2024-06

PREDICTING DIABETES BASED ON RISK FACTORS AND ASSOCIATED DISEASES USING ENSEMBLE MACHINE LEARNING

HAILEMARIAM, MULUALEM SIMENEH

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

FACULTY OF COMPUTING

DEPARTMENT OF COMPUTER SCIENCE

MSC THESIS ON:

PREDICTING DIABETES BASED ON RISK FACTORS AND ASSOCIATED DISEASES USING ENSEMBLE MACHINE LEARNING

BY:
HAILEMARIAM MULUALEM SIMENEH

JUNE, 2024

BAHIR DAR, ETHIOPIA

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

Faculty of computing

Department of computer science

Predicting diabetes based on risk factors and associated diseases using ensemble machine learning

BY

Hailemariam Mulualem Simeneh

A thesis submitted to the school of graduate studies of Bahir Dar Institute Of technology, BDU in partial fulfilment of the requirement of the degree Of masters in computer science in the faculty of computing

Advisor: Belete Biazen (Assistant professor)

JUNE, 2024

Bahir Dar, Ethiopia

BAHIR DAR UNIVERSITY
BAHIR DAR INSTITUTE OF TECHNOLOGY
SCHOOL OF GRADUATE STUDIES
FACULITY OF COMPUTING

Approval of thesis for defense result

I hereby confirm that the changes required by the examiners have been carried out and incorporated in the final thesis.

Name of Student Hailemariam Mulualem Signature _____ Date 30/12/2016 E.C

As members of the board of examiners, we examined this thesis entitled "Predicting diabetes based on risk factors and associated diseases using ensemble machine learning" by Hailemariam Mulualem. We hereby certify that the thesis is accepted for fulfilling the requirements for the award of the degree of Masters of science in " Computer Science".

Board of Examiners

Name of Advisor	Signature	Date
Mr. Belete Biazen(asst.prof)		12/12/2016 E.C
Name of External examiner	Signature	Date
Getinet Yilma(Ph.D)		10/01/2017 E.C
Name of Internal Examiner	Signature	Date
Tamir Anteneh (asst.prof)		Oct 01/2024
Name of Chairperson	Signature	Date
Dr. Mekonen Wagaw		Sep. 30, 2024
Name of Chair Holder	Signature	Date
Kidist Meshesha		oct 01/2024
Name of Faculty Dean Tesfa Tegegne (Ph.D)	Signature	Date Oct 01/2024

Faculty Stamp

DECLARATION

This is to certify that the thesis entitled "Predicting diabetes based on risk factors and associated diseases using an ensemble machine learning", submitted in partial fulfillment of the requirements for the degree of Master of Science in computer science under the Faculty of Computing, Bahir Dar Institute of Technology, is a record of original work carried out by me and has never been submitted to this or any other institution to get any other degree or certificates. The assistance and help I received during this investigation have been duly acknowledged.

Hailemariam Mulualem Simeneh		JUNE , 2024 G.C
Name of the candidate	signature	Date

ACKNOWLEDGEMENTS

First and foremost, praise and thanks to God, the Almighty, for His showers of blessings throughout my life. Secondly, I would like to express my deep gratitude to my advisor Belete Biazen (Assistant Professor), for his advice and assistance me. Finally, I am grateful to my family and friends for their unwavering support during the challenging times. Their encouragement, love, and motivation have kept me focused and driven. I would also like to recognize the research participants who generously gave their time and shared their experiences, making this study possible. Their insights and perspectives have broadened my understanding of the subject matter and helped me to see new possibilities.

ABSTRACT

The aim of this research is to develop predictive model for diabetes based on risk factors and associated diseases using ensemble machine learning. The problem addressed in this research is to enhance the public health and take the correct action. The research emphasizes the need for timely detection and prediction of diabetes to prevent complications and improve public health. The study was conducted by using experimental research. The data source for this research is the CDC, which was collected by BRFSS. The dataset was 253680 and there is imbalanced. After applying the data pre processing tasks and class balance using random under sampling majority class there is 70692 instances were used for the model. The attribute was reduced to 18 from their original 21features, by using feature selection technique wrapper method (recursive feature elimination)). To construct the best proposed model six experiments were conducted by splitting the dataset in to train, validation and test set with the ratio of 80%, 10%, 10% respectively using Random forest, Catboost, bagging decision tree, AdaBoost, XGBoost and Extra tree algorithms. The performance of the model were evaluate using different evaluation parameters such as precision, recall, accuracy, F1 score, AUC and confusion matrix. The overall accuracy of Random forest, Catboost, bagging decision tree, AdaBoost, XGBoost and Extra tree are 90.16%, 88.94%, 88.97%, 87.87%, 88.81% and 89.86% respectively. Random forest is the best predictive model with an accuracy of 90.16% and ROC of 96% from the others. Model explainability is made to understand and interpret how a machine learning model makes predictions or decisions using local interpretable model explanations (lime).

Key words: diabetes, risk factors, associated diseases, lime, ensemble machine learning, predictive model.

Tables of content

List of tables

List of figures

LIST OF ABBREVIATIONS

AdaBoost-------Adaptive boosting

AI----------Artificial intelligence

AUC----------Area under the Curve

BMI----------body mass index

BRFSS----------The Behavioural Risk Factor Surveillance System

CDC----------Centres for Disease Control and Prevention

CART---------- Classification and Regression tree

CFS----------Correlation-based Feature Selection

DM----------Diabetes mellitus

ID3---------- Iterative Dichortmister 3

HMO----------Health maintenance organization

HPO----------hyper parameter optimization

IG- ----------information gain

K-NN----------K-nearest neighbours

MAR----------Missing at Random

MCAR----------Missing Completely At Random

ML ----------machine learning

NHANES -------National Health and Nutrition Examination Survey

NMAR----------Not Missing at Random

PCA----------principal component analysis

PIDD ----------Pima Indian Diabetes Dataset

ROC- ----------receiver operating characteristic

SFS----------Sequential forward feature selection WHO----------World Health Organization

XGBoost------- Extreme gradient boosting

CHAPTER ONE

INTRODUCTION

1.1 Background

Diabetes is a group of metabolic diseases characterized by hyperglycaemia resulting from defects in insulin secretion, insulin action, or both. The chronic hyperglycaemia of diabetes is associated with long-term damage, dysfunction, and failure of different organs, especially the eyes, kidneys, nerves, heart, and blood vessels [1]

According to the World Health Organization (WHO), diabetes stands as one of the prominent contributors to global mortality [2]. It is estimated that around 422 million individuals, primarily residing in low- and middle-income nations, are affected by this condition. Consequently, diabetes leads to approximately 1.5 million deaths annually on a global scale. These statistics highlight the significant impact of diabetes on public health and emphasize the need for effective prevention and management strategies to combat this widespread disease. Throughout the past few decades, both the number of diabetes cases and its overall occurrence have consistently increased.

Diabetes, also referred to as Diabetes mellitus (DM), is a persistent condition that remains a major and worldwide issue due to its impact on the overall health of the population [3]. It is a long-term health condition distinguished by the insufficient production of insulin in the pancreas or the ineffective utilization of insulin in the body. Insulin, a crucial hormone responsible for managing blood sugar levels, plays a vital role in sustaining overall wellbeing. If left unmanaged diabetes results in hyperglycaemia which can progressively lead to substantial harm to numerous bodily systems especially affecting the nerves and blood vessels. This damage can bring about consequential effects for an individual's health and quality of life.

For this research we use ensemble machine learning because of Machine learning is a branch of artificial intelligence (AI) that focuses on developing algorithms and models to enable computers to learn and make predictions without explicit programming [4]. It involves using

statistical techniques and computational models to analyse complex data, identify patterns, and make informed decisions. Key points about machine learning include learning from data, training algorithms on large datasets, and developing models for learning from data.

Machine learning techniques include supervised learning, unsupervised learning, Semisupervised and reinforcement learning algorithms [5]. The primary goal of supervised learning is to acquire mapping knowledge between the input and the output whose correct values are provided by a supervisor.

There are two main types of supervised learning, classification, and regression, where there are input and output, and the main role is to find a mapping between the input and the output. Ensemble learning is grouped under supervised machine learning algorithms. Numerous students in an ensemble are typically referred to as base learners. Unlike standard machine learning techniques, which aim to learn a single hypothesis from training data, ensemble methods aim to create multiple hypotheses and combine them for practical applications [6]. It is a form of the hybrid learning system in which multiple analytics are combined intelligently in a homogeneous or heterogeneous way to obtain better (more accurate, more robust) results.

The generalization ability of an ensemble is usually much stronger than that of base learners. Ensemble learning is appealing because it can boost weak learners which are slightly better than a random guess to strong learners which can make very accurate predictions [6]. Most ensemble methods use a single base learning algorithm to produce homogeneous base learners, but some methods use multiple learning algorithms to produce heterogeneous learners.

1.2. Motivation

The motivation to develop a prediction model based on risk factors and associated diseases for diabetes using ensemble machine learning arises from several factors. Which means diabetes is a group of metabolic diseases characterized by hyperglycaemia resulting from defects in insulin secretion, insulin action, or both. The chronic hyperglycaemia of diabetes is associated with long-term damage, dysfunction, and failure of different organs, especially the eyes, kidneys, nerves, heart, and blood vessels[1]. According to the World Health Organization (WHO), diabetes stands as one of the prominent contributors to global mortality [2]. It is estimated that

around 422 million individuals, primarily residing in low- and middleincome nations, are affected by this condition. Consequently, diabetes leads to approximately
1.5 million deaths annually on a global scale.

1.3. Statement of the Problem

Diabetes is a group of metabolic diseases characterized by hyperglycaemia resulting from defects in insulin secretion, insulin action, or both. The chronic hyperglycaemia of diabetes is associated with long-term damage, dysfunction, and failure of different organs, especially the eyes, kidneys, nerves, heart, and blood vessels [1]. Over the years, a number of researchers conduct researches by using risk factors to create diabetes prediction models for type 1 and type 2. Risk factors for diabetes can vary depending on the type of diabetes. Risk factors for type 2 diabetes are: age, family history, physical activity, blood pressure etc. for type 1 diabetes risk factors also age, lifestyle, family history and others. In general there are different types of risk factors for diabetes such as age, BMI, Smoking, physical activity, and so on. In the other way, diabetes can directly affect different organs of the human body like the kidney, brain, liver, etc [1] and diabetes is a creator of different diseases [7]. Then diabetes leads to different types of diseases such as heart disease, hypertension, mental disease, and stroke. Due to this a person with those diseases may be becomes diabetic. Then, we understand that disease. So we develop diabetes prediction based on risk factors and associated disease. This is an open research idea because there are no any researches that make diabetes predictions using both risk factors for diabetes and associated disease. Having heart disease means you are more likely to develop diabetes. People with diabetes are also more likely to have certain risk factors, such as high blood pressure or high cholesterol that increase their chances of having a heart attack or a stroke. Then it is very important to study the relationship between diabetes with these associated diseases in addition to risk factors for diabetes.

Generally, this research answers the following research questions.

- ✓ What are the most determinants attributes for determining diabetes?
- ✓ How to select an ensemble machine learning algorithm for diabetes predictive model?
- ✓ To what extent does the proposed predictive model accurately identify the diabetes?

1.4. Objectives of the study

1.4.1. General Objective

The general objective of this research is to develop predictive model for diabetes based on risk factors and associated diseases using ensemble machine learning.

1.4.2. Specific Objectives

To achieve the general objective, the following specific objectives are identified ✓

Identify the key (main) risk factors for diabetes.

- ✓ Design a predictive model for diabetes.
- ✓ Identify which ensemble machine learning algorithm is suitable for diabetes prediction.
- ✓ Evaluate the performance of the predictive model.

1.5. Scope of the study

The coverage of this research is limited to investigating the possibility of designing a predictive model of knowing diabetes using ensemble machine learning approaches. The study focuses on diabetes prediction based on risk factors and associated diseases. The study does not consider the types of diabetes. It predicts the general diabetes. The dataset we use for this research is CDC dataset The Behavioural Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Our dataset contains both risk factors and associated diseases. This data is sufficient for conducting experiments to predict diabetes. To train, validate, test, and analyse the results, only ensemble machine learning algorithms were utilized. This is due to ensemble machine learning being considered best. Because, it combines the strengths of multiple models, improves predictive accuracy, reduces over fitting, handles biases, provides robustness and stability, offers flexibility, and can provide interpretability [8].

1.6. Significance of the study

This research has scientific, methodological, and practical significance by works worth better results in the prediction of diabetes using Ensemble Machine Learning techniques. There is no previous work done using both risk factors and associated diseases for the prediction of diabetes. So, the results of the research can be used as input to the next researchers in the area of diabetes. In the use of diabetes prediction using ensemble machine learning, different stakeholders such as healthcare professionals and affected populations can benefit from more effective and timely diabetes prediction. Accurate diabetes prediction models can assist healthcare systems in allocating resources more efficiently.

1.7. Organization of the thesis

The rest of the paper includes the following chapters. Chapter two focuses on the discussion of related literature. Chapter three provides insight into the methodology, which includes details on the model architecture, dataset collection, pre-processing, model development, and prediction of diabetes based on ensemble machine learning algorithms. The fourth chapter shows the experimental results of the model, highlighting the evaluation metrics, model results, overall model comparison, sample predictions and model explainability. The conclusion and future works are presented in the last chapter.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview

A literature review is essential for any research undertaking to review previous studies in the area of investigation and sum up the trends in research practices and the direction of the findings [9]. In this chapter, different kinds of literature conceptually relevant and related to this study were reviewed. In addition, different literature has been consulted in the area of machine learning (ML) techniques to construct predictive models, as well as different researches done to investigate the problem related to the diabetes.

Diabetes is a group of metabolic diseases characterized by hyperglycaemia resulting from defects in insulin secretion, insulin action, or both. The chronic hyperglycaemia of diabetes is associated with long-term damage, dysfunction, and failure of different organs, especially the eyes, kidneys, nerves, heart, and blood vessels [1].

2.1.1 Types of diabetes

Diabetes is a chronic health condition characterized by either the insufficient production of insulin by the pancreas (Type 1 diabetes) or the ineffective utilization of insulin by the body (Type 2 diabetes). Insulin is a hormone that helps regulate blood sugar levels and allows cells to utilize glucose for energy [1].

In Type 1 diabetes, the immune system mistakenly attacks and destroys the insulinproducing cells in the pancreas, leading to a lack of insulin production. People with Type 1 diabetes require regular insulin injections or the use of insulin pumps to manage their blood sugar levels.

Type 2 diabetes: is the most common form of diabetes, accounting for the majority of cases [10]. It occurs when the body becomes resistant to the effects of insulin or fails to produce enough insulin to meet the body's needs. Risk factors for Type 2 diabetes include obesity, sedentary lifestyle, unhealthy diet, family history, and certain ethnic backgrounds. It can often

be managed through lifestyle modifications such as a healthy diet, regular physical activity, weight management, and, in some cases, medication or insulin therapy. Both types of diabetes can lead to high blood sugar levels (hyperglycaemia), which, if left uncontrolled, can cause various complications such as cardiovascular disease, kidney damage, nerve damage, and eye problems**.**

2.1.2 Risk factors of diabetes

The risk factors for diabetes include a combination of non-modifiable and modifiable factors. Non-modifiable risk factors that increase the risk of developing diabetes include family history, race or ethnicity (such as African American, Hispanic/Latino, American Indian, Asian American, or Pacific Islander descent), age (especially over 45 years old), and a history of gestational diabetes. On the other hand, modifiable risk factors that can be controlled through lifestyle changes include being overweight or obese, physical inactivity, high blood pressure, abnormal cholesterol levels, smoking, unhealthy diet, excessive alcohol consumption, stress, and certain medical conditions associated with insulin resistance [11].

2.1.3 Associated diseases of diabetes

Associated diseases are health conditions that are often observed to coexist with diabetes or are more common in individuals with diabetes. These diseases can be a consequence or a manifestation of the underlying mechanisms that contribute to both diabetes and the associated disease. For example, cardiovascular diseases, hypertension, mental health disorders, Kidney Disease and stroke are commonly seen in individuals with diabetes [1]. These diseases highlight the systemic impact of diabetes on various organs and systems in the body, emphasizing the importance of managing diabetes effectively to prevent complications and maintain overall health.

2.2 Ensemble Machine Learning

Ensemble Machine Learning is a multimodal machine learning technique in which individual learners are combined (e.g., neural network Random Forest, support vector machine Nave Bayes, decision tree) to the predictive model to form one strong model [12]. It combines the output from a set of different algorithms to correctly predict a new dataset. Each singlegenerated model predicts new, unseen data and assigns it to the class value with the highest number of votes [13]. Constructing ensemble classifiers is useful to increase the accuracy of the model over a single machine-learning algorithm [14].

Ensemble learning is divided into two broad categories, namely sequential ensemble and parallel ensemble methods [15]. The sequential ensemble method generates data-dependent sequential base learners. Every data point in the base learner having dependence allows for improving the performance of the model by correcting mislabelled data based on its weight, e.g., adaptive boosting (AdBoost). The successive generations of base learners improve the performance of the model by assigning a higher weight to the previously misrepresented learners. But in the parallel method, the base learner is generated in parallel order, and the data generated in the base learner is independent of each other, e.g., in a random forest [16] [17]. The independence of the base learner is used to reduce the error due to averaging the model's output.

Most of the ensemble method applies a single machine learning algorithm to base learning that gives homogeneity to all base learners that contain the same type of base learners with the same quality [18] The homogeneous ensemble method is a combination of similar types of machine learning algorithms with different datasets for each algorithm that are generated randomly from the original dataset. Other ensemble methods apply heterogeneous base learners with different types and qualities of machine learning algorithms [16].

2.3 Main types of Ensemble machine learning

Ensemble machine learning approaches, like combining boosting algorithms and soft voting classifiers are employed to leverage the collective intelligence of multiple individual classifiers. These techniques address biases, errors, and challenges posed by imbalanced data and missing attribute values in diabetes prediction, ultimately improving overall performance and accuracy [19], [20], [21]. Some type's ensemble learning is discussed in the following sections:

2.3.1 Bagging Ensemble Learning

Bagging ensemble learning combines two machine learning models into a single ensemble model to reduce the high variance of the model [14]. The ensemble model from week learner in bagging is built on each sub-sample of training data through the decision tree by reducing the variance. The reduction of variance improves the performance of the model and reduces the overfitting of the training data. hence, eliminating overfitting and variance is a challenging task for many predictive models [13]. Using bagging is advantageous since each of the individual week learners developed using different sub-sampled training data are combined to form a single strong learner that is more stable than a single machine learning model. It also avoids variance by reducing the overfitting of the model. But bagging is computationally expensive. It can lead to more bias in the model when the proper step in bagging is ignored. The results of each week's learners are aggregated to get the final result of the prediction model [13].
The figure below2.1 shows how bagging ensemble learning works.

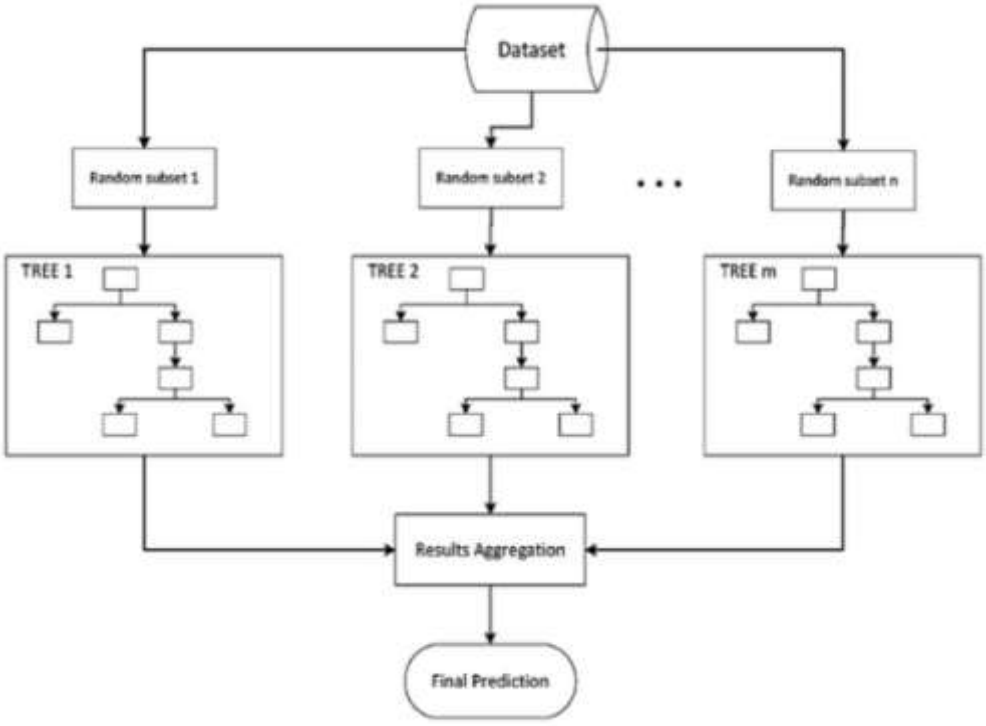

Figure2. 1 Architecture model of Bagging [13]

2.3.2 Boosting ensemble learning

Boosting ensemble is a type of parallel ensemble method that combines the same type of machine learning algorithm [18]. Boosting learns from previous predictor mistakes to make better predictions in the future. Boosting is an ensemble modelling technique that aims to create a strong classifier by combining multiple weak classifiers in a series. It involves building models sequentially, where each subsequent model corrects the errors of the previous one until the training data is predicted accurately or a stopping criterion is met [22]. The combination of different weak learners gives a strong learner, which is well correlated with true classification and prediction accuracy [14] and the boosting method, which addresses the problem of noisy instances, allows for the development of a stronger model from weak learners by averaging their weights.

2.3.3 Stacking ensemble learning

Stacking, another ensemble method, is often referred to as stacked generalization [23]. This technique works by allowing a training algorithm to ensemble several other similar learning algorithm predictions. In addition to this Stacking has been successfully implemented in regression, density estimations, distance learning, and classifications. It can also be used to measure the error rate involved during bagging.

2.4 Feature Selection Techniques

The main aim of feature selection techniques in machine learning is to choose a subset of features by eliminating features. The importance of feature selection is reducing dimensionality, removing irrelevant and redundant data, facilitating data understanding, and improving the accuracy of the predictive algorithm [24]. Three different models deal with feature selection, such as filters, wrappers, and embedded methods [25].

Filter models select the feature based on the characteristics of the data without utilizing a learning algorithm and consist of two steps: in the first step, order the feature, and in the second step, select the feature with the highest rank. They ignore the effect of the selected feature subset on the performance of classification [24].

Wrappers model works on predictive performance to evaluate the quality of selected features. The wrapper model performs the searching of a subset of features to produce a set of features, and the feature evaluation component uses the predefined learning algorithm to evaluate the performance, which has been returned to the feature search component for the next iteration [26].

Embedded models select the feature during the process of model development to perform feature selection without further evaluation of the feature [27]. Feature selection is the most important and frequently used in data processing to maximize classification performance in terms of speed, learning, and accuracy and to gain a better understanding of the underlying process that generates important data [28].

2.5 Related Works

In this study, the researcher analyses and evaluates various literatures from journals and conference papers to attain knowledge about the state of the problem. Also, we conduct a comprehensive literature review to gather existing knowledge on the risk factors influencing diabetes and different associated disease that caused by diabetes**.**

An article by [29] studied the importance of early detection in preventing severe complications of diabetes. They propose a framework that utilizes ensemble learning methods to improve the accuracy of diabetes diagnosis. The study demonstrates the superiority of ensemble methods over individual models, with the stacking method achieving the highest accuracy of 97.50%. However, a limitation of the study is the use of the Pima Indians Diabetes Database, which may limit the generalizability of the findings to other populations.

The study by [3] presents a comprehensive approach to predict diabetes using machine learning techniques. The study utilizes logistic regression and four classifiers (naïve Bayes, decision tree, Adaboost, and random forest) to identify risk factors and predict diabetic patients. The performance of these models is evaluated using accuracy and area under the curve (AUC). The authors employ a diabetes dataset from the National Health and Nutrition Examination Survey (NHANES), consisting of 6561 respondents, including 657 diabetic and 5904 control subjects. The logistic regression model identifies seven factors (age, education, BMI, systolic BP,

diastolic BP, direct cholesterol, and total cholesterol) as risk factors for diabetes. The ML-based system achieves an overall accuracy of 90.62%.

The study conducted by [30] presents a framework for predicting diabetes using machine learning techniques. They utilize decision tree-based random forest and support vector machine learning models and also it used Pima Indian Diabetes Database is a familiar and commonly used data set for the prediction of diabetes. This data set consists of 768 rows and 9 columns. The proposed framework offers 83% accuracy with a minimum error rate, demonstrating its potential for improving diabetes prediction and healthcare outcomes. The Limitation of the study is the use of the Pima Indians Diabetes Database, which may limit the generalizability of the findings to other populations.

The research by[19] a new ensemble learning-based framework for early prediction of diabetes using lifestyle indicators. Ensemble learning techniques such as Bagging, Boosting, and Voting are employed. The dataset used in the study was 27050 and it collects from 10 institutions. The study includes exploratory data analysis to assess the quality of the dataset and uses the synthetic minority oversampling technique for class balancing. The K-fold cross-validation technique is employed to validate the results. Feature engineering is applied to calculate the contribution of lifestyle parameters.

The research by [31] propose an ensemble approach using four different algorithms namely Random Forest, KNN, Naïve Bayes, and J48. They use two datasets, namely the Pima Indian Diabetes Dataset (PIDD) and the 130_US hospital diabetes dataset, for analysis. The paper's strengths include the utilization of multiple datasets for analysis and the achievement of high accuracy in diabetes prediction. However, the paper has weaknesses such as a limited discussion of methodology, lack of comparative analysis with existing methods, insufficient dataset information, and a lack of discussion on the limitations of the proposed approach.

The study conducted by[32] employs Principal component analysis (PCA) and information gain (IG) are two methods of double feature selection to enhance prediction accuracy. 738 records with ML algorithms, namely decision tree, random forest, support vector machine, logistic regression, and KNN were analysed. The study's results demonstrate an accuracy level of above 82.2%. The amount of the participant is too small.

Table 2. 1 Summary of related work

Author and (year)	Title	Methods and dataset	results	Gap
Maniruzzaman et al., (2020)	Classification and prediction of diabetes disease using machine learning paradigm	logistic regression and four classifiers (naïve Bayes, decision tree, Adaboost, and random forest) use 6561 respondents	The overall ACC of ML-based system is 90.62%	a single dataset from NHANES limits the generalizability of the findings
Krishnamoorthi et al., (2022)	A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques	decision tree-based random forest and support vector machine learning models and also it used Pima Indian Diabetes Database (768)	The proposed work gives 83% accuracy.	Use small amount of dataset
Saihood, Qusay Sonuç, Emrullah (2023)	A practical framework for early detection of diabetes using ensemble machine learning models	stacking, boosting, and bagging and used PIDD dataset	Achieves a score of 97.50%, 97.20%, 97.10%, respectively	the use of the Pima, which may limit the generalizability of the findings to other populations.
Ganie & Malik, (2022),	An ensemble Machine Learning approach for	Bagging, Boosting, and Voting are employed, The dataset used in the	the bagged decision tree achieved the highest accuracy	The number of institution where the dataset collected is small.

	predicting Type-II diabetes mellitus based on lifestyle indicators	study was 27050 and it collects from 10 institutions	rate 99.41%	
Alehegn et al.,(2019)	Diabetes analysis and prediction using random forest, KNN, Naïve Bayes, and J48: An ensemble approach	They use two datasets, namely the Pima Indian Diabetes Dataset (PIDD) and the 130_US hospital diabetes dataset, for analysis	The accuracy of proposed ensemble approach is 93.62% for PIDD and 88.56% for 130_US hospital dataset	a limited discussion of methodology, lack of comparative analysis with existing methods

Different researchers conduct Diabetes prediction based on different risk factors. A researcher conduct Diabetes prediction based on lifestyle indicators (like Age, Sex, Height, Weight, Thirst, Fatigue, etc.) [33], [21]. The paper recommended further works to be done with additional indicators. So, we develop Diabetes prediction based on both risk factors and associated diseases.

2.6 summary

Diabetes is a chronic health condition characterized by either the insufficient production of insulin by the pancreas (Type 1 diabetes) or the ineffective utilization of insulin by the body (Type 2 diabetes). Insulin is a hormone that helps regulate blood sugar levels and allows cells to utilize glucose for energy [1]. Machine learning is used to discover hidden patterns or features through historical learning and trends in data [34].The main goal of machine learning is to create models that can train themselves to improve and find solutions to new problems using the training dataset [35].

Machine learning technology is used in many different application areas that deal with large amounts of data and is used to train machines how to handle the data more efficiently and learn from the data [36]. Machine learning is examined in different types including supervised learning, unsupervised learning, Semi supervised learning, and reinforcement learning [5]. Ensemble machine learning approaches, like combining boosting algorithms and soft voting classifiers are employed to leverage the collective intelligence of multiple individual classifiers.

CHAPTER THREE

MODEL ARCHITECTURE AND METHODOLOGY

3.1 Overview

The purpose of this research is to develop a predictive model that will help professionals and consumers identify diabetes based on risk factors and associated diseases. The suggested model architecture is shown in this chapter. The data collection, data preparation, feature selection, train-test split, model training and testing and model explanaibility are all based on the suggested architecture.

3.2 Model Architecture

This research aims to develop a predictive model for diabetes based on risk factors and associated diseases using ensemble machine learning. The architecture of our model is depicted in Figure 3.1 below. The proposed architecture utilizes several phases for constructing a model for predicting the diabetes using CDC data. As it indicates, the model architecture has different components such as data pre-processing, model development (during the training phase) and model evaluation (during testing). The general architecture of the predictive model is from the CDC dataset shown in Figure 3.1 below.

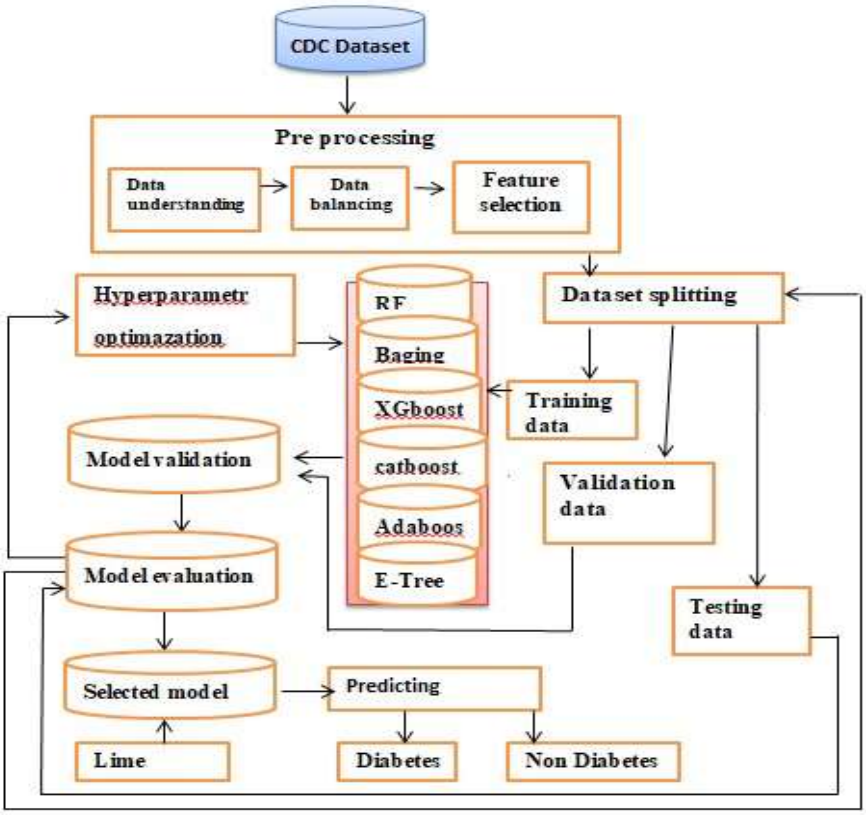

Figure 3. 1 Proposed Model Architecture For Diabetes Prediction

As shown in Figure 3.1, the data collected from the Centres for Disease Control And Prevention (CDC) first passes through data pre-processing. The pre-processing phase makes the input data suitable for the ensemble algorithm. After pre-processing, the preprocessed data splits in to train data, validation data and test data used to train machine learning algorithms to create the prediction model adding hyperparametr optimization and cross validation. Then, after having model validation using validation data, we measure the performance of the model by using test data to select the best-performing model. From the best-performing model, there is a model explainability using Lime.

3.3 Data Source and Description

For this research our dataset source is an online dataset containing both risk factors for diabetes and associated diseases. The Behavioural Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. For this research, the dataset available on Kaggle for the year 2015 was used [37] [38].

Table 3. 1 Description of Features

NO.	Features	Description	Data types	Features Value
1	HighBP	Have you EVER been told by a doctor, nurse or other health professional that you have high blood pressure? (0 = no high BP 1 = high BP)	nominal	O or 1
2	HighChol	Adults who have had their cholesterol checked and have been told by a doctor, nurse, or other health professional that it was high (0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years)	nominal	O or 1
3	CholCheck	Cholesterol check within past five years (0 = no high cholesterol 1 = high cholesterol)	nominal	O or 1
4	BMI	Body Mass Index (BMI)	numerical	12-98
5	Smoker	Do you now smoke cigarettes (Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes)	nominal	O or 1
6	Stroke	Ever told you had a stroke (Ever told you had a stroke. 0 = no 1 = yes)	nominal	O or 1
7	HeartDiseaseorAttack	Ever told you had a heart attack, also called a myocardial infarction?	nominal	O or 1

		(coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes)		
8	PhysActivity	During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise? (physical activity in past 30 days - not including job 0 = no 1 = yes)	nominal	O or 1
9	Fruits	Consume Fruit 1 or more times per day Consume Fruit 1 or more times per day 0 = no 1 = yes	nominal	O or 1
10	Veggies	Consume vegetables 1 or more times per day (Consume Vegetables 1 or more times per day 0 = no 1 = yes)	nominal	O or 1
11	HvyAlcohol Consump	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) (Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per)	nominal	O or 1

12	AnyHealthcare	Do you have any kind of health care coverage, including health insurance, prepaid plans such as	nominal	

		HMOs, or government plans such as Medicare, or Indian Health Service? (Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes)		O or 1
13	NoDocbcCost	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? (Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes)	nominal	O or 1
14	GenHlth	Would you say that in general your health is (Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor)	nominal	1 to 5

15	MentHlth	Thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (Thinking about your mental health, which includes stress, depression, and problems with emotions)	nominal	1 to 30
16	PhysHlth	Thinking about your physical	nominal	
		health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (Thinking about your physical health, which includes physical illness and injury, for how many days during the past 30)		1 to 30
17	DiffWalk	Do you have serious difficulty walking or climbing stairs? (Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes)	nominal	0 or 1
18	Sex	Indicate sex of respondent (0 = female 1 = male)	nominal	1=male, 0=femal
19	Age	Thirteen-level age category (13-level age category age 5 years for each level 1 = 18-24 9 = 60-64 13 = 80 or older)	numerical	1-13

20	Education	Level of education completed (Education level scale 1-6 1 = Never attended school or only kindergarten)	nominal	1 to 6
21	Income	Is your annual household income from all sources (Income scale 1-8, 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more)	numerical	1 to 8

The table 3.1 above shows the general 21 features, their descriptions, data types and the overall feature values.

3.4 Target Class

The purpose of this research is to predict the diabetes based on risk factors and associated disease. The target attribute selected in this study is binary diabetes. The target variable Diabetes binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. The class attribute is considered a dependent variable, while the rest of the variables indicated are the independent variables for this particular study.

3.5 Data Pre-processing

To ensure the quality and suitability of the collected data for analysis, it is important to perform data cleaning and pre-processing [4]. This involves various steps such as handling missing values, removing outliers, and transforming variables as necessary. Pre-processing, which includes feature selection, data transformation, data understanding, handling imbalanced and feature selection data sets is the final stage before to data analysis and modelling [8].

The data understanding phase primarily focuses on clearly understanding all the features of the dataset, how it can be applied for this research, and creating a target dataset with selected sets of variables that are related to the discovery process [39]. The data understanding phase primarily focuses on creating a target dataset with selected sets of variables that are related to the discovery process.

Data integration allows the combining of data from different sources and provides the data to the user with a unified view of the data [40]. Data cleaning deals with identifying and removing (correcting) errors and inconsistencies (noise) from the data to improve the quality of the data [41]. Handling missing values, incorrectness, inaccuracy, and irrelevancy is the major task of data cleaning [42]. Data transformation is a part of the data preparation phase that plays a basic role in ensuring data quality and representation, better data visualization, and data volume reduction before data analysis is conducted [43].

Feature selection is an essential data processing step to apply a learning algorithm[44]. In feature selection, relevant features are selected and irrelevant and redundant data are eliminated from the dataset with a minimum loss of data related to the outcome by using different feature selection methods. As a result, the researcher identified the most relevant feature as input to construct a predictive model using ensemble machine learning algorithms. Data pre-processing plays a great role in the performance of a machine learning algorithm by removing noise and irrelevant features from the dataset [41]. If there is irrelevant, redundant, noisy, and unreliable data in the data we used to develop the model, then discovering the desired outcome during the training phase is more difficult. So, data pre-processing is necessary and includes data cleaning, integration, normalization, transformation, and feature selection [45].

3.5.1 Data understanding

The data understanding phase primarily focuses on clearly understanding all the features of the dataset, how it can be applied for this research, and creating a target dataset with selected sets of variables that are related to the discovery process[46]. The data understanding phase primarily focuses on creating a target dataset with selected sets of variables that are related to the discovery process. Without understanding the existing data, it is difficult to describe the target dataset from the source since the real-world data is unclean and not suitable at the source to run the machine learning process. To understand the datasets in this study, we have used visualization techniques.

3.5.2 Data cleaning

Data pre-processing involves handling missing values, duplicates, and outliers [47]. There are three mechanisms for missing values [48]. Missing Completely At Random (MCAR), which occurred when the probability of a missing value was not related to the estimated value, and Missing at Random (MAR), which occurred when the probability of a missing value was unrelated to its value but depended on other aspects of the observed data, Not Missing at Random (NMAR) is the probability of a missing observation related to its value. In this phase of data pre-processing, we clean data by removal of redundant data (duplicates).

Table 3. 2 Missing value after cleaning

```
Missing values after cleaning:
HighBP                  0
HighChol                0
CholCheck               0
BMI                     0
Smoker                  0
Stroke                  0
HeartDiseaseorAttack    0
PhysActivity            0
Fruits                  0
Veggies                 0
HvyAlcoholConsump       0
AnyHealthcare           0
NoDocbcCost             0
GenHlth                 0
MentHlth                0
PhysHlth                0
DiffWalk                0
Sex                     0
Age                     0
Education               0
Income                  0
Diabetes                0
dtype: int64
```

As shown from the above table 3.2 there is no missing value after cleaning.

3.5.3 Feature Importance

A variable's significance is a measure of how much of the model it adds. The utility of a given variable for the current model and prediction is ascertained [49]. To indicate the overall relevance of a characteristic, we utilize a numerical value called the score; the greater the score

value, the more significant the feature. Having a feature significance score has many benefits. For instance, it is possible to ascertain the link between independent and dependent variables. Using variable relevance scores as a guide, we find and remove features that are not relevant. It is possible to speed up or enhance the model's performance by reducing the amount of irrelevant variables. As illustrated in figure 3.2 below, we compare the significance of the attributes utilized to determine which predictors are the most significant.

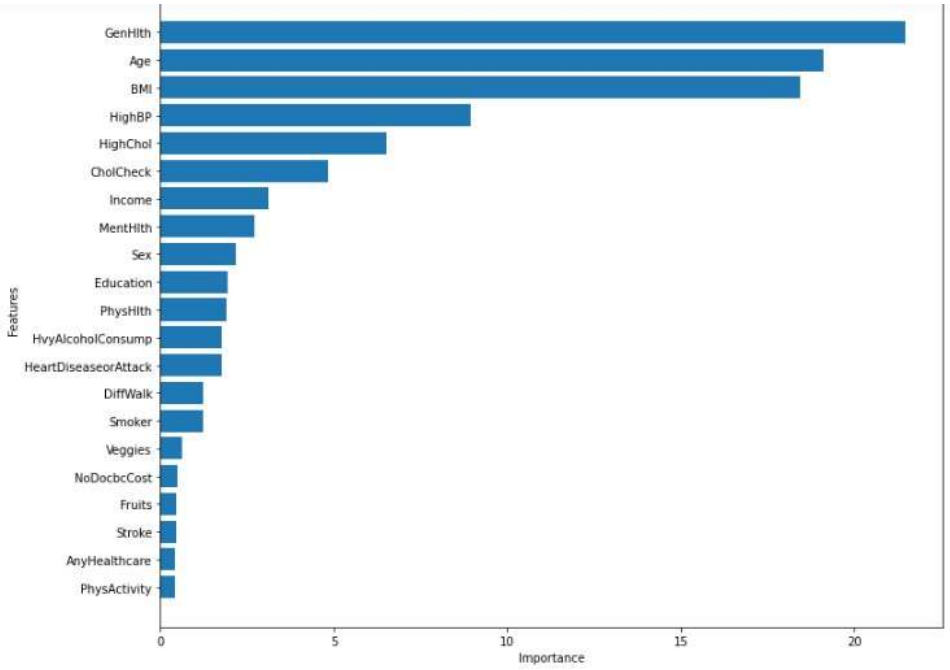

Figure 3. 2 Feature importance

3.5.4 Data Transformation

To make the dataset appropriate for this study, the data discretization and normalization techniques are applied to continuous attributes to minimize distinct values of attributes. Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals [50]. The feature that needs discretization is the body mass index attribute[51]. The attribute has continuous values in the original dataset, but this attribute was reduced as follows: Here, the researcher considers obesity, overweight, underweight and normal weight to reduce complexity BMI as shown in figure 3.3 below.

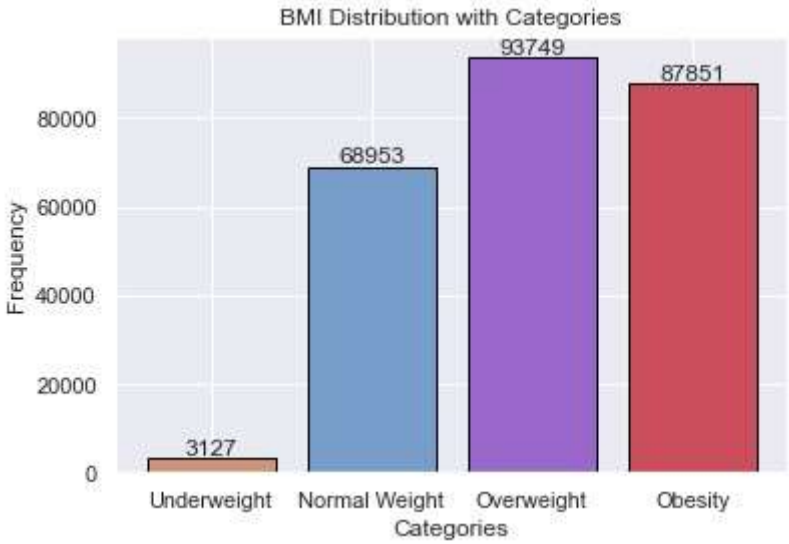

Figure 3. 3 BMI distribution after transformation

3.5.5 Data balancing

The class level of the collected data is imbalanced (see Figure 3.4 below). By adding or removing samples from the dataset, the imbalanced class distribution problems can be solved [52] . As you can see from figure 3.4 below the distribution of the class label was imbalanced. Random under sampling is a technique used to address class imbalance in datasets by reducing the number of instances in the majority class. This method is particularly important where the minority class is underrepresented, which can lead to biased models that favor the majority class. The primary goal of random under sampling is to create a more balanced dataset by reducing the size of the majority class to match that of the minority class. This helps prevent machine learning algorithms from being biased towards the majority class, which could lead to poor predictive performance, especially for the minority class[53]. Random under sampling reduces the dataset size, which can lead to faster training times for machine learning models. This is particularly beneficial when working with large datasets, as it can significantly decrease the computational resources required for model training and evaluation[53]. To get the best-performing model, the data must be balanced see figure 3.5. So, we used Random Under-sampling the majority class techniques to handle the class imbalance of the class levels of the dataset due to much amount of data and in order to save resources.

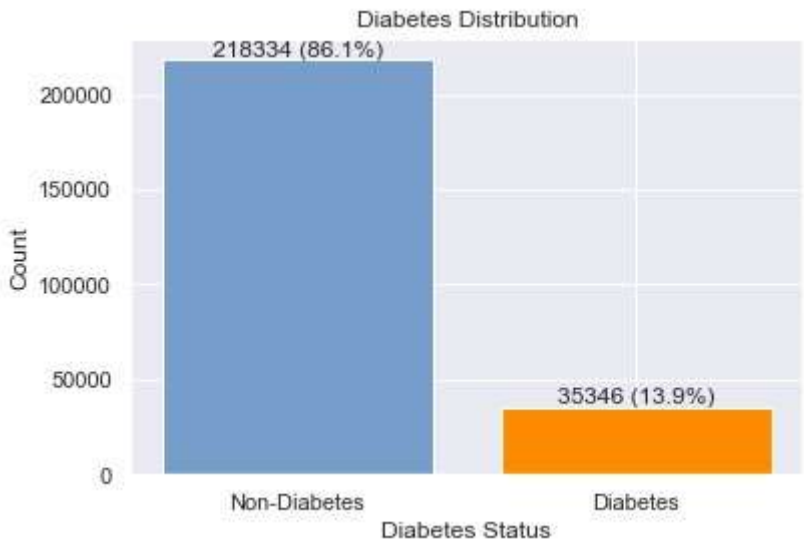

Figure 3. 4 Class distribution before balancing

The data which collect from CDC was 253680. After applying the data pre-processing methods, the remaining dataset is 226927 with 21 attributes including 1 target class and after applying Random Under-sampling techniques to the pre-processed data; the data becomes 70692 instances with 21 attributes.

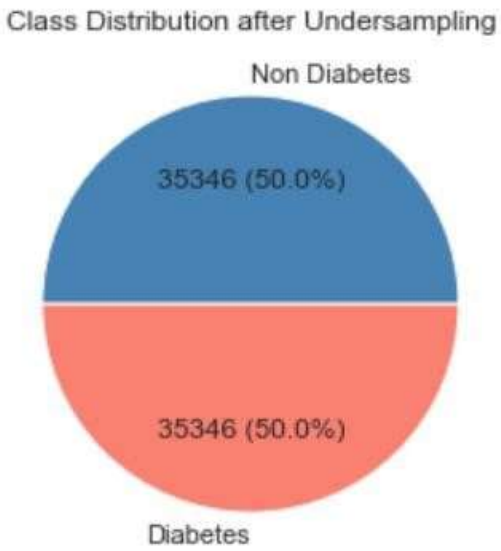

Figure 3. 5 Class distribution after under sampling

As we see from the above figure 3.4, the proportion of diabetes class is much smaller than non-diabetes class. But after applied random under sampling majority class the proportion of the class is equals. See figure 3.5. We used this data to an experiment for feature selection.

3.5.6 Feature selection

The process of selecting the most relevant features and creating a logical model with increased prediction power for signs is known as feature selection [54]. Feature selection is a pre-processing technique used to reduce the dimensionality of a dataset by removing redundant and irrelevant features [55]. In addition, feature selection improves the performance of classification algorithms by reducing the number of features, reducing model complexity with a lower computational cost to construct, and using the classification model by reducing over-fitting [56]. In this experiment, we used the two types of feature selection methods (filter and wrapper) to see which one could give us better performance, and we chose the best one. In both filter and wrapper feature selection methods, we use 70692 training sets with 21 features to compare the result with the best-performing attributes.

Filter method

In filter methods, features are selected based on the characteristics of the data without utilizing learning algorithms[57]. However, filter methods don't focus on the biases and heuristics of the learning algorithms[58]. Due to this, the filter may lose features that are relevant to the target learning algorithm. In the filter method, mutual information and a chisquare test were conducted. Mutual information is the measure of the amount of information between two random variables that are always symmetric and non-negative. It could be zero if and only if the variable is independent. A chi-square test determines the independence of two variables, such as the observed count and the expected count. When two features are independent, the observed count is close to the expected count, so we have a smaller ChiSquare value [58]. As a result, the greater the Chi-Square value, the more dependent the feature is on the response, and it can be chosen for model training. A filtering method consists of two steps[58]. In the first step, features are ranked based on certain criteria [58]. In the second step, features with the highest rankings are chosen [57]. The researcher determined threshold values for future experiments using a "Random Forest Classifier" through an iterative feature evaluation process for accuracy. Then, using all 21 features, we ran filter feature selection methods.

Wrapper methods

The major drawback of the filter method is that it ignores the effects of the selected feature subset on the performance of the clustering or classification algorithm[58]. The wrapper methods need classifiers for selecting relevant features[59]. So, for the step-forward and stepbackward sequential feature selection, we used random forest classifiers. Sequential forward feature selection (SFS) is a feature selection technique where features are added one after the other to an empty candidate set till the criterion is not lowered by adding further features[60]. Sequential backward is another feature selection technique in which features are sequentially removed from a full candidate set until the removal of further features increases the criterion [60]. Wrapper methods use a specific learning algorithm to evaluate the quality of the selected features [57] and also Common wrapper methods are forward feature selection, backward feature elimination, and recursive feature elimination. The wrapper methods need classifiers for selecting relevant features [57]. To conduct the study, we used random forest classifiers with parameters tuned by grid search, bagging classifier, adaboost classifier, XGBoost classifier, cat boost classifier and extra tree classifier.

3.5.7 Train-Test Split

The researcher attempted to prepare a dataset for training, validating and testing. The ratiobased random splitting technique is employed to split the whole dataset to train validate and test data. because, with a random split, one can run as many experiments as required [61]. The main objective of any machine learning model-building process is to develop a generalizable model on the available dataset that has performed well on predictions based on unseen new data [36].

In this research, to estimate a model's performance on unseen new data, we need to have a separate dataset. This is achieved by splitting our 70692 available datasets into training validating and testing sets. The training set is used for the ensemble algorithm to learn and create prediction models. The validating data is used to validate the model. The testing dataset is used to measure the accuracy (predictive capability) of the model for the unseen new datasets. It is important that the training set the validation set and testing set are independent of each other and do not overlap [62]. This is because if we use the testing set as part of our training data, then the classifier's generalizability has been low since it has already seen the testing examples before and learned from them [63]. We have to keep this testing set separate from the training process and use it only to evaluate the model.

3.5.8 Creating a predictive model

After completing data pre-processing and splitting the data into training validation and test sets, ensemble machine learning algorithms, namely, Random Forest, Extreme Gradient Boosting (XGBoost), Adaboosting, CatBoost, Bagging decision trees and Extra trees are used to build a predictive model.

A. Random forest

Random Forest is an ensemble machine-learning technique that supports multi-class prediction with multiple decision trees [14]. The prediction is robust against noise with high performance and is capable of training the dataset and classifying the new dataset with high speed. Also, a random selection of features to be used at splitting nodes enables fast training, even if the dimensionality of the feature vector is large[64]. Random forests use bagging to create sample subsets from the original dataset by random sampling from the training sample. For each sample random forest, construct a decision tree, and the class that has large average class probabilities is obtained as the classification output. However, a random forest requires many decision trees, as using fewer decision trees reduces the performance of the model. The beginning of random forest algorithm starts with randomly selecting "k" features out of a total of "m," where k is less than m. Second, among the "k" features, calculate the node "d" using the best-split point. Third, split the node into daughter nodes using the best split. Fourth, repeat the first through third steps as needed to create an "n" number of trees. Fifth, build a forest by repeating the first four steps an "n" number of times to create an "n" number of trees.

B. Bagging Decision Trees (Bootstrap Aggregating)

Bagging is short for bootstrap aggregating, which involves creating multiple subsets of the original dataset through random sampling with replacement. Each subset is used to train a separate decision tree model. During the training process, each decision tree is exposed to different subsets of the data, allowing them to capture different patterns and variations. The final prediction is made by aggregating the predictions of all the individual decision trees, typically through majority voting (for classification) or averaging (for regression).

C. Extra Trees (Extremely Randomized Trees)

Extra Trees, also known as Extremely Randomized Trees, is an extension of the random forest algorithm. Like random forest, it builds an ensemble of decision trees, but with a slight difference in the tree construction process. In Extra Trees, the splitting of nodes is done randomly, without considering the optimal split point. Instead of finding the best split among a subset of features, Extra Trees randomly selects split points. This randomness further reduces the variance of the model but may increase bias compared to traditional decision trees or random forests. Similar to random forest, predictions are made by aggregating the predictions from all the individual trees in the ensemble.

D. Extreme gradient boosting (XGBoost) algorithm

Extreme gradient boosting is preferred by data scientists because of its high execution speed outside of core computation [65]. Regression trees serve as the weak learners in gradient boosting regression; an input data point is mapped to one of the leaves of a regression tree, which provides a continuous score. By combining a convex loss function (derived from the difference between the target and projected outputs) with a penalty term for model complexity (i.e., the regression tree functions), XGBoost minimizes a regularized objective function. Iteratively adding new trees that forecast the residuals or errors of previous trees, which are then integrated with earlier trees to produce the final prediction, is how the training process is carried out. The author employed a gradient descent approach to minimize the loss when adding new models, which is why it's termed "gradient boosting" [65].

E. Adaboosting algorithm

AdaBoost, also called "adaptive boosting," is a technique in machine learning used as an ensemble method. The most common algorithm used with AdaBoost is a decision tree with one level, which means decision trees with a single split. AdaBoost works by putting more weight on difficult-to-classify instances and less on those already handled well. AdaBoost algorithms can be used for both classification and regression problems [66]. This algorithm creates a model and gives equal weights to all the data points. Then, it assigns higher weights to points that are

wrongly classified. Now all the points that have higher weights are given more importance in the next model. It has continued to train models until a smaller error is observed[66].

F. Cat Boost Algorithm

Cat Boost ensemble learning is efficient in predicting categorical features, and it is an implementation of gradient boosting, which makes use of decision trees as base predictors [67]. Cat Boosting is a high-performance decision tree-based gradient boosting algorithm with a light and accurate framework that has built-in support for categorical features. To solve categorical problems, Cat Boosting uses ordered-based boosting that builds an oblivious tree model that prevents over-fitting and greedy target statistics methods on randomly shuffled training datasets to improve model robustness. During training, a set of decision trees is built consecutively. Each successive tree is built with reduced loss compared to the previous trees. The number of trees is controlled by the starting parameters [68].

3.5.6 Model Explainability

Model explainability refers to the ability to understand and interpret how a machine learning model makes predictions or decisions. It involves gaining insights into the factors and patterns that contribute to the model's outputs. Explanation is important for transparency, trust, debugging, and compliance with regulations. Techniques such as assessing feature importance, extracting decision rules, generating local explanations, using visualizations, and creating simplified models can be employed to achieve explainability. While full explainability may not always be attainable, efforts are made to provide insights into model behavior without sacrificing performance [69].

3.6.1 LIME

LIME is a technique used to provide local, interpretable explanations for individual predictions made by any machine learning model [70]. It approximates the decision boundary of the model around a specific instance by sampling the input space and training a local surrogate model. This surrogate model approximates the original model's predictions in the vicinity of the

instance and its coefficients indicate feature importance. LIME is modelagnostic, applicable to any model type and input data. It has gained popularity for explaining black-box models, offering transparency where it's lacking. By providing local explanations, LIME builds trust, aids in identifying biases or errors, and evaluates the impact of input parameters on output predictions for specific instances [71]. The researcher used lime for model explainability.

3.7 Model Evaluation

The prediction model is evaluated using objective-based evaluations. The most common evaluation metrics used in most machine learning classification applications are accuracy, confusion matrix, precision, recall, and fl-score [72].

3.7.1 Confusion Matrix

A confusion matrix is used to evaluate the performance of the models. A confusion matrix is a table that summarizes how successful the classification model is at predicting examples of various classes [73]. One axis of the confusion matrix represents the label that the model predicted, and the other axis is the actual label. Each cell in the confusion matrix represents one of the True Positive (TP), True Negative (TN), False positive (FP), or False-Negative (FN) outcomes of the prediction results of a model. The TP and TN decisions are correct, while the FN and FP decisions are incorrect (Beauxis-aussalet, 2014). In this study, the class level has two classes. We have calculated all four possible decisions for the two classes. Each column in the confusion matrix represents classified instance counts based on predictions from the model, and each row of the matrix represents instance counts based on the actual class labels.

3.7.2 Accuracy

Accuracy is the percentage of correct predictions made by the model. This metric is commonly used for assessing classification models. It measures how often the classifier makes the correct predictions [73]. This metric is useful when there are equal numbers of observations in each class and all predictions are important. Accuracy has been valued in the range of [0, 1] [73]. If

accuracy is equal to 1, that means all samples in the dataset are correctly classified. In contrast, if accuracy is equal to 0, that means none of the samples in the dataset are classified correctly.
Accuracy= $(TP + TN)/(TP + TN + FP + FN)$ ---------- (3.1)

3.7.3 Precision

Precision measures the number of actual positive cases out of all the positive cases predicted by the model based on the positive class [73]. Precision refers to the closeness of two or more measurements to each other. The precision of measured values refers to how close the agreement is between repeated measurements[75]. The precision of a measuring tool is related to the size of its measurement increments. The smaller the measurement increment, the more precise the tool is [76].
Precision = $(TP)/(TP + FP)$ ---------- (3.2)

3.7.4 Recall

The recall is the percentage of real positive instances that can be efficiently predicted as positive [77]. This measures the coverage of the real positive instances through the +P (predicted positive) rule. Its suitable feature is reflected in how most of the relevant instances of the +P rule choices and recall tend to be neglected or averaged away in machine learning and computational linguistics, where the focus is on how confident we can be in the rule or classifier[77]. Recall= $(TP)/(TP + FN)$ ---------- (3.3)

3.7.5 F1-score

The F1 score is a weighted average score of the true positive (recall) and precision, as well as the F1-Score, which assesses the classification model's performance starting from the confusion matrix. It aggregates precision and recalls measures under the concept of harmonic mean. It reaches its best value at 1 and its worst score at 0 [78].
F1-score= (2*precision*Recall)/(precision + Recall) ---------- (3.4)
The true positives, true negatives, false positives, and false negatives are also useful in assessing the costs and benefits (or risks and gains) associated with a classification model.

3.7.6 ROC curve

A receiver operating characteristics (ROC) curve graph shows the true positive rates against the false positive rate at various cut points and visualizes, organizes, and selects classifiers based on their performance. In essence, it is another performance evaluation technique for classification models and also a useful tool for comparing two or more classification models. It also demonstrates a trade-off between sensitivity (recall) and specificity (the true negative rate). ROC graphs are two-dimensional graphs in which the TP rate is plotted on the Y-axis and the FP rate is plotted on the X-axis. A ROC graph depicts relative tradeoffs between benefits (true positives) and costs (false positives). To plot a ROC curve for a given classification model, the true positive (TP) rate is plotted on the Y-axis, and the false positive (FP) rate is plotted on the X-axis. In the process of drawing the Roc curve, we start at the bottom left-hand corner (where the true positive rate and false-positive rate are both 0), and we check the actual class label of the tuple at the top of the list. If we have a true positive (that is, a positive tuple that was correctly classified), then on the ROC curve, we move up and plot a point. If, instead, the tuple belongs to the "no" class, we have a false positive. On the ROC curve, we move right and plot a point. This process is repeated for each of the test tuples, each time moving up on the curve for a true positive or toward the right for a false positive.

The researcher considered the accuracy, precision, f1-score, recall, and ROC area when the classifier performance is evaluated to select the best model. The tuning parameter is set values of the parameter before the training process [55]. In machine learning, it is as important as data cleaning and feature extraction [79]. The hyper-parameter is very sensitive to a small change in the learning rate or the estimators, which leads to a great change in the accuracy of the model. In this study, we used a grid search parameter tuning method to optimize model performance. Grid search is the traditional method for tuning hyper parameters. We made a grid search for the best score by joining the values and finding the best combination of values. Grid search always results in an optimal solution, but it is timeconsuming and, because of the large combination, requires high computational power, which makes it expensive [79].

3.8 Development Tools

In this study, we used different hardware and software tools to develop a binary class diabetes prediction model based on risk factors and associated diseases using an ensemble machine

learning algorithm. As a hardware tool, we use an Intel(R) Core(TM) i3-7200U CPU @ 2.50 GHz or 2.70 GHz, 8 GB of memory, and a 1TB hard disk. Similarly, as software tools, we used Microsoft Office Word 2010 for writing documentation, Microsoft Office Presentation 2010 for thesis presentations and Microsoft Office Excel 2010 for understanding the datasets manually. This research was implemented using Anaconda 3 on the Jupyter Notebook.

3.9 Summary

This research aims to develop a predictive model for diabetes based on risk factors and associated diseases using ensemble machine learning. To ensure the quality and suitability of the collected data for analysis, it is important to perform data pre-processing .Feature selection is a pre-processing technique used to reduce the dimensionality of a dataset by removing redundant and irrelevant features. The data understanding phase primarily focuses on clearly understanding all the features of the dataset, how it can be applied for this research, and creating a target dataset with selected sets of variables that are related to the discovery process. The purpose of this research is to predict the diabetes based on risk factors and associated disease. The target attribute selected in this study is binary diabetes. For this research our dataset source is an online dataset containing both risk factors for diabetes and associated diseases. The Behavioural Risk Factor Surveillance System (BRFSS) is a healthrelated telephone survey that is collected annually by the CDC. For this research, the dataset available on Kaggle for the year 2015 was used. The most common evaluation metrics used in most machine learning classification applications are accuracy, confusion matrix, precision, recall, and fl-score.

CHAPTER FOUR

EXPERIMENTATION, RESULT AND DISCUSSION

4.1 Over view

In this chapter, the experiments carried out to develop the proposed model are discussed. The experiments were conducted to predict and characterize the contributing factors. The results of the developed model are explained and presented, taking into account the training, validation, and test outcomes based on the specified parameter considerations. The experimental evaluation confirms the effectiveness and feasibility of the proposed model architecture. Furthermore, the chapter thoroughly describes the influence of hyper parameters and the implementation details of the proposed model.

4.2 Descriptive analysis

The data collected from the CDC contains 253680 records with 21 attributes to predict the diabetes. But after applying the data pre-processing methods there is 226927 instances with 21 attributes including 1 target class and after applying class balancing using random under sampling technique, the remaining datasets are 70692 with 21 attributes, and 1 target variable.

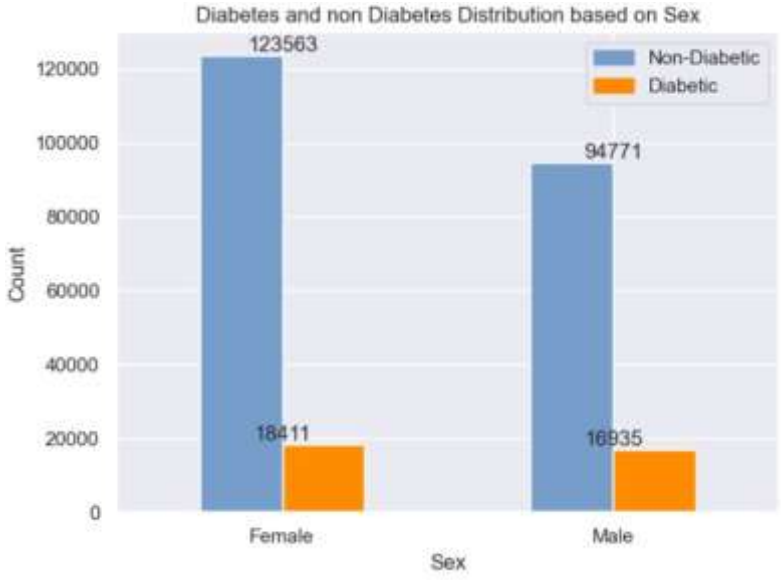

Figure 4. 1 Diabetes distribution based on sex

As the figure 4.1 above shows out of 253680 collected dataset there are 11794 females and are 111706 males. As the graph shows the proportion of females are greater than males. There are

123563 non diabetes and 18411 diabetes in females. Out of 111706 males, 16935 are diabetes and 94771 are non-diabetes.

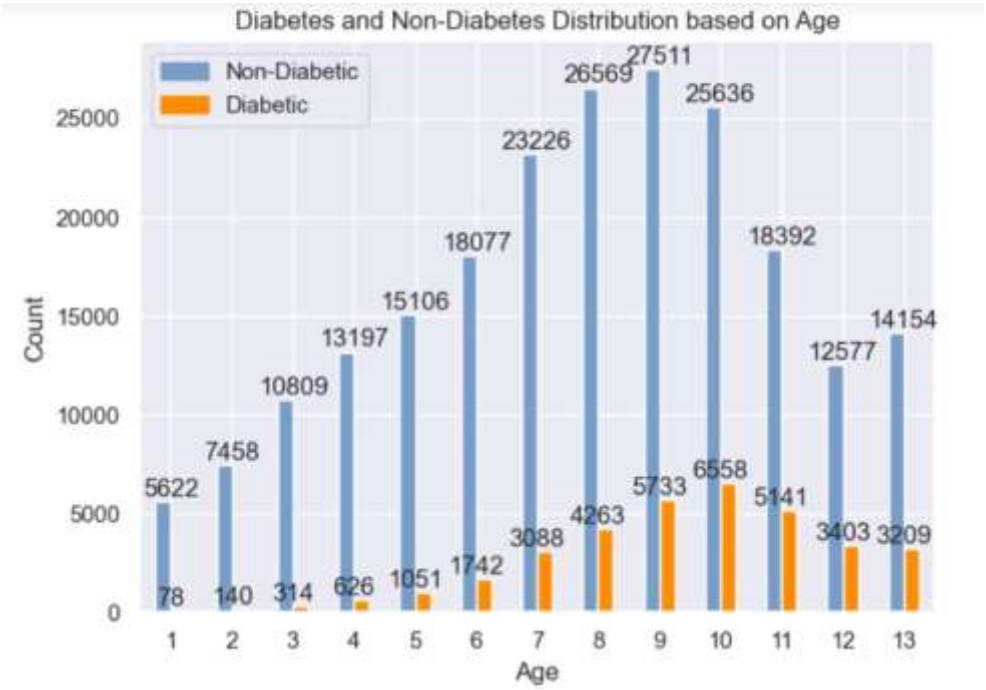

Figure 4. 2 Diabetes distribution based on age

As shown in the figure 4.2 there are 13 age categories in the dataset. As the figure shows when the age increase the probability of diabetic also increases in the age categories of one to ten.

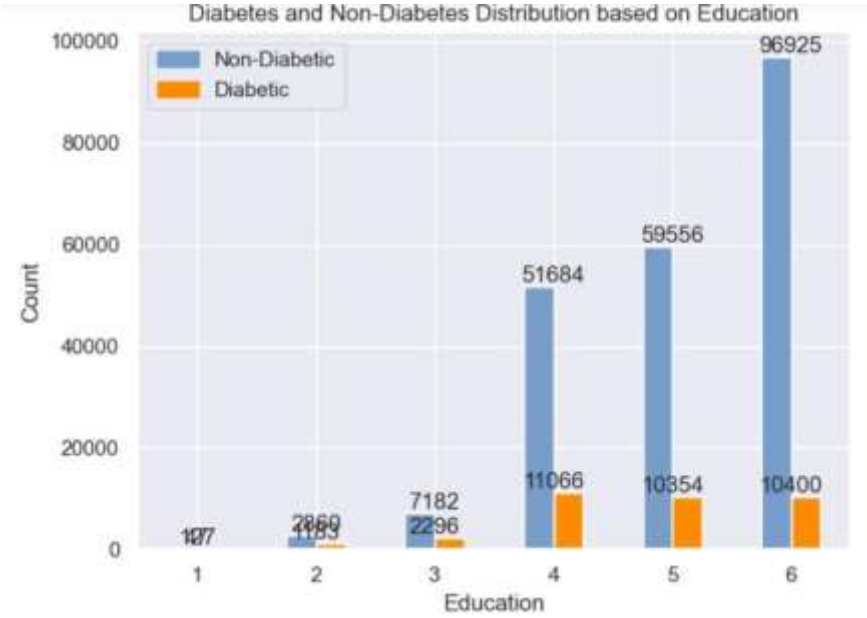

Figure 4. 3 Diabetes distribution based on education

As we see in fig 4.3 the number of non-diabetes is increase as increase education level and in college four years is greater than all education levels. As compared the diabetes class also,

grade 12 or high school graduates is greater than all the other education levels and there is a very small in non-diabetes and diabetes or both class who never attend school.

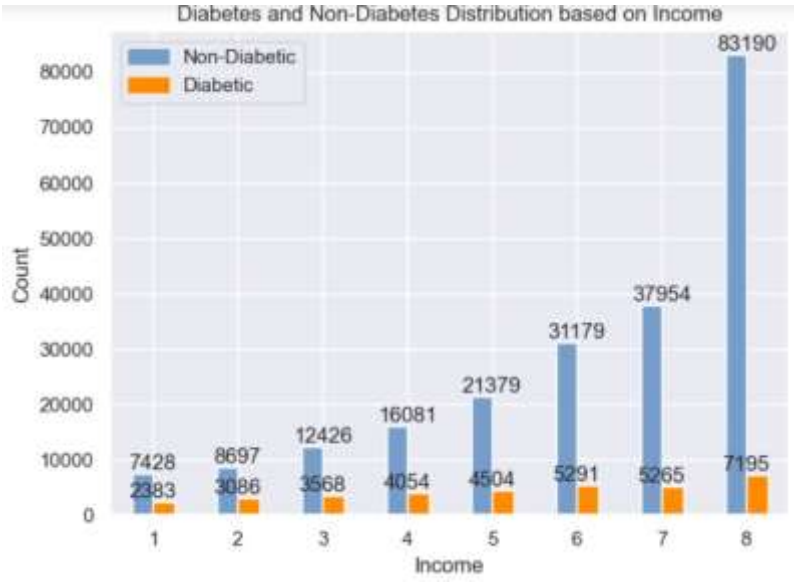

Figure 4. 4 Diabetes distribution based on income

As you see in fig 4.4 above, there are eight income levels based on annual income from <= $10000 to >= $75000. In income eight there is high non diabetes and diabetes class as compared from other income groups and there is 2383 diabetes and 7428 non diabetes in income less than or equals to 10000 annual income.

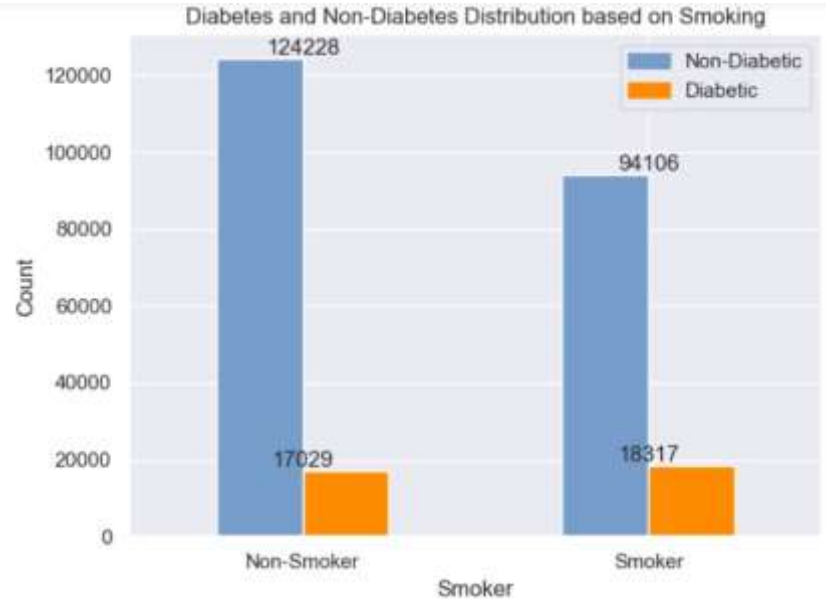

Figure 4. 5 Diabetes distribution based on smoker

As we see in figure 4.5 when the number of people who smoke increase, the probability of diabetes also increase. In non smoker class there are 17029 diabetes and 124228 non diabetes and also there are 18317 diabetes and 941106 non diabetes people in smokers.

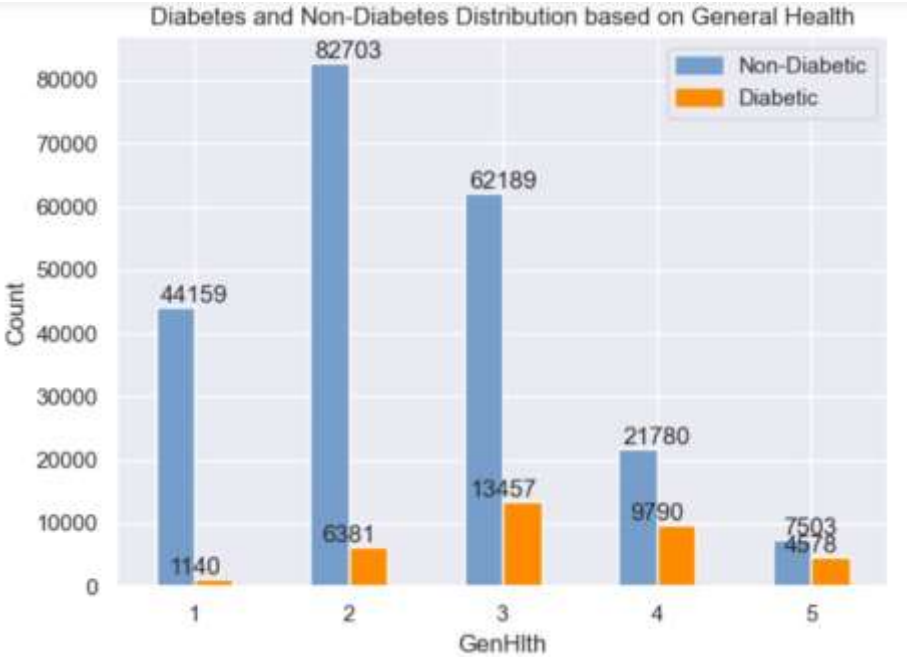

Figure 4. 6 Diabetes distribution based on general health

From fig 4.6 when the general health is excellent (1) there is less infected in diabetes and when the general health is poor (5) there is high probability of diabetes. Generally we conclude that when the general health is decrease there is a high probability of diabetes.

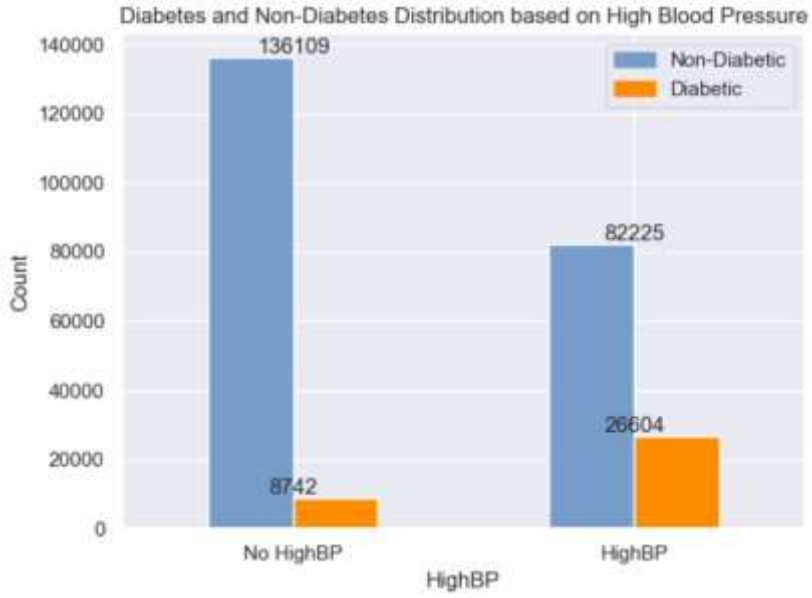

Figure 4. 7 Diabetes distribution based on High BP

As we see from fig 4.7 when there is high BP there is high diabetes (26604) and when no high BP (8742) is decrease the number of people who affected in diabetes is decrease and there is high non diabetes in no high BP (136109) as compared high BP (82225).

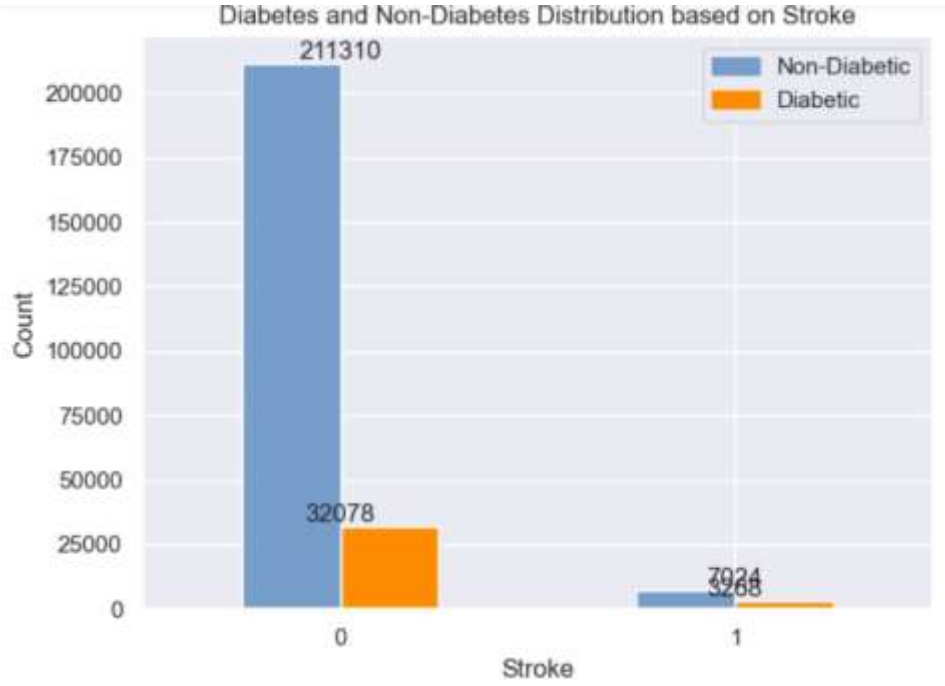

Figure 4. 8 Diabetes distribution based on stroke

As shown in the figure 4.8 there are 7924 non diabetes people and 3268 diabetes people who have stroke and in non diabetes class there are 32078 diabetes and 211310 non diabetes people who have no stroke. From here we conclude that the probability of diabetes is increased as increased stroke.

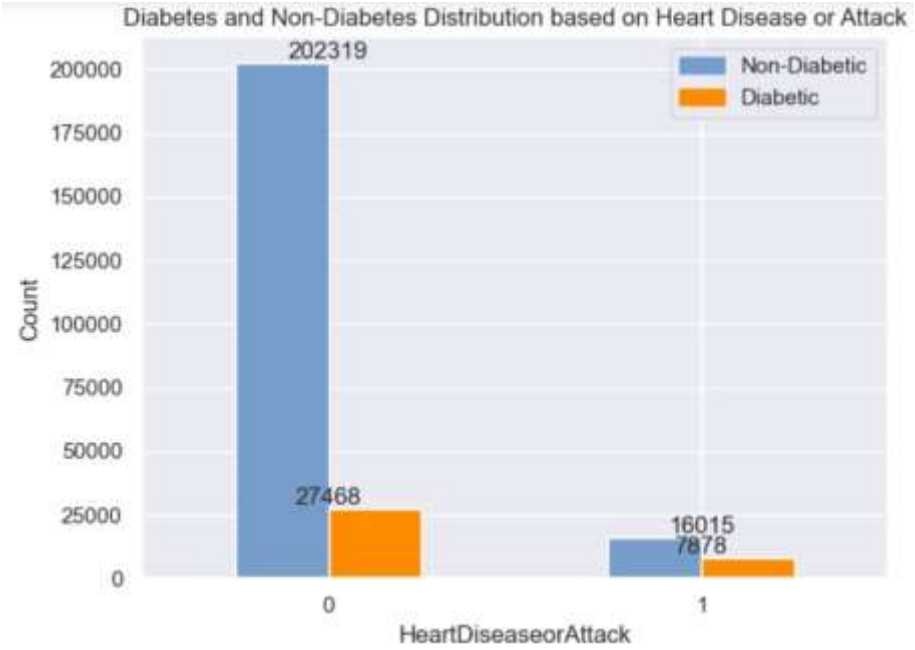

Figure 4. 9 Diabetes distribution based on Heart disease or attack

As you see the figure 4.9 above there are 16017 non diabetes people and 7878 diabetes people who have heart disease attack and also there are 27468 people have diabetes and 202319 non diabetes that haven't heart disease attack.

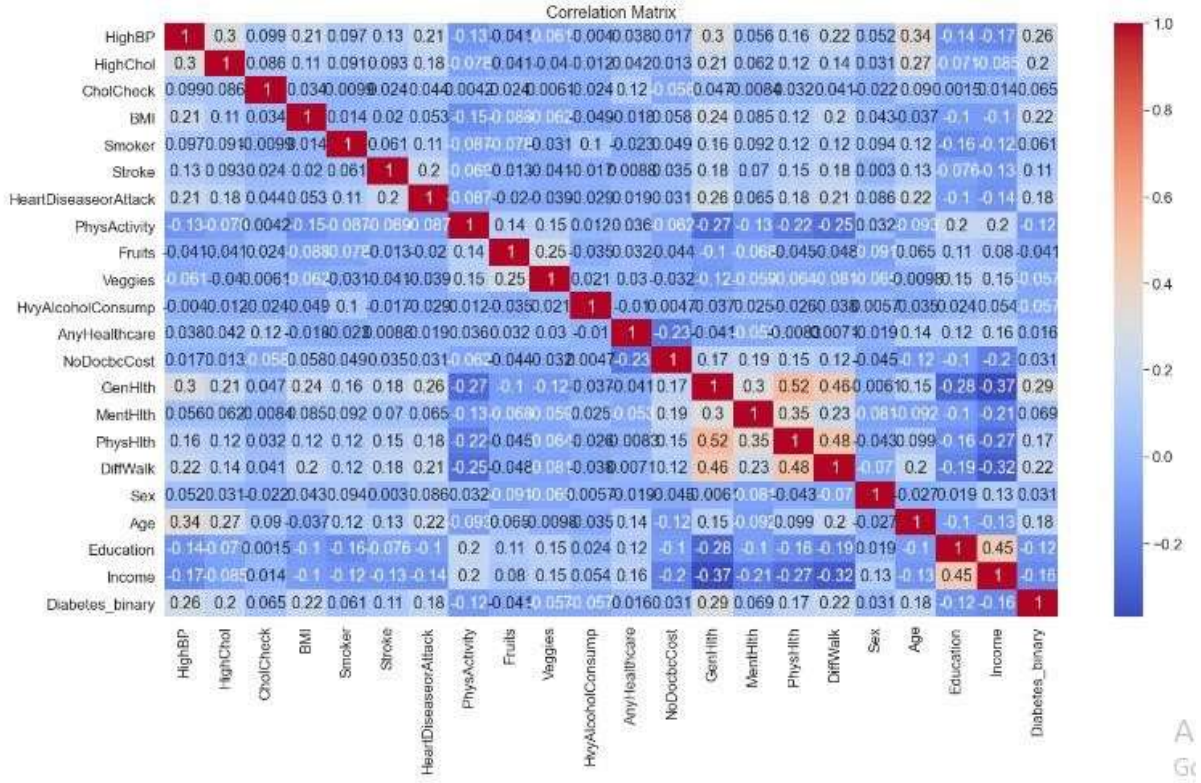

Figure 4. 10 Correlation matrix

From the above fig 4.10, the correlation matrix shows that the relationship between the indep endent features with each other and independent features with dependent feature or attributes. And also the highest values of the features are more important and the less value of features a re less important for the target class. Those more less significant attributes are removed.

4.3 Implementation Details

In this section, we introduced how the experiment is conducted. The first task done after data pre-processes and class balancing is relevant feature selection to identify the best prediction model and model expnaninaibility. To construct a prediction model for diabetes based on risk factors and associated diseases, we use relevant features and the best appropriate ensemble machine learning algorithm. In this research, we conducted six experiments using ensemble machine learning algorithms with relevant features selected by feature selection.

4.3.1 Dataset splitting

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. In this study, first, we split the 70692 total instances of the dataset into a train, validation and test size split proportion, such as a 70%, 15%, 15% to 80%, 10%, 10% test split size proportion, with the help of a ratio-based random splitting technique. Because of the lower performance in the train test split with a proportion of 70%, 15%, 15%, we used only a proportion of 80%, 10%, 10% for the train test split to develop the proposed model. In 80%, 10%, 10% proportion, out of 70692 instances, 56554 (80%) instances are training datasets that are used for building a prediction model, and 7070 (10%) instances are test datasets that are used to test the performance of the predictive model of the diabetes and 7070 (10%) for validation.

4.3.2 Hyper parameter Tuning

In the process of machine learning, the performance of the algorithm highly depends on the selection of parameters, which has always been a crucial step [80]. Grid search is a method of hyper-parameter tuning that builds and evaluates a model methodologically for each

combination of algorithmic parameters supplied in a grid [81]. Here, we used grid-search with Grid Search cross validation (CV) to select tuning parameters for an ensemble machinelearning algorithm. The tuning parameters for each ensemble machine learning algorithm are identified by grid search for all experiments (see Table 4.1 below).

Table 4. 1 Hyper parameter Tuning

No	Algorithms	parametres	valeus
1	Random forest	n_estimators	1000
		max_depth	10
		random_state	42
		criterion	gini
2	AdaBoost	n_estimators	500
		learning_rate	1.5
		base estimator	3
	Extreme	n_estimators	500
3	gradient boosting	max_depth	200
		learning_rate	0.5
		Max_depth	10
4	Bagging decision tree	n_estimators	500
		max_samples	0.5
		Random state	20
5	Catboost	Learning_rate	0.001
		iterations	3000
		depth	10
		Random strangth	1.2
6	Extra tree	n_estimators	500
		max_depth	5
		Minimum sample split	3

4.4 Feature Selection Method and Result

To get the best features two experiments were conducted. Those are:

Experiment #1 Using Filter Method

We performed filter feature selection methods such as mutual information and chi-square test techniques using the selected best K features. In this experiment, we set parameters such as score_func=”mutual_info_classifiers” for mutual information that takes the pair of dependent and independent features and returns an array of scores and values, score_func=”chi2” for chi-square feature selection, and selected features = 18 and Similarly, in the chisquare test techniques, instead of mutual_info_classifiers. We used “chi2” for score_func, which is score_func=”chi2” and selected features = 18. We have get and select the same relevant features with equal accuracy in both mutual information and chi-square test techniques from the filter feature selection techniques. Table 4.2 shows the experimental results

Experiment #2 Wrapper Method

Wrapper feature selection approach uses a classifier to select the important feature for the model. We used recursive feature elimination with a Random Forest Classifier, gradient boosting classifier, AdaBoost classifier, extra tree classifier, cat boost classifier and bagging classifier to select the number of relevant features based on the accuracy of the classifier.

Based on the experiments done, the classifier selected 18 features from the total features with different accuracy. The wrapper technique achieves better overall performance than filter-out techniques because the feature selection within the wrapper technique is optimized for the classification algorithm to be used and measures the usefulness of a subset of features by training a model on them [82]. But, wrapper methods are too expensive for a large dimensional dataset in terms of computational complexity and much slower than filters in finding sufficiently good subsets because they depend on the resource demands of the modelling algorithm [83].

Table 4. 2 Experimental results of feature selection

NO	**Using filter method**		**Using wrapper method**
	mutual information	**chi-square**	**RFE**
1	HighBP	HighBP	HighBP
2	HighChol	HighChol	HighChol
3	CholCheck	CholCheck	CholChek
4	BMI	BMI	BMI
5	Smoker	Smoker	Smoker
6	Stroke	Stroke	Stroke
7	HeartDisease orAttack	HeartDisease orAttack	HeartDise aseorAttk
8	PhysActivity	PhysActivity	PhysActiv ity
9	Fruits	Fruits	Fruits
10	Veggies	Veggies	Veggies
11	HvyAlcoholCons ump	HvyAlcoholConsup	HvyAlcoh olConsup
12	GenHlth	GenHlth	GenHlth
13	MentHlth	MentHlth	MentHlth
14	PhysHlth	PhysHlth	PhysHlth
15	DiffWalk	DiffWalk	DiffWalk
16	Age	Age	Age
17	Income	Income	Education
18	sex	sex	Income
Accuracy of Random forest	90.01	90.01	90.04
Accuracy of Bagging	88.85%,	88.85%,	87.76%

Accuracy of Adaboost	87.61%,	87.61%,	87.81%
Accuracy of EXGboost	88.67%,	88.67%,	88.19%
Accuracy of Cat boost	88.82%	88.82%	89.75%
Accuracy of Extratree	89.55%,	89.55%,	88.77%

Generally from the above, experiment #1 **Filter Method** and experiment#2 **Wrapper Method**, We conclude both in mutual information and chi-square selects the same features in different accuracy for all classifiers. And also when we compared the accuracy of the classifiers, random forest classifier is the highest value in wrapper method with an accuracy of 90.04%. So, we selected those attributes (features) for model building.

Finally, the data set is saved in CSV format for experimentation using classification algorithms. Table below 4.3 shows sample data used for the experiment in CSV format.

Table 4. 3 Sample dataset used for experiment

Out[36]:

	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	GenHlth
0	0	1	1	35	1	0	0	0	0	1	1	2
1	1	0	1	31	1	0	0	1	0	1	0	2
2	0	0	1	27	0	0	0	1	1	1	0	2
3	0	1	1	26	0	0	0	1	0	1	0	1
4	0	0	1	21	0	0	0	1	1	1	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...
70687	0	0	1	32	1	0	0	1	1	1	0	2
70688	0	0	1	23	0	0	0	1	1	1	0	2
70689	0	0	1	29	0	0	0	1	1	1	0	2
70690	0	1	1	24	1	0	1	1	1	1	0	5
70691	1	1	1	42	1	0	0	1	1	1	1	1

70692 rows × 19 columns

MentHlth	PhysHlth	DiffWalk	Age	Education	Income	Diabetes_binary
0	1	0	4	5	8	0
30	0	0	7	6	1	0
2	0	0	8	6	6	0
0	0	0	11	6	7	0
0	0	0	6	6	7	0
...	...	...	...	...	...	...
30	4	0	5	4	4	0
0	0	0	12	6	8	0
0	0	0	6	6	8	0
24	28	0	10	5	8	0
30	0	1	5	3	3	1

4.5 Proposed model result and analysis

In this phase, we applied and tested Random Forest, Ada Boosting, XGBoost, Cat Boost, extra trees and bagging decision tree to predict diabetes. To build a predictive model, 70692 instances and 18 features were used. To evaluate the performance of the model, 5-fold crossvalidation is used due to its relatively low bias and variations. This means that the data is randomly divided into five equal parts. That is four folds for training and one fold for validation.

Experiment# 1: Random Forest

In this experiment, the Random Forest is built from an ensemble of decision trees and is usually trained with the bagging method. We used 18 relevant feature as input chosen from the above experiment to develop a predictive model using a random forest algorithm and obtained an accuracy of 90.16%, precision of 91.5%, recall of 88.49%% and f1_score
89.97% and AUC of 96%. Look the confusion matrix fig 4.11 below.

Accuracy: 0.9015558698727015
Precision: 0.9150556531927357
Recall: 0.884985835694051 F1
Score: 0.8997695852534561

```
Confusion Matrix:
[[3250  290]
 [ 406 3124]]
```

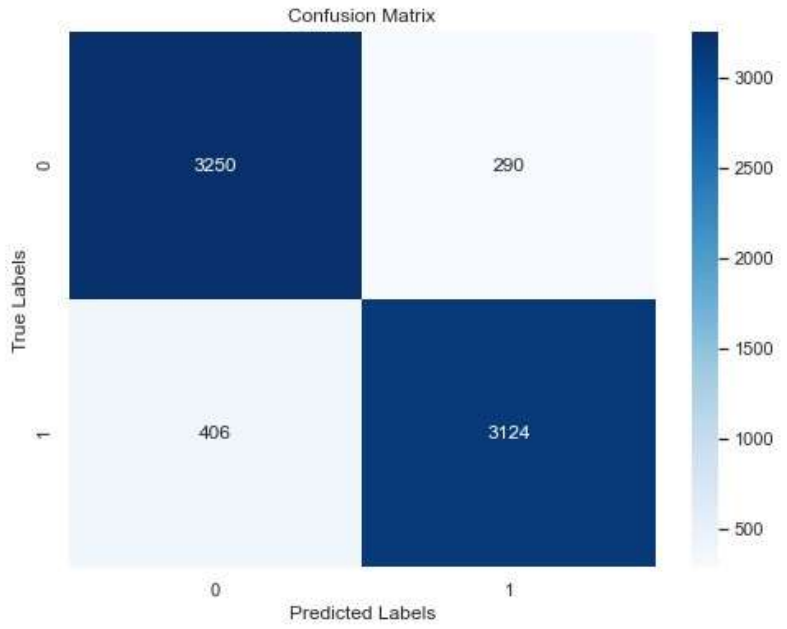

Figure 4. 11 Confusion matrix of Random forest model

Table 4. 4 Confusion matrix of Random Forest

		Predicted class	
		Non diabetes	Pridiabetes or diabetes
Actual class	Non diabetes	3250	290
	Pridiabetes or diabetes	406	3124

Based on table 4.4 above, in Random forest there are 7070 samples of data were used for test the model. Among the test sample data 6374 were correctly predicts and the remaining 696 samples of data were incorrectly predicts. Out of 3540 sample 3350 instances are correctly predicts as having non diabetes (true positive), whereas 290 instances is incorrectly predicts or false positive as having diabetes. 3124 instances are correctly predicts as having diabetes or true positive, whereas 406 instances are incorrectly predicts false positive non diabetes from the total of 3530 diabetes. All the correctly instances are predict with an accuracy of 90.16%, and the incorrectly instances are predict with an accuracy of 9.84%.

Experiment# 2: bagging decision tree

We build a predictive model using bagging decision tree ensemble classifier using 18 important features to evaluate the performance of the model. As illustrated in figure 4.12 below, this model has ROC curve (AUC) of 96%, a performance of 89.54% accuracy, precision of 89.97%, recall of 90.31% and f1_score 89.63%.

Accuracy: 0.8978783592644979
Precision: 0.9086146682188592
Recall: 0.8844192634560907
F1 Score: 0.8963537180591444

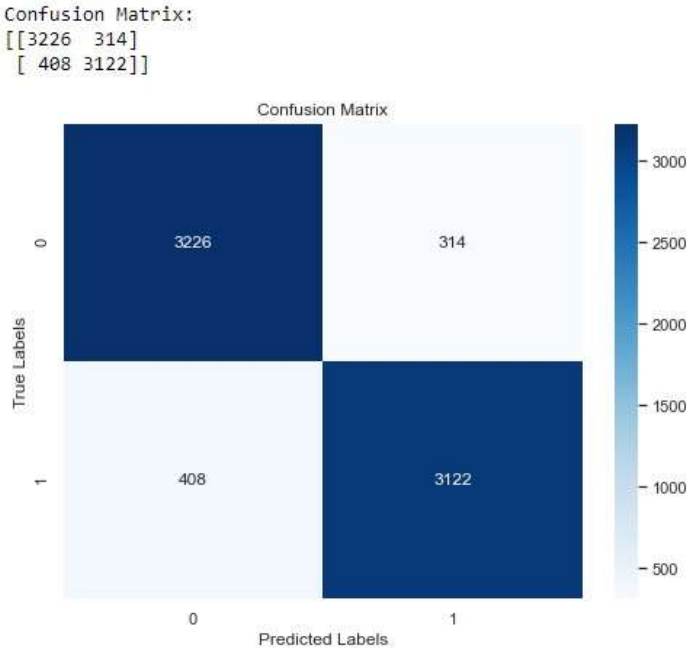

Figure 4.12 confusion matrix of bagging decision tree model

Table 4. 5 confusion matrix of bagging decision tree

		Predicted class	
		No diabetes	Pridiabetes or diabetes
Actual class	No diabetes	3226	314
	Pridiabetes or diabetes	408	3122

Based on table 4.5 here above, in Bagging there are 7070 samples of data were used for test the model. Among the test sample data 6348 were correctly predicts and the remaining 722 samples

of data were incorrectly predicts. Out of 3540 sample 3226 instances are correctly predicts as having non diabetes (true positive), whereas 314 instances is incorrectly predicts or false positive as having diabetes. 3122 instances are correctly predicts as having diabetes or true positive, whereas 408 instances are incorrectly predicts false positive non diabetes from the total of 3530 diabetes. All the correctly instances are predict with an accuracy of 89.78%, and the incorrectly instances are predict with an accuracy of 10.32%.

Experiment# 3: AdaBoosting

We build a predictive model using an AdaBoosting ensemble classifier using 18 important features to evaluate the performance of the model. As illustrated in figure 4.13 below, this model has a performance of ROC 94%, 88.26% of accuracy, precision of 89.12%, recall of 87.08% and f1_score 88.09%.

Accuracy: 0.8824611032531825
Precision: 0.8912728327051319
Recall: 0.8708215297450425
F1 Score: 0.8809284997850695

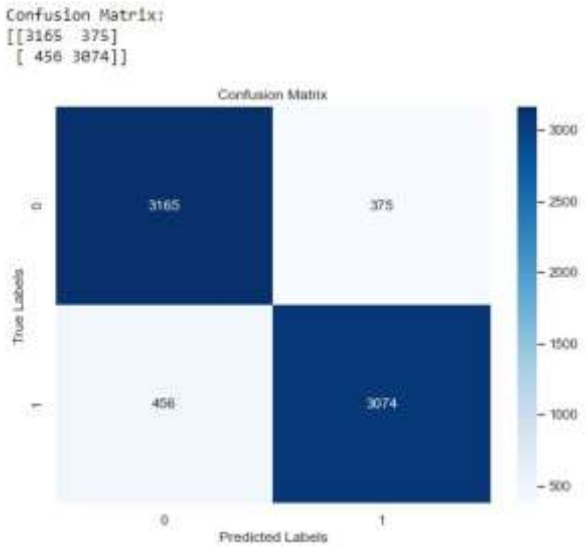

Figure 4. 12 Confusion matrix of AdaBoosting model

Table 4. 6 Confusion matrix of AdaBoosting model

		Predicted class	
		No diabetes	Pridiabetes or diabetes
Actual class	No diabetes	3165	375
	Pridiabetes or diabetes	456	3074

Based on table 4.6 here above, in AdaBoosting there are 7070 samples of data were used for test the model. Among the test sample data 6239 were correctly predicts and the remaining 831 samples of data were incorrectly predicts. Out of 3540 sample 3165 instances are correctly classified as having non diabetes (true positive), whereas 375 instances is incorrectly predicts or false positive as having diabetes. 3074 instances are correctly predicts as having diabetes or true positive, whereas 456 instances are incorrectly predicts false positive non diabetes from the total of 3530 diabetes. All the correctly instances are predict with an accuracy of 88.24%, and the incorrectly instances are predict with an accuracy of
11.76%.

Experiment# 4: Extreme gradient boosting

Similarly, we have also built a predictive model using an extreme gradient boosting (XGBoost) classifier, by using 18 relevant features. As illustrated in figure 4.14 below, this algorithm has a performance 95% of ROC, 89.53% accuracy, precision of 90.15%, recall of 88.72% and f1_score 89.43%.

Accuracy: 0.8953323903818954
Precision: 0.9015544041450777
Recall: 0.8872521246458923
F1 Score: 0.8943460879497429

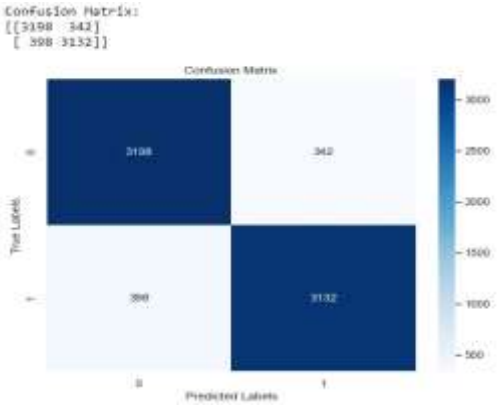

Figure 4. 13 Confusion matrix of XGBoost

Table 4. 7 Confusion matrix of XGBoost model

		Predicted class	
		No diabetes	Pridiabetes or diabetes
Actual class	No diabetes	3198	342
	Pridiabetes or diabetes	398	3132

Based on table 4.7 above, in XGBoost there are 7070 samples of data were used for test the model. Among the test sample data 6330 were correctly classified and the remaining 740 samples of data were incorrectly classified. Out of 3540 sample 3198 instances are correctly predicts as having non diabetes (true positive), whereas 342 instances is incorrectly predicts or false positive as having diabetes. 3132 instances are correctly predicts as having diabetes or true positive, whereas 398 instances are incorrectly predicts false positive non diabetes from the total of 3530 diabetes. All the correctly instances are predict with an accuracy of 89.53%, and the incorrectly instances are predict with an accuracy of 10.47%.

Experiment# 5: Ensemble Cat Boost

Similarly, we have also built a predictive model using a Cat Boost classifier by using relevant features that are selected for building a model by using sequential step-backward feature selection. In this experiment, as you can see figure 4.15 below, this model has the best performance of 95% ROC , accuracy of 89.53%, precision of 90.62%, recall of 88.15% and f1_score 89.37%.

Accuracy: 0.8953323903818954
Precision: 0.9062317996505533
Recall: 0.8815864022662889
F1 Score: 0.893739230327398

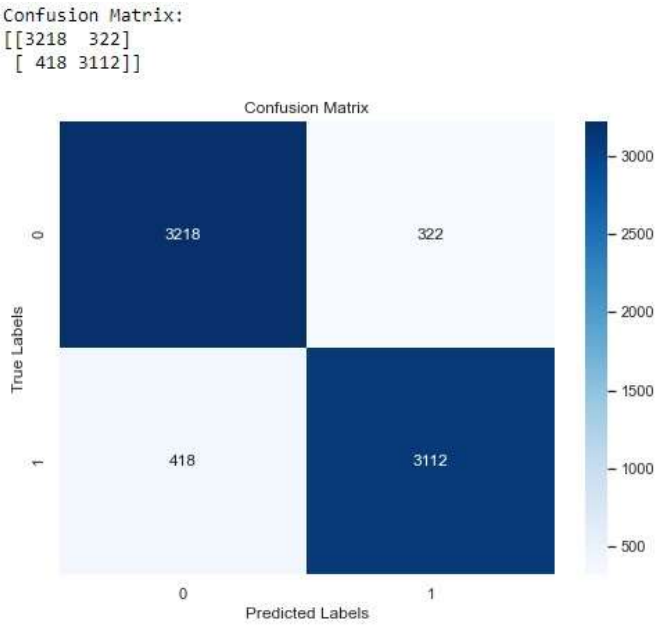

Figure 4. 14 Confusion matrix of Cat Boost

Table 4. 8 Confusion matrix of Cat Boost model

		Predicted class	
		No diabetes	Pridiabetes or diabetes
Actual class	No diabetes	3218	322
	Pridiabetes or diabetes	418	3112

Based on table 4.8 here above, in ensemble cat boost there are 7070 samples of data were used for test the model. Among the test sample data 6330 will correctly predict and the remaining 696 samples of data also incorrectly predicts. Out of 3540 sample 3218 instances are correctly classified as having non diabetes (true positive), whereas 322 instances is incorrectly predicts or false positive as having diabetes. 3112 instances are correctly predicts as having diabetes or true positive, whereas 418 instances are incorrectly predicts false positive non diabetes from the total of 3530 diabetes. All the correctly instances are predict with an accuracy of 89.5%, and the incorrectly instances are predict with an accuracy of
10.5%.

Experiment# 6: Extra tree

We have built a predictive model using Extra tree by using relevant features. In this experiment, as you can see figure 4.16 below, this model has its own performance ROC is 97% followed by accuracy of 89.85%, precision of 91.37%, recall of 88.21% and f1_score 89.76%.

Accuracy: 0.8995756718528995
Precision: 0.9137323943661971
Recall: 0.8821529745042493
F1 Score: 0.8976650331507641

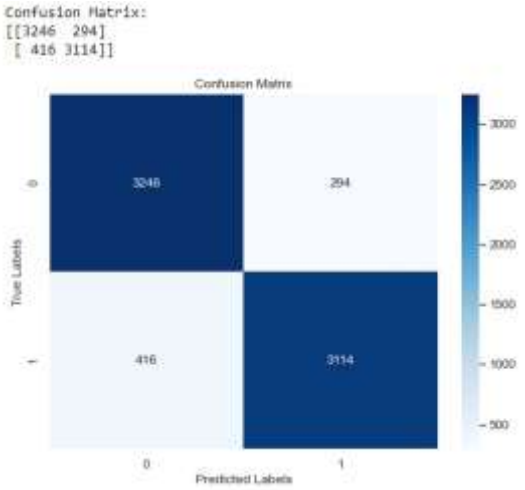

Figure 4. 15 Confusion matrix of Extra tree model

Table 4. 9 Confusion matrix of Extra tree

		Predicted class	
		No diabetes	Pridiabetes or diabetes
Actual class	No diabetes	3246	294
	Pridiabetes or diabetes	416	3114

Based on table 4.9 above, in extra tree there are 7070 samples of data were used for test the model. Among the test sample data 6360 were correctly predicts and the remaining 710 samples of data were incorrectly classified. Out of 3540 sample 3246 instances are correctly predicts as having non diabetes (true positive), whereas 294 instances is incorrectly predicts or false positive as having diabetes. 3114 instances are correctly predicts as having diabetes or true positive, whereas 416 instances are incorrectly predicts false positive non diabetes from the total of 3530 diabetes. All the correctly instances are predict with an accuracy of

89.95%, and the incorrectly instances are predict with an accuracy of 10.05%.

4.6 Model Comparison

To build a predictive model for diabetes prediction based on risk factors and associated diseases, we carried out a total of six experiments using the features that were chosen as input. We performed six experiments, experiment #1 to experiment #6, using the ensemble machine learning techniques such as random forest, bagging, AdaBoost, extreme gradient boosting, CatBoost and extra tree to create a prediction model. As a result, we compared the performance of algorithms to predict the diabetes using ensemble machine learning. We used overall accuracy and the ROC curve as an evaluation for predictive model comparison. According to the overall accuracy, the classification algorithm that registered the highest performance is used as the best model for predicting the diabetes and non-diabetes.

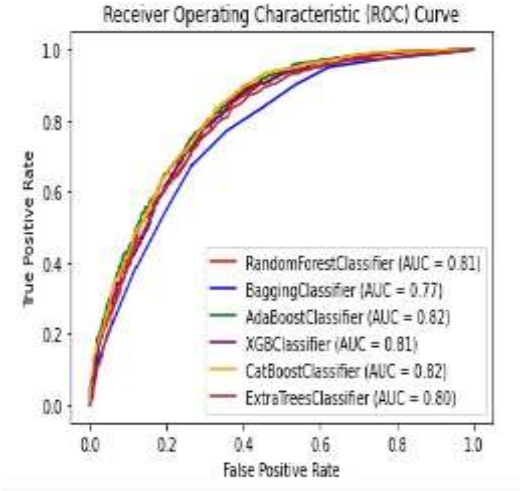

Figure 4. 16 ROC-AUC before HPO

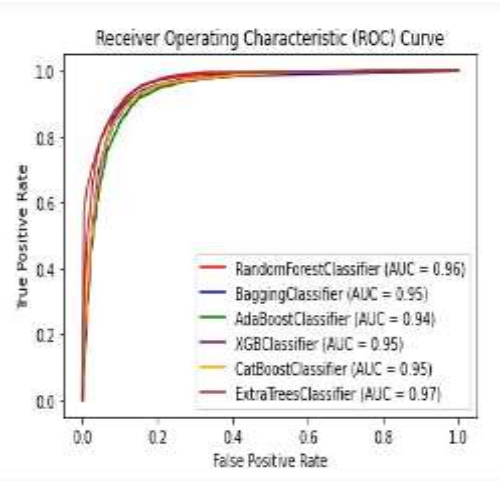

Figure 4. 17 ROC -AUC after HPO

As the figure 4.16 indicates the models of ROC-AUC values before hyper parameter optimization are different and smaller than that of ROC -AUC after HPO with their corresponding models See Figure 4.17.

Table 4. 10 The overall model performance of an experiment

Evaluation metrics	Random forest	Bagging decision tree	AdaBoost	XGBoost	CatBoost	Extra tree
Before HPO						
Accuracy	74%	70%	68%	75%	76%	73%
Precision	72%	71%	68%	73%	73%	71%
Recall	78%	69%	67%	79%	80%	77%
F1_score	75%	70%	67%	76%	77%	74%
ROC-AUC	81%	77%	82%	81%	82%	80%
After HPO						
Accuracy	90.16%	88.97%	87.87%	88.81%	88.94%	89.86%
Precision	88.85%	88.68%	87.88%	87.48%	87.46%	88.51%
Recall	91.82%	89.29%	87.81%	90.53%	90.87%	91.56%
F1_score	90.31%	90.31%	90.31%	90.13%	89.13%	90.13%
ROC-AUC	96%	95%	94%	95%	95%	97%
Rank	1th	3rd	6th	5th	4th	2nd

From the table 4.10 above shows that the two experiments of model training performance. Those are model training before hyper parameter optimization (HPO) and model training after hyper parameter optimization (HPO). In both case the researcher evaluates all the models using evaluation matrix. As compare the two experiments model training before hyper parameter optimization and model training after hyper parameter optimization the performance of the models are difference values. The performance the models before hyper parameter optimization are smaller than that of after hyper parameter optimization. This result indicates that hyper parameter optimization is essential to enhance the models performance. So the research is conducted by using after HPO.

Table 4.10 shows that the Random forest had the best-predicted performance, with an accuracy of 90.16%. While the AdaBoost model had the worst performance, with an accuracy of 87.87%. Random forest registered the highest accuracy of 90.16% and 96% for the ROC curve, and the 2nd mode having the highest accuracy is Extra tree with an accuracy of 89.86% and a ROC curve of 97%. Bagging decision tree is the 3rd mode with accuracy of 88.97% and 95%

ROC curve. CatBoost mode is the 4th with the accuracy of 88.94% and 95% ROC curve. The 5th mode also XGBoost with an accuracy of 88.81% and 95% ROC curve. AdaBoost is the last with accuracy of 87.87% and 94% ROC curve. So, Random forest is selected as the best ensemble machine learning as compared to another ensemble machine learning according to the results of Table 4.10.

4.7 Model explainability using lime

The researcher uses an instance to make and explain lime (local interpretable model explanations) for predictions made by random forest model see in figure 4.19 below.

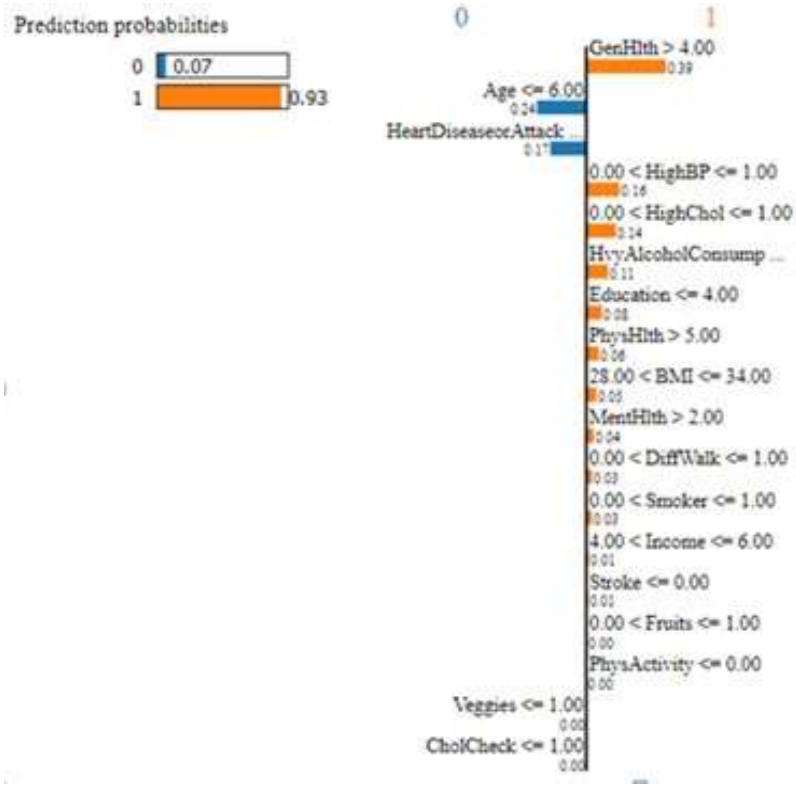

Figure 4. 18 prediction probability of diabetes and non diabetes class

As the figure 4.19 shows above the diabetes probability of an instance is 93% which is indicated by red colour and non diabetes probability also 7% indicated by blue colour.

Table 4. 11 Feature and value of an instance

Feature	Value
GenHlth	5.00
Age	4.00
HeartDiseaseorAttack	0.00
HighBP	1.00
HighChol	1.00
HvyAlcoholConsump	0.00
Education	2.00
PhysHlth	10.00
BMI	33.00
MentHlth	5.00
DiffWalk	1.00
Smoker	1.00
Income	6.00
Stroke	0.00
Fruits	1.00
PhysActivity	0.00
Veggies	0.00
CholCheck	1.00

Table 4.10 indicates a certain instance features has its own value to be diabetic or non diabetic. As the table shows when GenHlth =5.00, Age = 4.00, Heart Diseases or Attack =0.00, HighBP =1.00, Highchol =1.00, HvyAlcoholconsupm =0.00, Education =2.00, physHlth= 10, BMI = 33.00, MentHlth = 5.00, DiffWalk =1.00, smoker = 1.00, Income = 6.00, stroke = 0.00, Fruits = 1.00, PhysActivity = 0.00, Veggies = 0.00 and CholCheck =1.00, the probability of diabetes is 93% and 7% of non diabetes.

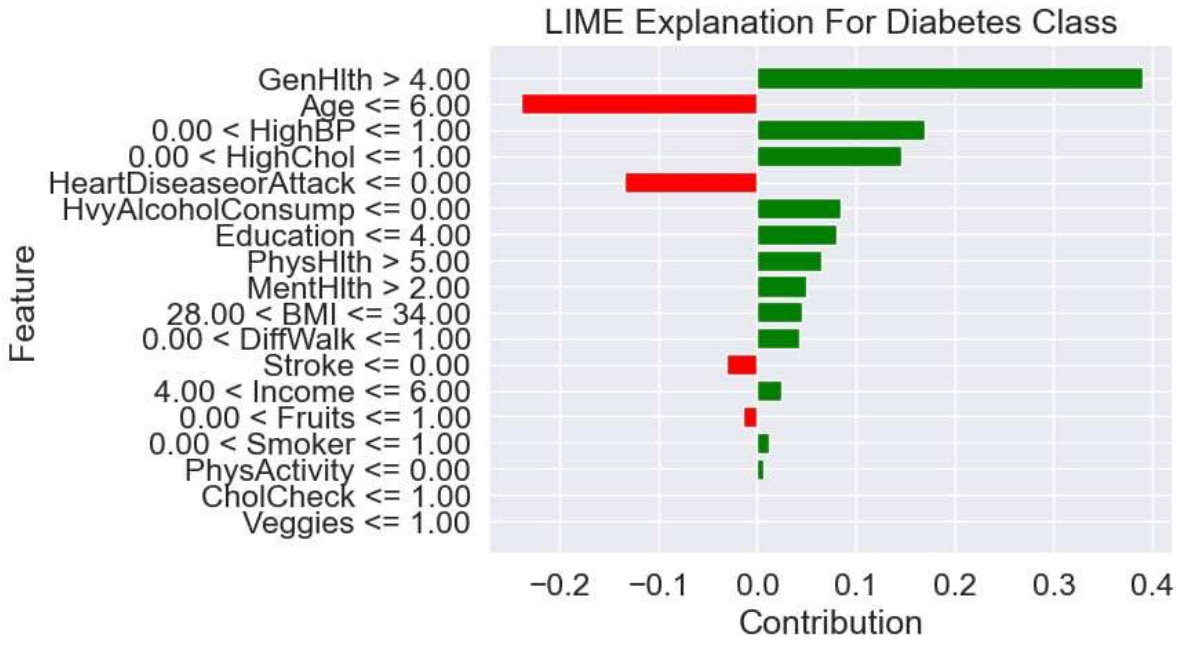

Figure 4. 19 lime explaination for diabetes class

As shown the above figure 4.20 when the features value of diabetes class of GenHlth>4.00, Age<= 6, 0.00< HighBP <=1.00, 0.00< Highchol <=1.00, Heart Diseases or Attack <=0.00, HvyAlcoholconsupm <=0.00, Education <=4.00, physHlth>= 5, MentHlth >= 2, 28.00 < BMI <= 34.00, 0.00 < DiffWalk <=1.00, stroke <= 0.00, 4.00<=Income <= 6.00, 0.00< Fruits <= 1.00, 0.00 < stroker <= 1.00, PhysActivity <= 0.00, CholCheck <=1.00 and Veggies <= 1.00. The figure of green colour of the features value have positive contribution for diabetes class of an instance and where as the red colour of the figure of the features value shows the negative contribution of an instance of diabetes class.

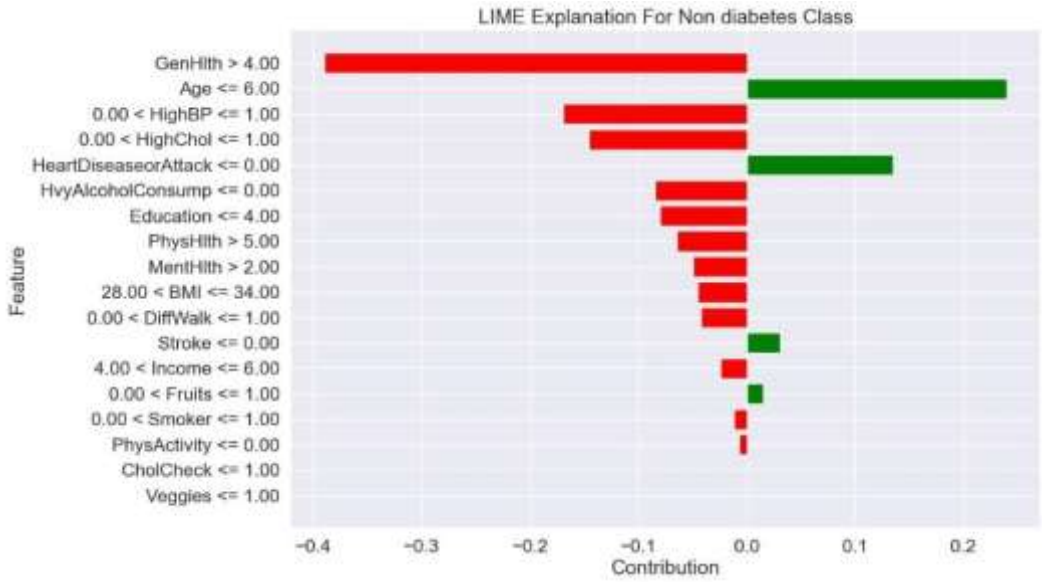

Figure 4. 20 lime explaination of for non diabetes class

As shown the above figure 4.21 when the features value of diabetes class of GenHlth>4.00, Age<= 6, 0.00< HighBP <=1.00, 0.00< Highchol <=1.00, Heart Diseases or Attack <=0.00, HvyAlcoholconsupm <=0.00, Education <=4.00, physHlth>= 5, MentHlth >= 2, 28.00 < BMI <= 34.00, 0.00 < DiffWalk <=1.00, stroke <= 0.00, 4.00<=Income <= 6.00, 0.00< Fruits <= 1.00, 0.00 < stroker <= 1.00, PhysActivity <= 0.00, CholCheck <=1.00 and Veggies <= 1.00. The figure of green colour of the features value have positive contribution for non diabetes class of an instance and where as the red colour of the figure of the features value shows the negative contribution of an instance of non diabetes class.

4.8 Risk factor analysis

In this study, we have generated all the most determinant factors by the best-performing algorithms with their feature importance (see table 4.12). As we have seen from the above six experiments, Random forest performs the best with different evaluation metrics, and due to its

better performance, we have used Random forest algorithms with feature importance techniques to determine the most determinant risk factors. The feature with the highest value is the most determinant factor, and the feature with the lowest value is the least determinant factor.

Table 4. 12 Feature importance rank

NO	features	values	NO	features	values
1	GenHlth	0.188	10	MentHlth	0.032
2	BMI	0.175	11	Smoker	0.018
3	HighBP	0.141	12	PhysActivity	0.018
4	Age	0.116	13	HeartDiseaseorAttack	0.018
5	Income	0.062	14	Fruits	0.017
6	HighChol	0.053	15	Veggies	0.013
7	PhysHlth	0.053	16	Stroke	0.007
8	DiffWalk	0.042	17	HvyAlcoholConsump	0.006
9	Education	0.037	18	CholCheck	0.004

From the above table you observe that general health(GenHlth) is the highest determinant from risk factors and cholesterol check is the least and also in associated diseases heart diseases is the higher and stroke is lower determinant for predicting diabetes.

4.9 Result and Discussion

As discussed in the previous sections, 70692 datasets with 18 features participated in constructing binary class predictive model for diabetes. The proposed model achieves accuracy of 90.16% performance and ROC of 96%.

In the beginning, this study has three research questions to answer. Let us discuss how these questions have been answered with this study.

- The first research question of this study is "What are the most determinant attributes for determining diabetes?"

To answer this question, feature importance techniques were used from all the features that we used to develop the predictive model, and the potential (top) attributes used for determining diabetes are most determined based on the best performed algorithms.

- The second research question was “How to select an ensemble machine learning algorithm for diabetes predictive model?’

To answer this question, six experiments for six ensemble machine learning algorithms namely random forest, bagging decision tree, extreme gradient boosting, Cat Boost, AdaBoost and extra decision tree are conducted. As the experiments showed that Random forest is the best ensemble machine learning algorithm to develop the predictive model for diabetes based on risk factors and associated diseases. Because, random forest algorithm achieves the best performance with an accuracy of 90.16% from other algorithms

- The third question of this study was “To what extent does the proposed predictive model accurately identify the diabetes?”

To answer the research question, the researcher compares the algorithms based on accuracy, precision, recall, F1-score and ROC values. So,

1. The proposed Random Forest algorithm achieves an accuracy of 90.16%, indicating that it correctly predicts diabetes cases in approximately 90.16% of instances. It also demonstrates a precision of 88.85%, meaning that 88.85% of the predicted diabetes cases are correct. The recall value of 91.82% suggests that the model identifies 91.82% of the actual diabetes cases in the test set. The F1-score of 90.31% combines precision and recall into a single value, considering both metrics. Additionally, the ROC value of 96% indicates that the model has a good ability to distinguish between diabetes and non-diabetes cases.
2. The Bagging Decision Tree algorithm achieves an accuracy of 88.97%. It shows a precision of 88.68% and a recall of 89.29% for diabetes prediction. The F1-score is 90.31%, and the ROC value is 95%.
3. The AdaBoost algorithm achieves an accuracy of 87.87%, with a precision of 87.88% and a recall of 87.81%. The F1-score is 90.31%, and the ROC value is 94%.
4. The XGBoost algorithm achieves an accuracy of 88.81%, with a precision of 87.48% and a recall of 90.53%. The F1-score is 90.13%, and the ROC value is 95%.
5. The Cat Boost algorithm achieves an accuracy of 88.94%, with a precision of 87.46% and a recall of 90.87%. The F1-score is 89.13%, and the ROC value is 95%.
6. The Extra Tree algorithm achieves an accuracy of 89.86%, with a precision of 88.51% and a recall of 91.56%. The F1-score is 90.13%, and the ROC value is 97%.

Based on these results, we can conclude that the proposed predictive models, particularly Random Forest, XGBoost, and Cat Boost, demonstrate relatively good performance in accurately identifying diabetes and the rests are relatively poor performance.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

Health is an essential thing for all living things especially for human beings. A wider range of people, including adolescents, adults, and kids, suffer from diabetes, which is a public health problem. According to a WHO report, most people affected and died in diabetes in the world increased in year to year.

To handle this problem we conducted this study. This study aimed to develop a predictive model for diabetes based on risk factors and associated diseases by using ensemble machine learning algorithms. The data source for this research is the CDC, which was collected by BRFSS. The dataset was 253680 and there is imbalanced. After applying the data pre processing tasks and class balance using random forest under sampling majority class there is 70692 instances were used for the model. The attribute was reduced to 18 from their original 21features, by using feature selection techniques wrapper (recursive feature elimination).

The proposed model was constructed using ensemble machine learning algorithms namely random forest, bagging decision tree, extreme gradient boosting, Cat Boost, extra tree and AdaBoost algorithms. To conduct this study we have done a total of six experiments. The performances of the models are evaluated using objective (such as confusion matrix, accuracy, precision, recall, f1_score, and ROC) evaluation metrics. In this study, the best ensemble machine learning algorithm is identified using objective-based evaluation metrics of the developed predictive model. Then, the best algorithm that predicts diabetes is constructed by random forest algorithms with 18 relevant features and has 90.16% of accuracy. After building a prediction model for diabetes, the researcher uses random forest algorithms with model explainability and identified the most determinant factors with feature importance technique.

Generally, in this study we identified the best appropriate ensemble machine learning algorithm to build a prediction model for diabetes based on risk factors and associated diseases. We identified the most determinant factors of the diabetes with feature importance techniques. The

study used only CDC data collected by The Behavioural Risk Factor Surveillance System (BRFSS) annually.

5.2 Strengths and Limitations of the Study

Concerning the strength of the study, although the intended classification accuracy was not achieved binary-class prediction of diabetes with good predictive performance accuracy, the study result ensures the potential capability of ensemble machine learning techniques in the field of health care. Further, since the objectives of this research were achieved, we can conclude that, overall the study was successful. Basically, in ensemble learning techniques, the intention is to build a predictive model with an accuracy of 90.16% However, the main limitation for the researcher is that the quality of the dataset is poor due to small features of associated diseases.

5.3 Contributions

The contributions of this research work are:

- Construct binary-class diabetes prediction model Identify risk factors and associated diseases for diabetes.
- Make model explainability.

5.4 Recommendation

The researcher forwards the following recommendations for further investigation based on the research's findings: To enhance the performance of the current study, there is a need to use other associated diseases, other class balanced techniques and use other algorithms like deep learning.

REFERENCE

[1] D. Of and D. Mellitus, "Diagnosis and classification of diabetes mellitus," *Diabetes Care*, vol. 37, no. SUPPL.1, pp. 81–90, 2014, doi: 10.2337/dc14S081.

[2] Q. Saihood and E. Sonuç, "A practical framework for early detection of diabetes using ensemble machine learning models," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 31, no. 4, pp. 722–738, 2023, doi: 10.55730/13000632.4013.

[3] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," *Heal. Inf. Sci. Syst.*, vol. 8, no. 1, pp. 1–14, 2020, doi: 10.1007/s13755-019-0095-z.

[4] D. Doreswamy and M. Nigus, "Feature Selection Methods for Household Food Insecurity Classification," *2020 Int. Conf. Comput. Sci. Eng. Appl. ICCSEA 2020*, 2020, doi: 10.1109/ICCSEA49143.2020.9132945.

[5] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[6] P. Pintelas and I. E. Livieris, "Special issue on ensemble learning and applications," *Algorithms*, vol. 13, no. 6, 2020, doi: 10.3390/A13060140.

[7] J. Thomas, A. Joseph, I. Johnson, and J. Thomas, "Machine Learning Approach For Diabetes Prediction," *Int. J. Inf. Syst. Comput. Sci.*, vol. 8, no. 2, pp. 55–58, 2019, doi: 10.30534/ijiscs/2019/13822019.

[8] M. Hasan *et al.*, "Ensemble machine learning-based recommendation system for effective prediction of suitable agricultural crop cultivation," *Front. Plant Sci.*, vol. 14, no. August, pp. 1–18, 2023, doi: 10.3389/fpls.2023.1234555.

[9] J. Paul and A. R. Criado, "The art of writing literature review: What do we know and what do we need to know?," *Int. Bus. Rev.*, vol. 29, no. 4, 2020, doi: 10.1016/j.ibusrev.2020.101717.

[10] S. Chatterjee, K. Khunti, and M. J. Davies, "Type 2 diabetes," *Lancet*, vol. 389, no. 10085, pp. 2239–2251, 2017, doi: 10.1016/S01406736(17)30058-2.

[11] K. Baptiste-Roberts *et al.*, “Family history of diabetes, awareness of risk factors, and health behaviors among African Americans,” *Am. J. Public Health*, vol. 97, no. 5, pp. 907–912, 2007, doi: 10.2105/AJPH.2005.077032.

[12] P. S. Kott, “3 B 3 B 0,” *Methods*, pp. 2241–2252, 2003.

[13] T. G. Dietterich, “<10.1.1.34.4718.Pdf>,” *Int. Work. Mult. Classif. Syst.*, pp. 1–15, 2000, [Online]. Available: http://www.cs.orst.edu/~tgd

[14] L. Rokach, “Ensemble-based classifiers,” *Artif. Intell. Rev.*, vol. 33, no. 1–2, pp. 1–39, 2010, doi: 10.1007/s10462-009-9124-7.

[15] I. Partalas, G. Tsoumakas, and I. Vlahavas, “A Taxonomy and Short Review of Ensemble Selection,” *...Ensemble Methods ...*, pp. 41–46, 2008, [Online]. Available: http://lpis.csd.auth.gr/publications/partalasecai08.pdf%5Cnhttp://lpis.csd.auth.gr/publications/partalas08-suema.pdf

[16] V. A. Boateng and B. Yang, “A Global Modeling Pruning Ensemble Stacking With Deep Learning and Neural Network Meta-Learner for Passenger Train Delay Prediction,” *IEEE Access*, vol. 11, no. May, pp. 62605–62615, 2023, doi: 10.1109/ACCESS.2023.3287975.

[17] E. Lutins, “Ensemble methods in machine learning: What are they and why use them?,” *Towar. Data Sci.*, 2017.

[18] R. Polikar, “Ensemble learning,” *Ensemble Mach. Learn. Methods Appl.*, pp. 1–34, 2012.

[19] S. M. Ganie and M. B. Malik, “An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators,” *Healthc. Anal.*, vol. 2, no. January, p. 100092, 2022, doi: 10.1016/j.health.2022.100092.

[20] S. Kumari, D. Kumar, and M. Mittal, “An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier,” *Int. J. Cogn. Comput. Eng.*, vol. 2, no. January, pp. 40–46, 2021, doi: 10.1016/j.ijcce.2021.01.001.

[21] K. Abnoosian, R. Farnoosh, and M. H. Behzadi, “Prediction of diabetes disease using an ensemble of machine learning multi-classifier models,” *BMC Bioinformatics*, vol. 24, no. 1, pp. 1–24, 2023, doi: 10.1186/s12859023-05465-z.

[22] S. S. Azmi and S. Baliga, “An Overview of Boosting Decision Tree Algorithms utilizing AdaBoost and XGBoost Boosting strategies,” *Int. Res. J. Eng. Technol.*, no. May, pp. 6867–6870, 2020, [Online]. Available: www.irjet.net

[23] R. Sikora and O. H. Al-laymoun, “A Modified Stacking Ensemble Machine Learning Algorithm Using Genetic Algorithms,” *J. Int. Technol. Inf. Manag.*, vol. 23, no. 1, 2014, doi: 10.58729/1941-6679.1061.

[24] C. Weihs, D. Jannach, I. Vatolkin, and G. Rudolph, *Music data analysis: Foundations and applications*. Chapman and Hall/CRC, 2016.

[25] M. Shardlow, “An Analysis of Feature Selection Techniques,” *Univ. Manchester*, vol. 14, no. 1, pp. 1–7, 2016.

[26] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, “Feature Subset Selection Problem using Wrapper Approach in Supervised Learning,” *Int. J. Comput. Appl.*, vol. 1, no. 7, pp. 13–17, 2010, doi: 10.5120/169-295.

[27] J. Li *et al.*, “Feature selection: A data perspective,” *ACM Comput. Surv.*, vol. 50, no. 6, 2017, doi: 10.1145/3136625.

[28] M. Ramaswami and R. Bhaskaran, “A Study on Feature Selection Techniques in Educational Data Mining,” vol. 1, no. 1, pp. 7–11, 2009, [Online]. Available: http://arxiv.org/abs/0912.3924

[29] E. Sonuç, “Turkish Journal of Electrical Engineering and Computer Sciences A practical framework for early detection of diabetes using ensemble machine learning models,” vol. 31, no. 4, 2023, doi: 10.55730/1300-0632.4013.

[30] R. Krishnamoorthi *et al.*, “Research Article A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques,” vol. 2022, 2022.

[31] M. Alehegn, R. R. Joshi, and P. Mulay, “Diabetes analysis and prediction using random forest, KNN, Naïve Bayes, and J48: An ensemble approach,” *Int. J. Sci. Technol. Res.*, vol. 8, no. 9, pp. 1346–1354, 2019.

[32] I. J. Kakoly, M. R. Hoque, and N. Hasan, “Data-Driven Diabetes Risk Factor Prediction Using Machine Learning Algorithms with Feature Selection Technique,” *Sustainability*, vol. 15, no. 6, p. 4930, 2023, doi: 10.3390/su15064930.

[33] S. M. Ganie and M. B. Malik, “An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators,” *Healthc. Anal.*, vol. 2, no. August, p. 100092, 2022, doi: 10.1016/j.health.2022.100092.

[34] Ö. ÇELİK, “A Research on Machine Learning Methods and Its Applications,” *J. Educ. Technol. Online Learn.*, vol. 1, no. 3, pp. 25–40, 2018, doi: 10.31681/jetol.457046.

[35] T. Choudhury, A. Katal, J.-S. Um, A. Rana, and M. Al-Akaidi, *Telemedicine: The Computer Transformation of Healthcare*, no. January. 2022. doi: 10.1007/978-3-030-99457-0.

[36] M. Z. I. Chowdhury *et al.*, “Prediction of hypertension using traditional regression and machine learning models: A systematic review and metaanalysis,” *PLoS One*, vol. 17, no. 4 April, 2022, doi: 10.1371/journal.pone.0266334.

[37] J. Gardner, Z. Popovic, and L. Schmidt, “Benchmarking Distribution Shift in Tabular Data with TableShift,” no. NeurIPS, 2023, [Online]. Available: http://arxiv.org/abs/2312.07577

[38] U.S. Center for Disease Control and Prevention, “Behavioral Risk Factor Surveillance System 2015 Codebook Report,” pp. 1–137, 2016, [Online]. Available: http://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf

[39] A. Amron, “The Influence of Brand Image, Brand Trust, Product Quality, and Price on the Consumer’s Buying Decision of MPV Cars,” *Eur. Sci. Journal, ESJ*, vol. 14, no. 13, p. 228, 2018, doi: 10.19044/esj.2018.v14n13p228.

[40] J. Colleoni Couto, O. Teixeira Borges, and D. Dubugras Ruiz, “Data integration in a Hadoop-based data lake: A bioinformatics case,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 12, no. 4, pp. 1–24, 2022, doi: 10.5121/ijdkp.2022.12401.

[41] A. Partovi, D. Lukose, and G. I. Webb, “MiPy: A Framework for Benchmarking Machine Learning Prediction of Unplanned Hospital and ICU Readmission in the MIMIC-IV Database,” *ResearchSquare*, pp. 0–22, 2022, doi: 10.21203/rs.3.rs-2100869.

[42] F. Ren and Z. Huang, “Automatic Facial Expression Learning Method Based on Humanoid Robot XIN-REN,” *IEEE Trans. Human-Machine Syst.*, vol. 46, no. 6, pp. 810–821, 2016, doi: 10.1109/THMS.2016.2599495.

[43] S. Manikandan, “Data transformation,” *J. Pharmacol. Pharmacother.*, vol. 1, no. 2, p. 126, 2010.

[44] I. Vatolkin, G. Rudolph, and C. Weihs, “Evaluation of album effect for feature selection in music genre recognition,” *Proc. 16th Int. Soc. Music Inf. Retr. Conf. ISMIR 2015*, no. 1, pp. 169–175, 2015.

[45] R. Karthikeyan and B. Selvanandhini, “Imputation Based Data PreProcessing in Machine Learning for Heart Disease Dataset,” *J. Pharm. Negat. Results*, vol. 13, no. 8, pp. 3218–3227, 2022, doi: 10.47750/pnr.2022.13.S08.397.

[46] B. Erickson and T. Nosanchuk, *Understanding data*. McGraw-Hill Education (UK), 1992.

[47] S. Zhang, C. Zhang, and Q. Yang, “Data preparation for data mining,” *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 375–381, 2003, doi: 10.1080/713827180.

[48] W.-C. Lin and C.-F. Tsai, “Missing value imputation: a review and analysis of the literature (2006–2017),” *Artif. Intell. Rev.*, vol. 53, pp. 1487–1509, 2020.

[49] B. Hoyle, M. M. Rau, R. Zitlau, S. Seitz, and J. Weller, “Feature importance for machine learning redshifts applied to SDSS galaxies,” *Mon. Not. R. Astron. Soc.*, vol. 449, no. 2, pp. 1275–1283, 2015.

[50] W.-H. Au, K. C. C. Chan, and A. K. C. Wong, “A fuzzy approach to partitioning continuous attributes for classification,” *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 5, pp. 715–719, 2006.

[51] F. Q. Nuttall, “Body mass index: obesity, BMI, and health: a critical review,” *Nutr. Today*, vol. 50, no. 3, pp. 117–128, 2015.

[52] M. S. Kumar, M. Z. Khan, S. Rajendran, A. Noor, A. S. Dass, and J. Prabhu, “Imbalanced Classification in Diabetics Using Ensembled Machine Learning,” *Comput. Mater. Contin.*, vol. 72, no. 3, pp. 4397–4409, 2022, doi: 10.32604/cmc.2022.025865.

[53] C. C. Tusell-Rey, O. Camacho-Nieto, C. Yáñez-Márquez, and Y.

Villuendas-Rey, “Customized Instance Random Undersampling to Increase Knowledge Management for Multiclass Imbalanced Data Classification,” *Sustain.*, vol. 14, no. 21, 2022, doi: 10.3390/su142114398.

[54] H. M. Deberneh and I. Kim, “Prediction of type 2 diabetes based on machine learning algorithm,” *Int. J. Environ. Res. Public Health*, vol. 18, no. 6, pp. 9–11, 2021, doi: 10.3390/ijerph18063317.

[55] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, “Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data,” *Ecol. Modell.*, vol. 406, pp. 109–120, 2019.

[56] S. H. Huang, “Supervised feature selection: A tutorial,” *Artif. Intell. Res.*, vol. 4, no. 2, 2015, doi: 10.5430/air.v4n2p22.

[57] M. Rostami, K. Berahmand, E. Nasiri, and S. Forouzande, “Review of swarm intelligence-based feature selection methods,” *Eng. Appl. Artif. Intell.*, vol. 100, 2021, doi: 10.1016/j.engappai.2021.104210.

[58] X. Zhu *et al.*, “Identification of immune-related genes in patients with acute myocardial infarction using machine learning methods,” *J. Inflamm. Res.*, pp. 3305–3321, 2022.

[59] Y. B. Wah, N. Ibrahim, H. A. Hamid, S. Abdul-Rahman, and S. Fong, “Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy,” *Pertanika J. Sci. Technol.*, vol. 26, no. 1, pp. 329–340, 2018.

[60] B. Remeseiro and V. Bolon-Canedo, “A review of feature selection methods in medical applications,” *Comput. Biol. Med.*, vol. 112, p. 103375, 2019.

[61] M. K. Uçar, M. Nour, H. Sindi, and K. Polat, “The Effect of Training and Testing Process on Machine Learning in Biomedical Datasets,” *Math. Probl. Eng.*, vol. 2020, 2020, doi: 10.1155/2020/2836236.

[62] P. Marcelino, M. de Lurdes Antunes, E. Fortunato, and M. C. Gomes, “Machine learning approach for pavement performance prediction,” *Int. J. Pavement Eng.*, vol. 22, no. 3, pp. 341–354, 2021.

[63] H. Margareth, “No Title طرق س دري ت لغة ال عرب ية ال,” *Экономика Региона*, p. 32, 2017.

[64] P. Kazienko, E. Lughofer, and B. Trawiński, “Hybrid and ensemble methods in machine learning J.UCS special issue,” *J. Univers. Comput. Sci.*, vol. 19, no. 4, pp. 457–461, 2013.

[65] T. Chen, T. He, and M. Benesty, “XGBoost : eXtreme Gradient Boosting,” *R Packag. version 0.71-2*, pp. 1–4, 2018.

[66] R. E. Schapire, “Explaining adaboost,” in *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, Springer, 2013, pp. 37–52.

[67] J. Ding, Z. Chen, L. Xiaolong, and B. Lai, “Sales forecasting based on catboost,” in *2020 2nd international conference on information technology and computer application (ITCA)*, 2020, pp. 636–639.

[68] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, “Comparison of the CatBoost classifier with other machine learning methods,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 11, 2020.

[69] A. Robert and K. Potter, “Explainable AI : Interpreting and Understanding Machine Learning Models,” no. January, 2024.

[70] S. Kirrane, “Why model why? Assessing the strengths and limitations of LIME ¨,” no. iii, 1998.

[71] R. O. Alabi, M. Elmusrati, I. Leivo, A. Almangush, and A. A. Mäkitie, “Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP,” *Sci. Rep.*, vol. 13, no. 1, pp. 1–14, 2023, doi: 10.1038/s41598-023-35795-0.

[72] S. Raschka, “Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning,” 2018, [Online]. Available: http://arxiv.org/abs/1811.12808

[73] R. Susmaga, “Confusion matrix visualization,” in *Intelligent Information Processing and Web Mining: Proceedings of the International IIS: IIPWM ‘04 Conference held in Zakopane, Poland, May 17–20, 2004*, 2004, pp. 107–116.

[74] E. Beauxis-Aussalet and L. Hardman, “Simplifying the visualization of confusion matrix,” *Belgian/Netherlands Artif. Intell. Conf.*, pp. 133–134, 2014.

[75] A. Turpin and F. Scholer, “User performance versus precision measures for simple search tasks,” in *Proceedings of the 29th annual international*

ACM SIGIR conference on Research and development in information retrieval, 2006, pp. 11–18.

[76] A. Turpin and F. Scholer, "User performance versus precision measures for simple search tasks," *Proc. Twenty-Ninth Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, vol. 2006, pp. 11–18, 2006, doi: 10.1145/1148170.1148176.

[77] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv Prepr. arXiv2010.16061*, 2020.

[78] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," *arXiv Prepr. arXiv2008.05756*, 2020.

[79] S. K. Smit and A. E. Eiben, "Parameter tuning of evolutionary algorithms: Generalist vs. specialist," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6024 LNCS, no. PART 1, pp. 542–551, 2010, doi: 10.1007/978-3-642-12239-2_56.

[80] R. G. Mantovani, A. L. D. Rossi, E. Alcobaça, J. C. Gertrudes, S. B.

Junior, and A. C. P. de L. F. de Carvalho, "Rethinking Default Values: a Low Cost and Efficient Strategy to Define Hyperparameters," 2020, [Online]. Available: http://arxiv.org/abs/2008.00025

[81] B. H. Shekar and G. Dagnew, "Grid search-based hyperparameter tuning and classification of microarray cancer data," *2019 2nd Int. Conf. Adv. Comput. Commun. Paradig. ICACCP 2019*, no. February, pp. 1–8, 2019, doi: 10.1109/ICACCP.2019.8882943.

[82] S. Banik, R. M. Rangayyan, and J. E. Leo Desautels, *Feature Selection and Pattern Classification*. 2013. doi: 10.1007/978-3-031-01656-1_6.

[83] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," *2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2015 - Proc.*, no. May, pp. 1200–1205, 2015, doi: 10.1109/MIPRO.2015.7160458.