UNIVERSITY of RWANDA

COLLEGE OF SCIENCE AND TECHNOLOGY

Regional Centre of Excellence in Biomedical Engineering and e-Health (CEBE)

DIABETES MONITORING AND PREDICTION USING MACHINE LEARNING TECHNIQUES

By:

KWIZERA MUNANA Dieudonné

Reference Number: 221028461

A Dissertation Submitted to the Regional Centre of Excellence in Biomedical Engineering and e-Health (CEBE), University of Rwanda as partial fulfilment of the requirements for the Master's Degree in Biomedical Engineering.

Supervised by:

Prof. Lee Chi Hwan

Dr. NZANYWAYINGOMA Frederic

February, 2023

DECLARATION

I, KWIZERA MUNANA Dieudonné, declare that this dissertation entitled "DIABETES MONITORING AND PREDICTION USING MACHINE LEARNING TECHNIQUES" is my original work based on research and prototype and has not been submitted for any other degree or professional qualification.

Student Name: KWIZERA MUNANA Dieudonné

Student Reference Number: 221028461

Student Signature: 

Date: 08/02/2023

UNIVERSITY of RWANDA

COLLEGE OF SCIENCE AND TECHNOLOGY

Regional Centre of Excellence in Biomedical Engineering and e-Health (CEBE)

CERTIFICATE

This is to certify that the project entitled "DIABETES MONITORING AND PREDICTION USING MACHINE LEARNING TECHNIQUES" is a record of original work done by KWIZERA MUNANA Dieudonné (Reference number: 221018461), a MSc. Degree student in Biomedical Engineering.

This work has been submitted under the guidance of Prof. Lee Chi Hwan and Dr. NZANYWAYINGOMA Frederic

Main Supervisor: Chi Hwan Lee

Prof. Lee Chi Hwan

Co-Supervisor:

Dr. NZANYWAYINGOMA Frederic

Biomedical Engineering Master's Program Coordinator

Dr. Gerard RUSHINGABIGWI

ACKNOWLEDGEMENTS

I would like to put my appreciation to UR/Centre of Excellence in Biomedical Engineering and e-Health (CEBE) for admitting me and providing all facilities for my further knowledge.

Deepest gratitude to my supervisors Prof. Lee Chi Hwan and Dr. NZANYWAYINGOMA Frederic for their unconditional assistance during this work. Many thanks to the CEBE community especial Prof. Celestin TWIZERE and Dr. Gerard RUSHINGABIGWI for their time, and guidance during this whole program.

Lastly, I wish to convey my special thanks to my lovely wife, to my families, colleagues, and friends for their encouragement and help.

God bless you all.

ABSTRACT

Diabetes is a chronic disease characterized by an increase in blood sugar levels. Diabetes, if left untreated, causes devastating body complications such as heart attacks, nerve damage, blindness, kidney failure, and limb amputations among others, which may lead to death. Detecting and treating diabetes at an early stage is critical for lowering the risk of serious complications and keeping diabetics healthy.

Various prediction algorithms were employed in this study to predict diabetes on a dataset containing 1000 rows and 8 features. We combined ensemble learning techniques such as Cat Boost Classifier and LGBM Classifier with K-Nearest Neighbor (KNN), Naive Bayes (GNB), Support vector machine (SVC), Logistic regression (LR), decision tree (DT), and Gaussian NB. Accuracy, recall, precision and f1 score were all used to evaluate each model. With an accuracy of 90%, the LGBM Classifier was the first model to perform well followed by Cat Boost Classifier (88.5%), SVC (86%), K Neighbors Classifier (85.5%), Decision Tree Classifier (83%), Logistic Regression (83%), and Gaussian NB (79%). Machine learning is helping to improve the health sector in a variety of ways, including disease prediction, which has helped to reduce death rates and complications. Using machine learning aids in identifying hidden information that traditional methods could not identify.

Screening people is critical for identifying people who are asymptomatic but at risk of developing diabetes. As a result, machine learning (ML) techniques can be used on new registered patients' data sets to detect disease at an early stage, assisting physicians in their decision making.

Keywords: Diabetes, Machine Learning, Classification, Detection

LIST OF ACRONYMS

ANN: Artificial Neural Network

AUC: Under the Curve

CDC: Centers for Disease Control & Prevention

CEBE: Centre of Excellence in Biomedical Engineering and e-Health

CPCSSN: Canadian primary care sentinel surveillance Network

DD: Diabetes Diseases

DT: Decision Tree

GNB: Gaussian Naïve Bayes

GPC: Gaussian Process Classification

IDF: International Diabetes Federation

IRB: Institutional Review Board

KFH: King Faisal Hospital

KNN: K-Nearest Neighborhood

LR: logistic regression

MDH: Munini District Hospital

ML: Machine Learning

NB: Naïve Bayes

REP: Reduces Error Pruning

RF: Random Forest

RL: Regression Model

SMO: Sequential minimal optimization

SVM: Support Vector Machine

UR: University of Rwanda

WEKA: Waikato Environment for Knowledge Analysis

LIST OF FIGURES

TABLE OF CONTENTS

CHAPTER 1. GENERAL INTRODUCTION

1.1 Introduction

Diabetes mellitus is a metabolic disorder in which the pancreas does not create enough insulin or the cells do not respond to the insulin that is produced, preventing glucose from being taken into the body's cells. Type 1, type 2, and gestational diabetes are the three forms of diabetes that have been recognized [1].

Diabetes mellitus is one of the fastest growing global health emergencies of the twenty-first century, according to the findings of the current International Diabetes Federation 10th edition published in 2021, 537 million people had diabetes worldwide number is expected to rise to 643 million by 2030 and 783 million by 2045 also over 6.7 million people aged 20–79 died from diabetes-related causes [2]. Another cause for concern is the continually high percentage of people with undiagnosed diabetes (45%), the majority of which is type 2 [3]. This emphasizes critical need to enhance ability to diagnose people with diabetes, many of whom are unaware of their condition, and to offer adequate and timely care to all persons with diabetes as soon as possible [4].

In Sub-Saharan Africa, like the rest of the world countries is experiencing an increasing prevalence of diabetes alongside other no communicable diseases. In 2045, they will be 40.7 million adults of 20 to 79 years old living with diabetes [5]. In Rwanda, International Diabetes Federation in 2009 estimated prevalence of diabetes in adults of 1.1 %, amounting to 88,000 people out of 10,473,282 total populations [6], in 2017 adults with diabetes (20-79 age) national prevalence was 3.4% and 4.5% in 2021 [7].

Diabetes has long been near the top of global rankings listing serious of diseases, many researchers and doctors have proposed algorithms and methods for its treatment and detection [8]. Several studies have recently been conducted in the subject of illness prediction, to the point where some clinicians now employ machine learning models to forecast certain diseases. Variety of different machine learning techniques have been developed for the prediction and diagnosis of diabetes disease such as: naïve Bayes (NB), support vector machine (SVM), artificial neural network (ANN), decision tree (DT), random forest (RF), Gaussian process classification (GPC), logistic regression (LR), and k-nearest neighborhood (KNN) [9]. Other researchers have utilized machine learning to develop predictive models of the transition from prediabetes to diabetes using algorithms such as gradient boosted trees, with the goal of providing early diagnosis for better

treatment and reducing subsequent risks. A modified support vector machine (SVM) algorithm was used in another study as an efficient technique for both linear and non-linear data.

Early detection of diseases such as diabetes can be controlled and human lives saved. To do so, this research looks into diabetes prediction using a variety of diabetes-related variables. We will collect Diabetes Dataset for this purpose, and we will use several Machine Learning classification and ensemble techniques to forecast diabetes. Various Machine Learning approaches are capable of making predictions, but selecting the optimum methodology is difficult. As a result, we use common classification and ensemble algorithms on the dataset to make predictions.

1.2 Problem statement

Diabetes is on the rise all over the world, and it is only going to get worse. The International Diabetes Federation predicts that there will be 578 million adults with diabetes by 2030, and 700 million by 2045. Other studies show that, not only are diabetes cases on the rise, but diabetic related deaths are as well [10] . According to new data, diabetes and its complications claimed the lives of 4.2 million adults in 2019 [4].

Many people are oblivious to the fact that they have diabetes nearly one in every four people with diabetes were unaware of their disease, according to the Centers for Disease Control & Prevention and if it goes untreated, high blood sugar levels will harm the body's cells and organs [10]. Kidney damage, which sometimes necessitates dialysis, eye damage, which can result in blindness, and an increased risk of heart disease or stroke are all possible complications [11]. This demonstrates how severe this disease can become if no steps are taken to stop it from progressing. Therefore, Machine learning (ML) techniques can be used on new registered patients' data set to efficiently detect the disease at its early stage, this will assist the physicians in their decision making.

1.3 Research Questions

- What extent can machine learning be used to predict diabetes mellitus?
- Which machine learning classifier has the best performance on prediction of diabetes mellitus?

1.4 Objectives

1.4.1 General Objective

The primary goal of this project was to build weak learners and ensembles that could be used to predict the presence of diabetes. In addition, we trained these underperforming learners and ensembles and evaluated their performance using various evaluations. As a result, we determined which machine learning algorithms are more effective at detecting diabetes.

1.4.2 Specific Objectives

The following specific goals were pursued in order to achieve the project's main goal:

- Develop five weak learners with two ensemble learning classifiers for diabetes prediction.
- To identify the best classifier in predicting diabetes
- Evaluating each model in terms of accuracy, precision, recall and f1 score.

1.5 Study Scope

Using patient datasets collected from various Rwandan hospitals, including King Faisal, Munini District Hospital, and Kibagabaga 2nd Teaching Hospital, this project develops high performance machine learning ensemble models for early diabetes prediction.

1.6 Significance of the Study

Many people are oblivious to the fact that they have diabetes and if it goes untreated, high blood sugar levels will harm the body's cells and organs. Kidney damage, which sometimes necessitates dialysis, eye damage, which can result in blindness, and an increased risk of heart disease or stroke are all possible complications. As a result, screening people is critical in order to locate individuals who are asymptomatic but are at risk of developing diabetes it is in that order machine learning can be an excellent tool for assisting healthcare practitioners in predicting diabetes based on clinical examination outcomes and can serve as a reference for medical professionals

1.7 Organization

There are five chapters in this research. The first chapter consists to a general introduction, while the second chapter is devoted to a literature review, which aids in the definition of the study's primary themes and attempts to build theoretical notions connected to the research issue. The technique is covered in the third chapter, while the data analysis, presentation and interpretation, and discussion of the findings are covered in the fourth chapter. Chapter 5 is Recommendation and conclusion which is the final chapters.

1.8 Summary

Every research project has a reason for being carried out. Diabetes was chosen as the focus of this work because it is a common disease that affects people of all ages all over the world. Diabetes has been studied for a variety of purposes, and its prediction has been done in machine learning using multiple classifiers, but our focus is primarily on outcomes. To provide a better indicator of diabetes, we want to build an ensemble learning classifier that can detect glucose levels in the blood.

CHAPTER 2. LITERATURE REVIEW

2.1 Introduction

Diabetes treatment is a major health problem for the world due to the growing number of diabetics, and we hope that early diagnosis will help patients control their blood sugar levels and therefore reduce the chance of severe complications. As a result, machine learning can be an excellent tool for assisting healthcare practitioners in predicting diabetes based on clinical examination outcomes and can serve as a reference for medical professionals. In this part, we will describe and discuss several publications in which machine learning and data mining have been used to diagnose diabetes.

2.2 Related work

Vispute et al. (2015), compared data mining classification techniques on a diabetes dataset using WEKA (Waikato Environment for Knowledge Analysis) with 10-folds cross validation to determine which algorithm performs best on the explorer. After using Naive Bayes, J48 Tree (algorithm used to create a decision tree), SMO (Sequential minimal optimization), REP (Reduces Error Pruning) Tree, and Random Tree to classify and extract useful knowledge from results, the best algorithm in terms of precision, mean absolute error, and time it takes to construct model by Explorer was Naive Bayes [12].

Aiswarya Iyer (2015), used classification technique to study hidden patterns in diabetes dataset. Naïve Bayes and Decision Trees were used in this model. Comparison was made for performance of both algorithms and effectiveness of both algorithms was shown as a result [13]

Sajida et al (2016), used the CPCSSN (Canadian primary care sentinel surveillance Network) dataset and three machine learning approaches to predict diabetes diseases (DD) in early stages to save human life from early death using the Modified training set. Bagging, Adaboost, and decision tree (J48) were used to predict diabetes in this study, and the results were compared. The researcher determined that the Adaboost approach was more effective and accurate than the other methods in the Weka data mining tools [14].

Steffi et al (2018), they tested the five models in terms of their accuracy, precision, sensitivity, specificity, and F1 Score measures, and they created five prediction models employing nine input variables and one output variable from the Dataset information. The goal of this study was to assess how well Nave Bayes, Logistic Regression, Artificial Neural Networks (ANNs), C5.0 Decision Tree, and Support Vector Machine (SVM) models predicted diabetes using common risk factors.

The logistic regression model, Nave Bayes, ANN, and the SVM had the best classification accuracy, followed by the decision tree model (C5.0) [15].

Faruque et al. (2019), used Performance Analysis of four machine learning algorithms, namely Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN) and C4.5 decision tree (DT) was evaluated based on precision, recall, f1-measure, and accuracy. According to their study C4.5 decision tree classifier achieved better results than other classifiers to predict diabetes mellitus, with 72% precision, 74% recall and 72% f-measure on the dataset used [16].

Naveen et al. (2020), used Pima Indians Dataset to train algorithms such as SVM, Decision Tree, KNN, Logistic Regression, and Random Forest. As a result, Random forest algorithm performed well with accuracy rate of nearly 74% [17].

Chawan, 2018 conducted a research aimed at developing a system which can predict diabetes at an early stage in patients with a high accuracy by combining the results of different machine learning techniques. The research predicts diabetes using two (2) different supervised machine learning methods including SVM and Logistic Regression. It considered seven (7) features of the patients. They reached a conclusion that SVM showed a better performance with accuracy of seventy-nine percent (79%) compared to logistic regression which had a performance accuracy of seventy-eight percent (78%) [18]

Sneha & Gangil, 2019 conducted a research that was aimed at selecting the attributes that aid in early detection of diabetes mellitus using WEKA which is a predictive analysis tool. They were able to reach a conclusion which shows that decision tree algorithm and Random Forest Algorithm has the highest predictive analysis by 98.20% and 98.00% respectively. While Naïve Bayesian outcomes states the best in performance accuracy with 82.30% [19].

Kaur & Kumari, 2022 developed five different models for the detection of diabetes using, linear kernel support vector machine (SVM-linear), radial basis kernel support vector machine (SVM-RBF), K Nearest Neighbour (k-NN), Artificial Neural Networks (ANN) and Multifactor Dimensionality Reduction (MDR) algorithms. Feature selection of dataset was done with the help of Boruta wrapper algorithm, considering some evaluation criteria namely; accuracy, recall, precision, F1 score, and Area Under the Curve (AUC). The experimental results indicated that all the models achieved good results with SVM-linear model providing a very good accuracy of 0.89 and precision of 0.88. From the results of this study, it can be concluded that on the basis of all the parameters linear kernel support vector machine (SVM-linear) and k-NN are the two (2) most accurate predictive models for diabetes. This work also suggested that Boruta wrapper algorithm can be used for feature selection as they were able to achieve a better accuracy with its use [20].

2.3 Taxonomy of machine learning algorithms

Machine learning has numerous algorithms which are classified into three categories: Supervised learning, Unsupervised learning, Semi-supervised learning [21].

The Supervised Learning/Predictive Models: Predictive models are built using supervised learning techniques. A predictive model uses other values in the dataset to forecast missing values. The supervised learning technique takes a set of input data and output data and creates a model to predict the response to a new dataset in a realistic way. Decision Tree, Bayesian Method, Artificial Neural Network, Instance Based Learning, and Ensemble Method are examples of supervised learning.

Unsupervised Learning / Descriptive Models: Unsupervised learning is used to create descriptive models. We have a known set of inputs in this model, but the outcome is uncertain. On transactional data, unsupervised learning is most commonly employed. Clustering algorithms such as k-Means clustering and k-Medians clustering are included in this strategy.

Semi-supervised Learning: On the training dataset, the semi-supervised learning approach uses both labeled and unlabeled data. Semi Supervised Learning includes techniques such as classification and regression. Regression techniques such as logistic regression and linear regression are examples.

2.3 Summary

This section summarizes relevant literature for this project. It discusses how technology affects various aspects of life and how it improves overall quality of life. This section highlights various health-related ML applications that have been introduced in the health sector to improve quality of life and predict diabetes.

From the above reviewed literatures, it is important to note that, although various research work have been carried out in the area of diabetes prediction in other countries using various risk factors that are peculiar to their environment but not much have been done in applying any of the machine learning techniques in diabetes prediction, using risk factors that are peculiar to the Rwandan environment. It is also evident from the reviewed literatures that supervised learning algorithms overtime, produced very good prediction accuracy in research works where they were applied though not much work have been done in comparing the prediction accuracy of the weak learner with ensembles learners. Therefore, this research develops the prediction model that is accurate, the proposed system does not replace the healthcare systems put in place but rather acts as a support tool for healthcare.

CHAPTER 3. RESEARCH METHODOLOGY

In this chapter the methods and approaches used to conduct the study are outlined. This includes the steps undertaken to complete the study, the system design methodology, and tools used in data collection.

3.1 System Components

The block diagram for this project's diabetes prediction model is shown below. There are four different modules in the model, including: Dataset Collection, Data Pre-processing, Clustering, Build Model and Evaluation

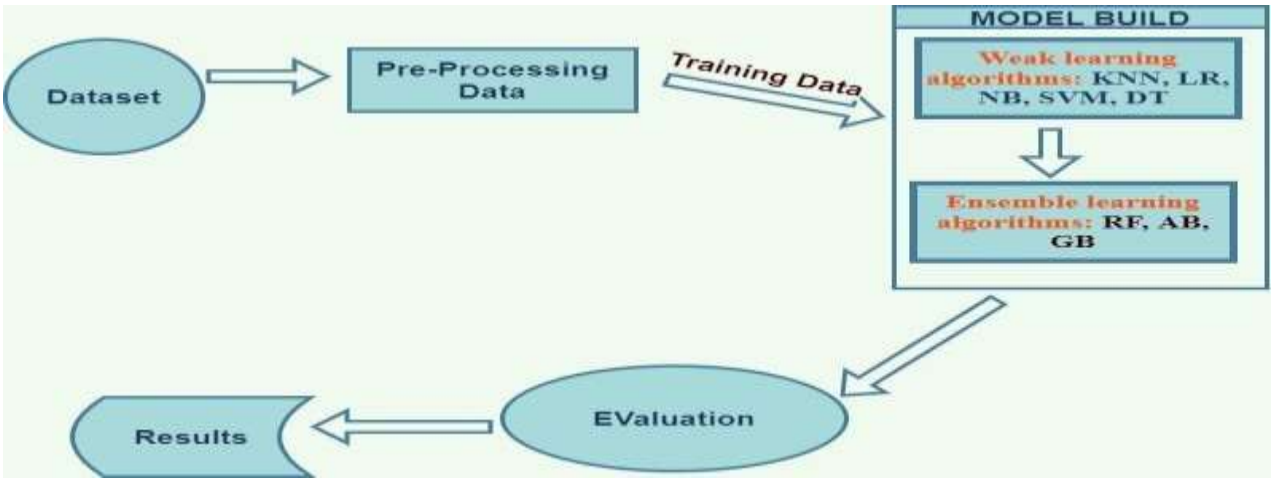


Figure 3. 1: Model Diagram

Collection of Datasets: The dataset for this study was created using prior clinical data or data gathered when diabetes patients were initially diagnosed. Goal features that distinguish between people with diabetes and those who do not should have two results (Yes and No). In addition to these factors, the target variable was compared to the patient's age, blood pressure, insulin, BMI, diabetes pedigree function, which determines the likelihood of diabetes based on family history, and glucose, which is the plasma glucose concentration over 2 hours in an oral glucose tolerance test.

Data Pre-processing: This step of the model deals with erroneous data in order to produce more precise and accurate findings.

Model Development: This is the most important step, which includes the creation of a diabetes prediction model. K-Nearest Neighbour, Gaussian Nave Bayes, Decision Tree, Logistic Regression, and Support Vector Machine for Weak learning algorithms, and Ada Boost and Gradient Boost for Ensemble learning algorithms were used to predict diabetes.

Evaluation: This is the final stage of the prediction model's development. To evaluate the prediction results, we used multiple assessment measures such as accuracy, precision, recall, and f1 score.

3.2 Tools and methods

3.2.1 Weak learning algorithms

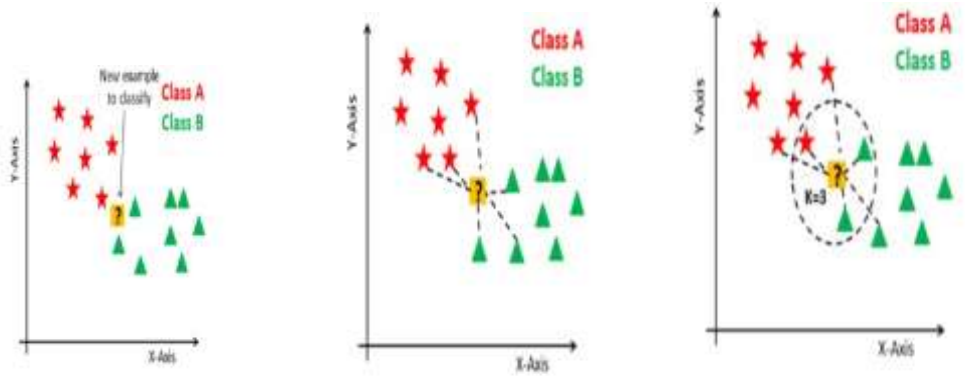
3.2.1.1 K-Nearest Neighbor (k-NN)

K-nearest neighbor is a supervised learning algorithm that is dependent on data similarities. It operates on the basis that the data can be in *n*-dimensions. The goal of the KNN classifier is to predict the target label using the neighboring class by calculating the Euclidian Manhattan distance, Minkowski distance, or Weighted distance.

The k-Nearest Neighbors algorithm operates in the following manner:

- Standardize the data
- Select the number, *k*, of the neighbors for each example in the data points to be classified
- Calculate of the distance between the test group and each train instance with the use of different distance measurements, such as the Euclidean distance which is the most popular one.

Below is the summarized algorithm's process (See figure 3.2).



(a) Initial data (b) Calculating the distance (c) Finding neighbors and voting for labels

Figure 3. 2: Steps of k-NN algorithm

It is worth noting that the number of nearest neighbors, K, is significant in the KNN algorithm and is determined as $k = \sqrt{N}$ where N is the number of samples in the training dataset. For a small number of K, there is a weak bias but a higher variance, i.e., there is a small similarity between the two groups, but this is not the case for a larger number of K. Furthermore, as the value of K

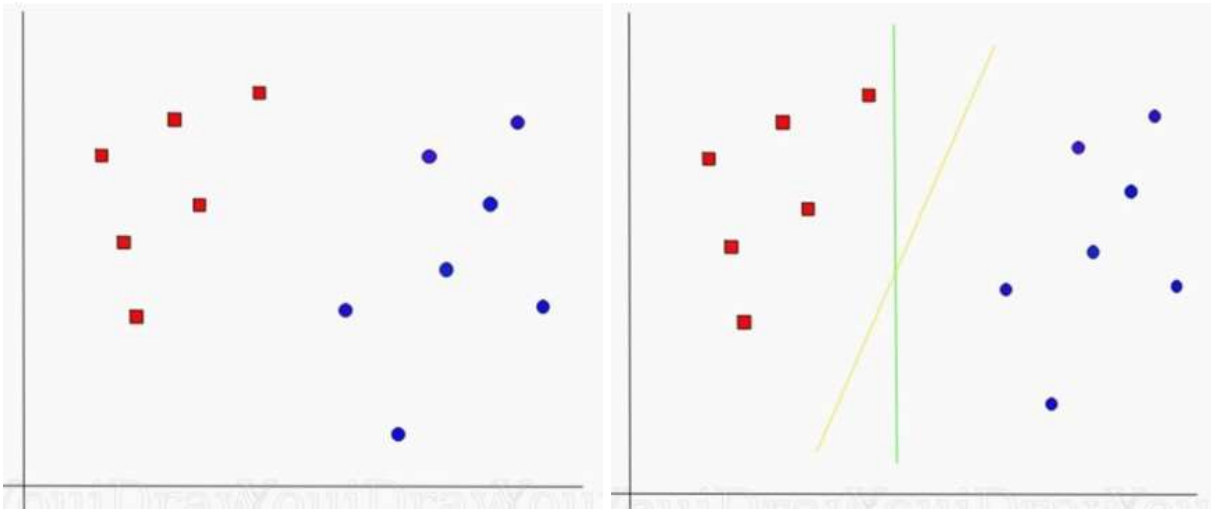
increases, the decision boundary that separates two classes changes. When this occurs, each data point seems to fall into its proper place. If the predictive class contains two components, the number of K to be chosen is an odd number.

3.2.1.2 Naive Bayes Classifier.

Naive Bayes is a classification algorithm that assumes the attributes are conditionally independent of the label. It means that any change to one attribute would have no effect on the others because they are independent. In practice, this is not necessarily the case when certain attributes are associated while others have no input or contribute differently to the target variable. This is not always the case in real-life since some attributes are correlated and others have no contribution or contribute differently to the target variable.

3.2.1.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a machine learning algorithm for classification and regression problem. The principle of SVM is to separate the data point into two classes using a line or a hyperplane. For illustration, Figure 3.3a below shows a group of data points on the left represented by a red square (class 1) and data point on the right represented by a circle in blue (class 2). Here, we can draw several lines or hyperplanes to classify these data points into two classes, Figure 3.3b. The problem arises when it comes to selecting the best line or hyperplane. As it is clearly shown in figure 3.3b, the green line is getting closer to the red data point. We can say that the brown line is optimal hyperplane but the challenge is how to get it.



(a) Data points within two classes (b) Data points with lines or hyperplanes

Figure 3. 3: Working principal of SVM.

3.2.1.4 Logistic Regression.

Like naive Bayes, logistic regression is a probabilistic classifier that makes use of supervised machine learning. It is a linear classification algorithm that is often used for classification problems with categorical target variable (Y). This is analogous to the dilemma we're attempting to tackle in that the outcome will be whether or not an individual is likely to develop diabetes. The Logistic Regression model (LR) is applied to two or more explanatory variables (dependent variables) and calculates the likelihood that an event will occur or not.

Logistic regression has two phases:

- Training: The model is trained (specifically the weights w and b) using stochastic gradient descent and the cross-entropy loss.
- Testing: Given a test example x, we compute p (y | x) and return the higher probability label y = 1 or y = 0.

The aim of binary logistic regression is to train a classifier that can make a binary judgment about the class of a new input observation. This section introduces the sigmoid classifier, which will assist us in making this decision.

Consider a single input observation x, represented by a vector of features $[x_1, x_2,...,x_n]$. The classifier output y can be 1 (meaning the observation is a member of the class) or 0 (the observation is not a member of the class). We want to know the probability $P(y = 1|x)$ that this observation is a member of the class. Logistic regression solves this task by learning, from a training set, a vector of weights and a bias term. Each weight w_i is a real number, and is associated with one of the input features x_i. The weight w_i represents how important that input feature is to the classification decision, and can be 1 or 0.

To make a decision on a test instance— after the model have learned the weights in training— the classifier first multiplies each x_i by its weight w_i, sums up the weighted features, and adds the bias term b.

3.2.1.5 Decision tree (DT).

The decision tree model has a tree structure that may be used to represent the process of categorizing instances based on feature characteristics. It may be regarded of as a set of if-then rules, as well as conditional probability distributions specified in feature and class space. The result of the evaluation determines if the attribute follows the True or False branch. The last nodes (leaves) decide which class the value belongs to. The degree of disorder in the two classes formed as a result of a separation indicates the quality of the split.

3.2.2 Ensemble learning algorithms

3.2.2.1 Adaptive Boosting (AdaBoost).

The Adaboost ensemble model reduces error during classification. It is adaptable in its application of various weak learners, such as the decision stump. The predictions made by a decision stump are taken into account by the next decision stump, and so on until the final classification is reached with a small error. This demonstrates how the following prediction improves the poor performance of a model until the majority of data points are correctly classified.

Consider the training set

$(x_1, y_1)...,(x_n,y_n)$

where $x_i \in D$ is an instance and $_{yi} \in \{0,1\}$ its class, for i = 1, 2...,n

The goal of implementing Adaptive boosting is to improve the performance of T low-performing learners. This entails assigning and adjusting weights vector wi to misclassified data points by the learners.

3.2.2.2 Gradient boosting

Boosting is a sequential model where the next model tries to improve (boost) the error (calculated by Gradient). Hence the name gradient boosting, where the primary goal is to reduce bias. It implies that we first construct a model, then determine its residual, then build another model based on the residual, and so on. It may be expressed mathematically as follows.

Given the dataset {(xi,yi)}in where x are the features and y is the target, we try to restore the function y = f(x) by approximately estimating f^(x) while measuring how good the mapping is using a loss function L(yi,f(xi)) and then take average over all the dataset points to get the final cost.

3.3 The Research Plan

The chart below shows the suggested research plan for this project.

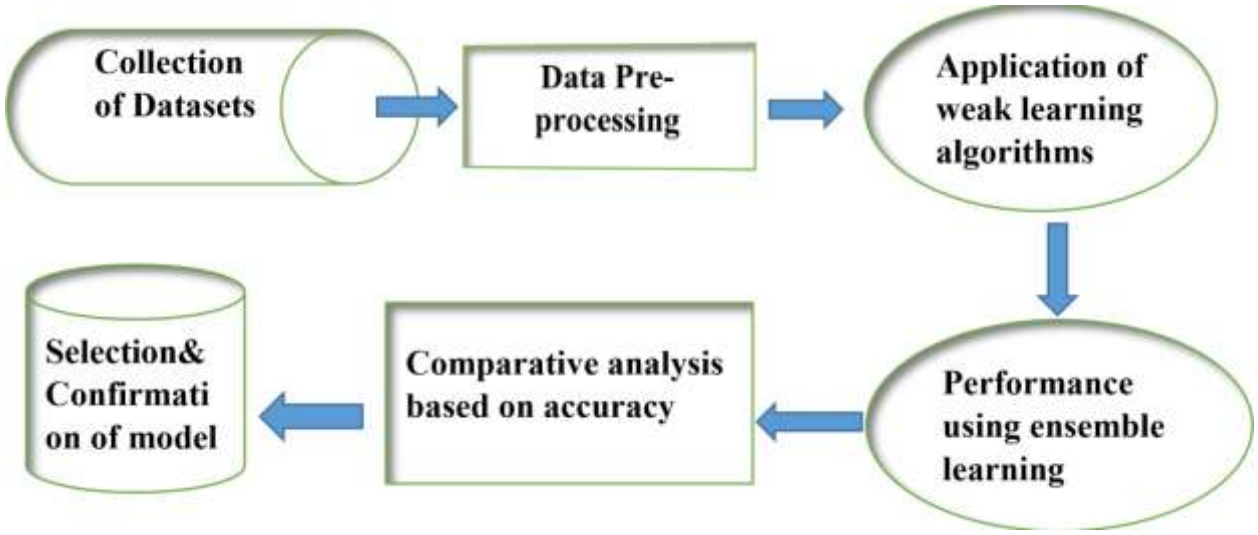


Figure 3. 4: Research plan

3.4 Research Setting and data collection method

This study was carried out in the non-communicable disease and laboratory departments of King Faisal Hospital, Munini District Hospitals, and Kibagabaga 2nd Teaching Hospital. Prior to developing the proposed model, quantitative data collection techniques were used to collect patient information for eight characteristics (columns). These are as follows: (1) the plasma glucose concentration at 2 hours in an oral glucose tolerance test, (2) diastolic blood pressure (mm Hg), (3) triceps skinfold thickness (mm), (4) 2-hour serum insulin (mu U/ml), (5) body mass index (weight in kg/height in m2), (6) age, and (7) diabetes pedigree function (function that scores the likelihood of diabetes based on familial history). (8) Diabetes Class Variable (0 or 1, where 0 indicates not diabetic and 1 indicates diabetic)

3.5 Target Population

This study's population will be made up of 500 diabetics and 500 non-diabetics who meet the requirements. The following criteria will be considered:

- Aged:18 to 50 years
- For diabetics must diagnosed according to the World Health Organization criteria where a patient is considered as diabetic if the 2-hour post-load plasma glucose is at least 200 mg/dl.

- For non-diabetics must diagnosed according to the World Health Organization criteria where a patient is considered as non-diabetic if the 2-hour post-load plasma glucose is under 200 mg/dl.
- Patients with complications like nephropathy, retinopathy, cardiovascular and other endocrinal disorders and patients already on antioxidant supplementation will be excluded.

3.6 Ethical Protection Plan

Ethical considerations

This project was approved by the regional Center of Excellence in Biomedical Engineering and EHealth (CEBE). The King Faisal Hospital (KFH), Munini District Hospital and Kibagabaga 2nd Teaching Hospital should be used at the very least for data collection because they have the necessary data. As the research moves into binary classification, there will be no harm done to the patient and the outcomes will be beneficial to future healthcare patients.

Protection of privacy and confidentiality

We intend to conduct research activities in accordance with all ethical standards to protect the confidentiality of patient information and enhance the quality of services provided to patient's in the future. we promise to protect the privacy of the supplied information as a researcher. Any identifying information, including name, address, or phone number, was removed from the data usage, and each patient was given a new file name. All given information was stored on password-protected laptops and in an encrypted document that is accessible only to myself, and two supervisors.

Safekeeping of data

In addition to the aforementioned procedures, the data was stored for period required before being deleted in accordance with KFH, MDH, Kibagabaga 2nd Teaching Hospital IRB regulations.

3.7 Summary

This chapter discussed the scientific research method that was used to ensure the success of this project as a scientific research work. It placed a strong emphasis on the sampling techniques used to design the study population as well as the analysis of the testing results using various ML models.

CHAPTER 4. THE PROJECT RESULTS

4.1 Dataset

The dataset used in this study was obtained from King Faisal Hospital, Munini District Hospital and Kibagabaga 2nd Teaching Hospital.

4.1.1 Features

The patients' information was gathered for eight features. These included: (1) the plasma glucose concentration at 2 hours in an oral glucose tolerance test, (2) diastolic blood pressure (mm Hg), (3) triceps skinfold thickness (mm), (4) 2-hour serum insulin (mu U/ml), (5) body mass index (weight in kg/ (height in m^2), (6) age, (7) diabetes pedigree function (function that scores the likelihood of diabetes based on familial history) and (8) diabetes status (being diabetic or not).

The models were trained using eight features. All of these attributes were included in the dataset with no missing values as the data was collected directly from the people, see Figure 4.1 for a screen shot of the dataset after inputting in python programming library.

```
In [21]: df.head()
```

Out[21]:

	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	148	72	35	68	33.6	0.627	50	1
1	190	66	29	43	26.6	0.351	31	1
2	183	64	25	44	23.3	0.672	32	1
3	89	66	31	94	28.1	0.167	21	0
4	137	40	35	168	43.1	2.288	33	1

```
In [4]: df.tail()
```

Out[4]:

	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
995	90	76	28	0	32.1	0.851	56	0
996	71	113	30	21	22.2	0.267	25	0
997	90	72	37	27	29.7	0.188	29	0
998	88	101	17	21	20.5	0.512	37	0
999	201	71	10	0	21.0	0.966	53	1

Figure 4. 1: Datasets after inputting in python programming library

4.1.2 Data Balance

Diabetes dataset had 1000 cases and was made up of a mixed sample of diabetic and non-diabetic people. However, diabetes patients made up just 48.3 percent of the whole sample, while non-diabetic patients made up 51.7 percent (see figure 4.2).

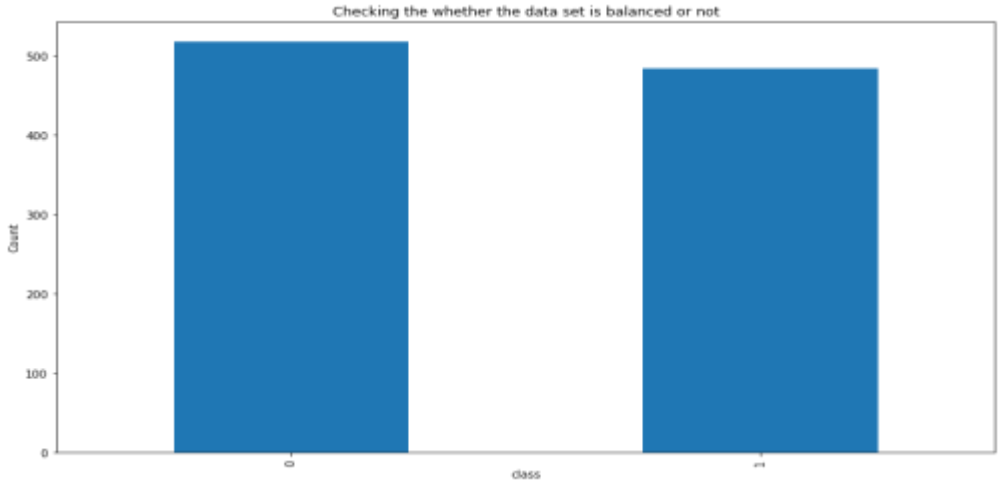


Figure 4. 2: Data Balance

Classification of unbalanced data, particularly in medical research, is difficult, which motivated the development of a classifier using a rebalancing technique. In fact, multiple studies such as Satoru et al., 2022, Alshammari et al., 2020, Laila et al., 2022 and Shamreen et al., 2022, have demonstrated that having balanced data leads to improved prediction accuracy; hence, a variety of well-known strategies have been created and employed in machine learning to address this issue in order to improve the models' performance [22], [23], [24], [25] .

4.2 Performance of the Classifiers

The results obtained from developing the models for the diabetes prediction were achieved using K-Nearest Neighbour, Gaussian Nave Bayes, Decision Tree, Logistic Regression, and Support Vector Machine for Weak learning algorithms, and Ada Boost and Gradient Boost for Ensemble learning algorithms were used to predict diabetes.

4.2.1 Performance of weak learners

4.2.1.1 Logistic Regression

Figure 4.3 shows that the logistic regression classifier has an accuracy of 83%. It means that 83% of cases predicted by the classifier as diabetic or non-diabetic were correct, with precision of 0.87 for ND and 0.81 for D, recall of 0.85 for ND and 0.81 for D, and F1 Score of 0.85 for ND and 0.81 for D.

	precision	recall	f1-score	support
0	0.85	0.85	0.85	111
1	0.81	0.81	0.81	89
accuracy			0.83	200
macro avg	0.83	0.83	0.83	200
weighted avg	0.83	0.83	0.83	200

Figure 4. 3: Performance of Logistic Regression classifier

4.2.1.2 Decision Tree Classifier

The Decision Tree Classifier Function from the Python programming library was used in this experiment as shown in the figure 4.4 below. During the experiment, 80% of the data was used for training and the remaining 20% for testing with a random sample size of 100. This resulted in 83% accuracy, with precision of 0.90 for ND and 0.77 for D, recall of 0.78 for ND and 0.82 for D, F1 Score of 0.84 for ND and 0.82.

	precision	recall	f1-score	support
0	0.90	0.78	0.84	111
1	0.77	0.89	0.82	89
accuracy			0.83	200
macro avg	0.83	0.84	0.83	200
weighted avg	0.84	0.83	0.83	200

Figure 4. 4: Performance of Decision Tree Classifier

4.2.1.3 Support Vector Machine

According to the results in table 4.5, the precision of the support vector machine classifier is 86%. It means that the classifier correctly predicted diabetic or non-diabetic status in 86 percent of cases, with precision of 0.90 for ND and 0.82 for D, recall of 0.85 for ND and 0.88 for D, and F1 Score of 0.87 for ND and 0.85.

	precision	recall	f1-score	support
0	0.90	0.85	0.87	111
1	0.82	0.88	0.85	89
accuracy			0.86	200
macro avg	0.86	0.86	0.86	200
weighted avg	0.86	0.86	0.86	200

Figure 4. 5: Performance of Support Vector Machine Classifier

4.2.1.4 K Neighbors Classifier

According to the figure 4.6, the precision of the k-nearest neighbor classifier is 91 percent for being Non Diabetic and 80 percent for being Diabetic. This means that the classifier predicted being non diabetes correctly in 91 percent of cases and being diabetic correctly in 80 percent of cases. A recall score of 82 and 90 percent, on the other hand, represents the proportion of cases classified by the classifier as ND and D, respectively. The 85 percent accuracy indicates that the model correctly predicted 85 percent of the cases as Non Diabetic or Diabetic.

	precision	recall	f1-score	support
0	0.91	0.82	0.86	111
1	0.80	0.90	0.85	89
accuracy			0.85	200
macro avg	0.85	0.86	0.85	200
weighted avg	0.86	0.85	0.86	200

Figure 4. 6: Performance of K Neighbors Classifier

4.2.1.5 Gaussian NB Classifier

The precision of the Gaussian NB classifier is 78%, according to the results in table 4.7. It means that in 86 percent of cases, the classifier correctly predicted diabetic or non-diabetic status, with precision of 0.76 for ND and 0.81 for D, recall of 0.86 for ND and 0.68 for D, and F1 Score of 0.81 for ND and 0.74.

	precision	recall	f1-score	support
0	0.76	0.86	0.81	109
1	0.81	0.68	0.74	91
accuracy			0.78	200
macro avg	0.78	0.77	0.77	200
weighted avg	0.78	0.78	0.78	200

Figure 4. 7: Performance of Gaussian NB classifier

4.2.2 Results of ensembles learners

4.2.2.1 Cat Boost Classifier

The Cat Boost Classifier has an accuracy of 89%, as shown in Figure 4.8. It means that the classifier correctly classified 89% of cases as diabetic or non-diabetic, with precision of 0.93 for ND and 0.85 for D, recall of 0.87 for ND and 0.92 for D, and F1 Score of 0.90 for ND and 0.89 for D.

	precision	recall	f1-score	support
0	0.93	0.87	0.90	111
1	0.85	0.92	0.89	89
accuracy			0.90	200
macro avg	0.89	0.90	0.89	200
weighted avg	0.90	0.90	0.90	200

Figure 4. 8: Performance of Cat Boost Classifier

4.2.2.2 LGBM Classifier

The accuracy of the LGBM Classifier is 92%, as shown in Figure 4.9. With precision of 0.96 for ND and 0.88 for D, recall of 0.89 for ND and 0.96 for D, and F1 Score of 0.93 for ND and 0.91 for D.

	precision	recall	f1-score	support
0	0.96	0.89	0.93	111
1	0.88	0.96	0.91	89
accuracy			0.92	200
macro avg	0.92	0.92	0.92	200
weighted avg	0.92	0.92	0.92	200

Figure 4. 9: Performance of LGBM Classifier

4.2 Model accuracy comparison

From the outcomes of each model's tests. The table 4.10 below illustrates how different performance levels were produced by the various performance evaluation techniques. With an accuracy of 90%, LGBM Classifier was the first model to perform well. It was followed by Cat Boost Classifier (88.5%), SVC (86%), K Neighbors Classifier (85.5%), Decision Tree Classifier (83%), Logistic Regression (83%), and Gaussian NB (79%).

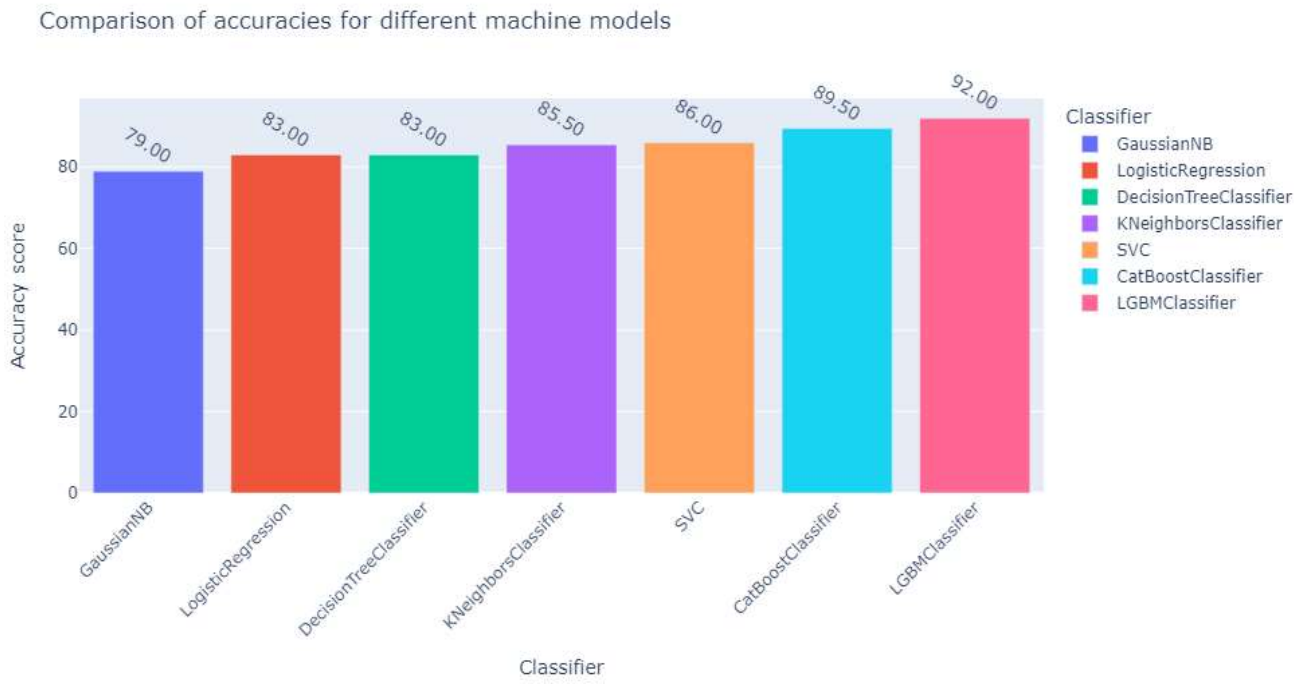


Figure 4. 10: Comparison of accuracy of different models

4.4 Summary

According to research, diabetes is now the leading cause of death in both developed and developing countries. This number is expected to more than double in the coming decades. Because of advanced machine learning techniques and the availability of large diabetes datasets, machine learning has a very good ability to change the risk of diabetes for the better. This would aid in prompt and precise prediction of the disease before it escalates. Diabetes detection at an early stage is critical for treatment.

Diabetes therapy is dependent on early identification. Various prediction algorithms were employed in this study to predict diabetes on a dataset containing 1000 rows and 8 features. We combined ensemble learning techniques such as LGBM Classifier and Cat Boost Classifier with K-Nearest Neighbour, Gaussian Nave Bayes, Decision Tree, Logistic Regression, and Support Vector Machine. Accuracy, recall, precision and f1 score were all used to evaluate each model. LGBM Classifier was the first model to perform well, with an accuracy of 90%. Cat Boost Classifier (88.5%), SVC (86%), K Neighbors Classifier (85.5%), Decision Tree Classifier (83%), Logistic Regression (83%), and the last was Gaussian NB (79%)

CHAPTER 5. CONCLUSION AND RECOMMENDATION

5.1 Conclusion

The aim of this study was to develop weak learners and ensembles algorithms that could be used to predict the presence of diabetes after discovering how prevalent it is today. To accomplish this general goal, the research concentrated on three specific goals: Create two ensemble learning classifiers and five weak learners to predict diabetes. To determine the most accurate diabetes predictor. Evaluating each model in terms of accuracy, precision, recall and f1 score.

For diabetes prediction, the researchers used five weak machine learning methods and two ensemble methods. The F1 score, precision, recall, and accuracy were used to evaluate the classifier's performance in order to select the best classifier. This study employed five weak classifiers and two ensemble classifiers. With an accuracy of 90%, the LGBM Classifier was the first model to perform well followed by Cat Boost Classifier (88.5%), SVC (86%), K Neighbors Classifier (85.5%), Decision Tree Classifier (83%), Logistic Regression (83%), and Gaussian NB (79%). Machine learning is helping to improve the health sector in a variety of ways, including disease prediction, which has helped to reduce death rates and complications.

5.2 Recommendations

The LGBM Classifier, which was identified as the best classifier in this study, should be used in the field and monitored to validate the findings for diabetes prediction. This will aid in the development of various interventions that will aid in the reduction of the diabetic problem.

This study focused on only five weak algorithms and two ensembles: Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Gaussian NB, Decision Tree, LGBM Classifier, and Cat Boost Classifier, but there are other machine learning classifiers with higher predicting power, such as XGboost and GradientBoosting. These algorithms are preferred when the data contains both categorical and numerical features.

The management of diabetes is a big health challenge to the world due to its increasing rate. Its early detection will be helpful for patients since their sugar blood level can be regulated. This can reduce the risk of serious complications, because physicians have past data on diabetics. Machine learning (ML) techniques can be used on new registered patients' data set to efficiently detect the disease at its early stage, this will assist the physicians in their decision making.

REFERENCES

[1] WHO, “Rwanda , www. who.int/diabetes/country-profiles/rwa_en.pdf(2016 accessed July 20, 2018).,” p. 2016, 2017.

[2] N. Barakat, A. P. Bradley, and M. N. H. Barakat, “Intelligible support vector machines for diagnosis of diabetes mellitus,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 4, pp. 1114–1120, 2010, doi: 10.1109/TITB.2009.2039485.

[3] Centers for Disease Control and Prevention, “National Diabetes Statistics Report, 2020,” *Natl. Diabetes Stat. Rep.*, p. 2, 2020.

[4] N. H. Cho *et al.*, “IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045,” *Diabetes Res. Clin. Pract.*, vol. 138, pp. 271–281, 2018, doi: 10.1016/j.diabres.2018.02.023.

[5] R. Williams *et al.*, “Global and regional estimates and projections of diabetes-related health expenditure: Results from the International Diabetes Federation Diabetes Atlas, 9th edition,” *Diabetes Res. Clin. Pract.*, vol. 162, 2020, doi: 10.1016/j.diabres.2020.108072.

[6] IDF Diabetes Atlas Group, *IDF Diabetes Atlas Fourth Edition 2009*. 2009.

[7] IDF, *Eighth Edition 2017*. 2017.

[8] R. Hooda, V. Joshi, and M. Shah, “A comprehensive review of approaches to detect fatigue using machine learning techniques,” *Chronic Dis. Transl. Med.*, vol. 8, no. 1, pp. 26–35, 2022, doi: 10.1016/j.cdtm.2021.07.002.

[9] S. Bashir, U. Qamar, and F. H. Khan, “IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework,” *J. Biomed. Inform.*, vol. 59, pp. 185–200, 2016, doi: 10.1016/j.jbi.2015.12.001.

[10] L. Guariguata, “By the numbers: New estimates from the IDF Diabetes Atlas Update for 2012,” *Diabetes Res. Clin. Pract.*, vol. 98, no. 3, pp. 524–525, 2012, doi: 10.1016/j.diabres.2012.11.006.

[11] M. S. Anjali D Deshpande, Marcie Harris-Hayes, “Diabetes-Related Complications,” vol. 88, no. 11, 2008.

[12] J. Vispute, Dinesh Kumar, and RajputAnil, “An Empirical Comparison by Data Mining Classification Techniques for Diabetes Data Set,” *Int. J. Comput. Appl.*, vol. 131, no. 2, pp. 6–11, 2015, doi: 10.5120/ijca2015907238.

[13] A. Iyer, J. S, and R. Sumbaly, “Diagnosis of Diabetes Using Classification Mining Techniques,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 1, pp. 01–14, 2015, doi: 10.5121/ijdkp.2015.5101.

[14] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, “Performance Analysis of Data Mining Classification Techniques to Predict Diabetes,” *Procedia Comput. Sci.*, vol. 82, no. March, pp. 115–121, 2016, doi: 10.1016/j.procs.2016.04.016.

[15] J. Steffi and D. R. B. 1M. Phil Student, “Predicting Diabetes Mellitus using Data Mining Techniques Comparative analysis of Data Mining Classification Algorithms,” *Int. J. Eng. Dev. Res.*, vol. 6, no. 2, pp. 460–467, 2018.

[16] M. F. Faruque, Asaduzzaman, and I. H. Sarker, “Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus,” *2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE 2019*, no. February, pp. 1–4, 2019, doi: 10.1109/ECACE.2019.8679365.

[17] T. R. S. R. Naveen Kishore G, V.Rajesh, A.Vamsi Akki Reddy, K.Sumedh and Abstract, “Prediction Of Diabetes Using Machine Learning Classification Algorithms,” *PINTERNATIONAL J. Sci. Technol. Res.*, vol. 9, no. 01, pp. 706–716, 2020, doi: 10.1016/j.procs.2020.03.336.

[18] T. N. Chawan, P.M. & Joshi, “Logistic Regression and Svm Based Diabetes,” *Int. J. Technol. Res. Eng.*, vol. 5, no. July, pp. 4347-4350., 2018.

[19] N. Sneha and T. Gangil, “Analysis of diabetes mellitus for early prediction using optimal features selection,” *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0175-6.

[20] H. Kaur and V. Kumari, “Predictive modelling and analytics for diabetes using a machine learning approach,” *Appl. Comput. Informatics*, vol. 18, no. 1–2, pp. 90–100, 2022, doi: 10.1016/j.aci.2018.12.004.

[21] M. Utmal, “Taxonomy on Machine Learning Algorithms,” vol. 10, no. 8, pp. 1–7, 2021.

[22] S. Kodama *et al.*, “Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: A meta-analysis,” *J. Diabetes Investig.*, vol. 13, no. 5, pp. 900–908, 2022, doi: 10.1111/jdi.13736.

[23] R. Alshammari, N. Atiyah, T. Daghistani, and A. Alshammari, “Improving Accuracy for Diabetes Mellitus Prediction by Using Deepnet,” *Online J. Public Health Inform.*, vol. 12, no. 1, pp. 1–12, 2020, doi: 10.5210/ojphi.v12i1.10611.

[24] U. e. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, “An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study,” *Sensors*, vol. 22, no. 14, pp. 1–15, 2022, doi: 10.3390/s22145247.

[25] B. S. Ahamed, M. S. Arya, and A. O. V. Nancy, “Diabetes Mellitus Disease Prediction Using Machine Learning Classifiers with Oversampling and Feature Augmentation,” *Adv. Human-Computer Interact.*, vol. 2022, 2022, doi: 10.1155/2022/9220560.

APPENDICES

Appendix 1: Ethical Clearance from UR/CMHS Institutional Review Board

UNIVERSITY of RWANDA

COLLEGE OF MEDICINE AND HEALTH SCIENCES

DIRECTORATE OF RESEARCH & INNOVATION

CMHS INSTITUTIONAL REVIEW BOARD (IRB)

Kigali, 28/12/2022
Ref: CMHS/IRB/597/2022

Kwizera Munana Dieudonne
Masters of Biomedical Engineering,
College of Sciences and Technology, CEBE, UR

Dear Kwizera Munana Dieudonne,

RE: ETHICAL CLEARANCE

Reference is made to your application for ethical clearance for the study entitled ***"Diabetes Monitoring and Prediction Using Machine Learning Techniques."***

Having reviewed your application and been satisfied with your protocol, your study is hereby granted ethical clearance. The ethical clearance is valid for one year starting from the date it is issued and shall be renewed on request. You will be required to submit the progress report and any major changes made in the proposal during the implementation stage. In addition, at the end, the IRB shall need to be given the final report of your study.

We wish you success in this important study.

Assoc. Prof. Stefan Jansen (PhD)
Acting Chairperson Institutional Review Board,
College of Medicine and Health Sciences, UR

Cc:

- Principal College of Medicine and Health Sciences, UR
- University Director of Research and Postgraduate studies, UR

Email: researchcenter@ur.ac.rw P.O Box 3286 Kigali, Rwanda www.ur.ac.rw

Appendix 2: King Faisal Data Collection Approval

KING FAISAL HOSPITAL RWANDA
INSTITUTIONAL REVIEW BOARD
Patient Centered Care

IRB Notification of Approval

Ref: KFH/2022/ 021/IRB

Date: December 23, 2022
Protocol Title: Diabetes monitoring and prediction using machine learning techniques.
Principal Investigator:
KWIZERA MUNANA Dieudonné **Email:** *munanakd@gmail.com*
Tel: + *0788265454*
Protocol Reference #: KFH/2022/ 021/IRB
Date of IRB Initial Review: November 15, 2022
Review Type: Expedited Review
IRB Review Decision: Approved
Date of Effectiveness: December 23, 2022
Date of Expiry: December 22, 2023

Dear KWIZERA MUNANA Dieudonné,

King Faisal Hospital Rwanda's Institutional Review Board (KFH IRB) reviewed your protocol resubmission. This letter is to notify you that the KFH IRB approved your resubmission, and this approval is valid for one (1) year and then must be renewed according to the KFH IRB Standard Operating Procedures.

Please note the following considerations:

1. Please review the KFH IRB Standard Operating Procedures and ensure compliance with all requirements, including participant content, changes or amendments to the protocol, and reporting requirements.
2. All project materials, including signed consent forms, must be retained and are subject to review in case of a routine audit
3. Notify the KFH Directorate of Research once data collection is completed
4. The Principal Investigator is requested to submit a hard copy of his/her final manuscript to the Directorate of Research upon completion.
5. Principal Investigators must follow the appropriate study continuing review and closure procedures as indicated in the Standard Operating Procedures Manual.

Please contact us at irb@kfhkigali.com in case of any questions or clarifications.

Sincerely,
Dr. Jean Marie Vianney Dushimiyimana
Consultant ENT Surgeon
Chair, Institutional Review Board

KING FAISAL HOSPITAL