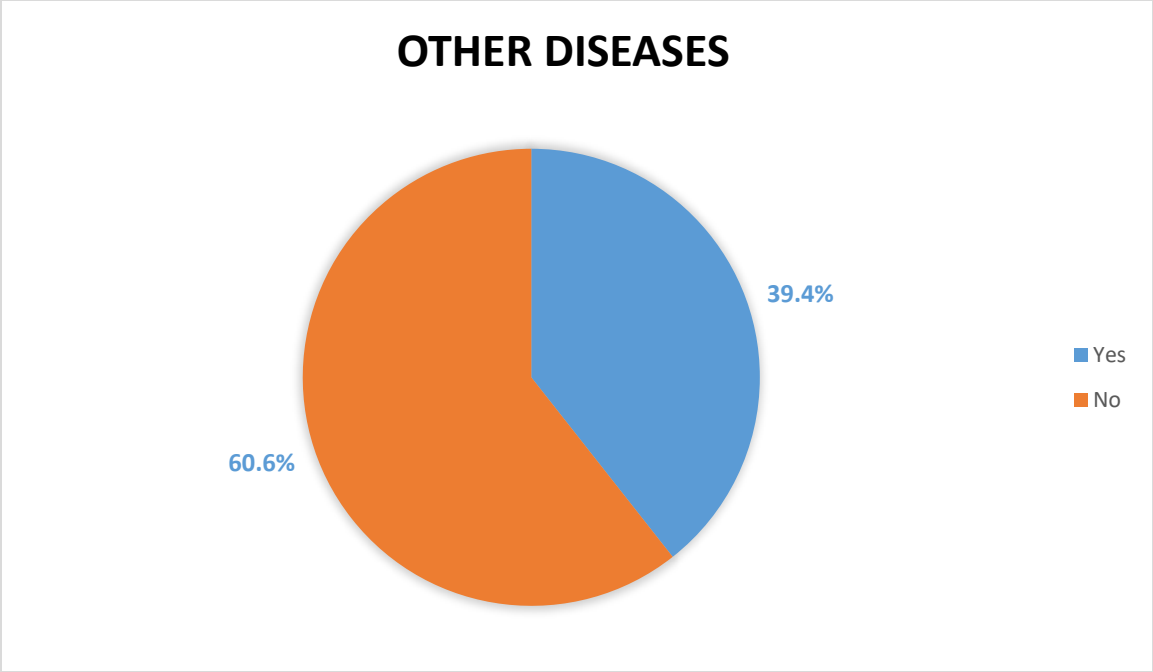


Figure 4.12: A pie chart representing the percentage of other diseases in the study

Table 15: Frequency distribution of other diseases

Other Disease	Frequency	Percentage (%)
Asthma	**6**	**2.3**
Cataract	**5**	**1.9**
Chronic Renal Failure (CRF)	**3**	**1.2**
Congenital Heart Defects (CHD)	**2**	**0.8**
Dental Caries	**6**	**2.3**
Dyslipidemia	**4**	**1.5**
Foot Ulcer	**1**	**0.4**
Hepatitis C	**1**	**0.4**
Migraine	**2**	**0.8**
Peripheral Vascular Disease (PVD)	**5**	**1.9**
Sinus	**3**	**1.2**
Skin Disease	**1**	**0.4**
Urinary Tract Infection (UTI)	**2**	**0.8**
Others	**61**	**23.6**

From observations, there are 102 people with other diseases. The table above simply summarizes the number of people with different diseases. Among them, 23.6% is the highest percentage of people who suffer from diseases other than those mentioned. There is a significant amount of people with dental caries and asthma as well, with 2.3% each. According to the data, with a percentage of 1.9% each, the next most prevalent diseases are Cataracts and PVD. Another 1.5% said they have dyslipidemia. This is followed by sinus and CRF, both comprising 1.2% of the population. Also, people who suffer from CHD, migraine, and UTI are only 0.8% each. The

remaining ones, each with a percentage of 0.4%, said they have foot ulcers, hepatitis C, and skin disease.

4.1.4 Family history data

Table 16: Frequency distribution of patient records of diabetes

Diabetes Record	Frequency	Percentage (%)
Yes	**161**	**62.2**
No	**98**	**37.8**

The diabetes record table shows the number of people from the study who actually have diabetes. It illustrates that 62.2% of the population do have diabetes while 37.8% said they do not. So, the focus is more on the diabetes patients.

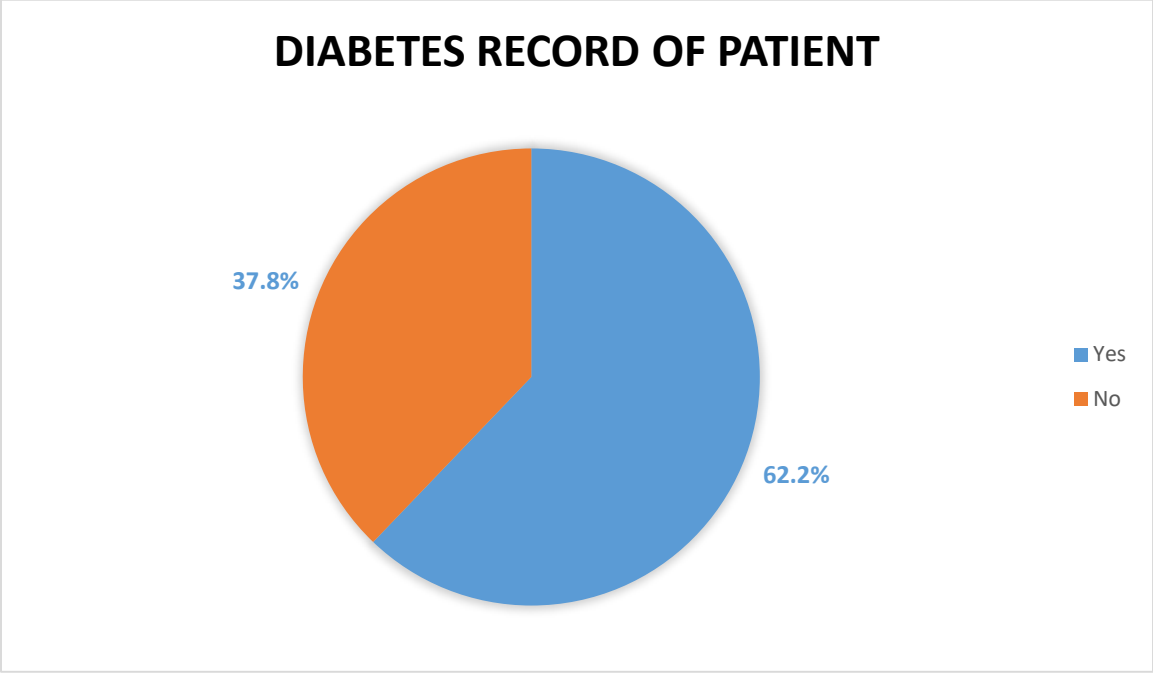


Figure 4.13: A pie chart indicating the percentage of diabetes patients in the study

Table 17: Frequency distribution of diabetes record in the family

Diabetes Record of Family	Father	Percentage	Mother	Percentage	Siblings	Percentage
Yes	**160**	**61.8%**	**135**	**52.1%**	**120**	**46.3%**
No	**99**	**38.2%**	**124**	**47.9%**	**139**	**53.7%**

From the total population, 61.8% said their father had diabetes, while only 38.2% of the people's fathers do not have diabetes. For the mother's record, the table shows that 52.1% of the individuals' mothers suffer from diabetes, and 47.9% are free from it. According to the siblings' records, only 46.3% said their brother/sister had diabetes. The remaining 53.7% do not have diabetes.

Table 18: Frequency distribution of diabetes record in the family of diabetic patients

Diabetes Record of Family	Father	Percentage	Mother	Percentage	Siblings	Percentage
Yes	**129**	**80.1%**	**121**	**75.2%**	**116**	**72.0%**
No	**32**	**19.9%**	**40**	**24.8%**	**45**	**28.0%**

Among the diabetes patients, 80.1% of the fathers had diabetes, and the remaining 19.9% had none. The record shows that 75.2% said their mothers had diabetes, but 24.8% did not have it. Of the siblings, 72% also had diabetes, leaving 28% with no problems with the disease.

4.1.5 Cross-tabulation

Table 19: Cross Table Analysis between diabetes patients and Body Mass Index

Diabetes	***Body Mass Index (BMI)***								***Total***	
	Underweight		**Healthy**		**Overweight**		**Obese**			
	Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**
Yes	**-**	**-**	**4**	**1.5**	**10**	**3.9**	**147**	**56.8**	**161**	**62.2**
No	**6**	**2.3**	**57**	**22.0**	**24**	**9.3**	**11**	**4.2**	**98**	**37.8**
Total	**6**	**2.3**	**61**	**23.5**	**34**	**13.2**	**158**	**61**	**259**	**100**

This table shows the BMI range of the diabetes patients along with those who do not have diabetes. So, based on the data presented above, only 2.3% of the people who do not have diabetes are underweight. Again, nobody in this range was a diabetes patient. In addition, of 23.5% of the patients who fall within the standard BMI limit, 1.5% said they had diabetes, while a huge 22% said they did not. Also, from the overweight category, only 3.9% had diabetes. Finally, it is evident that a disproportionate number of people with diabetes are obese, with a percentage of 56.8% as opposed to 4.2%.

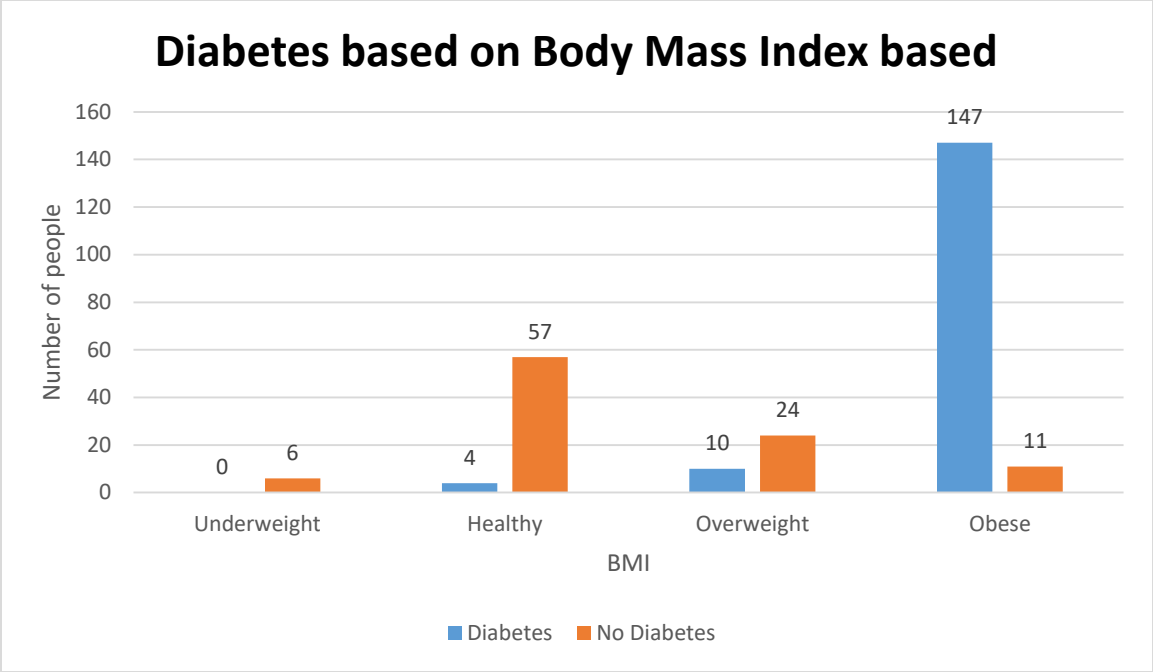


Figure 4.14: A bar graph representation of the diabetes patients based on their Body Mass Index in the population

Table 20: Cross Table Analysis between Gender and Body Mass Index among diabetes patients

Gender	***Body Mass Index (BMI)***								***Total***	
	Underweight		**Healthy**		**Overweight**		**Obese**			
	Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**
Male	**-**	**-**	**1**	**0.6**	**6**	**3.7**	**33**	**20.5**	**40**	**24.8**
Female	**-**	**-**	**3**	**1.9**	**4**	**2.5**	**114**	**70.8**	**121**	**75.2**
Total	**-**	**-**	**4**	**2.5**	**10**	**6.2**	**147**	**91.3**	**161**	**100**

This table groups the male and female diabetes patients into their appropriate BMI categories. From the total population, 161 people had diabetes which is 62.2%. Among these individuals, only 2.5% are healthy, of which females cover a large portion of 1.9% versus 0.6% males. However, of the 6.2% of overweight people, 3.7% are males. Moreover, a substantial amount of female participants were found to be obese, with an alarming percentage of 70.8% as compared to 20.5% males. The table further shows that none of the patients are underweight.

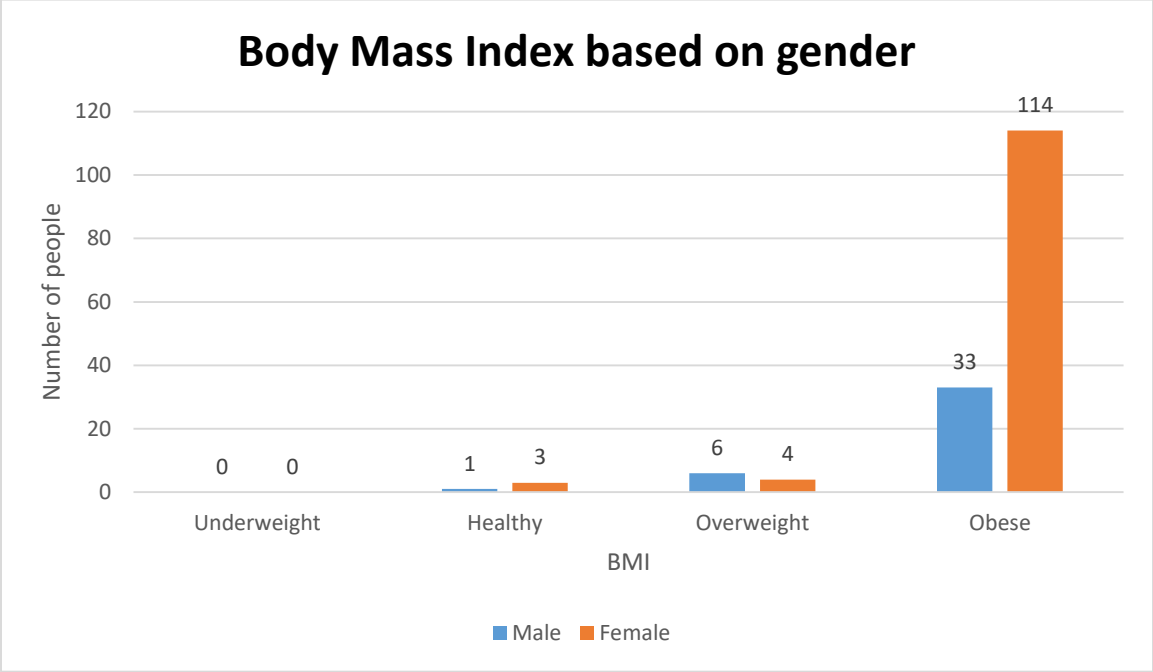


Figure 4.15: A bar graph representation of Body Mass Index among diabetes patients based on their gender in the population

Table 21: Cross Table Analysis between gender and the occurrence of diabetes

Gender	***Diabetes***				***Total***	
	Yes		**No**			
	Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**
Male	**40**	**15.4**	**45**	**17.4**	**85**	**32.8**
Female	**121**	**46.7**	**53**	**20.5**	**174**	**67.2**
Total	**161**	**62.1**	**98**	**37.9**	**259**	**100**

The above table separates the diabetes patients from those who do not have diabetes based on their gender. It follows that among the 32.8% of males, 15.4% said they had diabetes, whereas 17.4% said they did not have diabetes. In comparison, of the 67.2% of females, a large proportion of 46.7% were diabetes patients, and only 20.5% responded with a no.

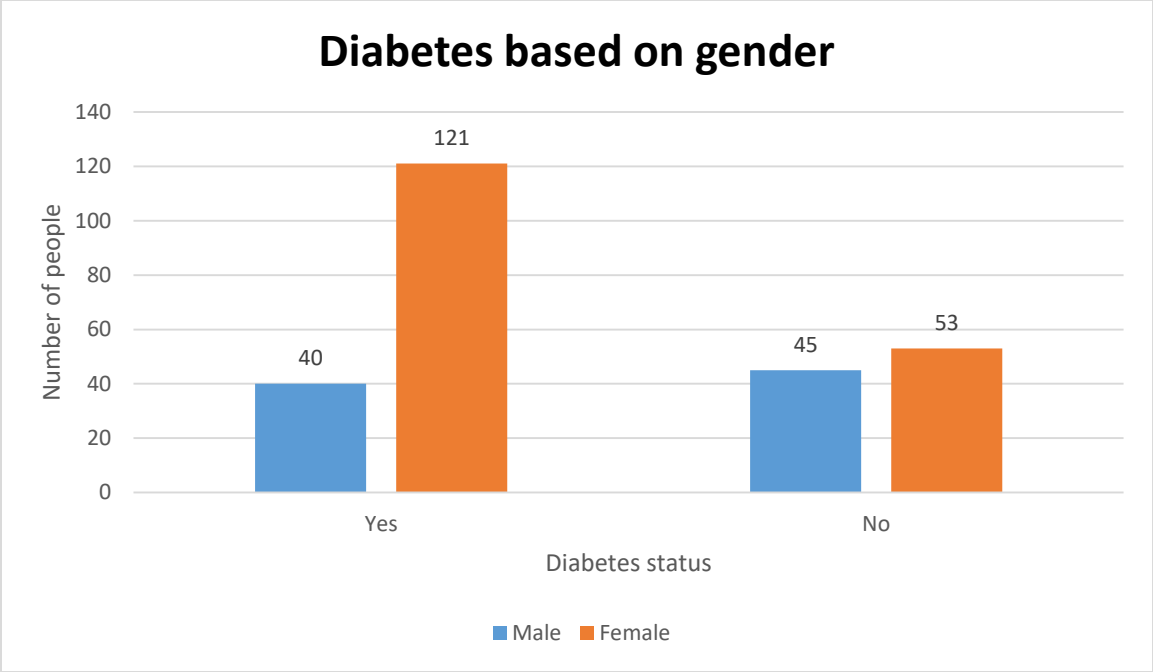


Figure 4.16: A bar graph representation of the diabetes patients based on gender in the population

Table 22: Cross Table Analysis between diabetes and the occurrence of heart diseases

Diabetes	***Heart Disease***				***Total***	
	Yes		**No**			
	Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**
Yes	**60**	**23.2**	**101**	**39.0**	**161**	**62.2**
No	**1**	**0.4**	**97**	**37.4**	**98**	**37.8**
Total	**60**	**23.6**	**101**	**76.4**	**259**	**100**

The table above studies the relationship between diabetes patients and heart diseases in the population. Of the 23.6% of heart disease patients, 23.2% also were diabetes patients, but only 0.4% had no diabetes. On the other hand, from the 76.4% of people with no heart disease, 39% said they had diabetes, as opposed to the 37.4% of people who do not have diabetes.

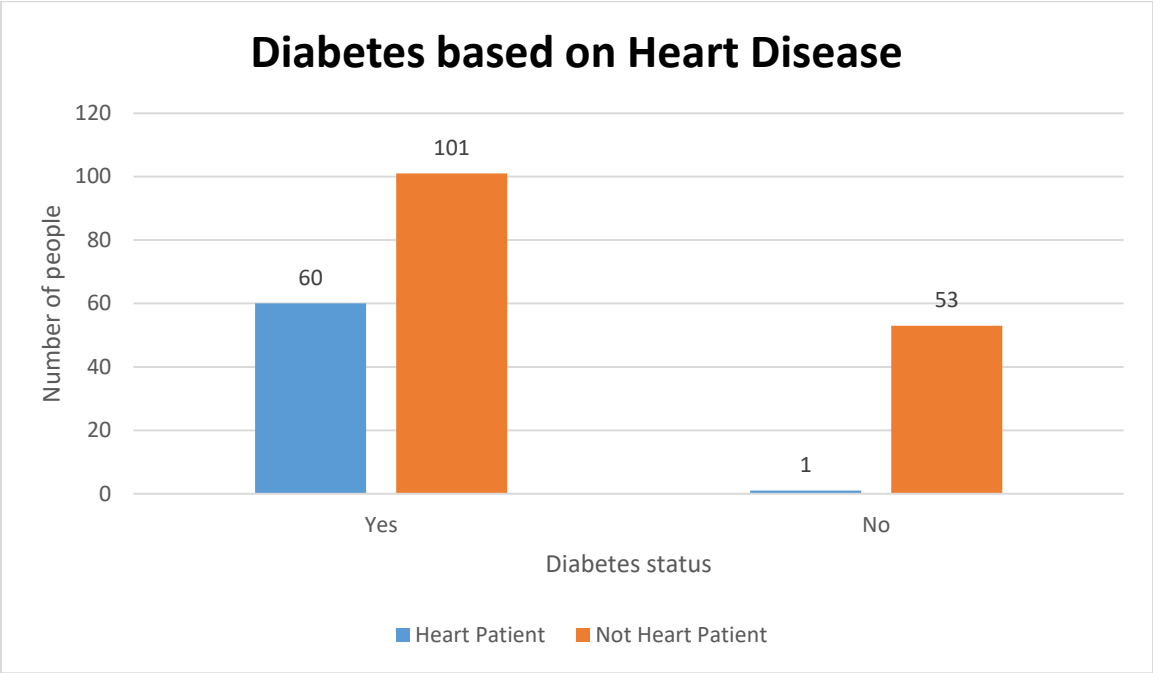


Figure 4.17: A bar graph representation of the prevalence of heart diseases among diabetes patients in the population

Table 23: Cross Table Analysis between gender and the occurrence of heart diseases among diabetes patients

Gender	***Heart Disease***		***Total***	
	Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**
Male	**19**	**31.7**	**19**	**31.7**
Female	**41**	**68.3**	**41**	**68.3**
Total	**60**	**100**	**60**	**100**

The table groups the diabetes patients with heart diseases into males and females. As seen clearly, there are 31.7% males and 68.3% females, indicating that more females took part in this study.

Table 24: Cross Table Analysis between diabetes and the occurrence of high blood pressure

Diabetes	***Blood Pressure***				***Total***	
	Yes		**No**			
	Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**
Yes	**79**	**30.5**	**82**	**31.7**	**161**	**62.2**
No	**31**	**11.9**	**67**	**25.9**	**98**	**37.8**
Total	**110**	**42.4**	**149**	**63.6**	**259**	**100**

This table studies the relationship between diabetes patients and the population's blood pressure patients. Around 42.4% of the people had blood pressure, of which 30.5% said they had diabetes, while 11.9% said they had no diabetes. Of the 63.6% of people with no blood pressure, 31.7% were diabetes patients, and 25.9% did not have diabetes.

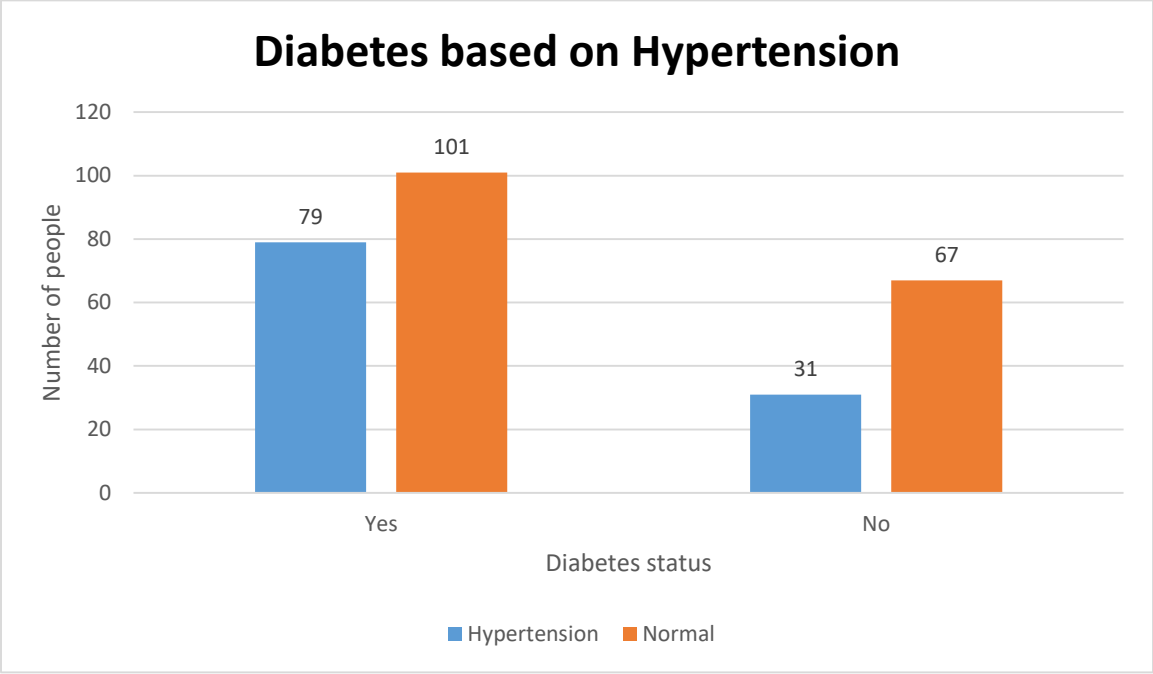


Figure 4.18: A bar graph representation of the prevalence of hypertension among diabetes patients in the population

Table 25: Cross Table Analysis between gender and the occurrence of high blood pressure among diabetes patients

Gender	***Blood Pressure***		***Total***	
	Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**
Male	**17**	**21.5**	**17**	**21.5**
Female	**62**	**78.5**	**62**	**78.5**
Total	**79**	**100**	**79**	**100**

This table groups the diabetes patients who have blood pressure into their gender. It shows that most people with blood pressure problems are females, with a percentage of 78.5%. In contrast, the males are as little as 21.5%. This is again in line with females being in more numbers than males in the study.

Table 26: Cross Table Analysis between diabetes and the occurrence of kidney diseases

Diabetes	***Kidney Disease***				***Total***	
	Yes		**No**			
	Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**
Yes	**15**	**5.8**	**146**	**56.4**	**161**	**62.2**
No	**-**	**-**	**98**	**37.8**	**98**	**37.8**
Total	**60**	**5.8**	**101**	**94.2**	**259**	**100**

The table demonstrates the relationship between diabetes patients and heart diseases in the population. From this table, only 5.8% of the people had both diabetes and kidney disease. For those with no kidney diseases, 56.4% said they had diabetes, whereas 37.8% had none.

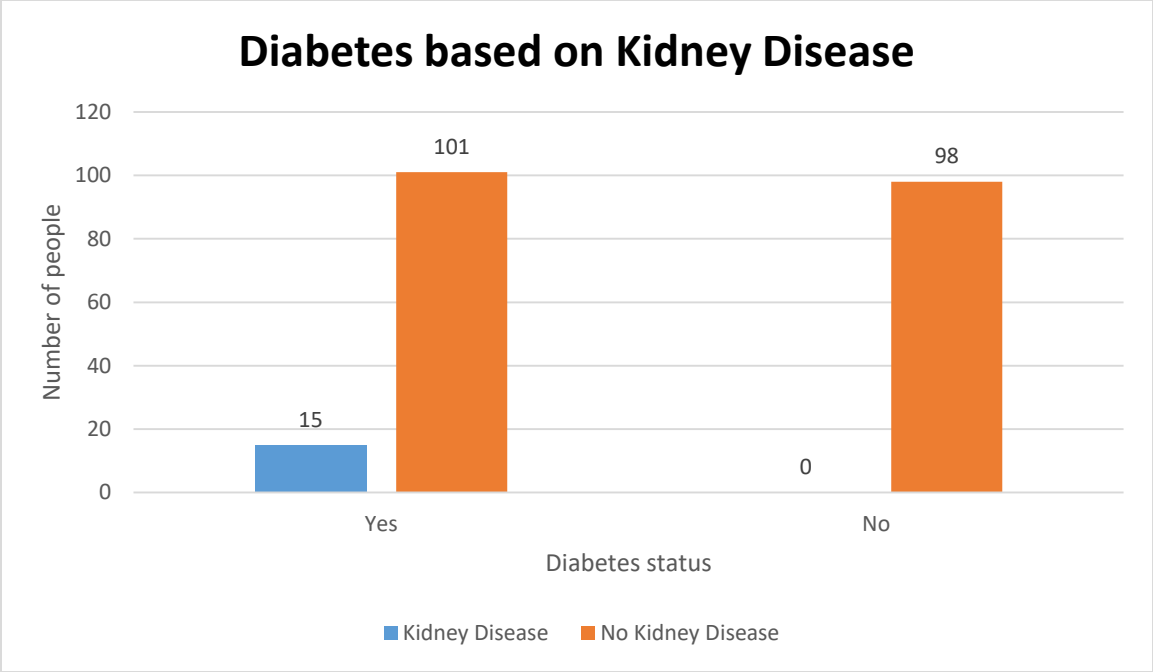


Figure 4.19: A bar graph representation of kidney diseases among diabetes patients in the population

Table 27: Cross Table Analysis between diabetes and smoking habit

Diabetes	***Smoking Habit***				***Total***	
	Yes		**No**			
	Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**
Yes	**16**	**6.2**	**145**	**56.0**	**161**	**62.2**
No	**12**	**4.6**	**86**	**33.2**	**98**	**37.8**
Total	**28**	**10.8**	**231**	**89.2**	**259**	**100**

The table above shows a relationship between smoking habits among people with no diabetes and diabetes patients. It can be seen that only 10.8% of the population smokes. Among them, 6.2% of people have diabetes. Nevertheless, for non-smokers, a large 56% of the people were found to have diabetes as compared to the 33.2% with no diabetes.

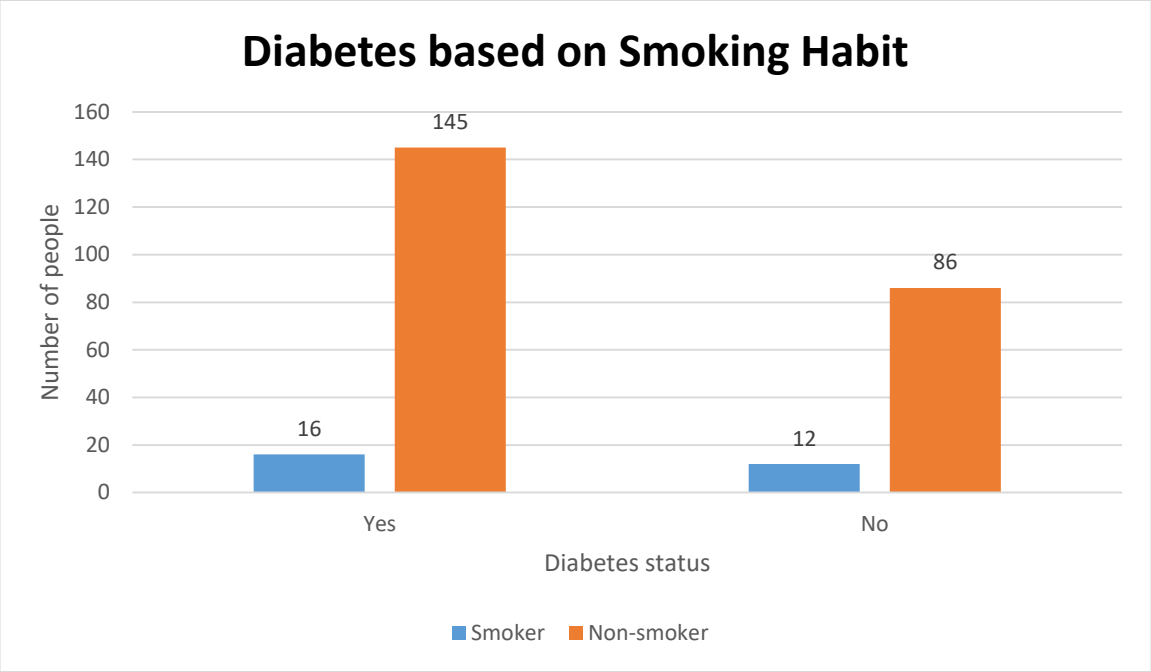


Figure 4.20: A bar graph representation of smoking habits among diabetes patients in the population

Table 28: Cross Table Analysis between diabetes and the living conditions of the people

Diabetes	***Living Condition***				***Total***	
	Urban		**Rural**			
	Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**
Yes	**125**	**48.3**	**36**	**13.9**	**161**	**62.2**
No	**83**	**32.0**	**15**	**5.8**	**98**	**37.8**
Total	**60**	**80.3**	**101**	**19.7**	**259**	**100**

The table demonstrates that many people with diabetes live in the city, with a percentage of 48.3%. In contrast, just 13.9% of diabetes patients are from the countryside. Only 5.8% and a moderate amount of 32% of people who do not have diabetes are from the rural area and urban regions, respectively.

Table 29: Cross Table Analysis of diabetes based on profession

Profession	***Diabetes***				***Total***	
	Yes		**No**			
	Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**
Business	**20**	**7.7**	**11**	**4.2**	**31**	**11.9**
Housewife	**95**	**36.7**	**4**	**1.5**	**99**	**38.2**
Service	**38**	**14.7**	**22**	**8.5**	**60**	**23.2**
Student	**1**	**0.4**	**59**	**22.8**	**60**	**23.2**
Retired	**7**	**2.7**	**2**	**0.8**	**9**	**3.5**
Total	**125**	**62.2**	**36**	**37.8**	**259**	**100**

From the data represented in the above table, the highest number of people with diabetes are housewives, with 36.7%. Only 1.5% of housewives said they do not have diabetes. Next, a fairly large quantity of diabetes patients was in service jobs, comprising 14.7% as opposed to 8.5%. The table also shows that out of the 11.9% of people in the business sector, those with diabetes outnumbered those with no diabetes, making up 7.7% and 4.2%, respectively. Since only a small number of retired people participated, comprising 3.5%, it covered a minority of people with and without diabetes. Even so, the diabetes patients in this category were moderately disproportionate, where 2.7% had diabetes. Also, according to the table, most students are not diagnosed with diabetes, and they cover 22.8%. This leaves only 0.4% of students having diabetes.

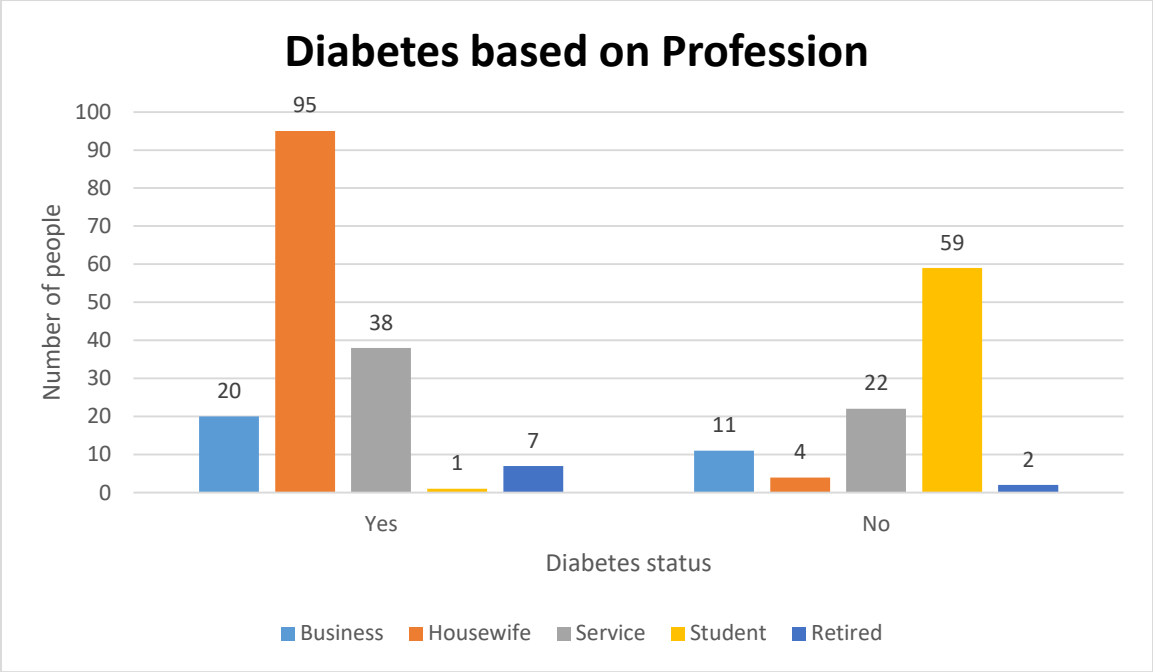


Figure 4.21: A bar graph representation of the prevalence of diabetes among the different professions in the total population

Table 30: Cross Table Analysis of diabetes based on sleep duration

Sleep Duration	***Diabetes***				***Total***	
	Yes		**No**			
	Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**	**Frequency**	**Percentage (%)**
4 hours	**-**	**-**	**1**	**0.4**	**1**	**0.4**
5 hours	**4**	**1.5**	**13**	**5.0**	**17**	**6.5**
6 hours	**34**	**13.1**	**23**	**8.9**	**57**	**22**
7 hours	**57**	**22.0**	**24**	**9.3**	**81**	**31.3**
8 hours	**52**	**20.1**	**26**	**10.0**	**78**	**30.1**
9 hours	**12**	**4.6**	**7**	**2.7**	**19**	**7.3**
10 hours	**2**	**0.8**	**2**	**0.8**	**4**	**1.6**
12 hours	**-**	**-**	**2**	**0.8**	**2**	**0.8**
Total	**161**	**62.2**	**98**	**37.8**	**259**	**100**

This table depicts how long a person sleeps affects their risk of getting diabetes. Out of the 62.2% of people with diabetes, it can be seen that 22% slept for 7 hours, 20.1% slept for 8 hours, and 13.1% slept for 6 hours. This shows that 6-8 hours of sleep is typical. In some cases, people slept for 9 hours, with 4.6%, followed by a 5-hour sleep comprising 1.5%. Very few even slept for 10 hours, which is only 0.8%. For the remaining 37.8% of the population with no diabetes, the greatest proportion of people slept for 8 hours, having a percentage of 10%. A fair amount of people slept

for 6 and 7 hours as well, making up 8.9% and 9.3%, respectively. Only 0.4% got 4 hours of sleep. Moreover, as the duration increases from 9 to 12 hours, the number of people reduces.

Table 31: Cross Table Analysis of diabetes based on income

Family/Own Income (Tk.)	***Diabetes***		***Total***	
	Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**
2000-5000	**22**	**13.7**	**22**	**13.7**
5001-10000	**24**	**14.9**	**24**	**14.9**
10001-20000	**40**	**24.8**	**40**	**24.8**
20001-30000	**22**	**13.7**	**22**	**13.7**
30001-40000	**15**	**9.3**	**15**	**9.3**
40001-50000	**16**	**9.9**	**16**	**9.9**
50001-100000	**19**	**11.8**	**19**	**11.8**
100000 above	**3**	**1.9**	**3**	**1.9**
Total	**161**	**100**	**259**	**100**

The above table shows the monthly income or (in the case of some students), the family income of diabetes patients. Most of the people earn 10001-20000Tk. comprising 24.8%. The next highest amount consists of 14.9% of people with a salary of 5001-10000Tk. There is an equal number of people with a salary range of 2000-5000Tk. and 20001-30000, having 13.7% each. From the data, it can also be seen that 11.8% of people have an income of 50001-100000Tk. Comparatively, the

patients with a salary range of 30001-40000Tk. and 40001-50000Tk. are small in number and almost equal with 9.3% and 9.9%, respectively. Only very few people have earnings of 100000Tk. above, which is 1.9%.

4.2 Model Analysis

We have used a data mining software called "RapidMiner" to generate the results from our collected data and build our predictive models. Below are the confusion matrices of each model, which are produced using our dataset. The performance metrics are calculated and compared below.

4.2.1 Logistic Regression

Table 32: Confusion Matrix of Logistic Regression

		ACTUAL VALUES		
	Test Result	Disease Present	Disease Absent	Total
PREDICTED VALUES	Positive	142	9	151
	Negative	19	89	108
	Total	161	98	259

Using the data, the first model, Logistic Regression, has the following score on each performance measure: Accuracy= 89.18%, Precision= 94.04%, Predicted Value Negative= 82.41%, Recall= 88.20%, and Specificity= 90.82%. Each of these numerical values is of some significance. We will now discuss what these quantities infer about each model.

Accuracy= 89.18%

Explanation: Having an 89.18% accuracy indicates this model has an overall success rate of 89.18%. This means out of all the predictions made, this model correctly identifies the results 89.18% of the time, or it accurately classified 89.18% of the cases within the dataset. In other words, the model correctly identified 89.18% of the data points within the dataset while incorrectly labelling the remaining 10.82%.

Precision= 94.04%.

Explanation: Having a 94.04% precision shows that 94.04% of the time, this model correctly predicts the positive patterns (TPs) from the total predicted positives, or 94.04% of the positive occurrences were correctly classified by the model. Simply put, 94.04% of the cases the model predicts as positive are true positives, whereas the remaining 5.96% are false positives, i.e., cases that were wrongly classified as positive but are actually negative. So, when the model predicts that a person has diabetes, it is correct around 94.04% of the time, or in other words, 94.04% of the people from this population whose test results are positive are diseased.

Predicted value negative (PVN)= of 82.41%

Explanation: Having a predicted value negative (PVN) of 82.41% suggests that 82.41% of the time, this model correctly predicts the negative patterns (TNs) from the total predicted negatives or 82.41% of the negative instances were correctly classified by the model. More specifically, 82.41% of the cases the model predicts to be negative are indeed negatives, while the remaining 17.59% are false positives, i.e., cases that were wrongly classified as negative but are actually positive. So, when the model predicts that a person has no diabetes, it is correct around 82.41% of the time, or in other words, 82.41% of the people from this population whose test results are negative, are healthy.

Recall= 88.20%

Explanation: A recall rate of 88.20% indicates the model predicts correct positive values from the actual positives, 88.20% of the time. This means for all the patients who actually have diabetes, the model correctly identified patients as having diabetes 88.20% of the time, or 88.20% of positive cases in the class of interest were successfully detected by the model. Put differently, the model did well identifying 88.20% of the situations where the outcome was negative, but it incorrectly identified 11.8% of the true negatives, resulting in false positives.

Specificity= 90.82%

Explanation: A specificity of 90.82% implies that the model predicts correct negative values from the actual negatives, 90.82% of the time. This means for all the patients who do not actually have diabetes, the model correctly identified them as not having diabetes 90.82% of the time, or 90.82% of the actual negative cases from the class other than the one of interest, were classified correctly by the model. To be clear, the model successfully recognized 90.82% of the situations where the outcome was negative, but it incorrectly identified 9.18% of the cases where the outcome was negative, resulting in false positives.

4.2.2 K-Nearest Neighbor

Table 33: Confusion Matrix of K-NN

		ACTUAL VALUES		
	Test Result	Disease Present	Disease Absent	Total
PREDICTED VALUES	Positive	141	35	176
	Negative	20	63	83
	Total	161	98	259

Using the data, the second model, K-NN, has the following score on each performance measure: Accuracy= 78.78%, Precision= 80.11%, Predicted Value Negative= 75.90%, Recall= 87.58%, and Specificity= 64.29%

Accuracy= 78.78%

Explanation: An accuracy of 78.78% shows that the model accurately classified the presence or absence of diabetes in 78.78% of the occurrences within the dataset. To be more precise, 78.78% of the data points in the sample are correctly identified by the model, while the remaining 21.22% are incorrectly classified.

Precision= 80.11%

Explanation: A precision of 80.11% means that among all the occurrences classified as positive (i.e., predicted to have diabetes) by the model, approximately 80.11% of them are accurately identified as true positives (i.e., correctly identified cases of diabetes). Put differently, within the set of instances that the model has identified as positive, 80.11% of them are indeed positive, while the remaining 19.89% are false positives, meaning they are mistakenly classified as positive when they are actually negative. So, the model's predictions for individuals diagnosed with diabetes are correct for about 80.11% of the cases.

Predicted Value Negative= 75.90%

Explanation: A PVN of 75.90% suggests that among all the occurrences classified as negative (i.e., predicted to not have diabetes) by the model, around 75.90% of them correspond to true negatives (i.e., correctly identified cases of being healthy). Simply put, within the set of instances that the model has predicted as negative, 75.90% of them are indeed negative, while the remaining 24.10%

are classified as false negatives, signifying instances that have been inaccurately labelled as negative despite being positive. So, the model's predictions for individuals not diagnosed with diabetes are correct for about 75.90% of the cases.

Recall= 87.58%

Explanation: An 87.58% recall rate indicates that the model accurately classified 87.58% of the positive instances of individuals with diabetes relative to the total number of positive instances. In other words, the model exhibited a 12.42% failure rate in detecting positive cases, thereby classifying them as false negatives, but it was effective in capturing 87.58% of the true diabetic cases present in the dataset.

Specificity= 64.29%

Explanation: A 64.29% specificity signifies that the model accurately classified 64.29% of the actual negative instances or individuals without diabetes as negatives. In simpler terms, the model correctly identified 64.29% of the negative instances or true non-diabetic cases in the dataset while incorrectly classifying 35.71% of the true negatives as positives, resulting in false positives.

4.2.3 Naïve Bayes

Table 34: Confusion Matrix of Naïve Bayes

		ACTUAL VALUES		
	Test Result	Disease Present	Disease Absent	Total
PREDICTED VALUES	Positive	157	11	168
	Negative	4	87	91
	Total	161	98	259

Using the data, the third model, Naïve Bayes, has the following score on each performance measure: Accuracy= 94.23%, Precision= 93.45%, Predicted Value Negative= 95.60%, Recall= 97.52%, and Specificity= 88.78%

Accuracy= 94.23%

Explanation: A prediction accuracy of 94.23% indicates that the model was able to properly identify the category (i.e., presence or absence of diabetes) of 94.23% of the total cases within the dataset. To clarify, out of all the data points in the dataset, the model got 94.23% right, while it misclassified the remaining 5.77%.

Precision= 93.45%

Explanation: A precision of 93.45% suggests that 93.45% of the cases the model identified as positive (i.e., predicted to have diabetes) were actually positive (i.e., cases of diabetes correctly diagnosed). It can be stated that 93.45% of the cases that the model predicts to be positive are truly positive, while the remaining 6.55% are false positives, i.e., cases that were wrongly classified as positive but are actually negative. So, 93.45% of the people from this population whose test results are positive, have the disease.

Predicted Value Negative= 95.60%

Explanation: A PVN of 95.60% shows that 95.60% of the model's negative cases (i.e., cases indicating an absence of diabetes) are properly categorized as true negatives (i.e., cases of correctly identifying people without diabetes). That is to say, 95.60% of the cases that the model labels as negative are truly negative, while the remaining 4.40% are false negatives, i.e., cases that were incorrectly classified as negative but are actually positive. So, 95.60% of the people from this population whose test results are negative, do not have the disease.

Recall= 97.52%

Explanation: Having a recall of 97.52% means that 97.52% of the people with diabetes, among all the actual positive cases, are correctly recognized as positive by the model. To put it another way, the model failed to identify only 2.48% of the true positive cases, which resulted in false negatives. So, 2.48% of the people who actually have diabetes was not detected by the model.

Specificity= 88.78%

Explanation: Having a specificity of 88.78% signifies that 88.78% of the people without diabetes, among all the actual negative cases, are accurately classified as negative by the model. Thus, 88.78% of the negative cases are properly labelled, but 11.20% are misclassified as positives, leading to false positives. So, 11.20% of healthy people were wrongly labelled as having diabetes.

4.2.4 Decision Tree

Table 35: Confusion Matrix of Decision Tree

		ACTUAL VALUES		
	Test Result	Disease Present	Disease Absent	Total
PREDICTED VALUES	Positive	139	13	152
	Negative	22	85	107
	Total	161	98	259

Using the data, the fourth model, Decision Tree, has the following score on each performance measure: Accuracy= 86.49%, Precision= 91.45%, Predicted Value Negative= 79.44%, Recall= 86.34%, and Specificity= 86.73%

Accuracy= 86.49%

Explanation: A model with an accuracy of 86.49% has successfully predicted the category (i.e., presence or absence of diabetes) of 86.49% of the total cases inside the dataset. To be more precise, the model correctly classified 86.49% of the data points in the dataset while incorrectly classifying the rest of the 13.51%.

Precision= 91.45%

Explanation: A model with a precision of 91.45% has rightly labelled 91.45% of the model's positive cases (indicating diabetes presence) as true positives or cases that the model successfully identified as having diabetes. That is to say, 91.45% of the model's projected positives are true positives, whereas the other 8.55% are false positives, i.e., incorrectly categorized as positive but actually negative (indicating diabetes absence). So, 91.45% of the people from this population whose test results are positive have the disease.

Predicted Value Negative= 79.44%

Explanation: A model with a PVN of 79.44% has correctly labelled 79.44% of the model's negative cases (indicating diabetes absence) as true negatives or cases that were properly identified as healthy. In other words, 79.44% of the model's projected negative cases are actually negative, whereas the rest 20.56%, are false negatives, i.e., incorrectly labelled as negative but actually

positive (indicating diabetes presence). So, 79.44% of the people from this population whose test results are negative do not have the disease.

Recall= 86.34%

Explanation: A model with an 86.34% recall rate has successfully detected 86.34% of the positive cases, accurately recognizing diabetes in people in the dataset. Thus, it can also be said that the model missed 13.66% of the actual positive cases (i.e., people who have diabetes), and these instances are the false negatives. So, the model misclassified 13.66% of the people as healthy.

Specificity= 86.73%

Explanation: A model with an 86.73% specificity has successfully classified 86.73% of the actual negative cases as negative, accurately classifying healthy people in the dataset. Simply put, the model did well identifying 86.73% of the situations where the outcome was negative (i.e., people who have no diabetes), but it incorrectly identified 13.27% of the true negative cases, leading to false positives. So, the model misclassified 13. 27% of the people as diseased.

4.2.5 Random Forest

Table 36: Confusion Matrix of Random Forest

		ACTUAL VALUES		
	Test Result	Disease Present	Disease Absent	Total
PREDICTED VALUES	Positive	156	9	165
	Negative	5	89	94
	Total	161	98	259

Using the data, the last model, Random Forest, has the following score on each performance measure: Accuracy= 94.62%, Precision= 94.55%, Predicted Value Negative= 94.68%, Recall= 96.89%, and Specificity= 94.68%

Accuracy= 94.62%

Explanation: With a 94.62% accuracy, the model properly identified the class (i.e., presence or absence of diabetes) for 94.62% of the dataset's occurrences. To rephrase, the model correctly categorized 94.62% of the data points in the sample while incorrectly labeling the other 5.38%.

Precision= 94.55%

Explanation: With a 94.55% precision, of all the instances that the model classified as positive (i.e., presence of diabetes), 94.55% are true positives, representing correctly identified cases of diabetes. More specifically, the model predicts 94.62% of the actual positive instances, and the rest 5.38%, are false positives indicating that 5.38% are wrongly labelled diabetes patients. So, 95.60% of the people from this population whose test results are positive do have the disease.

Predicted Value Negative= 94.68%

Explanation: With a 94.68% PVN, of all the instances that the model classified as negative (i.e., absence of diabetes), 94.68% are true negatives, correctly identifying people without diabetes. This means that 94.68% of the instances predicted by the model to be negative are indeed negative, and the remaining 5.32% are false negatives, i.e., indicating that 5.32% are misclassified as healthy. So, 94.68% of the people from this population whose test results are negative do not have the disease.

Recall= 96.89%

Explanation: With a recall of 96.89%, the model successfully recognized 96.89% of the positive instances (i.e., diabetes patients). To put it another way, only 3.11% of the true positives went unnoticed by the model, resulting in false negatives. So, the model missed out on 3.11% of the diabetes patients.

Specificity= 94.68%

Explanation: With a specificity of 94.68%, the model correctly identified 94.68% of all the actual negative instances (i.e., individuals who are disease free). In simpler terms, only 5.32% of the true negatives were erroneously labelled as positive, leading to false positives. So, the model misclassified only 5.32% of the people as diseased.

4.3 Comparing the models

Now, each model is compared on the basis of their performance measure scores. All the values are analyzed, and the highest percentage recorded is determined from all the models. The performance comparison of all the classifiers based on their individual scores is demonstrated below in separate bar charts.

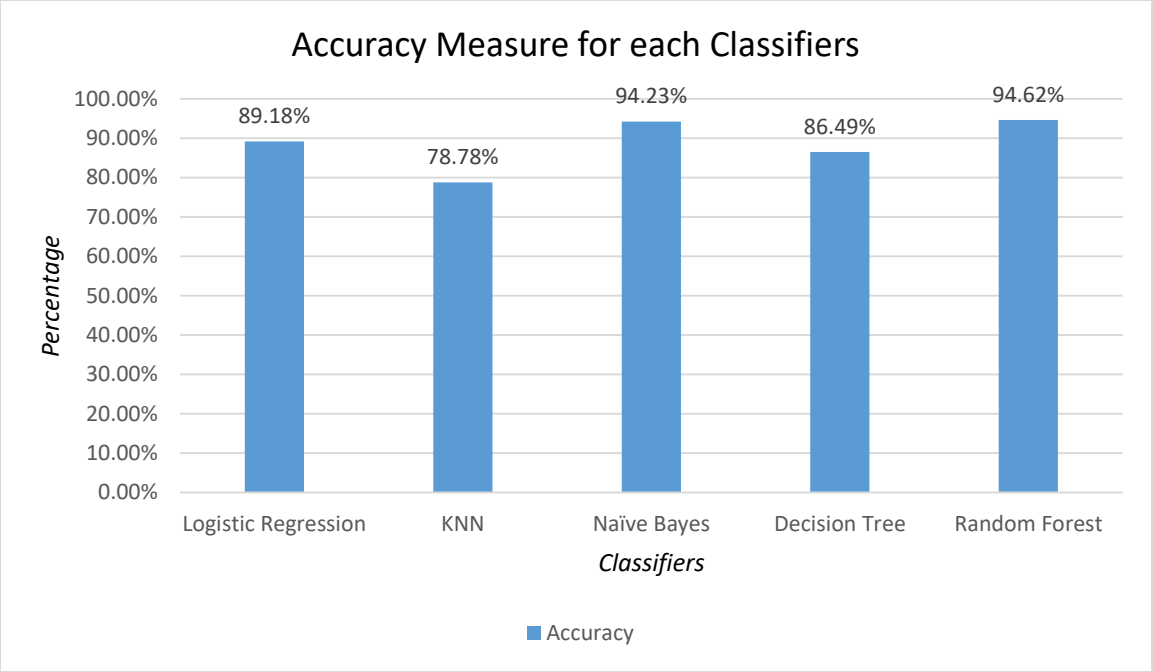


Figure 4.22: Accuracy comparison of each model

Comparing the accuracy score of all the algorithms, the algorithm with the highest accuracy is Random Forest**,** followed by Naïve Bayes and Logistic Regression. Next comes the Decision Tree, with K-NN having the lowest accuracy score.

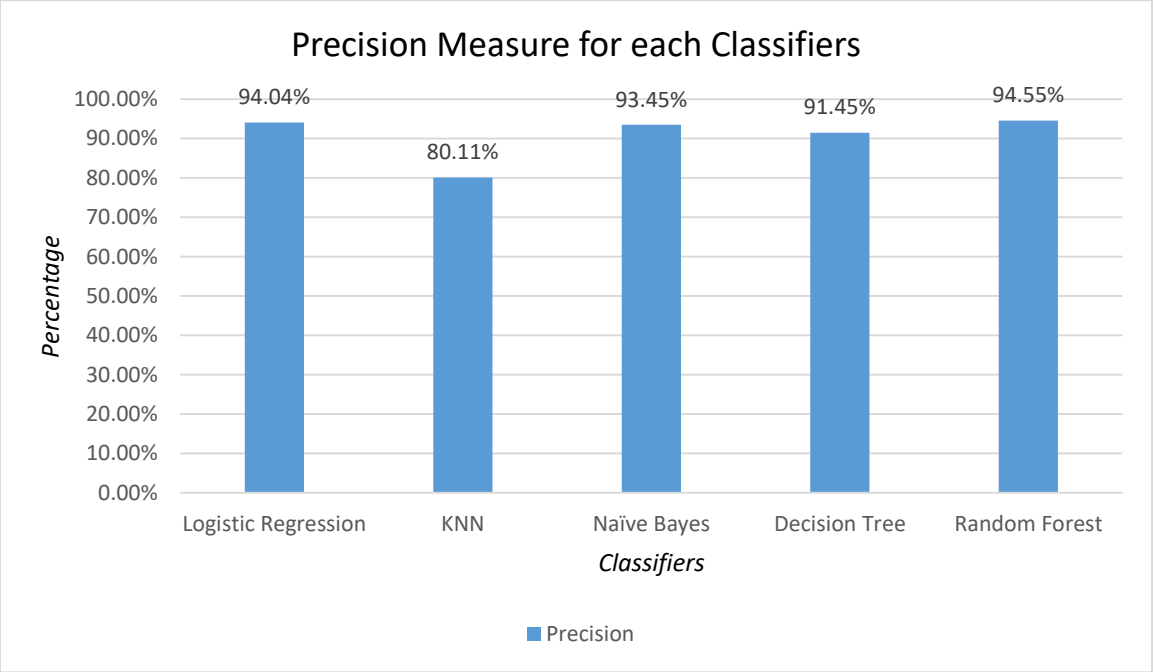


Figure 4.23: Precision comparison of each model

Comparing the precision of all the algorithms, it is found again that Random Forest has the highest precision score. Logistic Regression scores a little less than Random Forest. This is followed by Naïve Bayes and Decision Tree. Once more, K-NN scored the lowest.

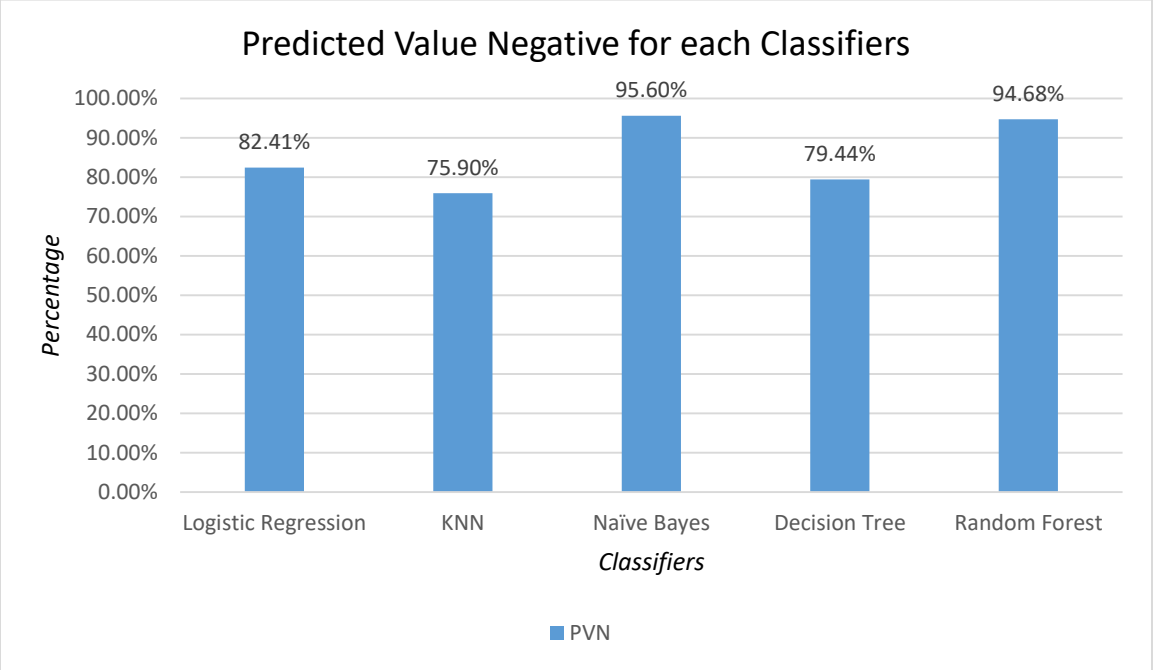


Figure 4.24: Predicted value negative comparison of each model

Comparing the PVN of all the algorithms, this time, Naïve Bayes had the highest score, coming ahead of Random Forest. Logistic Regression follows next with a fair percentage. Decision Tree and K-NN are at the bottom, respectively, for their low scores.

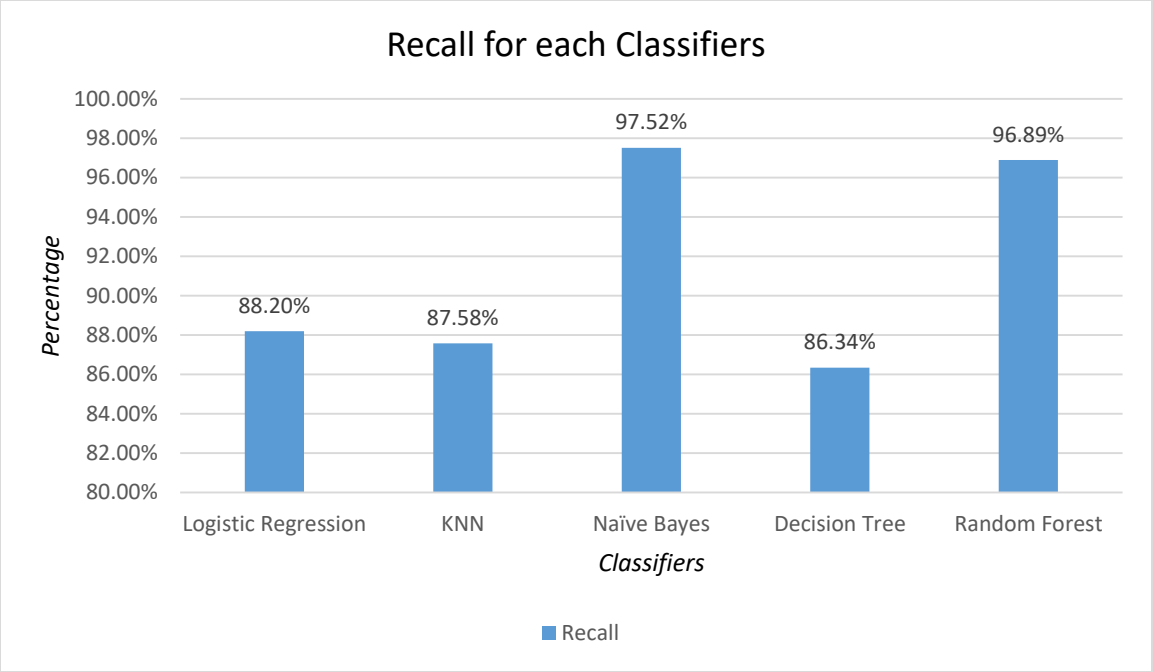


Figure 4.25: Recall comparison for each model

Comparing the recall of all the algorithms, again, Naïve Bayes scored the highest, followed by Random Forest. Logistic Regression and K-NN come next, while the Decision Tree had the lowest percentage.

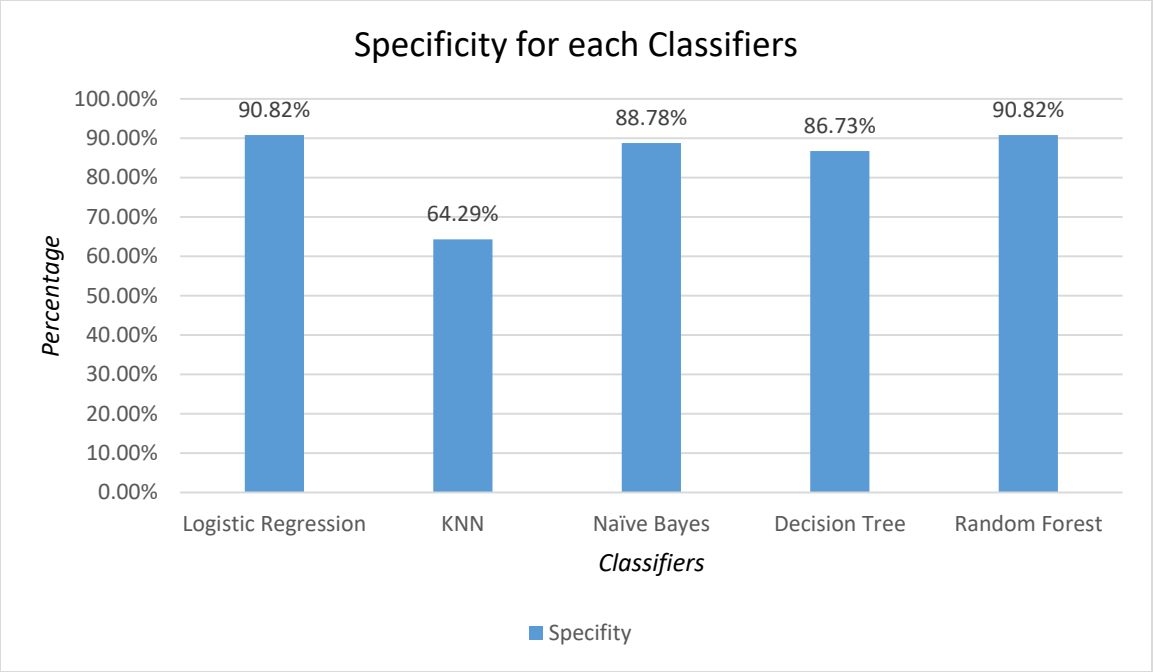
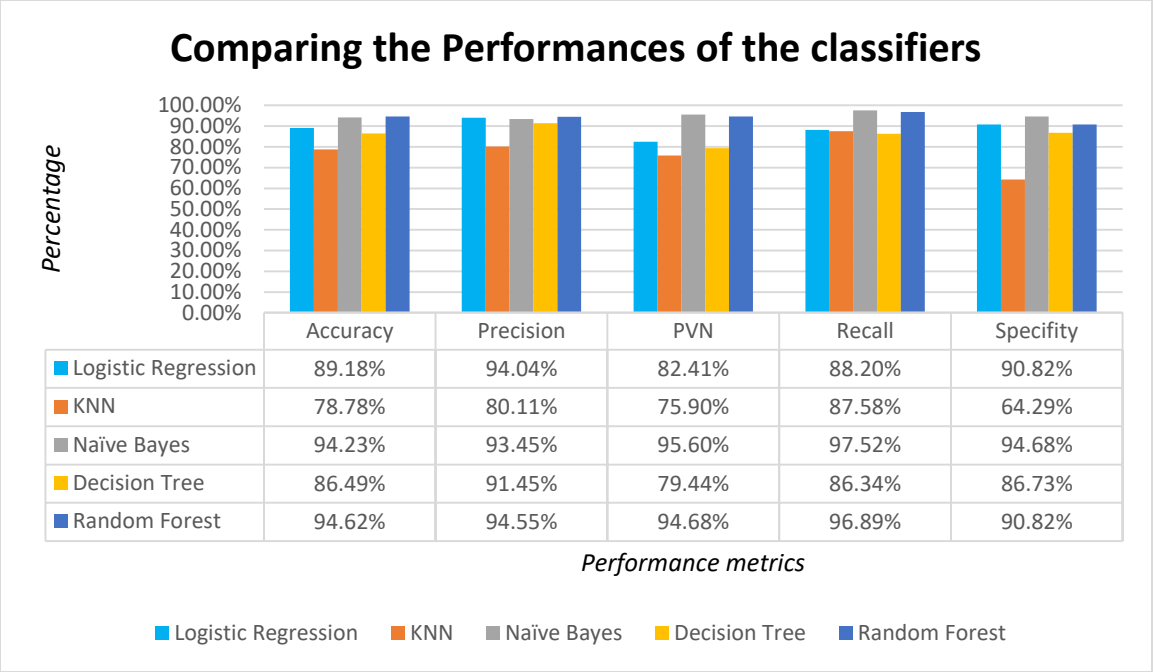


Figure 4.26: Specificity comparison of each model

Comparing the Specificity of all the algorithms, both Random Forest and Logistic Regression are found to have performed exceptionally well, with each scoring 90.82%. Then comes Naïve Bayes and Decision Tree, and one more time, K-NN has the lowest score.



	Accuracy	Precision	PVN	Recall	Specifity
Logistic Regression	89.18%	94.04%	82.41%	88.20%	90.82%
KNN	78.78%	80.11%	75.90%	87.58%	64.29%
Naïve Bayes	94.23%	93.45%	95.60%	97.52%	94.68%
Decision Tree	86.49%	91.45%	79.44%	86.34%	86.73%
Random Forest	94.62%	94.55%	94.68%	96.89%	90.82%

Figure 4.27: Performance comparison of each classifier based on several metrics

Finally, comparing the algorithms' performance, it can be seen that only Random Forest and Naïve Bayes demonstrated the best performance. However, regardless of all the other scores, the ideal model will be selected depending on the accuracy measure. Thus, having the highest accuracy score, Random Forest and Naïve Bayes are the best choices for predicting diabetes.

Chapter 5 Conclusion

In conclusion, this study examined the crucial topic of diabetes forecast through the utilization of supervised machine learning algorithms, aiming to ensure faster and more accurate detection of this prevalent health issue. Using a mixed method strategy and incorporating primary data collected from the adult population, the study was conducted. It was found that in comparison to men, the majority of the participants were women who were housewives. As a result, there were more non-smokers. The maximum number of people who took part were 45-55 years old and resided in cities. Most of their income was between 10001-20000 Tk. or 50001-100000 Tk. In addition, most people said they slept 7 hours a day. Furthermore, a large proportion of the participants were observed to be obese, most of whom were evidently women. There were only a few people with kidney diseases in the study, whereas a fair amount of individuals had heart diseases, other diseases, and blood pressure problems like hypertension. Due to the dominance of the female population in the study, there were more women than men who were affected by diabetes. Also, a huge number of these diabetes patients belonged to the obese category. Only a few amount of heart patients also had diabetes. Again, the number of diabetes patients was comparatively less among people with hypertension than those who were normal. Since kidney patients were very few in number in this study, the number of diabetes patients who have kidney problems was almost negligible. Likewise, non-smoker diabetes patients were recorded in bulk. Essentially, it was observed that some people diagnosed with diabetes already had a history of diabetes in their family, where the fathers mostly have the disease, followed by the mothers and the siblings. The diabetes status, i.e., whether a person has diabetes or not (yes/no), is dependent on the other factors, i.e., age, weight, BMI, smoking habit, etc., which remain independent.

The study also showcased the effectiveness of some selected machine learning algorithms, such as Logistic Regression, K-Nearest Neighbor, Naïve Bayes, Decision Tree, and Random Forest, in accurately categorizing individuals with diabetes. The RapidMiner tool was used to construct and assess the model. Confusion matrices for each model were also generated. Furthermore, the performance evaluation and comparative analyses enabled us to learn about the models' advantages and limitations. This assessment was done by comparing the values of several performance measures, namely, accuracy, precision, predicted value negative, recall, and specificity, also called true negative rate. Consequently, the appropriate algorithm could be chosen. From the results, it was found that Random Forest performed the best with the highest accuracy (94.62%). This implies it has the highest success rate and properly predicts the positive and negative results of the test. Naïve Bayes comes after, scoring the second highest accuracy (94.23%). The findings of this study can facilitate earlier diagnosis in the healthcare system, and treatment can be started immediately. Hence, patients can maintain their diabetes before it becomes serious to cause any potential harm.

References

Al-Aidaroos, K. M., Bakar, A. A., & Othman, Z. (2010). Naïve Bayes Variants in Classification Learning. *2010 International Conference on Information Retrieval & Knowledge Management (CAMP)*, 276–281. https://doi.org/10.1109/INFRKM.2010.5466902

Amit, Y., & Geman, D. (1997). Shape Quantization and Recognition with Randomized Trees. *Neural Computation*, *9*(7), 1545–1588. https://doi.org/10.1162/neco.1997.9.7.1545

Babcock University, F.Y, O., J.E.T, A., O, A., J. O, H., O, O., & J, A. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology*, *48*(3), 128–138. https://doi.org/10.14445/22312803/IJCTT-V48P126

Bangladesh diabetes report 2000—2045. (n.d.). Retrieved July 15, 2023, from https://diabetesatlas.org/data/

Berry, M. W., Mohamed, A., & Yap, B. W. (Eds.). (2020). *Supervised and Unsupervised Learning for Data Science*. Springer International Publishing. https://doi.org/10.1007/978-3-030-22475-2

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). [No title found]. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Carbonell, J. G., Michalski, R. S., & Mitchell, T. M. (1983). AN OVERVIEW OF MACHINE LEARNING. In *Machine Learning* (pp. 3–23). Elsevier. https://doi.org/10.1016/B978-0-08-051054-5.50005-4

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

Choudhary, R., & Gianey, H. K. (2017). Comprehensive Review On Supervised Machine Learning Algorithms. *2017 International Conference on Machine Learning and Data Science (MLDS)*, 37–43. https://doi.org/10.1109/MLDS.2017.11

Confusion Matrix in Machine Learning—Javatpoint. (n.d.). Retrieved July 5, 2023, from https://www.javatpoint.com/confusion-matrix-in-machine-learning

Cox, M. E., & Edelman, D. (2009). Tests for Screening and Diagnosis of Type 2 Diabetes. *Clinical Diabetes*, *27*(4), 132–138. https://doi.org/10.2337/diaclin.27.4.132

Dāsa, R. (2016). What Is Unit Testing? In R. Dāsa, *Learn CakePHP* (pp. 8–13). Apress. https://doi.org/10.1007/978-1-4842-1212-7_2

DeMaris, A. (1995). A Tutorial in Logistic Regression. *Journal of Marriage and the Family*, *57*(4), 956. https://doi.org/10.2307/353415

Dhingra, D. (2021, May 18). *In-depth understanding of Confusion Matrix*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/in-depth-understanding-of-confusion-matrix/

Diabetes. (n.d.-a). Retrieved July 5, 2023, from https://www.who.int/news-room/fact-sheets/detail/diabetes

Diabetes. (n.d.-b). Retrieved July 6, 2023, from https://www.who.int/health-topics/diabetes#tab=tab_1

Diabetes Testing. (2023, February 28). Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/getting-tested.html

Dietterich, T. G. (n.d.). *An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization*.

Edeh, M. O., Khalaf, O. I., Tavera, C. A., Tayeb, S., Ghouali, S., Abdulsahib, G. M., Richard-Nnabu, N. E., & Louni, A. (2022). A Classification Algorithm-Based Hybrid Diabetes Prediction Model. *Frontiers in Public Health*, *10*, 829519. https://doi.org/10.3389/fpubh.2022.829519

Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Random forests: From early developments to recent advancements. *Systems Science & Control Engineering*, *2*(1), 602–609. https://doi.org/10.1080/21642583.2014.956265

Foster, I. (Ed.). (2021). *Big data and social science: Data science methods and tools for research and practice* (Second edition). CRC Press.

Gupta, B., Rawat, A., Jain, A., Arora, A., & Dhami, N. (2017). Analysis of Various Decision Tree Algorithms for Classification in Data Mining. *International Journal of Computer Applications*, *163*(8), 15–19. https://doi.org/10.5120/ijca2017913660

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction: with 200 full-color illustrations*. Springer.

Imandoust, S. B., & Bolandraftar, M. (2013). *Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background*. *3*(5), 7.

Jadhav, S. D., & Channe, H. P. (2016). Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. *International Journal of Science and Research (IJSR)*, *5*(1), 1842–1845. https://doi.org/10.21275/v5i1.NOV153131

Kotsilieris, T., Pintelas, E., Livieris, I. E., & Pintelas, P. (n.d.). *REVIEWING MACHINE LEARNING TECHNIQUES FOR PREDICTING ANXIETY DISORDERS*. 22.

Kumari, A. (n.d.). Study on Naive Bayesian Classifier and its relation to Information Gain. *International Journal on Recent and Innovation Trends in Computing and Communication*, *2*(3).

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In C. Nédellec & C. Rouveirol (Eds.), *Machine Learning: ECML-98* (Vol. 1398, pp. 4–15). Springer Berlin Heidelberg. https://doi.org/10.1007/BFb0026666

Liu, Y., Wang, Y., & Zhang, J. (2012). New Machine Learning Algorithm: Random Forest. In B. Liu, M. Ma, & J. Chang (Eds.), *Information Computing and Applications* (Vol. 7473, pp. 246–252). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-34062-8_32

Lowd, D., & Domingos, P. (2005). Naive Bayes models for probability estimation. *Proceedings of the 22nd International Conference on Machine Learning - ICML '05*, 529–536. https://doi.org/10.1145/1102351.1102418

Mahesh, B. (2018). *Machine Learning Algorithms—A Review*. *9*(1), 7.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Prasath, V. B. S., Alfeilat, H. A. A., Hassanat, A. B. A., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., & Salman, H. S. E. (2019). Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier—A Review. *Big Data*, *7*(4), 221–248. https://doi.org/10.1089/big.2018.0175

Rokach, L., & Maimon, O. (2005). Decision Trees. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 165–192). Springer-Verlag. https://doi.org/10.1007/0-387-25465-X_9

Samuel, A. L. (1959). *Some Studies in Machine Learning Using the Game of Checkers*. 21.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Sharma, G. (2021, June 5). Confusion Matrix: What is it and its Applications. *Medium*. https://geetanshsharma2018.medium.com/confusion-matrix-what-is-it-and-its-applications-4d8b0b958edb

Silverman, B. W., & Jones, M. C. (1989). E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951). *International Statistical Review / Revue Internationale de Statistique*, *57*(3), 233. https://doi.org/10.2307/1403796

Software, C. H. I. (2019, May 20). Supervised vs. Unsupervised Machine Learning. *Medium*. https://chisoftware.medium.com/supervised-vs-unsupervised-machine-learning-7f26118d5ee6

Song, Y., & Lu, Y. (2015). *Decision tree methods: Applications for classification and prediction*. *27*(2), 7.

Sunasra, M. (2019, February 28). Performance Metrics for Classification problems in Machine Learning. *Medium*. https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b

Taheri, S., & Mammadov, M. (2013). Learning the naive Bayes classifier with optimization models. *International Journal of Applied Mathematics and Computer Science*, *23*(4), 787–795. https://doi.org/10.2478/amcs-2013-0059

Tin Kam Ho. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282. https://doi.org/10.1109/ICDAR.1995.598994

Tin Kam Ho. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844. https://doi.org/10.1109/34.709601

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, New Series*, *59*(236), 433–460.

Webb, G. I. (2016). Naïve Bayes. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 1–2). Springer US. https://doi.org/10.1007/978-1-4899-7502-7_581-1

What is a Confusion Matrix in Machine Learning? (n.d.). Simplilearn.Com. Retrieved June 13, 2023, from https://www.simplilearn.com/tutorials/machine-learning-tutorial/confusion-matrix-machine-learning

What is Diabetes? (2023, April 24). Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/diabetes.html

What is Machine Learning? | *IBM*. (n.d.). Retrieved July 27, 2023, from https://www.ibm.com/topics/machine-learning

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics*, *9*, 515. https://doi.org/10.3389/fgene.2018.00515

Appendix A

1. Name (You can keep it blank if you don't want to share your name):

2. Email:

3. Gender:
   - Male
   - Female

4. Age:

5. Weight in kg:

6. Height in feet:

7. Blood Pressure in mmHg (optional):

8. How many hours do you sleep in a day?

9. Does your mother have diabetes?
   - Yes
   - No

10. Does your father have diabetes?
    - Yes
    - No

11. Does any of your siblings have diabetes?
    - Yes
    - No

12. Do you have diabetes?
    - Yes
    - No

13. Do you have any heart disease?
    - Yes
    - No

14. Do you have kidney disease?
- Yes
- No

15. Family/Own Income:
- Below Tk. 5000
- Tk. 5001-10000
- Tk. 10001-20000
- Tk. 20001-30000
- Tk. 30001-40000
- Tk. 40001-50000
- Tk. 50001-100000
- Above Tk. 100000

16. Your living condition:
- Rural
- Urban

17. Smoking Habit:
- Yes
- No

18. What is your profession?
- Service
- Business
- Housewife
- Retired
- Student

19. Do you have any other disease?
- Yes
- No