



# PREDICTING DIABETES PATIENTS BASED ON MACHINE LEARNING TECHNIQUES

By

Student #1 Raghad Alanazi

Student #2 Sara Alsibgh

Student #3 Atheer Almutairi

Student #4 Danah Alanazi

Student #5 Sarah Alhussein

Supervised By

Dr Hoda Ahmed Abdelhafez

A Graduation Project Report Submitted to  
College of Computer Sciences and Information at PNU  
in Partial Fulfillment of the Requirements for the  
Degree of  
Bachelor of Science  
in  
Information Technology

CCIS, PNU

Riyadh, KSA

1446H

# Table of Contents

List of Tables .....	iv
List of Figures .....	v
List of Symbols and Abbreviations.....	vi
Acknowledgments.....	viii
Abstract.....	ix
Chapter 1: Introduction.....	1
1.1 Introduction.....	2
1.2 Problem Statement & Significance .....	3
1.3 Proposed Solution .....	4
1.4 Project Domain & Limitations.....	5
1.4.1 Domain.....	5
Chapter 2: Background Information & Related Work .....	6
2.1 Background Information.....	7
2.1.1 Diabetes.....	7
2.1.2 Machine Learning .....	10
2.2 Related Work Survey .....	13
2.3 Proposed & Similar Systems Comparison.....	16
Chapter 3: System Analysis.....	18
3.1 Requirement Specification.....	19
3.1.1 Input Data Collection.....	19
3.1.2 Machine Learning Algorithms .....	24
3.1.3 Feature Selection.....	29
3.1.4 Evaluation .....	29
3.2 Requirements Analysis .....	36
3.2.1 Description of Diabetes Data .....	36
3.2.2 Data Flow Diagram.....	37

Chapter 4: System Design .....	38
4.1 System Architecture.....	39
4.1.1 Hardware:.....	39
4.1.2 Software .....	39
4.2 User Interface.....	41
Chapter 5: Implementation .....	46
5.1 Implementation Requirements .....	47
5.2 Implementation Details .....	47
5.3 I/O Screens.....	57
Chapter 6: Testing.....	63
6.1 Test plan.....	64
6.2 Test cases .....	65
6.3 Test results .....	66
Chapter 7: Conclusion .....	74
7.1. Evaluation .....	75
7.2. Future work.....	خطأ! الإشارة المرجعية غير معروفة.
References:.....	76

## List of Tables

Table 2.1: Similar Systems Comparison.....	16
Table 3.1: Confusion matrix.....	29
Table 3.2: Data Description about health indicators, their normal ranges, and thresholds indicating diabetes risk .....	36
Table 5.1: Model Performance of Naïve Bayes.....	48
Table 5.2: Performance of Naïve Bayes for each class .....	49
Table 5.3: Model Performance of Naïve Bayes.....	49
Table 5.4: Performance of Naïve Bayes for each class .....	49
Table 5.5: Model Performance of Logistic Regression .....	50
Table 5.6: Performance of Logistic Regression for each class .....	50
Table 5.7: Model Performance of Logistic Regression .....	51
Table 5.8: Performance of Logistic Regression for each class .....	51
Table 5.9: Model Performance of Neural Networks.....	51
Table 5.10: Performance of Neural Networks for each class .....	52
Table 5.11: Model Performance of Neural Networks.....	52
Table 5.12: Performance of Neural Networks for each class .....	52
Table 5.13: Model Performance of Random Forest.....	53
Table 5.14: Performance of Random Forest for each class .....	53
Table 5.15: Model Performance of Random Forest .....	54
Table 5.16 Performance of Random Forest for each class .....	54
Table 5.17: Model Performance of Decision Tree.....	54
Table 5.18: Performance of Decision Tree for each class .....	55
Table 5.19: Model Performance of Decision Tree.....	56
Table 5.20: Performance of Decision Tree for each class .....	56
Table 6.1: summarizes all the activities that had done for testing the project.....	64
Table 6.2: Test cases and their expected outcome .....	65
Table 6.3: Performance by machine learning algorithms .....	70

## List of Figures

<i>Figure 3.1: Example on Decision Tree</i> .....	27
<i>Figure 3.2: Random Forest Algorithm approach</i> .....	28
<i>Figure 3.3: Machine Learning Workflow for diabetes prediction</i> .....	37
<i>Figure 4.1: Flowchart of the diabetes prediction model</i> .....	40
<i>Figure 4.2: Gender Attribute in WEKA</i> .....	41
<i>Figure 4.3: Age Attribute in WEKA</i> .....	41
<i>Figure 4.4: Urea Attribute in WEKA</i> .....	42
<i>Figure 4.5: Cr Attribute in WEKA</i> .....	42
<i>Figure 4.6: HbA1c Attribute in WEKA</i> .....	42
<i>Figure 4.7: Chol Attribute in WEKA</i> .....	43
<i>Figure 4.8: TG Attribute in WEKA</i> .....	43
<i>Figure 4.9: HDL Attribute in WEKA</i> .....	43
<i>Figure 4.10: LDL Attribute in WEKA</i> .....	44
<i>Figure 4.11: VLDL Attribute in WEKA</i> .....	44
<i>Figure 4.12: BMI Attribute in WEKA</i> .....	44
<i>Figure 4.13: Class Attribute in WEKA</i> .....	45
<i>Figure 5.1: Decision Tree visualization for all features</i> .....	55
<i>Figure 5.2: Decision Tree visualization for four features</i> .....	57
<i>Figure 6.1: Naïve Bayes Model Performance</i> .....	66
<i>Figure 6.2: Random Forest Model Performance</i> .....	67
<i>Figure 6.3: Decision Tree Model Performance</i> .....	68
<i>Figure 6.4: Neural Networks Model Performance</i> .....	69
<i>Figure 6.5: Logistic Regression Model Performance</i> .....	70
<i>Figure 6.6: Accuracy performance for five ML algorithms</i> .....	71
<i>Figure 6.7: F-measure for five ML algorithms</i> .....	72
<i>Figure 6.8: MCC for five ML algorithms</i> .....	72
<i>Figure 6.9: ROC for five ML algorithms</i> .....	73

## List of Symbols and Abbreviations

<b>ML</b>	Machine Learning
<b>NN</b>	Neural Networks
<b>T1DM</b>	Type 1 Diabetes Mellitus
<b>T2DM</b>	Type 2 Diabetes Mellitus
<b>BMI</b>	Body Mass Index
<b>WEKA</b>	Waikato Environment for Knowledge Analysis
<b>RF</b>	Random Forest
<b>HbA1c</b>	Hemoglobin A1c
<b>TP</b>	True Positive
<b>TN</b>	True Negative
<b>FP</b>	False Positive
<b>FN</b>	False Negative
<b>HDL</b>	High-Density Lipoprotein
<b>LDL</b>	Low-Density Lipoprotein
<b>VLDL</b>	Very Low-Density Lipoprotein
<b>TG</b>	Triglycerides
<b>F1-Score</b>	Mean of Precision and Recall
<b>FPG</b>	Fasting Plasma Glucose
<b>Cr</b>	Creatinine Ratio
<b>NB</b>	Naïve Bayes
<b>LR</b>	Logistic Regression
<b>DT</b>	Decision Tree
<b>NB</b>	Naïve Bayes
<b>HDL-C</b>	High-Density Lipoprotein Cholesterol
<b>LDL (+)</b>	Positively Charged Low-Density Lipoprotein
<b>LDL (-)</b>	Negatively Charged Low-Density Lipoprotein
<b>Chol</b>	Cholesterol
<b>Urea</b>	Blood Urea Level
<b>RAE</b>	Relative Absolute Error
<b>MAE</b>	Mean Absolute Error
<b>RMSE</b>	Root Mean Squared Error

<b>RRSE</b>	Root Relative Squared Error
<b>MCC</b>	Matthews Correlation Coefficient
<b>ROC</b>	Receiver Operating Characteristic
<b>PRC</b>	Precision-Recall Curve

## **Acknowledgments**

First and foremost, we thank Allah for guiding us in every stage of our life, for giving us strength through our education, and for helping us complete this project. Throughout the days we worked on graduation project 1, our supervisor provided us with unconditional support as well as valuable and golden advice. We would like to take this opportunity to express our heartfelt gratitude and unceasing gratitude to our supervisor Dr. Hoda Ahmed Abdelhafez, for guiding us to success on this journey. As well as our university Princess Noura bint Abdulrahman, who provided us with the wonderful opportunity to work on this project. Finally, we are deeply grateful to the College of Computer Sciences and Information for equipping us with the knowledge and skills that were fundamental to complete this project.



## **Abstract**

Our graduation project focused on developing a robust diabetes prediction model using machine learning techniques, with an emphasis on implementation through the WEKA platform and performance optimization via feature selection. The dataset comprised 1,000 patient records, featuring clinical attributes such as age, gender, BMI, cholesterol, HbA1c, triglycerides, and creatinine levels. To enhance prediction accuracy, five machine learning algorithms Naïve Bayes, Logistic Regression, Neural Networks, Random Forest, and Decision Tree, were implemented, trained, and evaluated.

The implementation process involved systematic data preparation, including data cleaning, feature selection using the Wrapper Method, and model training with a 70/30 training-test split. Feature selection identified four key predictors Gender, HbA1c, Triglycerides, and BMI that significantly improved model efficiency. Each model was tested with and without feature selection to compare predictive performance.

A detailed evaluation using accuracy, precision, recall, F1-score, ROC area, and MCC showed that the Decision Tree algorithm achieved the highest accuracy at 99% without feature selection. Neural Networks and Random Forest also performed very well when feature selection was applied. The Wrapper Method especially improved the performance of Naïve Bayes and Logistic Regression, increasing both their accuracy and ability to make general predictions.

Our results highlight the powerful impact of machine learning in the healthcare field, particularly in automating complex diagnostic processes and facilitating early detection of diabetes. The developed system assists healthcare professionals by transforming patient data into meaningful insights, enabling timely interventions and the creation of personalized treatment plans.

# **Chapter 1: Introduction**

## 1.1 Introduction

Diabetes is one of the most common powerful chronic diseases around the world, as it significantly affects the lives of individuals and the quality of health care. To reduce its negative effects, early prediction of the disease is a crucial step. Machine learning techniques offer an innovative and effective solution in this field, using advanced algorithms to analyze huge amounts of data related to patients' health, such as age, gender, health habits, and clinical data, with the aim of predicting an individual's diabetes with high accuracy. Through these predictive models, it is possible to enhance the accuracy of early diagnosis and the provision of personalized health care, helping to improve the quality of life of individuals and reduce the possible complications of the disease. Diabetes is one of the most prominent chronic health challenges facing health systems and communities around the world. Early diagnosis and preventive intervention play a crucial role in improving the quality of life of patients and reducing health complications associated with the disease. In this context, machine learning is a revolutionary technique that contributes to the prediction of diabetes through the analysis of patients' medical and personal data. Using intelligent algorithms capable of detecting complex patterns in data, machine learning can make accurate predictions about the risk of diabetes, enabling physicians to make proactive treatment decisions and enhancing the effectiveness of healthcare provided, in 2023, more than half a billion people worldwide are living with diabetes, impacting men, women, and children across all age groups in every country. This number is expected to more than double to 1.3 billion over the next 30 years, with an increase anticipated in every nation, Currently, the global prevalence rate of diabetes stands at 6.1%, making it one of the top ten leading causes of death and disability. Among different regions, North Africa and the Middle East exhibit the highest rate at 9.3%, projected to rise to 16.8% by 2050. Latin America and the Caribbean are also expected to see an increase, with rates projected to reach 11.3%, Diabetes is particularly prevalent among individuals aged 65 and older, with a global prevalence exceeding 20% in this demographic. The highest rates are found among those aged 75 to 79, reaching 24.4%. When examining regional differences, North Africa and the Middle East report the highest prevalence in this age group at 39.4%, while Central Europe, Eastern Europe, and Central Asia show the lowest rate at 19.8% [1]

In 2024, diabetes prevalence in Saudi Arabia has reached significant levels, with recent statistics indicating that 18.7% of adults aged 20-79 are affected by the disease. This high prevalence is driven by factors such as lifestyle changes, dietary habits, and a high rate of obesity. The International Diabetes Federation (IDF) projects that the number of diabetes cases in Saudi Arabia could nearly double by 2045 if current trends continue. In response, the Saudi Ministry of Health has implemented various preventive and educational programs to address and manage diabetes more effectively [2].

## **1.2 Problem Statement & Significance**

Diabetes has recently spread widely in Saudi Arabia, as statistics published in 2024 indicate that approximately 18.7% of adults between the ages of 20 and 79 were diagnosed with the disease [2]. Diabetes also contributed to an additional 460,000 deaths from kidney disease and was responsible for approximately 20% of cardiovascular disease-related deaths. These data underscore the significant global health challenge posed by diabetes, with millions of individuals worldwide at risk of serious complications and reduced quality of life. Diabetes impacts daily life in many ways, from managing dietary restrictions and regular medications to dealing with potential risks associated with sudden symptoms, including hypoglycemia or hyperglycemia, which can lead to emergency situations. Furthermore, the financial burden on individuals and families increases as they bear the costs of treatment, monitoring devices, and potential caregiver support, which can make life difficult for some people. Today's modern technologies can offer a variety of solutions to help manage diabetes not only by making life easier but also by enabling them to take proactive control of their health. The diabetes prediction project leverages machine learning algorithms to enhance the lives of individuals with diabetes by offering user-friendly and smart technology. It involves implementing algorithms that predict blood sugar levels through continuous glucose monitoring and provide alerts for fluctuations, helping to better manage and reduce the impact of diabetes.

### 1.3 Proposed Solution

The proposed solution is to develop a model based on machine learning algorithms to predict diabetes based on a set of medical laboratory tests of patients, such as body mass index, cholesterol, urea, LDL, TG, HDL, and VLDL. The project aims to enable doctors and healthcare providers to predict diabetes early, which helps in making early treatment decisions.

Several machine learning algorithms are used to implement this solution, namely Decision Trees, Random Forest, Logistic Regression, Support Vector Machines (SVM), and Neural Networks. Each algorithm has specific features; the steps of this solution are:

1. **Data collection and processing:** Data is collected from medical records and analyzed, missing values are processed, and impurities are removed to ensure data quality.
2. **Data analysis and feature selection:** Using the Wrapper approach to feature selection, the effectiveness of each set of features is tested, and the most influential ones are identified to ensure improved model accuracy.
3. **Model building:** Training and testing the model using different algorithms and selecting the best model in terms of accuracy and sensitivity.
4. **Performance evaluation:** Comparing the performance of models using performance metrics such as accuracy, specificity, and F1 to determine the most effective algorithm.

We expect the predictive model to be able to provide an accurate assessment that help doctors diagnose diabetes early. We also expect that the project provides a deeper understanding of the key factors that influence the onset of the disease.

## **1.4 Project Domain & Limitations**

### **1.4.1 Domain**

Our project focuses on predicting diabetes patients using machine learning techniques to analyze patient medical data from the Iraqi Medical City Hospital and the Specialized Center for Endocrinology and Diabetes at Al-Kindy Teaching Hospital. This dataset includes 1,000 patient entries with medical attributes such as age, gender, BMI, cholesterol levels, and creatinine ratio, allowing for early detection and intervention. The domain of our project is doctors or healthcare professionals through providing them with accurate predictions of diabetes status and enabling them to receive quick insights that assist in diagnosis and treatment planning.

### **Limitations**

1. The project is exclusively designed to predict diabetes patients and does not cover other health conditions or all Diseases.
2. The dataset is sourced from specific hospitals in Iraq, which may not be representative of other populations or regions, affecting generalizability.
3. Performance may vary when applied to datasets from different demographics or clinical backgrounds.

## **Chapter 2: Background Information & Related Work**

## **2.1 Background Information**

### **2.1.1 Diabetes**

Diabetes is a chronic condition that occurs when blood glucose levels are too high, and the body cannot produce enough insulin to regulate them. Insulin, a hormone produced by the pancreas, allows cells to absorb glucose from the bloodstream to use as energy. People with prediabetes have higher-than-normal blood glucose levels, but not high enough to be classified as diabetes. Prediabetes is a warning sign that indicates an increased risk of developing diabetes in the future [3]. Diabetes is more common in individuals over the age of 45 and can be influenced by genetic factors. People with diabetes are at a higher risk of developing cardiovascular disease and complications affecting the feet and eyes compared to those with normal glucose levels. Proper management of blood glucose, blood pressure, and cholesterol can help reduce the risk of these health problems. In 2023, more than half a billion people worldwide are living with diabetes, impacting men, women, and children across all age groups in every country. This figure is expected to more than double to 1.3 billion over the next 30 years, with an increase anticipated in every nation. Currently, the global prevalence rate of diabetes stands at 6.1%, making it one of the top ten leading causes of death and disability. Among different regions, North Africa and the Middle East exhibit the highest rate at 9.3%, projected to rise to 16.8% by 2050. Latin America and the Caribbean are also expected to see an increase, with rates projected to reach 11.3%. Diabetes is particularly prevalent among individuals aged 65 and older, with a global prevalence exceeding 20% in this demographic. The highest rates are found among those aged 75 to 79, reaching 24.4%. When examining regional differences, North Africa and the Middle East report the highest prevalence in this age group at 39.4%, while Central Europe, Eastern Europe, and Central Asia show the lowest rate at 19.8% [2].



## **Types of diabetes**

Type 1: autoimmune diabetes (T1DM), formerly known as insulin-dependent or juvenile diabetes, is an autoimmune condition characterized by the destruction of pancreatic  $\beta$ -cells, leading to insufficient insulin production and resulting in hyperglycemia. This condition necessitates daily insulin administration. While symptoms typically manifest during childhood or adolescence, they can occasionally develop much later in life. The pathogenesis of T1DM can be divided into three stages, based on the presence or absence of hyperglycemia and associated symptoms such as polyuria and excessive thirst. Currently, there is no cure for T1DM, and patients rely on lifelong insulin injections. However, innovative approaches to insulin management are being developed, including insulin pumps, continuous glucose monitoring systems, and hybrid closed-loop systems [4].

Type 2: non-insulin-dependent diabetes (T2DM), occurs due to insufficient insulin production by the pancreatic beta cells or when the muscle, liver, and fat cell receptors do not respond properly to normal insulin levels, a condition known as insulin resistance. In type 2 diabetes, however, symptoms are often mild and may go unnoticed for many years. Common symptoms of diabetes include excessive thirst, frequent urination, blurred vision, fatigue, and unintentional weight loss. Over time, diabetes can damage blood vessels in the heart, eyes, kidneys, and nerves. People with diabetes are more likely to experience heart attacks, strokes, and kidney failure. Diabetes can also lead to permanent vision loss by damaging the blood vessels in the eyes, and many individuals with diabetes develop foot problems due to nerve damage and poor circulation, which can result in ulcers and amputations [5]. Almost all global cases (96%) are type 2 diabetes (T2DM). Research has identified 16 risk factors associated with T2DM, with high body mass index (BMI) being the primary contributor, responsible for 52.2% of T2DM-related disability and mortality. Other significant risk factors include dietary risks, environmental and occupational hazards, tobacco use, physical inactivity, and alcohol consumption. The rapid increase in diabetes cases poses significant challenges for healthcare systems worldwide. While many associate the causes of type 2 diabetes with obesity, lack of exercise, and poor nutrition, the prevention and management of diabetes are complex issues. They are influenced by genetic, social, and financial factors, particularly in low- and middle-income countries. Addressing these challenges requires a multifaceted approach that considers not only

lifestyle changes but also the broader socio-economic context in which individuals live. Using data from the Global Burden of Disease (GBD) 2021 study, researchers analyzed the prevalence, morbidity, and mortality of diabetes across 204 countries and territories from 1990 to 2021, and projected future prevalence rates through 2050. This comprehensive study also provided estimates for both type 1 diabetes (T1D) and type 2 diabetes (T2D) and assessed the proportion of the T2D burden attributable to the identified risk factors. The research team included experts from the Institute for Health Metrics and Evaluation (IHME) and collaborators worldwide [6].

### **Main Diagnostic Tests**

1-Fasting Plasma Glucose (FPG) Test: The FPG test measures blood glucose levels after an overnight fast. A fasting plasma glucose level of 126 mg/dL (7.0 mmol/L) or higher is indicative of diabetes. This test is widely used due to its simplicity and reliability [7].

2-Oral Glucose Tolerance Test (OGTT): The OGTT assesses the body's response to glucose. After fasting, a patient consumes a glucose-rich beverage, and blood glucose levels are measured at intervals. A two-hour plasma glucose level of 200 mg/dL (11.1 mmol/L) or higher confirms a diabetes diagnosis. This test is particularly useful for diagnosing gestational diabetes and insulin resistance [8].

3-Hemoglobin A1c (HbA1c) Test: The HbA1c test reflects average blood glucose levels over the past two to three months. An HbA1c level of 6.5% (48 mmol/mol) or higher indicates diabetes. This test is valuable for monitoring long-term glucose control and can be performed without fasting, making it convenient for patients [9].

4-Random Plasma Glucose Test: This test measures blood glucose at any time of day, regardless of when the patient last ate. A random plasma glucose level of 200 mg/dL (11.1 mmol/L) or higher suggests diabetes, particularly in symptomatic individuals [10].

### **Treatment for diabetes**

Lifestyle Modifications: dietary changes, increased physical activity, and weight management. A Mediterranean-style diet and regular aerobic exercise have been shown to improve glycemic control and reduce cardiovascular risk [10].

Pharmacological Treatments: Several classes of medications are available for diabetes: Metformin the first-line treatment for Type 2 diabetes, metformin improves insulin

sensitivity and lowers hepatic glucose production. Recent studies indicate that metformin also offers cardiovascular benefits [11]. **Insulin Therapy:** For individuals with Type 1 diabetes and some with Type 2 diabetes, insulin therapy is essential for managing blood glucose levels. Various insulin formulations are available, including rapid-acting, long-acting, and pre-mixed insulin [12].

**Continuous Glucose Monitoring (CGM):** The use of CGM devices allows for real-time monitoring of glucose levels, facilitating better glycemic control and reducing the risk of hypoglycemia. Recent advancements in CGM technology have improved accuracy and user-friendliness, enhancing patient engagement in diabetes management.

**Behavioral Interventions and Education:** Structured diabetes education programs have been shown to improve self-management skills, leading to better glycemic control and health outcomes. These interventions empower patients to make informed decisions regarding their diet, exercise, and medication adherence [13].

### **2.1.2 Machine Learning**

Machine Learning (ML) is a subset of artificial intelligence (AI) that allows systems to learn and improve from experience without being explicitly programmed. It involves the development of algorithms that can identify patterns in large data sets, make decisions, and provide predictions based on these patterns. The core idea is to develop models that generalize well from sample data (training data) to unseen situations (test data or real-world applications). In contrast to traditional programming, where instructions are coded explicitly, ML models identify these instructions by analyzing data [14]. This discipline is rooted in statistics, computer science, and optimization theory. Its methods often overlap with data mining and pattern recognition, focusing on prediction, classification, and knowledge discovery from data. Machine learning provides powerful tools for predicting diseases like diabetes by analyzing large and complex medical datasets. Supervised learning, particularly classification algorithms such as logistic regression, decision trees, and neural networks, is commonly used for diabetes prediction. By leveraging these algorithms, healthcare professionals can identify at-risk individuals early, potentially preventing the onset of diabetes through timely interventions.

## Types of Machine Learning

Machine learning can be classified into three main types based on how they handle data:

1. **Supervised learning:** This type of learning occurs when an algorithm is trained on a labeled dataset where the matching of inputs and outputs is known. The goal of the model is to learn the relationship between the input variables (features) and the target variable (label) to predict new, unseen data. Common algorithms are linear regression, logistic regression, decision trees, random forests, support vector machines (SVMs), neural networks. Predicting the onset of diabetes by learning patterns between medical test results and patient outcomes (i.e., diabetes diagnosis) is an example of ML application.
2. **Unsupervised learning:** Unsupervised learning deals with unlabeled data. The goal of the algorithm is to find hidden patterns or underlying structures in the input data. Since there is no specific goal, the algorithm explores the data to cluster, aggregate, or reduce its dimensionality. Common algorithms are K-means clustering, hierarchical clustering, principal component analysis (PCA), autoencoders. Example of Unsupervised learning is grouping patients into different risk profiles based on health metrics without prior knowledge of their diabetes status.
3. **Reinforcement Learning:** In reinforcement learning, an agent interacts with an environment and learns to take actions to maximize a cumulative reward. The feedback provided is in the form of rewards or penalties, not direct instruction. Common Algorithms are Q-Learning, Deep Q-Networks (DQN), Proximal Policy Optimization (PPO). Example of reinforcement learning is optimization of personalized treatment plans where each decision maximizes a patient's long-term health outcome.

Classification models are used when the outcome is a discrete label (e.g., predicting whether a person has diabetes or not). Binary classification (e.g., diabetic vs. non-diabetic) and multi-class classification (if different subtypes of diabetes are to be predicted) are common tasks. For instance, logistic Regression is a binary classification algorithm useful for predicting probabilities of disease occurrence.

## **Machine Learning for predicting diabetes patients.**

Using machine learning to predict diabetes predictive models for diabetes are widely used in clinical research and healthcare to identify high-risk individuals and provide early interventions. For example, researchers use Random Forest algorithm on patient data (e.g., blood pressure, glucose levels, and BMI) to predict the likelihood of developing diabetes. Regression models are used for predicting continuous outcomes. In the context of diabetes prediction, regression models could predict continuous measures such as blood sugar levels based on a patient's medical features [15]. Support Vector Machines (SVM) are a class of supervised learning algorithms adept at handling high-dimensional datasets. They work by identifying the optimal hyperplane that separates different classes, thereby facilitating the classification of individuals as at risk or not at risk for developing diabetes [16]. Decision tree: These models are used to generate decision rules based on patient metrics such as glucose levels, blood pressure, and age, making them particularly useful for healthcare professionals who need transparency into predictive outcomes. Decision trees have been integrated into decision support systems for diabetes management, allowing clinicians to create personalized treatment plans based on a patient's unique risk profile. When used in ensemble methods, such as random forests or gradient boosted trees, they dramatically improve predictive power while mitigating the risk of overfitting that individual trees often face.

Neural networks: are increasingly being used in diabetes care for tasks such as image analysis (e.g., detecting diabetic retinopathy from retinal images) and time series forecasting (e.g., predicting blood sugar fluctuations). Convolutional neural networks (CNNs) are being used to analyze medical images, identifying subtle changes in eye or nerve health that may indicate diabetes complications. Recurrent neural networks (RNNs) and long-short-term memory (LSTM) models are applied to continuous glucose monitoring (CGM) data to predict future glucose levels, giving patients and physicians real-time insights for insulin management [17].

## 2.2 Related Work Survey

Khanam and Foo [18] compared machine learning algorithms for diabetes prediction using the Pima Indian Diabetes (PIDD) dataset. They evaluated seven models including Decision Tree (DT), K Nearest Neighbor (KNN), Random Forest (RF), Naïve Bayes (NB), Adaptive Boosting (AB), Logistic Regression (LR), and Support Vector Machine (SVM) all models provided an accuracy greater than 70% with LR and SVM showing accuracies around 77%–78%. They built a Neural Network (NN) model with varying hidden layers and epochs, finding that the NN with two hidden layers achieved the highest accuracy of 88.6. The NN model had weaknesses as it used only five input features (Glucose, BMI, Insulin, Pregnancy, and Age) for prediction, which limited the model's potential. Additionally, it relied solely on the PIDD dataset, which was small and might not have been representative of other populations.

Butt et al. [19] used machine learning techniques to enhance diabetes detection, employing Random Forest (RF), multilayer perceptron (MLP), and Logistic Regression (LR) for classification, and Long Short-Term Memory (LSTM), and Linear Regression for prediction, using the Pima Indian Diabetes dataset (PIDD). They developed an MLP-based algorithm for diabetes classification and an LSTM-based model for prediction. They also proposed an IoT-based real-time monitoring system using a smartphone, (Bluetooth Low Energy) BLE sensor, and machine learning. Their models, tested on the PIMA Indian Diabetes dataset, achieved accuracies of 86.08% for MLP and 87.26% for LSTM, outperforming existing methods. However, the models' effectiveness relied heavily on the quality and balance of the training data. Imbalanced, noisy, or insufficiently detailed data can lead to suboptimal performance of both the MLP and LSTM models.

Tan et al. [20] proposed a GA-stacking model that combines Genetic Algorithm (GA) with convolutional neural networks and Support Vector Machines to improve diabetes risk prediction. They used two datasets, the desensitization physical examination data from Qingdao CDC with 8,787 records including health metrics, and the UCI early-stage diabetes risk prediction dataset containing 520 instances with 16 diabetes-related attributes. The GA-stacking model excelled in diabetes risk prediction with an accuracy of 85.88%, outperforming KNN, SVM, Logistic Regression, Naive Bayes, and CNN. Its effective feature selection using Genetic Algorithms and strong generalization

through stacking CNN and SVM contributed to its high performance. Yet the model struggled with small or imbalanced datasets.

Krishnamoorthi et al. [21] proposed an intelligent diabetes mellitus prediction framework (IDMPF) using machine learning techniques, specifically Decision Tree-based Random Forest (RF) and Support Vector Machine (SVM) models, to predict diabetes. They used the Pima Indian Diabetes dataset (PIDD). The framework developed after a comprehensive review of existing models and achieved 83% accuracy with a minimum error rate, emphasizing its potential in early diagnosis. However, they relied solely on the PIDD dataset, which was small and might not have been representative of other populations.

Kaur and Kumari. [22] examined five machine learning models using the Pima Indian Diabetes dataset (PIDD), consisting of linear kernel Support Vector Machine (SVM-linear), Radial Basis Function (RBF) kernel Support vector machine, K Nearest Neighbor (KNN), Artificial Neural Network (ANN), and Multifactor Dimensionality Reduction (MDR). The results indicated that SVM-linear model achieved the highest accuracy (0.89) and precision (0.88), while KNN excelled in recall and F1 score (0.90 and 0.88). This highlights that both models are effective for diabetes prediction. However, a significant limitation of the study is the imbalanced dataset, which contained more non-diabetic cases than diabetic ones. This imbalance can bias the models toward favoring the majority class, potentially inflating accuracy metrics and diminishing the model's ability to identify diabetic cases effectively. Although the F1 score addresses the trade-off between precision and recall, the dataset imbalance remains a constraint in accurately detecting diabetes.

Quan Zou et al. [23] explored predicting diabetes mellitus using machine learning techniques, including Decision Trees (DT), Random Forests (RF), and Neural Networks (NN), on a dataset of hospital physical examination data from Luzhou, China, it contains 14 attributes, such as age and glucose levels, and consists of 151,598 diabetic and 69,082 healthy instances after data cleaning. The study utilized principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) for dimensionality reduction, which helped enhance model performance and ensure robustness across different datasets. The Random Forest (RF) model outperformed other techniques, achieving an accuracy of 80.84%, demonstrating its potential for

practical applications. Despite achieving a high accuracy, the study's focus on limited set of features, such as fasting glucose, indicates that the models may not fully capture the complexity of diabetes.

Rajput and Khedgikar. [24] explored diabetes prediction using the Mendeley Diabetes types of dataset, which contains data for 1,000 patients with various attributes such as age, gender, Body Mass Index (BMI), and blood sugar levels. They highlighted the importance of careful feature selection and outlier removal to improve model accuracy. Employing multiple machine learning algorithms including Stochastic Gradient Boosting (SGD), Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF), and Multinomial Logistic Regression (MLR) along with a robust data preprocessing strategy, they achieved impressive results, with Stochastic Gradient Boosting reaching an accuracy of 97.04% and the Decision Tree model achieving 95.07%. However, there is a potential for overfitting that arises from the inclusion of numerous features in the model. which can lead to a scenario where the model learns noise in the training data rather than the underlying patterns. This is particularly concerning when the dataset contains many attributes but a relatively small number of samples.



## 2.3 Proposed & Similar Systems Comparison

Table 2.1: Similar Systems Comparison

	Your system	System1 [24]	System 2[54]
<b>Problem solved</b>	The project solved the problem of predicting diabetes risk by analyzing health data with machine learning algorithms, helping healthcare professionals or doctors with early detection and prevention.	Prevention and prediction of diabetes based on human body traits and machine learning algorithms.	The project focuses on diabetes prediction and prevention using machine learning algorithms to support patients and health care
<b>Domain &amp; Users</b>	Healthcare providers and Doctors	Diabetes and researchers and health care.	Healthcare and Medical professionals.
<b>Design methods</b>	<p><b>Data Collection and Processing:</b> Dataset collected from the Iraqi Medical City Hospital and the Endocrinology Specialized Center, analyzed, and missing and outlier values were treated to ensure data accuracy.</p> <p><b>Feature Selection by wrapper:</b> The wrapper approach selected to fit the goals of our project. The goal is to select the feature that has the greatest impact on the accuracy of the model.</p> <p><b>Model Building:</b> five different algorithms were selected based on performance and the most accurate for predicting diabetes, which are: SVM, NN, Logistic Regression, Decision Tree, Random Forest.</p> <p><b>Evaluation:</b> To evaluate the model's performance, we chose 5 measures in the project: recall, precision, accuracy, specificity, and F1-score.</p>	<p><b>Data processing:</b> The dataset utilized in the system is retrieved from Mendeley Diabetes dataset, A box plot was applied to the database to remove any outliers to ensure the quality of the data used for analysis.</p> <p><b>Correlation Analysis:</b> Pearson's correlation analysis was performed to check the relationship between different features and their suitability for predicting diabetes.</p> <p><b>Feature Selection by ANOVA:</b> The study used the ANOVA F-test to determine which features had a significant impact on the diabetes class and should be included in model building. The result was features are age, HbA1c, VLDL, and BMI.</p> <p><b>Model Building and Deployment:</b> The proposed methodology implemented five different machine learning algorithms for diabetes prediction: Multinomial Logistic Regression, DT, RF, Stochastic Gradient Boosting, Naïve Bayes.</p> <p><b>Performance Evaluation:</b> The performance of the classifiers were evaluated using various evaluation metrics.</p>	<p><b>Data Collection:</b> The dataset was obtained from Kaggle, consisting of 768 instances and 9 attributes.</p> <p><b>Data Preprocessing:</b> null values were checked and removed, and categorical values are converted into numerical ones.</p> <p><b>Exploratory Data Analysis:</b> Correlation matrices and visualizations like box plots were used to analyze the data and select key features.</p> <p><b>Train-Test Split:</b> The dataset was divided into 75% training data and 25% testing data.</p> <p><b>Algorithms Used:</b> Four algorithms were implemented: Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Gradient Boost.</p>

<b>Software &amp; hardware</b>	WEKA tool, PC.	Building models using the R programming environment, PC.	WEKA tool, PC.
<b>Output</b>	Accurate prediction using 5 machine learning algorithms: DT, Random Forest, Logistic Regression, SVM, NN to determine the best algorithm in terms of accuracy and performance	The results of the experiments showed that Stochastic gradient boost and decision tree algorithms achieved a high accuracy of 97.04 compared to other machine learning algorithms. BMI and HbA1c were key factors in increasing the risk of diabetes.	Gradient Boost algorithm provided the best accuracy of around 81.25% and KNN had lowest accuracy, which was 78%.
<b>features</b>	The system is cost-effective, reducing healthcare expenses. Also, Early detection of diabetes leads to better health outcomes.	The study identified the main risk factors. This helps in taking preventive measures and changing lifestyle. Timely identification of diabetes improved health outcomes and lowers healthcare costs.	It provided early detection with High Precision and could manage large datasets.
<b>limitations</b>	The project focuses specifically on diabetes and certain features related to it and does not cover all diseases. It may miss important factors like lifestyle or rare conditions.	The study was limited by a small sample size, which restricts its ability to be generalized.	The study relied on specific features, and the selection process might not capture all relevant factors influencing diabetes.

## **Chapter 3:   System Analysis**

## 3.1 Requirement Specification

### 3.1.1 Input Data Collection

The dataset collected from the Iraqi society, specifically from the laboratory of Medical City Hospital and the Specialized Center for Endocrinology and Diabetes at Al-Kindy Teaching Hospital. This dataset included medical information and laboratory analyses. The dataset contains data for 1000 patients includes 103 (no-diabetes), 53 (pre-diabetic) and 844 (diabetic) patients. The data attributes for our model are Age, Gender, Creatinine Ratio (Cr), Body Mass Index (BMI), Urea, Cholesterol, Low-Density Lipoprotein (LDL), Very Low-Density Lipoprotein (VLDL), Triglycerides (TG), High-Density Lipoprotein (HDL) Cholesterol, HBA1C. The patient's diabetes disease classification contained three classes: Diabetic, Non-Diabetic, or Predict-Diabetic) [25].

#### **Body Mass Index (BMI)**

Lifetime diabetes risk increases with baseline BMI in both sexes and at every baseline age. Obesity, or excessive weight gain, recognized as the most significant risk factor for the development and progression of type 2 Diabetes Mellitus (DM) across all age groups. A higher BMI and increased abdominal fat distribution is linked to an elevated risk of type 2 DM, primarily due to changes in adipose tissue biology that contribute to insulin resistance and beta cell dysfunction. According to the World Health Organization (WHO), obesity accounts for 44% of diabetes cases, with projections indicating that the incidence of obesity-related diabetes is expected to double, reaching 300 million by 2025 [26]. At the age of 18 the diabetes risk among males exhibits considerable variation based on BMI, with a risk of 7.6% for those with a BMI below 18.5 kg/m<sup>2</sup> and a substantial increase to 70.3% for those with a BMI exceeding 35 kg/m<sup>2</sup>. for females, the risk ranges from 12.2% for baseline BMIs below 18.5 kg/m<sup>2</sup> to 74.4% for those with BMIs greater than 35 kg/m<sup>2</sup>. However, the relationship between increasing BMI and diabetes risk becomes less pronounced with advancing age. By the age of 65 the risk for males diminishes to 2.2% for those with a BMI under 18.5 kg/m<sup>2</sup>, while increasing to 34.7% for those with a BMI above 35 kg/m<sup>2</sup>. In comparison, females at the same age face a risk ranging from 3.7% for a BMI below 18.5 kg/m<sup>2</sup> to 36.0% for a BMI above 35 kg/m<sup>2</sup> [27].

## **Cholesterol**

Cholesterol is a waxy substance produced by the body and found in certain animal-based foods. Blood cholesterol levels refer to a group of fats, also known as lipoproteins or lipids, which include High-Density-Lipoprotein-Cholesterol (HDL-C), often referred to as "good" cholesterol, and Low-Density-Lipoprotein-Cholesterol (LDL-C), known as "bad" cholesterol. In individuals with diabetes, there is a tendency for "good" cholesterol levels to decrease while triglycerides and "bad" cholesterol levels rise. This imbalance significantly increases the risk of developing heart disease and stroke, highlighting the importance of monitoring, and managing cholesterol levels in diabetic patients [28]. The National Library of Medicine (NLM) provides guidelines for healthy cholesterol levels in adults. According to their recommendations, total cholesterol should be between 125 and 200 milligrams per deciliter (mg/dl). Monitoring these cholesterol levels is crucial for managing overall health, especially in individuals with diabetes [29].

## **Creatinine ratio (CR)**

The creatinine ratio is a valuable biomarker in assessing kidney function and has significant implications in diabetes prediction and management. Elevated levels of creatinine in the urine can indicate kidney damage, which is particularly relevant for individuals with diabetes, as they are at a higher risk of developing diabetic nephropathy. The kidneys play a crucial role in filtering waste products from the blood, including creatinine. Monitoring urinary creatinine levels can help detect early signs of kidney dysfunction, which is critical for diabetic patients who may be prone to kidney-related complications. Research indicates that alterations in the creatinine ratio may correlate with the onset of type 2 diabetes (T2DM). Regular monitoring of urinary creatinine levels can aid in early detection of kidney dysfunction, facilitating timely intervention and improved patient outcomes [30].

## **Urea**

In the clinical setting, urea levels are primarily used to assess kidney function. The relationship between urea and diabetes is particularly important because of the risk of diabetic nephropathy, a common complication of chronic hyperglycemia. Diabetic

nephropathy results in structural and functional changes in the kidneys, including thickening of the glomerular basement membrane. These changes impair the kidneys' ability to filter blood, leading to retention of nitrogenous waste products such as urea in the bloodstream. Chronic hyperglycemia leads to oxidative stress and the formation of advanced glycation end products (AGEs), which damage kidney tissue, worsening kidney dysfunction. Elevated serum urea levels (azotemia) in patients with diabetes often reflect worsening kidney function and are associated with an increased risk of chronic kidney disease (CKD). In fact, diabetes is the leading cause of chronic kidney disease worldwide, and elevated urea levels may be one of the early signs of kidney dysfunction in this population. Assessment of urea, along with other renal markers such as creatinine and estimated glomerular filtration rate (eGFR), is critical for early diagnosis and management of renal complications in diabetes. Regular monitoring allows for early intervention to slow the progression of kidney damage, often through strict glycemic control, blood pressure management, and the use of renal protective agents such as angiotensin-converting enzyme (ACE) inhibitors and angiotensin receptor blockers (ARBs) [31].

### **Low-Density Lipoprotein (LDL)**

low-density lipoprotein, is a type of lipoprotein that carries cholesterol in the blood. LDL is a major component of the body's lipid profile and is the lipoprotein most closely associated with atherosclerosis and coronary heart disease (CHD). It is known as "bad cholesterol" because it can build up in the walls of the arteries, narrowing them and increasing the risk of CHD. It is the main target of lipid-lowering therapy. [32], [33]

There are two different types of LDL (low-density lipoprotein). They are identified based on their electrical properties:

LDL (+): This type carries a positive electrical charge.

LDL (-): This type carries a negative electrical charge and considered to have higher atherogenic properties, making it more associated with heart disease risk.

LDL is considered an independent risk factor for heart disease, but its effect on blood sugar levels is not direct. However, diabetes can lead to increased levels of LDL (low-density lipoprotein), which increases the risk of heart disease. High levels of LDL, such as negatively charged LDL (-), which is more strongly associated with heart disease

risk, have been linked to abnormally high levels in people with diabetes. They have been linked to poor blood sugar control in people with type 1 diabetes, but not in people with type 2 diabetes. Therefore, it is important to monitor LDL levels in people with diabetes, as controlling these levels can help reduce cardiovascular risk.[32], [33]

### **Triglycerides (TG)**

Triglycerides (TG) are a type of fat found in the blood and are an important indicator of metabolic health, especially for individuals with diabetes. High triglyceride levels are a common feature of diabetes and can significantly increase the risk of cardiovascular disease. People with diabetes are more likely to have higher triglyceride levels than the general population. This condition, known as diabetic dyslipidaemia, typically involves not only high triglycerides but also low levels of high-density lipoprotein (HDL) cholesterol and small, dense low-density lipoprotein (LDL) particles. The American Diabetes Association (ADA) provides specific guidelines for triglyceride levels in individuals with diabetes: normal: less than 150 mg/dL, high risk (moderately elevated): 150–199 mg/dL, high: 200–499 mg/dL, very high: 500 mg/dL and above. Studies show that a significant percentage of people with diabetes have high triglyceride levels. According to various reports: About 30-50% of individuals have high triglyceride levels (above 150 mg/dL) [34]. This percentage can increase to 60% or higher in people with poorly controlled diabetes. These high triglyceride levels are due to insulin resistance, which disrupts the normal process of fat metabolism. As a result, the liver increases the production of triglycerides, leading to higher concentrations in the blood. High triglycerides are of particular concern for people with diabetes because they are associated with an increased risk of Cardiovascular disease. High triglyceride levels can lead to hardening of the arteries, increasing the risk of heart attacks and strokes. Pancreatitis: Very high triglyceride levels (above 500 mg/dL) can lead to acute pancreatitis, a potentially life-threatening condition. Insulin resistance: High triglyceride levels worsen insulin resistance, creating a vicious cycle that makes it difficult to control blood sugar levels.

### **HDL (High-density lipoprotein)**

high-density lipoprotein, is a type of cholesterol that considered beneficial for heart health. It is known as the “good cholesterol” because it helps remove cholesterol from cells, especially from artery walls, through a process called “reverse cholesterol

excretion.” This means that HDL picks up cholesterol from peripheral cells and returns it to the liver, where it can reuse or dispose of. This process helps reduce the buildup of cholesterol in the arteries, reducing the risk of atherosclerosis and cardiovascular disease. [35]

HDL-C refers to the level of cholesterol contained within high-density lipoprotein. HDL-C uses as a marker and to measure and assess the risk of heart disease. Studies suggest that low levels of HDL-C are one of the strongest independent risk factors for heart disease. As for the effect of HDL on diabetes, individuals with type 2 diabetes exhibit several lipid disorders, in which low levels of HDL-C are a prominent feature. Therefore, improving HDL levels may have a positive impact on the health of individuals with diabetes. [36]

### **VLDL (Very Low-Density Lipoprotein)**

Fats like cholesterol and triglycerides can’t travel solo through your blood, they need lipoproteins to carry them to various organs and tissues throughout your body. VLDL “very-low-density lipoprotein” is a type of lipoprotein that your liver creates and sends out into your bloodstream. It helps your body gain energy, store energy and regulate blood pressure. They’re important for your overall body function but having too many VLDLs in your blood can be dangerous and raise your risk of cardiovascular disease and diabetes. There is no direct way to measure VLDL in your body. but blood tests will show you the level of triglycerides in your blood and Your doctor can then use the triglyceride level to determine if your VLDL levels are average, below, or above normal [37]. On average, VLDL higher than (often above 30 mg/dL) can be indicative of an increased risk of developing diabetes, high triglycerides is a risk factor for Type 2 diabetes because it often causes the body to produce more VLDL due to insulin resistance.[38].



### **3.1.2 Machine Learning Algorithms**

#### **Naïve Bayes (NB)**

Naïve Bayes (NB) is a powerful statistical classification algorithm that leverages Bayes' Theorem for predictive modeling. It assumes conditional independence between features, meaning that the effect of a particular feature on the class label is independent of other features. This simplifying assumption reduces the complexity of the model, making it faster and computationally efficient, especially when dealing with high-dimensional data. In practice, Naïve Bayes works by calculating the prior probability of each class and the likelihood of each feature given a specific class label, applying Bayes' Theorem to compute the posterior probability of each class given the input features. The class with the highest posterior probability is then selected as the predicted class. This approach is especially beneficial for multi-class classification tasks, where the goal is to assign an instance to one of several potential categories based on its features. Despite the simplifying assumption of independence, which may not always hold in real-world data, Naïve Bayes classifiers have been shown to perform surprisingly well in many practical scenarios[39]. One of the key advantages of Naïve Bayes is its ability to handle large datasets with relatively little computational overhead. It requires minimal training data to estimate the necessary parameters and can be used effectively in cases where other, more complex algorithms may struggle with time and resource constraints. The algorithm's ability to classify multi-class data efficiently and deliver quick, interpretable results further highlights its versatility. Naïve Bayes has been shown to perform well across a variety of domains due to its simple yet effective approach. Its computational speed, combined with a strong track record of real-world success, makes it a reliable option for classification tasks, particularly in environments where fast and efficient predictions are necessary [40][41].

#### **Neural Networks (NN):**

A neural network is a computational model inspired by how the human brain processes information. It is composed of interconnected layers of nodes (or neurons), where each node is a simple processor that works in parallel with others. Neural networks are designed to recognize patterns, classify data, and make decisions.

In the context of diabetes prediction, a neural network takes in various patient health parameters (such as Urea, BMI, age, and Creatinine ratio, etc.), processes them through its layers of neurons, and predicts whether the individual is likely to have diabetes. This process works in the following way:

1. *Input Layer*: The patients' health data input into the neural network input layer. Each feature represented as a separate input.
2. *Hidden Layers*: The inputs are processed through the hidden layers, where each neuron performs computations based on weighted inputs. The weights represent the strength or importance of each input, and the weighted sum is passed through an activation function (often a sigmoid function, which outputs values between 0 and 1).
3. *Output Layer*: The final layer produces the prediction, typically a probability indicating the likelihood of diabetes.

During training, the network adjusts the weights and biases based on how well the predicted outputs match the actual outcomes. This process is repeated to allow the network to learn patterns associated with diabetes data, improving its predictive accuracy over time, neural networks help in diabetes prediction by analyzing complex relationships between various health factors, thus enabling early detection and potentially better patient outcomes [42].

Neural networks in diabetes prediction have several advantages and disadvantages. One of the key advantages is their ability to recognize complex patterns that might not be apparent with traditional models. Because they can process multiple variables simultaneously (such as Urea, BMI, age, and Creatinine ratio, etc.), they are highly effective at detecting subtle relationships between these factors. Neural networks also tend to achieve high accuracy, especially when trained on large and diverse datasets, as they refine their internal weights to improve predictions.

However, neural network has challenges. A significant drawback is their dependence on data quality and quantity. For accurate diabetes prediction, the network requires comprehensive and well-labeled health data. If the data is incomplete or contains errors, the performance of the model will suffer. Furthermore, training neural networks can be computationally expensive, especially with deep architectures that have many hidden

layers. This means that training might require considerable computational resources and time [43].

### **Logistic Regression Algorithm**

Logistic regression is a statistical method for binary classification problems in which the outcome is usually one of two options (e.g., yes/no, true/false). It calculates the likelihood that a given input point belongs to a specific category using one or more predictor factors.

How a Logistic Regression works: Logistic regression determines if a given input belongs to a specific category. It applies a logistic function to the training data, transforming the input features into a probability score between 0 and 1. To perform classification, the model is trained on labeled data, using maximum likelihood estimation to determine the parameters of the logistic function. When a new instance is introduced, the algorithm calculates the probability of it belonging to the positive class using the learned function, then assigns the class with the highest probability [44].

### **Decision Tree Algorithm**

In decision tree technique, each path from the root to the leaves is described through a series of tests until a binary outcome is reached. A decision tree contains nodes that represent the properties being classified, making it easy to understand and analyze. A decision tree is simple and easy to understand and can be used to classify both categorical and numerical outcomes. However, the complexities of a decision tree can lead to incorrect decisions, especially as the number of layers increases. There are several types of decision tree algorithms such as ID3, C4.5, CART, and CHAID, each with their own advantages and disadvantages.

How a decision tree works: Building the tree: The process begins by identifying a set of data, then the data is divided into subsets based on the property values using algorithms such as ID3, C4.5, and CART to determine how to split the data.

Selecting properties: The properties that are most important for splitting the data are selected, usually using measures such as the acquired information or the Gini index. The goal is to reduce the uncertainty in the data after each split.

Iteration: The process continues to split the data until a state is reached where all the data is in one group or until a certain depth in the tree is reached. At this stage, the final results are assigned to each group.

Testing: After building the tree, its performance evaluated using a test dataset. The accuracy of the model is measured through comparing the predicted results with the actual results.

The decision tree is a powerful, accurate, and popular tool in many fields of data classification, as studies have shown that it outperforms many other algorithms in classification accuracy, achieving the highest accuracy of up to 99.93% in some applications as shown in figure 3.2. Various optimization techniques have also used to enhance the performance of the decision tree, leading to better results in many fields.[45]

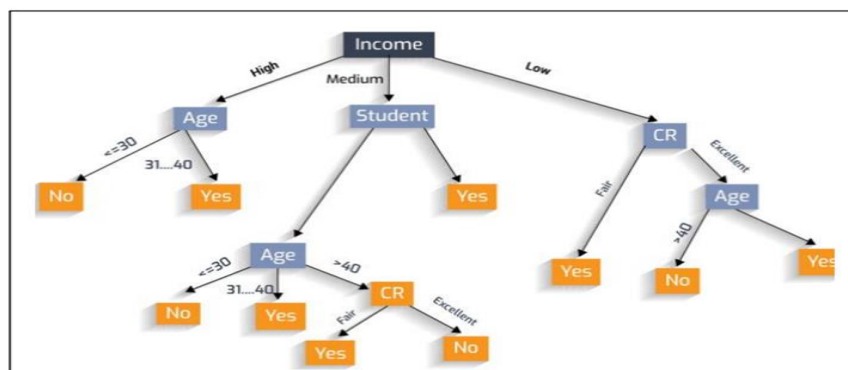


Figure 3.1: Example on Decision Tree

## Random Forest Algorithm

Random Forest algorithm is a powerful tree learning technique in Machine Learning, It works by creating many decision trees during the training phase as shown in figure 3.3. Each tree is constructed using a random subset of the data set to measure a random subset of features in each partition. This randomness creates variability among individual trees, reducing the risk of overfitting and improving accuracy of predictions In prediction. The algorithm combines the results of all trees either by voting (for

classification tasks) or by averaging (for regression tasks), This collaborative decision-making process supported by multiple trees with their insights leads to more stable and accurate results. Random forests are widely used for classification and regression functions because of their ability to handle complex data, reduce overfitting, and provide reliable forecasts in different environments.

Preparing Data for Random Forest Modeling: *Handling Missing Values*: addressing any missing values in the dataset and handle them.

*Encoding Categorical Variables*: Random Forest requires numerical inputs so categorical variables need to be encoded.

*Scaling and Normalization*: Random Forest is not sensitive to feature scaling but normalizing numerical features can contribute to a more efficient training process.

*Feature Selection*: Assess the importance of features within the dataset. Random Forest inherently provides a feature importance score, aiding in the selection of relevant features for model training.

*Addressing Imbalanced Data*: If dealing with imbalanced classes, implement techniques like adjusting class weights or employing resampling methods to ensure a balanced representation during training.[46]

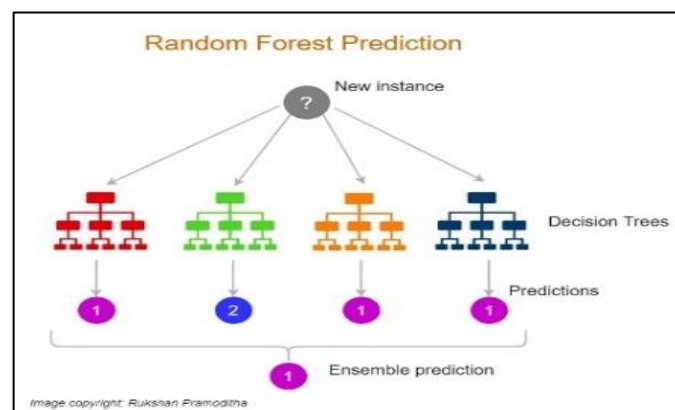


Figure 3.2: Random Forest Algorithm approach

### 3.1.3 Feature Selection

Feature selection is a critical step in dimensionality reduction with wide-ranging applications. It involves selecting relevant features while removing irrelevant, redundant, and noisy ones, with the goal of obtaining the best-performing subset of original features without transformation.[47] This process is essential for developing machine learning models, as it enables the design of simpler models with lower costs and reduced computational complexity, ultimately improving predictive performance [48]. Feature selection can be conducted in either supervised or unsupervised modes. In supervised feature selection, the primary objective is to reduce the dimensionality of the feature space while preserving the discriminative power of features for classification. Feature selection methods are categorized into three main types: filter, wrapper, and embedded approaches. In our model we will use the wrapper methods, also known as greedy algorithms, which construct models from scratch for each generated feature subset. The wrapper approach uses the model's prediction performance as the criterion to evaluate the efficiency of each subset. This approach considers the interactions between features. Wrapper methods evaluate features within the learning process, these methods assess the feature subsets to determine the best values according to their predictive power [49].

### 3.1.4 Evaluation

The performance of classification models is assessed using evaluation metrics to determine how accurately the model predicts outcomes. A confusion matrix is a table that summarizes the model's performance by displaying the counts of true positives, true negatives, false positives, and false negatives as shown in table 3.1. This matrix is essential for calculating various evaluation metrics, such as Precision, Recall, and F1-score [50].

Table 3.1: Confusion matrix

	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

**Recall (Sensitivity):** Recall, also known as sensitivity, is a metric defined as the ratio of correctly classified diabetic patients (TP) to the total number of patients who actually have the diabetes.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (1)$$

TP (True Positive): When the actual value is positive, the model correctly predicts it as positive. FN (False Negative): When model predicts a negative result, but the actual value is positive.

Recall measures how effectively patients are classified as having diabetes. A higher recall means most positive cases (True Positives) are accurately identified, even if it results in more false positives and lower overall accuracy. In contrast, low recall indicates many false negatives, implying that when a positive case is identified, it is more likely to be accurate [50]. Recall is important in diabetes prediction because it shows how well the model identifies actual diabetic patients. Failing to recognize these patients (false negatives) can lead to serious consequences, so maximizing recall helps ensure most cases are detected.

### **Precision:**

Precision measures the proportion of true positive predictions among all instances predicted as positive. This metric is particularly valuable when the concern for false positives is greater than that for false negatives, as it helps assess the reliability of the model [50].

$$\text{Precision} = \frac{TP}{TP + TN} \quad (2)$$

### **Accuracy:**

Accuracy is the most used measure of performance. It measures the ratio of correctly predicted instances to the total number of instances [50]. The formula for calculating accuracy is:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Where:

TP (True Positive): The number of instances correctly classified as positive.

TN (True Negative): The number of instances correctly classified as negative.

FP (False Positive): The number of instances incorrectly classified as positive.

FN (False Negative): The number of instances incorrectly classified as negative.

### **Specificity (true negative rate):**

Specificity (true negative rate) is the ability of a test to accurately identify individuals who do not have a condition, minimizing false positives.

*True Negatives (TN)*: These are the people who do not have diabetes, and the test correctly says they are healthy (no diabetes).

*False Positives (FP)*: These are the people who do not have diabetes, but the test incorrectly says they do (it falsely flags them as diabetic) [51].

$$Specificity = \frac{TN}{FP+TN} \quad (4)$$

### **F1-Score**

F1-Score It is a tool used in cases of classifying an unbalanced data set and it is the harmonic mean between precision and recall allowing for balance between them. It is calculated using this equation:

$$F1-Score = \frac{1}{\frac{1}{recall} + \frac{1}{precision}} \quad (5)$$

When precision and recall are equal, the F1 score will be high, indicating good performance. However, F1 score alone may have low interpretability. Therefore, it's recommended to use a combination of additional metrics alongside the F1 score to provide a more comprehensive evaluation of the model's performance [50].



## Kappa Statistic

The **Kappa Statistic** ( $\kappa$ ) is used to measure the degree of agreement between two raters or classifiers, correcting for the agreement that could occur by chance. It is particularly useful in classification tasks, as it provides a more accurate evaluation than simple percent agreement by accounting for random chance in the classification [52].

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (6)$$

$P_o$  = **Observed Agreement**: The proportion of instances where the predicted and true labels agree.

$$P_o = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$P_e$  = **Expected Agreement**: The proportion of agreement expected by chance.

$$P_e = \left( \frac{(TP + FP)(TP + FN)}{N^2} \right) + \left( \frac{(TN + FP)(TN + FN)}{N^2} \right) \quad (8)$$

## Relative Absolute Error (RAE)

Measures the ratio of the model's absolute error to the absolute error obtained when using the mean of the actual values as a predictor. It provides a relative measure of how well the model performs compared to a naive baseline (mean) [53].

Formula:

$$RAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{\sum_{i=1}^n |y_i - \bar{y}|} \quad (9)$$

- If **RAE = 0%**, the model is perfect (no error).
- If **RAE = 100%**, the model performs no better than simply using the mean as a predictor.
- Higher values indicate worse model performance.

### Mean Absolute Error (MAE)

is a common metric used to measure the accuracy of a model by calculating the average absolute difference between the actual values and the predicted values.

**Formula:**

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (10)$$

A lower MAE indicates better model performance, as it means predictions are closer to the actual values, while a higher MAE suggests larger errors. MAE is easy to interpret since it uses the same units as the target variable and doesn't penalize large errors more than small ones. However, it doesn't account for the direction of errors or give extra weight to larger mistakes like other metrics such as MSE [54].

### Root Mean Squared Error (RMSE)

It is a common metric used to evaluate the accuracy of a model's predictions, especially in machine learning and data analysis. It measures how close the predicted values are to the actual values. A lower RMSE value indicates a more accurate model [55].

How to calculate RMSE:

1. Calculate the difference between each actual value and its predicted value
2. Square each difference to eliminate negative values:  $e_i^2$ .
3. Calculate the mean of these squared differences:

$$MSE = \frac{1}{n} \sum \left( \underbrace{y - \hat{y}} \right)^2$$

4. Take the square root of this mean to get the RMSE:

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}} \quad (11)$$

The smaller the RMSE value, the more accurate the model.

### Root Relative Squared Error (RRSE)

It is used to evaluate the accuracy of predictive models, especially in regression problems. It measures how well a model's predictions match the actual values, relative to a baseline model (the mean of the actual values) [56].

#### Formula:

$$RRSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}} \quad (12)$$

#### Steps to Calculate:

1. Find the mean of actual values ( $\bar{y}$ ).
2. Compute the squared errors:  $(y_i - \hat{y}_i)^2$ .
3. Compute the squared deviations from the mean:  $(y_i - \bar{y})^2$ .
4. Divide the sum of squared errors by the sum of squared deviations.
5. Take the square root to get RRSE.

#### Interpretation:

- If  $RRSE=0$ , the model makes perfect predictions.
- If  $RRSE=1$ , the model performs the same as predicting the mean of the actual values.
- If  $RRSE>1$ , the model performs worse than just using the mean.

### Matthews Correlation Coefficient (MCC)

The Matthews Correlation Coefficient (MCC) is a statistical metric used to assess the performance of binary classification models. It provides a score between -1 and +1, where +1 represents a perfect classification, 0 indicates a prediction no better than random chance, and -1 signifies a completely incorrect classification [57].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (13)$$

### ROC (Receiver Operating Characteristic) Area (AUC-ROC)

The Receiver Operating Characteristic (ROC) curve is a graphical representation of a binary classifier's performance across different threshold values. The Area Under the Curve (AUC-ROC) measures the classifier's ability to distinguish between positive and negative classes. AUC values range from 0 to 1, where 1 indicates a perfect classifier, 0.5 represents random guessing, and 0 signifies completely incorrect predictions. AUC-ROC is particularly useful for evaluating models on imbalanced datasets, as it considers both the true positive rate (TPR) and false positive rate (FPR) [58]. The AUC can be mathematically expressed as:

$$AUC = \int_0^1 TPR(FPR) dFPR \quad (14)$$

Where:

TPR (True Positive Rate):

$$= \frac{TP}{TP+FN}$$

FPR (False Positive Rate):

$$= \frac{FP}{FP+TN}$$

### PRC (Precision-Recall Curve) Area

The Precision-Recall Curve (PRC) is a graphical representation used to evaluate the performance of classification models. It plots precision against recall at various threshold values. The PRC is valuable in scenarios where the positive class is rare, as it focuses on the performance of the model in predicting the minority class. The Area Under the Precision-Recall Curve (AUPRC) is a single scalar value that summarizes the model's performance across all thresholds. A higher AUPRC indicates better performance, as it shows how well the model can correctly identify positive instances while minimizing false positives and negatives [59].

$$AUPRC = \sum_{i=1}^{n-1} (R_{i+1} - R_i) \times \frac{P_i + P_{i+1}}{2} \quad (15)$$

Where  $P_i$  and  $R_i$  are the precision and recall values at each threshold.

## 3.2 Requirements Analysis

### 3.2.1 Description of Diabetes Data

Table 3.2 shows key health indicators, their normal ranges in adults, and levels that may signal potential risks of being a pre-diabetes or having diabetes in either of the stages [60].

Table 3.2: Data Description about health indicators, their normal ranges, and thresholds indicating diabetes risk.

Variable Name	Description	Average Range (Adults)	Range where it could be an indicator for diabetes (Adults)
<b>Urea</b>	Blood urea level, used to assess kidney function.	2.5 - 7.0 mmol/L	> 7.0 mmol/L indicates possible kidney issues, often seen in diabetics.
<b>Cr (Creatinine)</b>	Creatinine level in the blood, an indicator of kidney health.	60 - 110 $\mu$ mol/L (Men); 45 - 90 $\mu$ mol/L (Women)	> 110 $\mu$ mol/L (Men); > 90 $\mu$ mol/L (Women) may indicate diabetic nephropathy.
<b>HbA1c</b>	Hemoglobin A1c level, showing average blood sugar levels over the past 2-3 months (indicator of diabetes).	5.4% - 5.6%	$\geq$ 6.5% is used to diagnose diabetes.
<b>Chol (Cholesterol)</b>	Cholesterol level in the blood.	4.0 - 5.2 mmol/L	> 5.2 mmol/L increases the risk of cardiovascular issues in diabetics.
<b>TG (Triglycerides)</b>	Triglyceride level in the blood, associated with fat and metabolism.	1.0 - 1.7 mmol/L	> 2.2 mmol/L may indicate metabolic syndrome, a precursor to diabetes.
<b>HDL</b>	High-density lipoprotein (good cholesterol), helps remove excess cholesterol.	1.3 - 1.7 mmol/L	< 1.0 mmol/L (Men); < 1.3 mmol/L (Women) may increase risk of diabetes.
<b>LDL</b>	Low-density lipoprotein (bad cholesterol), linked to heart disease.	2.0 - 3.5 mmol/L	> 3.5 mmol/L increases cardiovascular risk, often elevated in diabetics.
<b>VLDL</b>	Very-low-density lipoprotein, linked to triglycerides and diabetes risk.	0.10 - 0.25 mmol/L	> 0.30 mmol/L may indicate insulin resistance or Type 2 diabetes risk.

<b>BMI</b>	Body Mass Index, a measure of body fat based on height and weight.	23 - 29 kg/m <sup>2</sup>	≥ 30 kg/m <sup>2</sup> (obesity) is a major risk factor for Type 2 diabetes.
------------	--	---------------------------	--

### 3.2.2 Data Flow Diagram

Figure 3.4 presents a workflow for developing a machine learning model to predict diabetes. It starts with defining the classification problem into Diabetic, Pre-Diabetic, and Non-Diabetic categories. Data from 1,000 patients at the Iraqi Medical City Hospital and Al-Kindy Teaching Hospital is used. After preprocessing the data various algorithms selected for training include Support Vector Machine (SVM), Decision Trees, Random Forest, Neural Networks, and Regression Models. The model is then tested and evaluated using metrics like accuracy, recall, precision, specificity, and F1-score to determine its effectiveness in diabetes prediction.

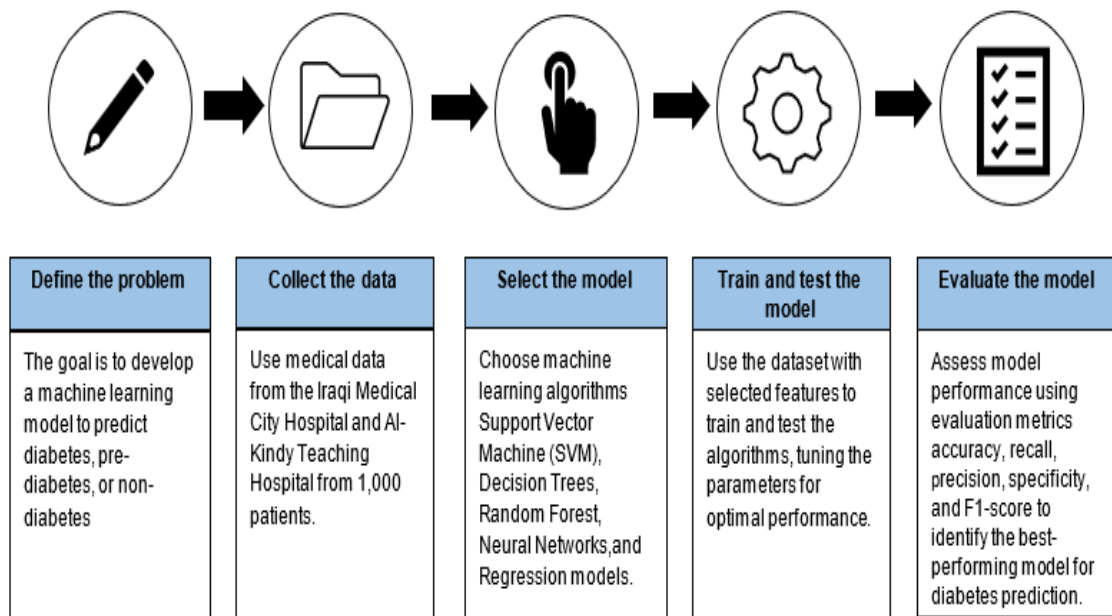


Figure 3.3 Machine Learning Workflow for diabetes prediction

## **Chapter 4:    System Design**

## 4.1 System Architecture

### 4.1.1 Hardware:

For optimal performance when using WEKA for machine learning tasks, a robust hardware setup is essential. The system must be equipped to handle computationally intensive processes, such as model training, feature selection, and validation.

**Central Processing Unit (CPU):** The CPU is fundamental for handling WEKA's most demanding tasks, including model training, cross-validation, and feature selection. A Quad-Core processor, such as the Intel Core i5 or AMD Ryzen 5, is suitable for small to medium datasets.

**Memory (RAM):** RAM plays a critical role in storing datasets and model parameters during processing. At least 8 GB of RAM is recommended to efficiently load datasets and run multiple models simultaneously.

**Storage (Disk):** In machine learning projects, disk space is essential for storing models, datasets, and temporary files generated during processing to ensure sufficient space and fast data access.

### 4.1.2 Software

#### WEKA Application

WEKA stands for Waikato Environment for Knowledge Analysis is an open-source tool developed for machine learning and data mining tasks. It supports various techniques including data preprocessing, classification, filtering, normalization, and transformation, making it a comprehensive tool for preparing data for analysis and developing predictive models [61].

Some of WEKA's features include wide range of machine learning algorithms: WEKA accommodates various types of algorithms for classification, clustering, regression, and more, making it versatile for different analytical needs.

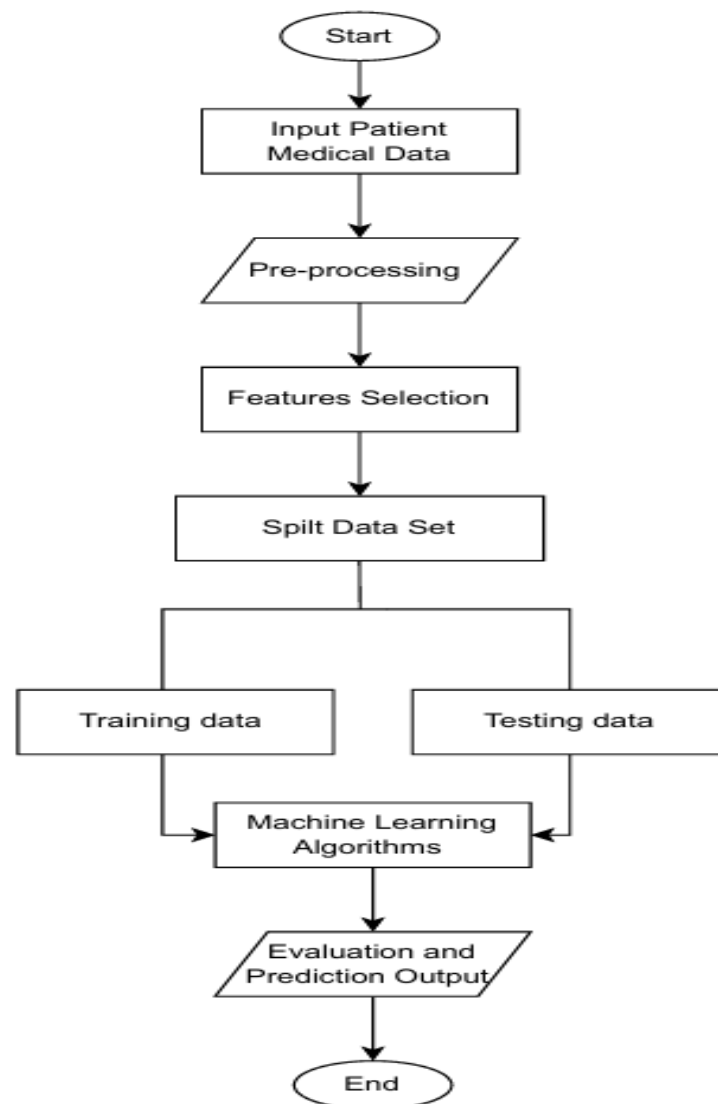
**Multi-format compatibility:** It supports several data formats, including ARFF, CSV, and JSON, making data import and manipulation straightforward.



Visualization tools: WEKA offers tools for visualizing datasets and analyzing model outputs, enabling users to gain deeper insights into the data and model performance.

Cross-platform compatibility: WEKA runs smoothly on different operating systems like Windows, macOS, and Linux.

Experiment management: Users can easily set up and run experiments to compare the performance of different models and algorithms, simplifying the evaluation process.



*Figure 4.1: Flowchart of the diabetes prediction model*

This flowchart in figure 4.1 represents a diabetes prediction workflow. It begins with collecting and pre-processing patient data, followed by feature selection to identify relevant attributes. The data is split into training and testing sets, and various machine learning algorithms are applied to the training data. The model’s performance is then evaluated using the testing set, with final predictions generated as output.

## 4.2 User Interface

### Dataset Attributes (Screenshots from WEKA)

These screenshots from figure 4.2 to figure 4.13 show the list of attributes loaded into WEKA, representing the medical dataset used for diabetes prediction.

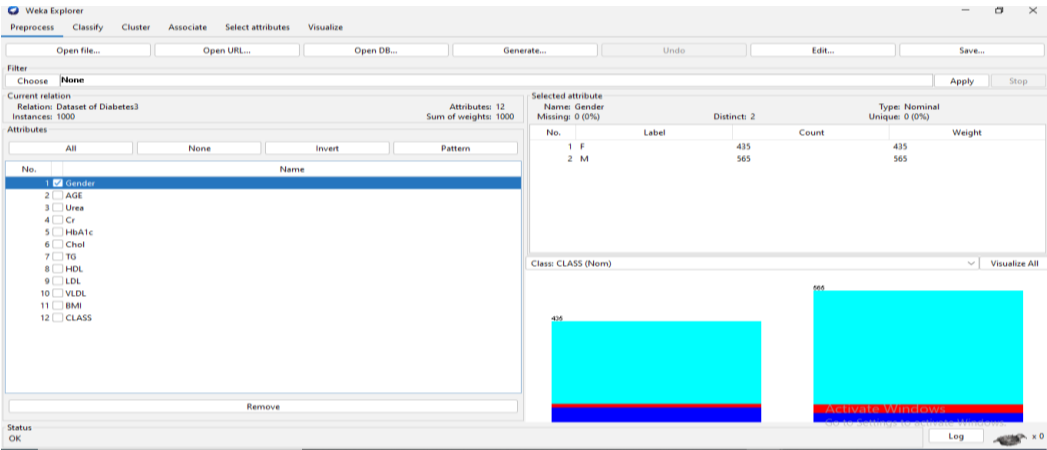


Figure 4.2: Gender Attribute in WEKA

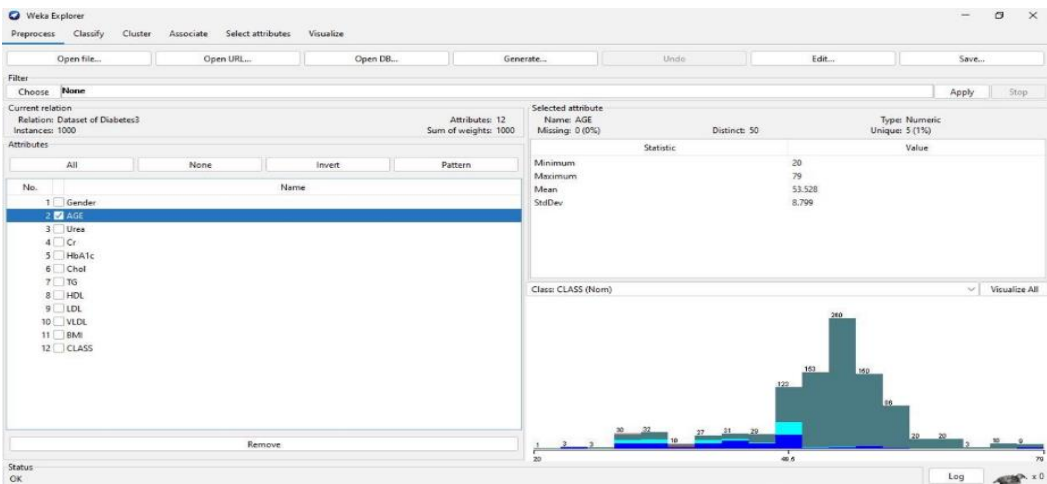


Figure 4.3: Age Attribute in WEKA

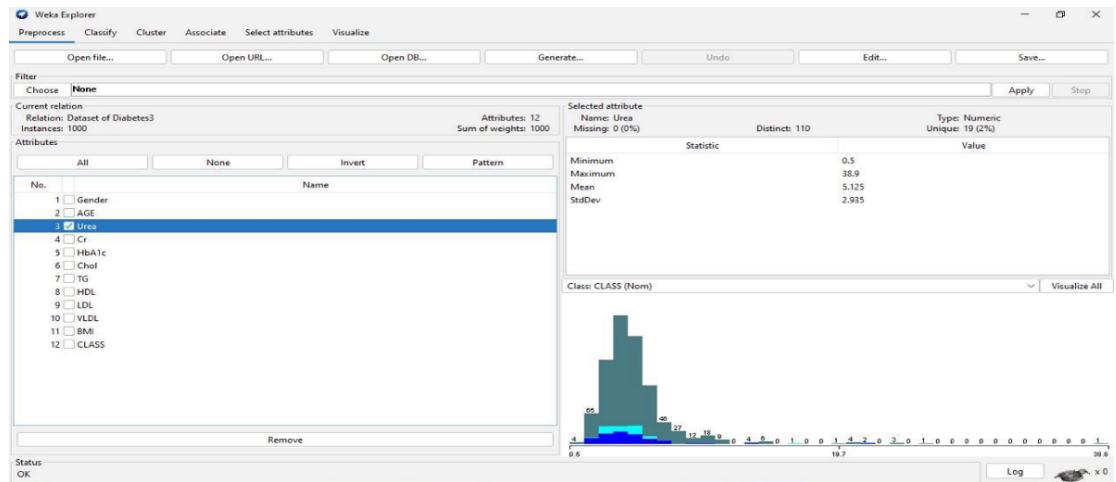


Figure 4.4: Urea Attribute in WEKA

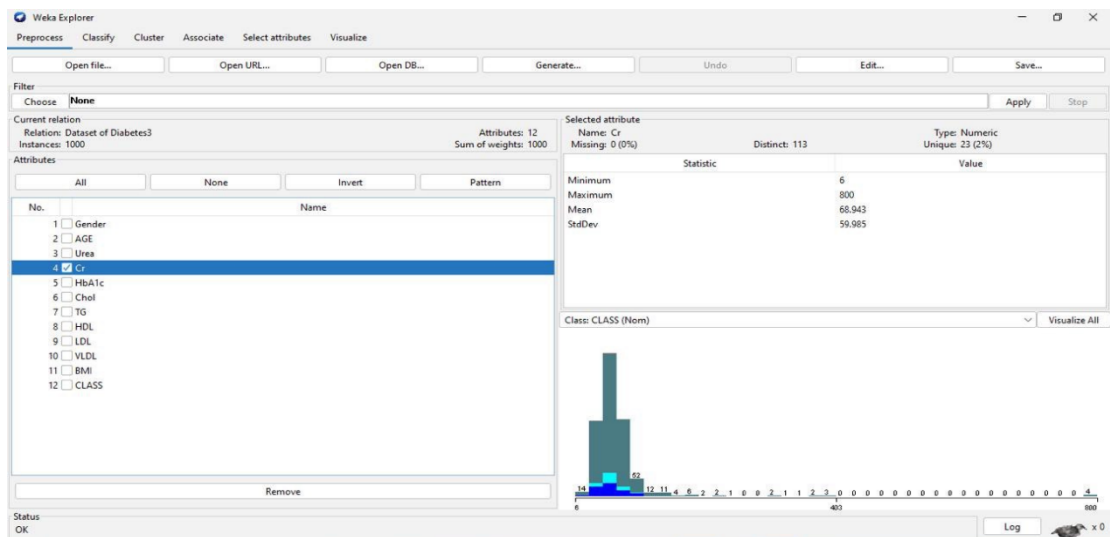


Figure 4.5: Cr Attribute in WEKA

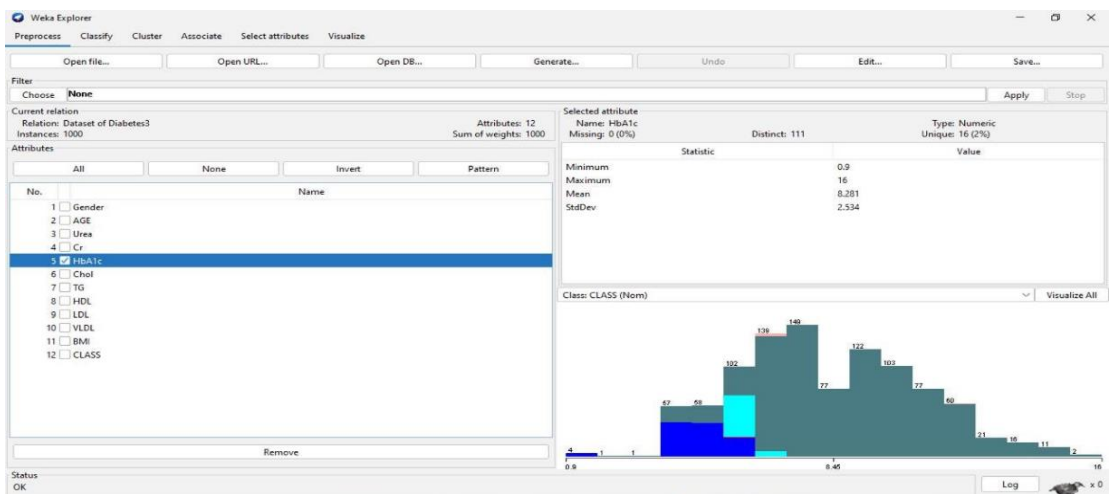


Figure 4.6: HbA1c Attribute in WEKA

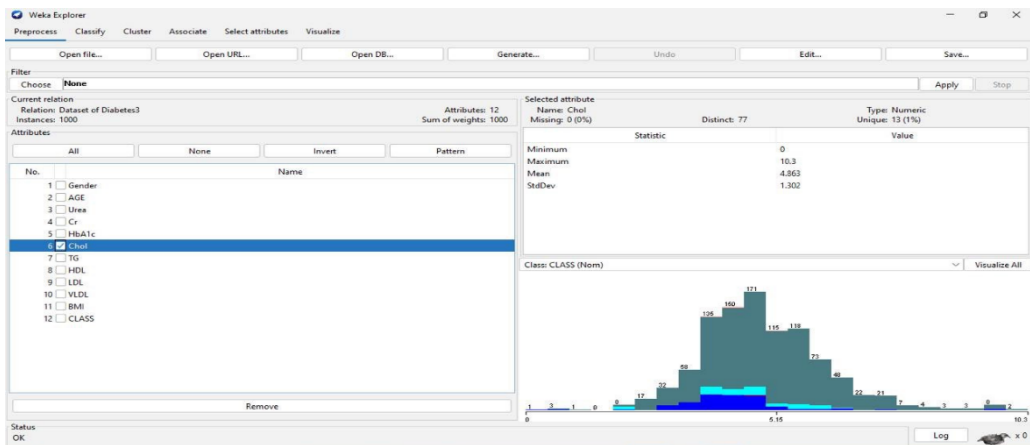


Figure 4.7: Chol Attribute in WEKA

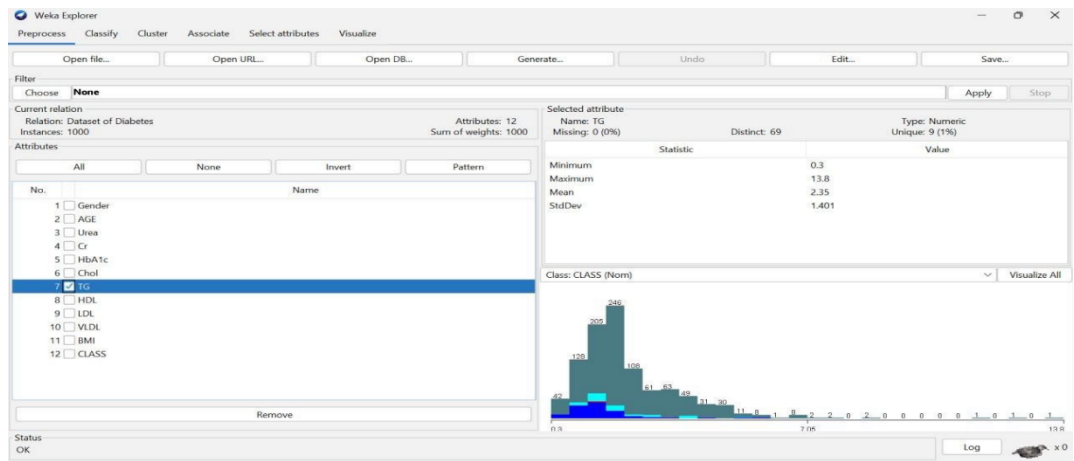


Figure 4.8: TG Attribute in WEKA

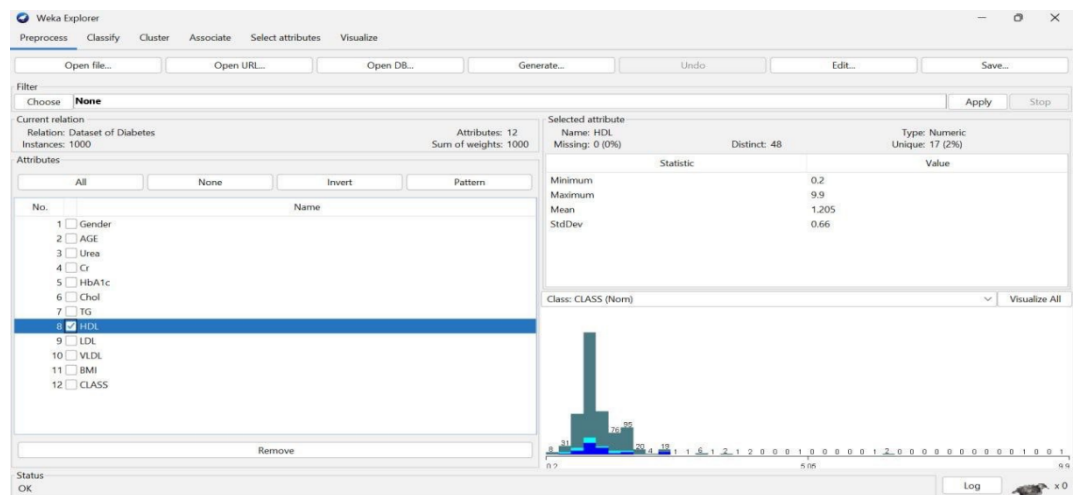


Figure 4.9: HDL Attribute in WEKA

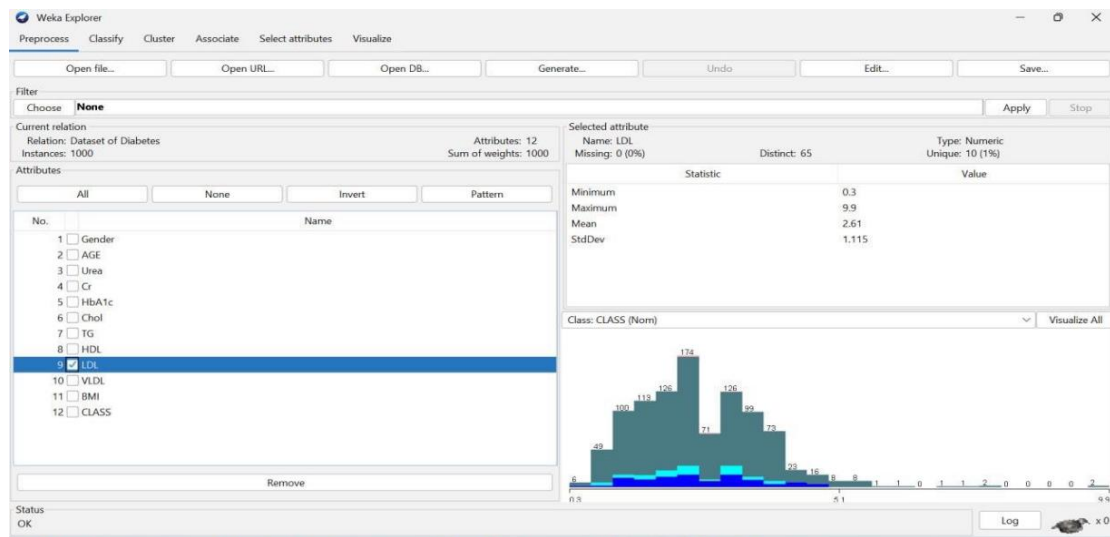


Figure 4.10: LDL Attribute in WEKA

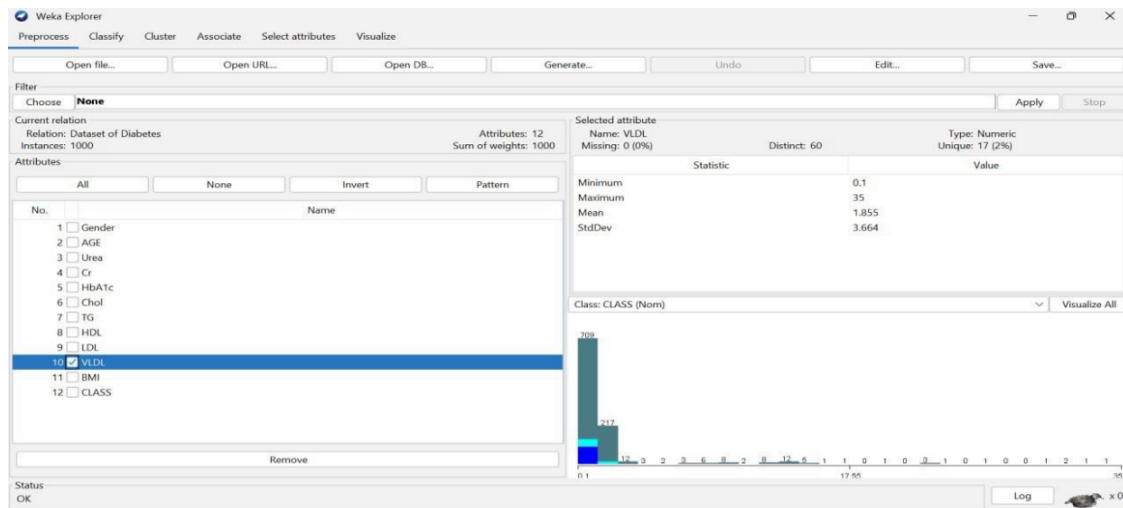


Figure 4.11: VLDL Attribute in WEKA

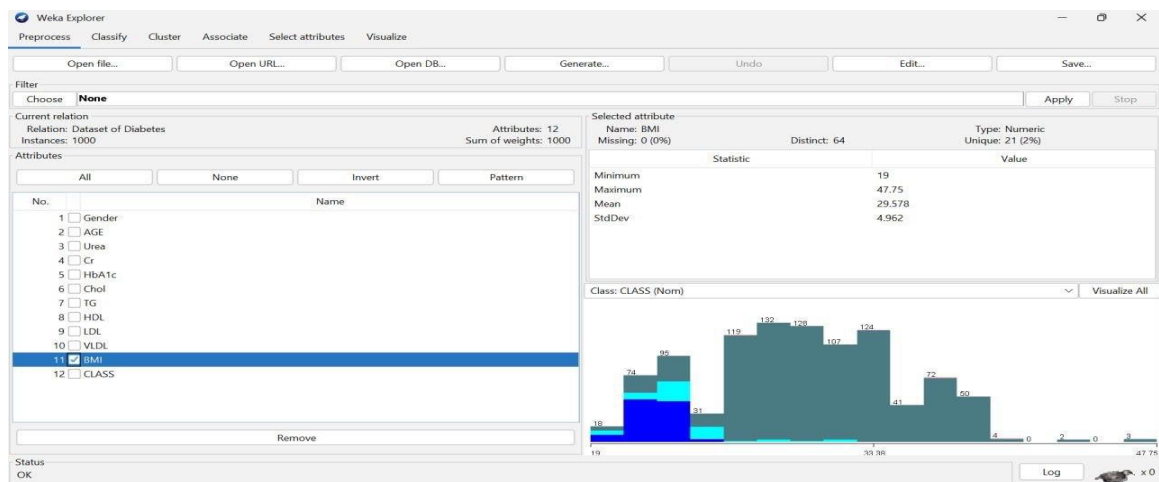


Figure 4.12: BMI Attribute in WEKA

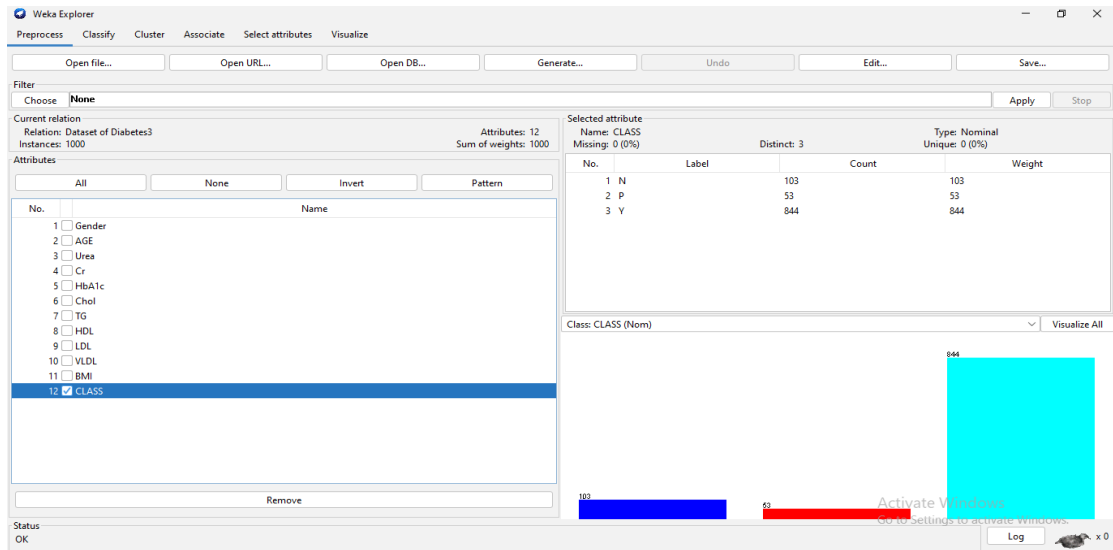


Figure 4.13: Class Attribute in WEKA

## **Chapter 5:   Implementation**

## 5.1 Implementation Requirements

### 1. Hardware Requirements

The system should be capable to handle machine learning workloads efficiently.

Therefore, the minimum requirements are:

Processor: Intel Core i5 (Quad-Core) / AMD Ryzen 5

RAM: 8 GB

Storage: 100 GB HDD/SSD

Graphics Card: Integrated GPU (for standard machine learning tasks)

Operating System: Windows 10/11.

### 2. Software Requirements

- WEKA for data processing and model training & testing

- Input CSV file
- Feature Selection: Using techniques like Wrapper Method
- Model Training and testing: Supporting algorithms like Random Forest and Decision Tree
- Evaluation Metrics: such as Accuracy, Precision, Recall, F1-Score

## 5.2 Implementation Details

This section describes the step-by-step implementation of the system using Weka, focusing on dataset processing, feature selection, model building, and performance evaluation.

**Data Preparation:** A dataset containing multiple features related to diabetes prediction was used. The dataset was imported into Weka and prepared for analysis and classification. The data was split into 70% for training and 30% for testing (300 test instances) using the Percentage Split method in Weka.

**Building the Model:** Five machine learning algorithms were applied to train predictive models. These algorithms are: (1) Naïve Bayes (NB), (2) Logistic Regression (LR), (3) Neural Network (NN), (4) Random Forest (RF), and (5) Decision Tree (DT). Each model was trained and evaluated using all available features before applying feature selection.



**Feature Selection Process:** To enhance model efficiency, the Wrapper Method was applied to identify the most important features for diabetes. Before applying the Wrapper method, models were trained using all available features. After applying the Wrapper method, four key features were selected across all models: Gender, HbA1c, Triglycerides (TG), Body Mass Index (BMI).

**Model Evaluation & Performance Analysis:** Each model was evaluated before and after feature selection using the following performance metrics: Correctly and incorrectly classified instances, Kappa Statistic, MAE & RMSE, Precision, Recall, and F-Measure as well as ROC and PRC Area

### 5.2.1 Applying the Five Algorithms

#### 1. Naïve Bayes

- Naïve Bayes with all features.

Table 5.1 Model Performance of Naïve Bayes

Metric	Value
Correctly Classified Instances	282 / 300 (94%)
Incorrectly Classified Instances	18 / 300 (6%)
Kappa Statistic	0.813
Mean Absolute Error (MAE)	0.0468
Root Mean Squared Error (RMSE)	0.1846
Relative Absolute Error (RAE)	25.0038%
Root Relative Squared Error (RRSE)	59.2213%

The table 5.1 presents the performance metrics of a Naïve Bayes model, which achieved 94% accuracy in classifying instances. It shows a strong Kappa statistic of 0.813 and relatively low error rates, with a Mean Absolute Error of 0.0468 and a Root Mean Squared Error of 0.1846.

Table 5.2 shows the performance metrics for Naïve Bayes across different classes (Non-Diabetes, Pre-Diabetes, Diabetes). The model demonstrates high precision and recall for all classes, with particularly strong results for Diabetes, achieving a Precision of 0.987 and an F-Measure of 0.963

Table 5.2 Performance of Naïve Bayes for each class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Non-Diabetes (N)	0.933	0.041	0.718	0.933	0.812	0.796	0.954	0.768
Pre-Diabetes (P)	0.950	0.014	0.826	0.950	0.884	0.877	0.951	0.936
Diabetes (Y)	0.940	0.060	0.987	0.940	0.963	0.810	0.946	0.980
Weighted Avg.	0.940	0.055	0.950	0.940	0.943	0.813	0.947	0.956

Table 5.2 shows the performance metrics for Naïve Bayes across different classes (Non-Diabetes, Pre-Diabetes, Diabetes). The model demonstrates high precision and recall for all classes, with particularly strong results for Diabetes, achieving a Precision of 0.987 and an F-Measure of 0.963

#### - Naïve Bayes with feature selection (Wrapper Method)

Table 5.3 Model Performance of Naïve Bayes

Metric	Value
Correctly Classified Instances	290 / 300 (96.67%)
Incorrectly Classified Instances	10/ 300 (8.33%)
Kappa Statistic	0.8882
Mean Absolute Error (MAE)	0.0401
Root Mean Squared Error (RMSE)	0.1484
Relative Absolute Error (RAE)	21.4252%
Root Relative Squared Error (RRSE)	47.5945%

Table 5.3 shows the performance metrics of a Naïve Bayes model after applying feature selection using the Wrapper Method. The model shows improved classification accuracy (96.67%) and a higher Kappa Statistic (0.8882), along with reduced error rates, indicating better predictive performance compared to the model without feature selection.

Table 5.4 Performance of Naïve Bayes for each class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Non-Diabetes (N)	0.967	0.026	0.806	0.967	0.879	0.868	0.987	0.865
Pre-Diabetes (P)	0.850	0.000	1.000	0.850	0.919	0.917	0.997	0.976
Diabetes (Y)	0.976	0.060	0.988	0.976	0.982	0.895	0.985	0.997
Weighted Avg.	0.967	0.053	0.970	0.967	0.967	0.894	0.986	0.983

Table 5.4 shows Naïve Bayes performance with feature selection using the Wrapper Method, highlighting improved precision, recall, and F-Measure across all classes. The model achieves strong overall metrics, with a weighted average F-Measure of 0.967 and ROC Area of 0.986.

## 2. Logistic Regression

### - Logistic Regression with all features.

Table 5.5 Model Performance of Logistic Regression

Metric	Value
Correctly Classified Instances	274 / 300 (91.3%)
Incorrectly Classified Instances	26 / 300 (8.7%)
Kappa Statistic	0.0676
Mean Absolute Error (MAE)	0.0438
Root Mean Squared Error (RMSE)	0.1582
Relative Absolute Error (RAE)	38.5941 %
Root Relative Squared Error (RRSE)	66.7036 %

Table 5.5 shows the performance of the logistic regression model when using all features, achieving a classification accuracy of 91.3% with an error rate of 8.7%. Other values such as MAE, RMSE, and RAE indicate the model's error magnitude, where lower values represent better performance.

Table 5.6 Performance of Logistic Regression for each class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Non-Diabetes (N)	0.851	0.973	0.771	0.794	0.833	0.759	0.030	0.833
Pre-Diabetes (P)	0.497	0.939	0.385	0.390	0.302	0.552	0.014	0.302
Diabetes (Y)	0.989	0.963	0.727	0.958	0.967	0.950	0.269	0.967
Weighted Avg.	0.944	0.961	0.709	0.907	0.913	0.904	0.230	0.913

Table 5.6 shows the model's performance for each class of data (Non-Diabetes, Pre-Diabetes, and Diabetes), displaying metrics such as True Positive Rate (TP Rate), False Positive Rate (FP Rate), Precision, and Recall, which assess the model's ability to correctly predict each class.

### - Logistic Regression with feature selection (Wrapper Method)

Table 5.7 Model Performance of Logistic Regression

Metric	Value
Correctly Classified Instances	276 / 300 (92%)
Incorrectly Classified Instances	24/ 300 (8%)
Kappa Statistic	0.7099
Mean Absolute Error (MAE)	0.0456
Root Mean Squared Error (RMSE)	0.1514
Relative Absolute Error (RAE)	39.3668%
Root Relative Squared Error (RRSE)	62.1233%

Table 5.7 shows the performance of the logistic regression model after applying the Wrapper method for feature selection. Reducing the number of input features led to a slight improvement in accuracy, increasing to 92% compared to 91.3%, while the error rate decreased, indicating enhanced model efficiency.

Table 5.8 Performance of Logistic Regression for each class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Non-Diabetes (N)	0.900	0.989	0.799	0.820	0.833	0.806	0.022	0.833
Pre-Diabetes (P)	0.434	0.941	0.107	0.087	0.050	0.333	0.007	0.050
Diabetes (Y)	0.993	0.967	0.622	0.944	0.976	0.914	0.451	0.976
Weighted Avg.	0.943	0.967	0.509	0.617	0.897	0.684	0.377	0.897

Table 5.8 shows the model's performance after feature selection for each class, where improvements in values such as TP Rate and MCC were observed, reflecting higher classification accuracy after eliminating unnecessary features.

### 3. Neural Networks

#### - Neural Networks with all features.

Table 5.9 Model Performance of Neural Networks

Metric	value
Correctly Classified Instances	288 /300 (%96)
Incorrectly Classified Instances	12 /300 (%4)
Kappa statistics	0.858
Mean Absolute Error (MAE)	0.0208
Root Mean Squared Error (RMSE)	0.1146
Relative Absolute Error (RAE)	%17.9707
Root Relative Squared Error (RRSE)	47.0283%

The table 5.9 summarizes the performance of a classification model, showing high accuracy and strong agreement based on the Kappa statistic. It also includes error metrics that indicate the model's reliability and the level of deviation in its predictions.

Table 5.10 Performance of Neural Networks for each class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Non-Diabetes (N)	0.840	0.007	0.913	0.840	0.875	0.865	0.997	0.969
Pre-Diabetes (P)	1.000	0.017	0.737	1.000	0.848	0.851	0.994	0.869
Diabetes (Y)	0.973	0.103	0.984	0.973	0.979	0.844	0.992	0.999
Weighted Avg.	0.963	0.091	0.967	0.963	0.964	0.846	0.992	0.990

The table 5.10 shows the model's classification performance for Non-Diabetes, Pre-Diabetes, and Diabetes, with high ROC and PRC Areas indicating strong accuracy. MCC, F-Measure, Recall, and Precision reflect balanced and reliable predictions across all classes.

#### - Neural Networks (NN) with feature selection (Wrapper Method)

Table 5.11 Model Performance of Neural Networks

Metric	Value
Correctly Classified Instances	293 / 300 (97.67%)
Incorrectly Classified Instances	7/ 300 (2.33%)
Kappa Statistic	0.9218
Mean Absolute Error (MAE)	0.0376
Root Mean Squared Error (RMSE)	0.1231
Relative Absolute Error (RAE)	20.094%
Root Relative Squared Error (RRSE)	39.4727%

The table 5.11 summarizes the model's performance, showing 97.67% accuracy with a Kappa statistic of 0.9218, indicating strong agreement. The error metrics (MAE, RMSE, RAE, and RRSE) suggest low prediction errors and good overall reliability.

Table 5.12 Performance of Neural Networks for each class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Non-Diabetes (N)	1.000	0.019	0.857	1.000	0.923	0.917	0.995	0.944
Pre-Diabetes (P)	0.900	0.000	1.000	0.900	0.947	0.945	0.993	0.949
Diabetes (Y)	0.980	0.040	0.992	0.980	0.986	0.919	0.994	0.999
Weighted Avg.	0.977	0.035	0.979	0.977	0.977	0.920	0.994	0.990

The table 5.12 presents classification performance metrics for Non-Diabetes, Pre-Diabetes, and Diabetes, with high PRC and ROC Areas indicating strong model accuracy. MCC, F-Measure, Recall, and Precision values show balanced and reliable predictions across all classes, with low false positive rates.

#### 4. Random Forest

##### - Random Forest with all features.

Table 5.13 Model Performance of Random Forest

Metric	Value
Correctly Classified Instances	264/300 (98%)
Incorrectly Classified Instances	6/ 300 (2%)
Kappa Statistic	0.9302
Mean Absolute Error (MAE)	0.0245
Root Mean Squared Error (RMSE)	0.0875
Relative Absolute Error (RAE)	21.1207%
Root Relative Squared Error (RRSE)	35.8785%

The table 5.13 summarizes the model's performance, showing 98% accuracy with a Kappa statistic of 0.9302, indicating strong agreement. The error metrics (MAE, RMSE, RAE, and RRSE) suggest low prediction errors and good overall reliability.

Table 5.14 Performance of Random Forest for each class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Non-Diabetes (N)	0.967	0.000	1.000	0.967	0.983	0.981	1.000	0.999
Pre-Diabetes (P)	0.850	0.000	1.000	0.850	0.919	0.917	0.999	0.986
Diabetes (Y)	0.996	0.098	0.980	0.996	0.988	0.928	0.986	0.995
Weighted Avg.	0.980	0.081	0.980	0.980	0.980	0.929	0.987	0.991

The table 5.14 presents the Random Forest model's performance by class. The results show high true positive rates (TP Rate) across all classes, indicating strong prediction accuracy. The precision, recall, and F-measure values are consistently high, confirming the model's reliability. The ROC and PRC areas are close to 1, demonstrating excellent discrimination ability. Overall, these metrics reflect a robust and effective model.

##### - Random Forest (RF) with feature selection (Wrapper Method)

Table 5.15 Model Performance of Random Forest

Metric	Value
Correctly Classified Instances	293/ 300 (97.67%)
Incorrectly Classified Instances	7/ 300 (2.33%)
Kappa Statistic	0.9148
Mean Absolute Error (MAE)	0.0234
Root Mean Squared Error (RMSE)	0.1039
Relative Absolute Error (RAE)	12.49%
Root Relative Squared Error (RRSE)	33.31%

The table 5.15 summarizes the performance of the Random Forest model, showing an accuracy of 97.67% with a Kappa statistic of 0.914, indicating strong agreement. The error metrics (MAE, RMSE, RAE, and RRSE) demonstrated low prediction errors and good overall reliability.

Table 5.16 Performance of Random Forest for each class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Non-Diabetes (N)	0.833	0.000	1.000	0.833	0.909	0.905	1.000	0.996
Pre-Diabetes (P)	0.900	0.000	1.000	0.900	0.947	0.945	0.999	0.991
Diabetes (Y)	1.000	0.140	0.973	1.000	0.986	0.915	0.999	1.000
Weighted Avg.	0.977	0.117	0.977	0.977	0.976	0.916	0.999	0.999

The table 5.16 presents the performance of the Random Forest model by class. It shows high true positive rates (TP Rate) across all classes, indicating strong prediction accuracy. The precision, recall, and F-measure values are consistently high, confirming the model's reliability. Overall, these metrics illustrated a robust and effective model.

## 5. Decision tree

### - Decision tree with all features.

Table 5.17 Model Performance of Decision Tree

Metric	Value
Correctly Classified Instances	296/300 (99%)
Incorrectly Classified Instances	4/ 300 (1%)
Kappa Statistic	0.955
Mean Absolute Error (MAE)	0.006
Root Mean Squared Error (RMSE)	0.0689
Relative Absolute Error (RAE)	5.2156 %
Root Relative Squared Error (RRSE)	28.2584 %

Table 5.17 shows the Decision Tree model's performance demonstrating 99% accuracy and a Kappa statistic of 0.955. The model exhibits minimal errors, with a Mean Absolute Error (MAE) of 0.006 and a Root Mean Squared Error (RMSE) of 0.0689, indicating high predictive reliability and precision.

Table 5.18 Performance of Decision Tree for each class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Non-Diabetes (N)	1.000	0.004	0.968	1.000	0.984	0.982	1.000	0.999
Pre-Diabetes (P)	0.950	0.000	1.000	0.950	0.974	0.973	0.975	0.953
Diabetes (Y)	0.992	0.039	0.992	0.992	0.992	0.953	0.978	0.991
Weighted Avg	0.987	0.033	0.987	0.987	0.987	0.954	0.978	0.983

Table 5.18 shows the Decision Tree model's performance by class. Non-Diabetes has a perfect TP Rate (1.000) and ROC Area (1.000). Pre-Diabetes achieves a Precision of 1.000 and Recall of 0.950, while Diabetes has high Precision and Recall (0.992). The weighted averages reflect strong overall performance with an F-Measure of 0.987 and ROC Area of 0.978. The visualization of the decision tree is shown in figure 5.1.

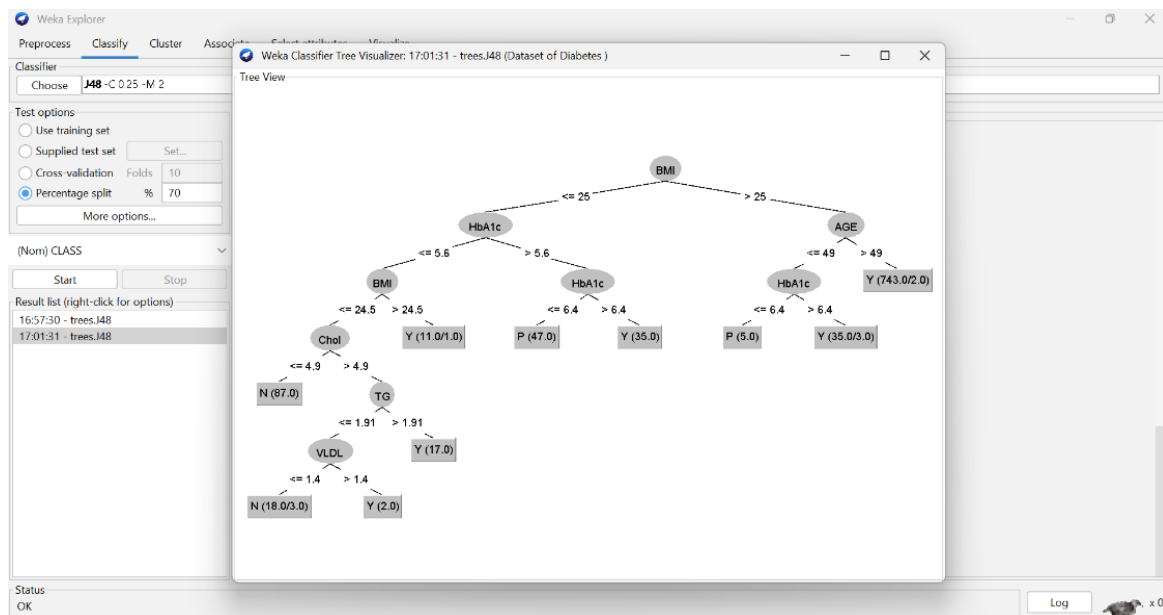


Figure 5.1: Decision Tree visualization for all features

- Decision tree with feature selection from Wrapper Method:



Table 5.19 Model Performance of Decision Tree

Metric	Value
Correctly Classified Instances	295 (98.33%)
Incorrectly Classified Instances	5 (1.67%)
Kappa Statistic	0.9402
Mean Absolute Error (MAE)	0.0208
Root Mean Squared Error (RMSE)	0.1005
Relative Absolute Error (RAE)	11.09%
Root Relative Squared Error (RRSE)	32.22 %

Table 5.19 presents the performance of the Decision Tree model with feature selection using the Wrapper Method. It correctly classifies 98.33% of instances, with a Kappa statistic of 0.9402. The model shows low error rates, with a MAE of 0.0208, RMSE of 0.1005, RAE of 11.09%, and RRSE of 32.22%.

Table 5.20 Performance of Decision Tree for each class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Non-Diabetes (N)	0.900	0.000	1.000	0.900	0.947	0.943	0.983	0.962
Pre-Diabetes (P)	0.900	0.000	1.000	0.900	0.947	0.945	0.959	0.908
Diabetes (Y)	1.000	0.100	0.980	1.000	0.990	0.939	0.961	0.985
Weighted Avg.	0.983	0.083	0.984	0.983	0.983	0.940	0.963	0.977

Table 5.20 shows the performance of the Decision Tree model by class. Non-Diabetes has high Precision (1.000) and Recall (0.900), while Pre-Diabetes also shows strong performance with Precision and Recall of 1.000 and 0.900, respectively. Diabetes achieves a perfect TP Rate (1.000) with Precision of 0.980. The weighted averages indicate solid overall performance with an F-Measure of 0.983 and ROC Area of 0.963. The visualization of the decision tree after applying feature selection is shown in figure 5.2.

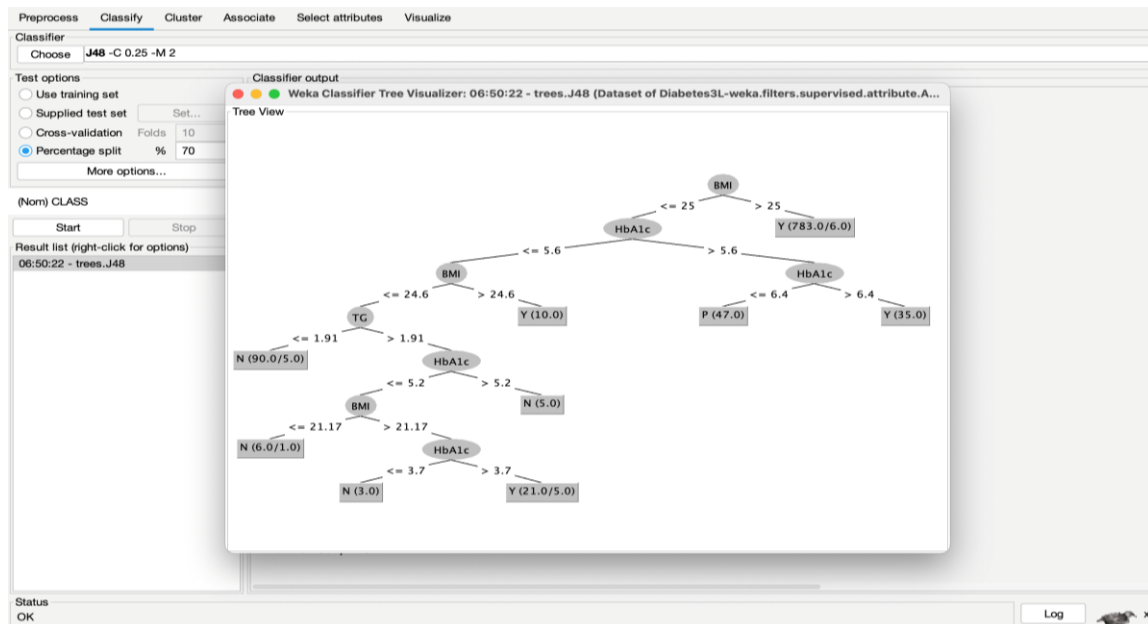
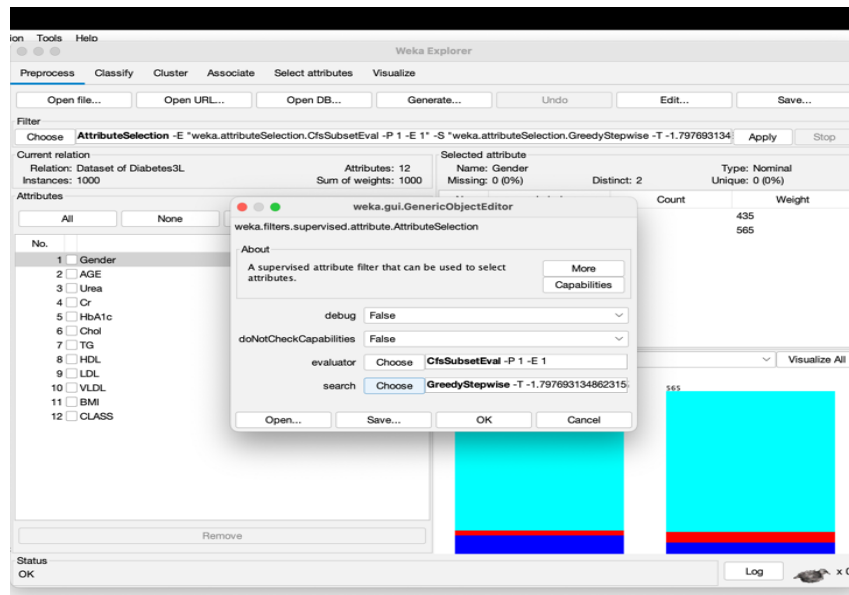


Figure 5.2: Decision Tree visualization for four features

### 5.3 I/O Screens

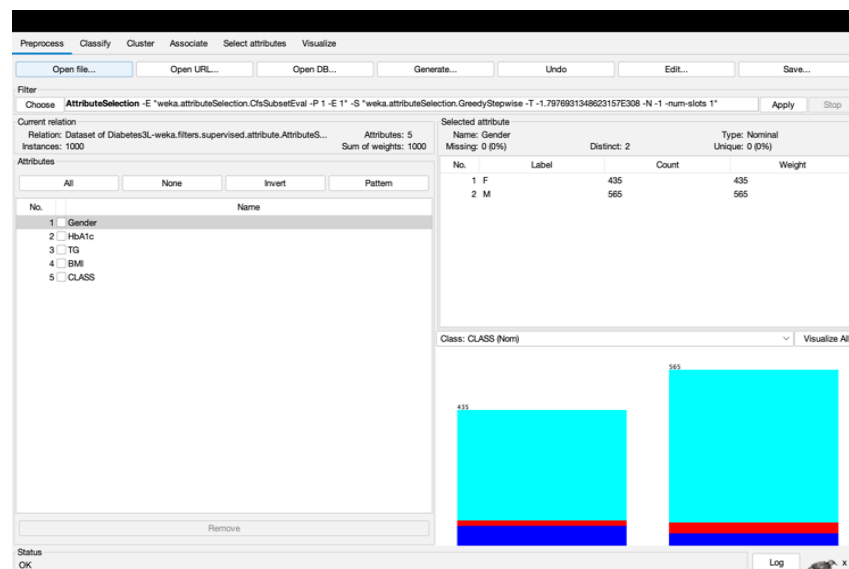
Input/Output	Screen	Description
Input	<p>The screenshot shows the WEKA Explorer interface. The 'Preprocess' tab is active. The 'Current relation' is 'Diabetes3L' with 1000 instances and 12 attributes. The 'Selected attribute' is 'Gender' with 2 distinct values. The 'Class' is 'CLASS (Nom)'. The interface displays a list of attributes (AGE, Urea, Cr, HbA1c, Cholesterol, TG, HDL, LDL, VLDL, BMI, CLASS) and a bar chart showing the distribution of the 'CLASS' attribute.</p>	<p>This screenshot shows the WEKA Explorer after the dataset has been uploaded. It displays the dataset attributes, before any preprocessing or feature selection is applied.</p>

## Input



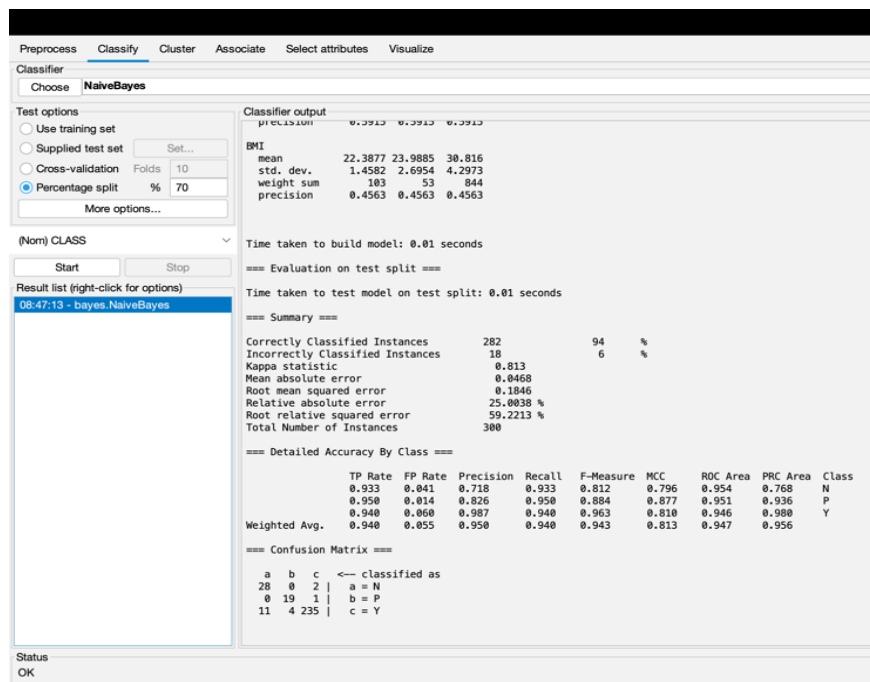
This screenshot shows the attribute selection process in WEKA using the Wrapper method to identify the most relevant features for diabetes prediction.

## Output



This screenshot shows the results of attribute selection using the wrapper method, identifying the most relevant features for diabetes prediction: **Gender, HbA1c, TG, BMI, and CLASS,**

## Output



Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier  
Choose **NaiveBayes**

Test options  
☐ Use training set  
☐ Supplied test set Set...  
☐ Cross-validation Folds 10  
☒ Percentage split % 70  
 More options...

(Nom) CLASS

Start Stop

Result list (right-click for options)  
08:47:13 - bayes.NaiveBayes

Status  
OK

Classifier output

precision 0.9913 0.9913 0.9913

BMI  
 mean 22.3877 23.9885 30.816  
 std. dev. 1.4582 2.6954 4.2973  
 weight sum 183 53 844  
 precision 0.4563 0.4563 0.4563

Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances 282 94 %  
 Incorrectly Classified Instances 18 6 %  
 Kappa statistic 0.813  
 Mean absolute error 0.0468  
 Root mean squared error 0.1846  
 Relative absolute error 25.0038 %  
 Root relative squared error 59.2213 %  
 Total Number of Instances 300

=== Detailed Accuracy By Class ===

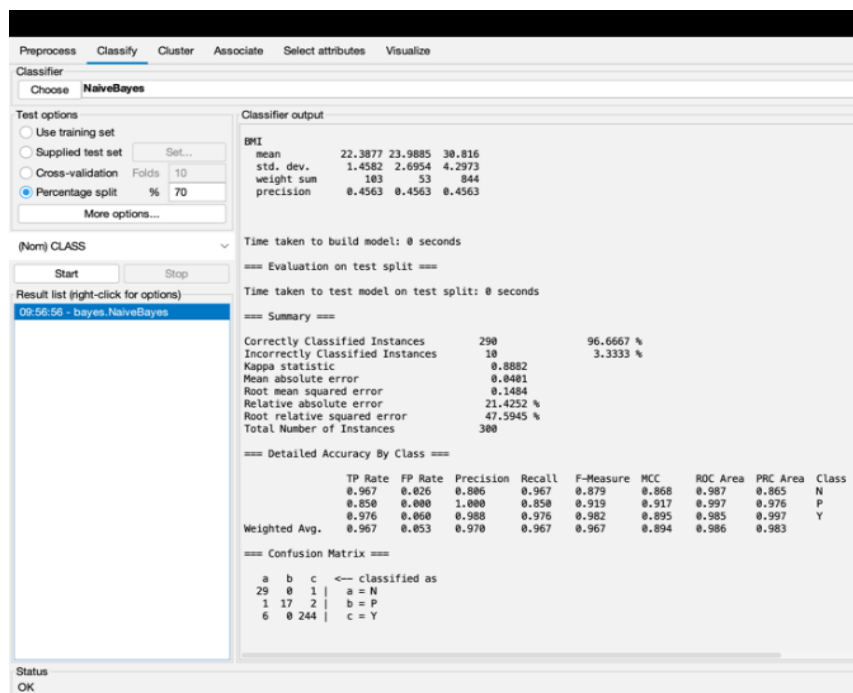
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.933	0.041	0.718	0.933	0.812	0.796	0.954	0.768	N
	0.950	0.014	0.826	0.950	0.884	0.877	0.951	0.936	P
	0.940	0.060	0.987	0.940	0.963	0.810	0.946	0.980	Y
Weighted Avg.	0.940	0.055	0.950	0.940	0.943	0.813	0.947	0.956	

=== Confusion Matrix ===

	a	b	c	<-- classified as
28	0	2		a = N
0	19	1		b = P
11	4	235		c = Y

This screenshot displays the classification results in WEKA using the **NaïveBayes** algorithm with **all** features.

## Output



Preprocess **Classify** Cluster Associate Select attributes Visualize

Classifier  
Choose **NaiveBayes**

Test options  
☐ Use training set  
☐ Supplied test set Set...  
☐ Cross-validation Folds 10  
☒ Percentage split % 70  
 More options...

(Nom) CLASS

Start Stop

Result list (right-click for options)  
09:56:56 - bayes.NaiveBayes

Status  
OK

Classifier output

BMI  
 mean 22.3877 23.9885 30.816  
 std. dev. 1.4582 2.6954 4.2973  
 weight sum 183 53 844  
 precision 0.4563 0.4563 0.4563

Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances 290 96.6667 %  
 Incorrectly Classified Instances 10 3.3333 %  
 Kappa statistic 0.8882  
 Mean absolute error 0.0401  
 Root mean squared error 0.1484  
 Relative absolute error 21.4252 %  
 Root relative squared error 47.5945 %  
 Total Number of Instances 300

=== Detailed Accuracy By Class ===

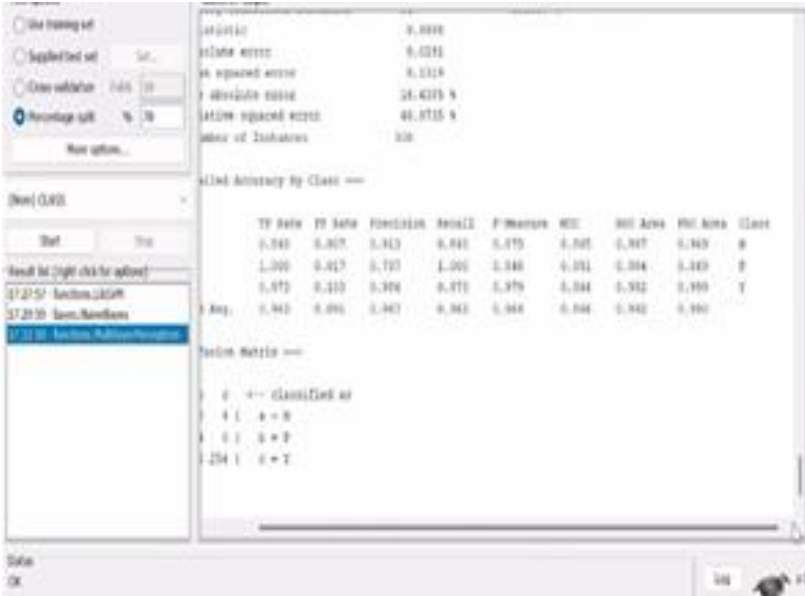
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.967	0.026	0.806	0.967	0.879	0.868	0.987	0.865	N
	0.850	0.000	1.000	0.850	0.919	0.917	0.997	0.976	P
	0.976	0.060	0.988	0.976	0.982	0.895	0.985	0.997	Y
Weighted Avg.	0.967	0.053	0.970	0.967	0.967	0.894	0.986	0.983	

=== Confusion Matrix ===

	a	b	c	<-- classified as
29	0	1		a = N
1	17	2		b = P
6	0	244		c = Y

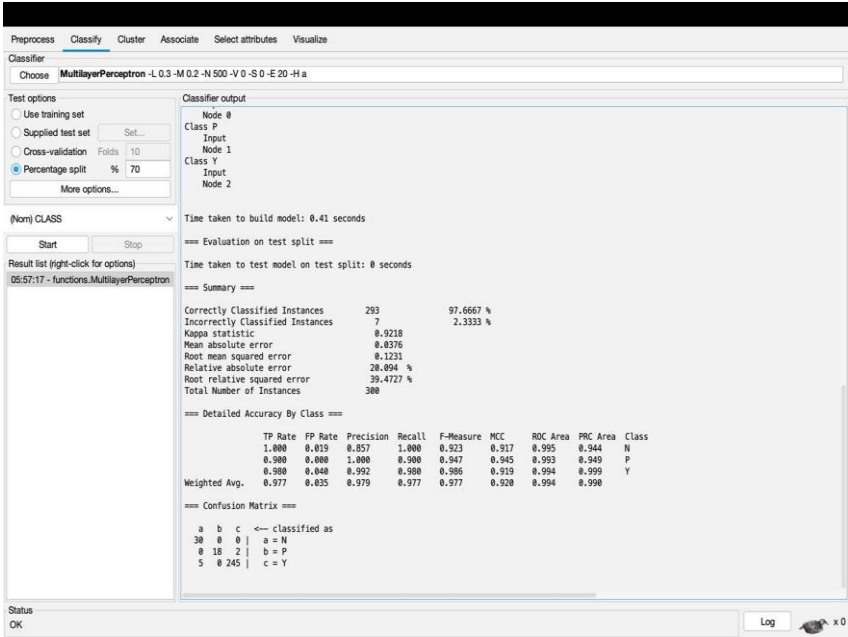
This screenshot displays the classification results in WEKA using the **NaïveBayes** algorithm with feature selection using **Wrapper** method.

output

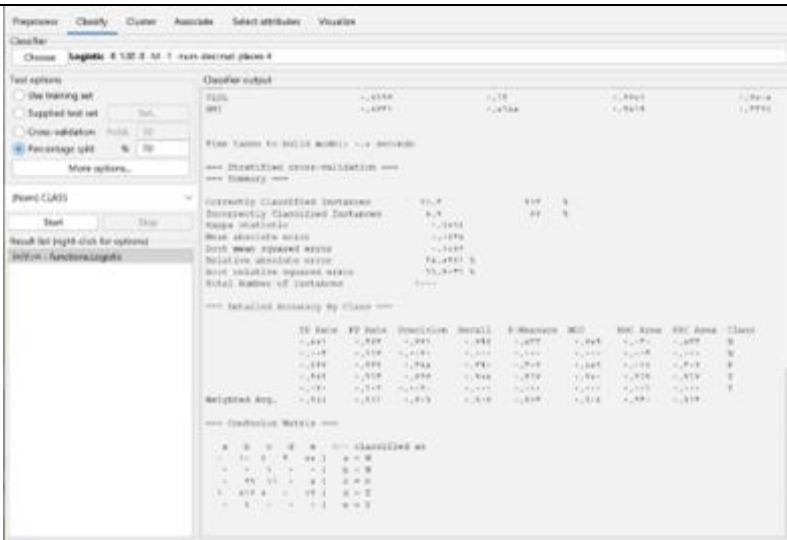
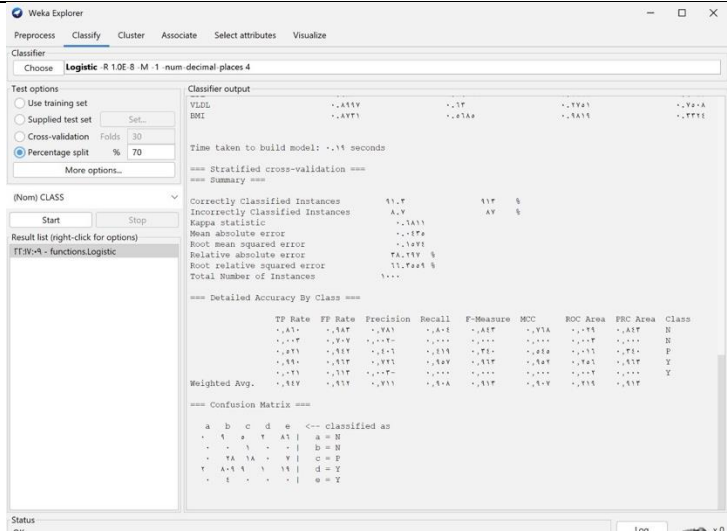
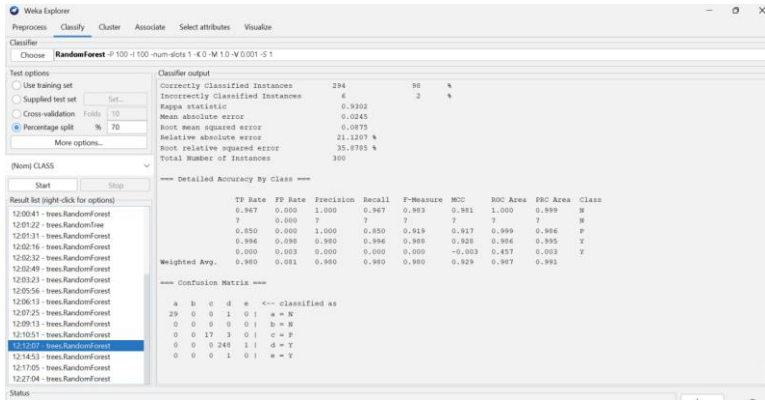


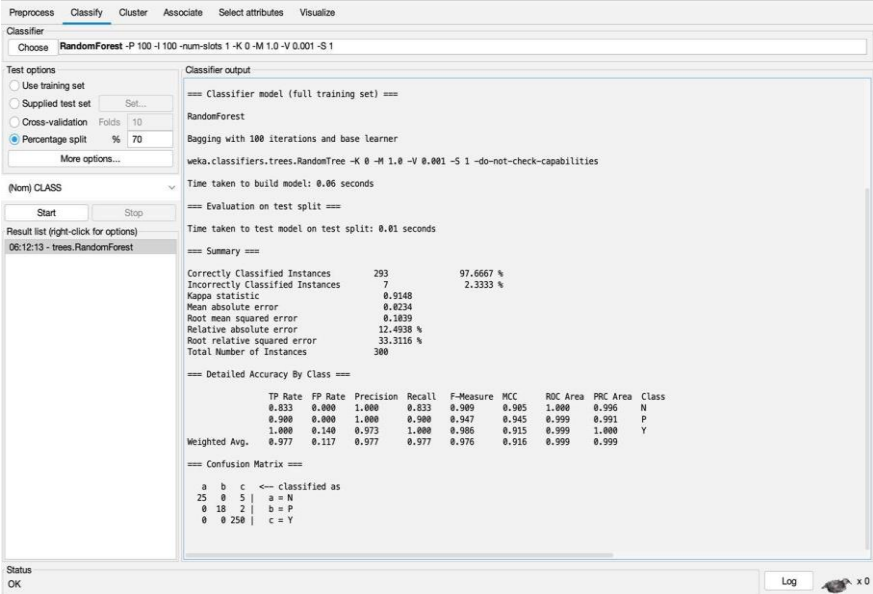
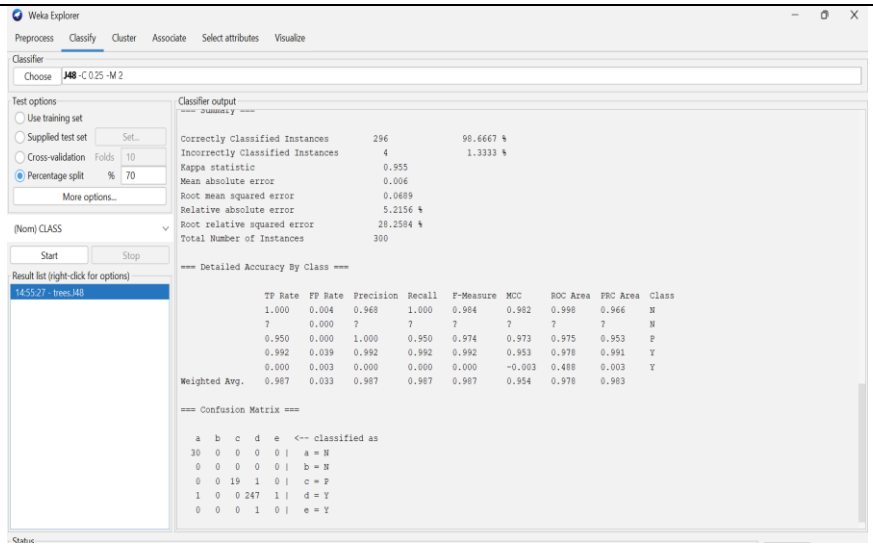
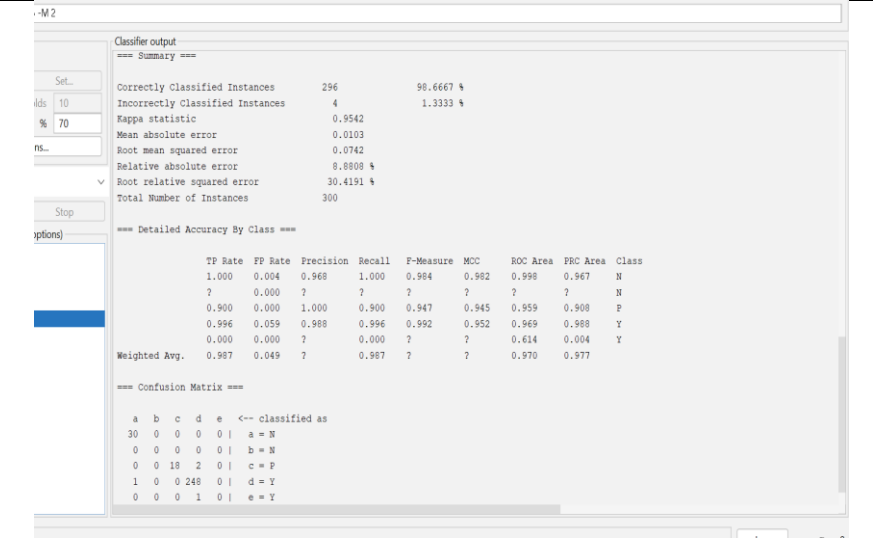
This screenshot displays the classification results in WEKA using the **Neural Networks** algorithm with **all** features.

Output



This screenshot displays the classification results in WEKA using the **Neural Networks** algorithm with feature selection using **Wrapper method**.

Output	 <p>The screenshot shows the WEKA Classifier window with the Logistic Regression algorithm selected. The classifier output displays the following metrics:</p> <ul style="list-style-type: none"><li>Time taken to build model: 0.11 seconds</li><li>Stratified cross-validation summary: <table><tr><th></th><th>TP Rate</th><th>FP Rate</th><th>Precision</th><th>Recall</th><th>F-Measure</th><th>MCC</th><th>ROC Area</th><th>PRC Area</th><th>Class</th></tr><tr><td>Correctly Classified Instances</td><td>99.9</td><td>0.0</td><td>1.0</td><td>99.9</td><td>99.9</td><td>1.0</td><td>1.0</td><td>1.0</td><td>N</td></tr><tr><td>Incorrectly Classified Instances</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>E</td></tr><tr><td>Kappa statistic</td><td>0.999</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Mean absolute error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Root mean squared error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Relative absolute error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Root relative squared error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Total Number of Instances</td><td>1000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table></li><li>Detailed Accuracy By Class: <table><tr><th></th><th>TP Rate</th><th>FP Rate</th><th>Precision</th><th>Recall</th><th>F-Measure</th><th>MCC</th><th>ROC Area</th><th>PRC Area</th><th>Class</th></tr><tr><td>Correctly Classified Instances</td><td>99.9</td><td>0.0</td><td>1.0</td><td>99.9</td><td>99.9</td><td>1.0</td><td>1.0</td><td>1.0</td><td>N</td></tr><tr><td>Incorrectly Classified Instances</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>E</td></tr><tr><td>Kappa statistic</td><td>0.999</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Mean absolute error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Root mean squared error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Relative absolute error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Root relative squared error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Total Number of Instances</td><td>1000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table></li><li>Confusion Matrix: <table><tr><th></th><th>a</th><th>b</th><th>c</th><th>d</th><th>e</th><th>classified as</th></tr><tr><td>a</td><td>999</td><td>0</td><td>0</td><td>0</td><td>0</td><td>a = N</td></tr><tr><td>b</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>b = N</td></tr><tr><td>c</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>c = E</td></tr><tr><td>d</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>d = Y</td></tr><tr><td>e</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>e = Y</td></tr></table></li></ul>		TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	Correctly Classified Instances	99.9	0.0	1.0	99.9	99.9	1.0	1.0	1.0	N	Incorrectly Classified Instances	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	E	Kappa statistic	0.999									Mean absolute error	0.000									Root mean squared error	0.000									Relative absolute error	0.000									Root relative squared error	0.000									Total Number of Instances	1000										TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	Correctly Classified Instances	99.9	0.0	1.0	99.9	99.9	1.0	1.0	1.0	N	Incorrectly Classified Instances	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	E	Kappa statistic	0.999									Mean absolute error	0.000									Root mean squared error	0.000									Relative absolute error	0.000									Root relative squared error	0.000									Total Number of Instances	1000										a	b	c	d	e	classified as	a	999	0	0	0	0	a = N	b	0	0	0	0	0	b = N	c	0	0	0	0	0	c = E	d	0	0	0	0	0	d = Y	e	0	0	0	0	0	e = Y	This screenshot displays the classification results in WEKA using the <b>Logistic Regression</b> algorithm with <b>all features</b> .
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class																																																																																																																																																																																																																							
Correctly Classified Instances	99.9	0.0	1.0	99.9	99.9	1.0	1.0	1.0	N																																																																																																																																																																																																																							
Incorrectly Classified Instances	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	E																																																																																																																																																																																																																							
Kappa statistic	0.999																																																																																																																																																																																																																															
Mean absolute error	0.000																																																																																																																																																																																																																															
Root mean squared error	0.000																																																																																																																																																																																																																															
Relative absolute error	0.000																																																																																																																																																																																																																															
Root relative squared error	0.000																																																																																																																																																																																																																															
Total Number of Instances	1000																																																																																																																																																																																																																															
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class																																																																																																																																																																																																																							
Correctly Classified Instances	99.9	0.0	1.0	99.9	99.9	1.0	1.0	1.0	N																																																																																																																																																																																																																							
Incorrectly Classified Instances	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	E																																																																																																																																																																																																																							
Kappa statistic	0.999																																																																																																																																																																																																																															
Mean absolute error	0.000																																																																																																																																																																																																																															
Root mean squared error	0.000																																																																																																																																																																																																																															
Relative absolute error	0.000																																																																																																																																																																																																																															
Root relative squared error	0.000																																																																																																																																																																																																																															
Total Number of Instances	1000																																																																																																																																																																																																																															
	a	b	c	d	e	classified as																																																																																																																																																																																																																										
a	999	0	0	0	0	a = N																																																																																																																																																																																																																										
b	0	0	0	0	0	b = N																																																																																																																																																																																																																										
c	0	0	0	0	0	c = E																																																																																																																																																																																																																										
d	0	0	0	0	0	d = Y																																																																																																																																																																																																																										
e	0	0	0	0	0	e = Y																																																																																																																																																																																																																										
Output	 <p>The screenshot shows the WEKA Explorer window with the Logistic Regression algorithm selected. The classifier output displays the following metrics:</p> <ul style="list-style-type: none"><li>Time taken to build model: 0.11 seconds</li><li>Stratified cross-validation summary: <table><tr><th></th><th>TP Rate</th><th>FP Rate</th><th>Precision</th><th>Recall</th><th>F-Measure</th><th>MCC</th><th>ROC Area</th><th>PRC Area</th><th>Class</th></tr><tr><td>Correctly Classified Instances</td><td>99.9</td><td>0.0</td><td>1.0</td><td>99.9</td><td>99.9</td><td>1.0</td><td>1.0</td><td>1.0</td><td>N</td></tr><tr><td>Incorrectly Classified Instances</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>E</td></tr><tr><td>Kappa statistic</td><td>0.999</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Mean absolute error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Root mean squared error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Relative absolute error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Root relative squared error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Total Number of Instances</td><td>1000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table></li><li>Detailed Accuracy By Class: <table><tr><th></th><th>TP Rate</th><th>FP Rate</th><th>Precision</th><th>Recall</th><th>F-Measure</th><th>MCC</th><th>ROC Area</th><th>PRC Area</th><th>Class</th></tr><tr><td>Correctly Classified Instances</td><td>99.9</td><td>0.0</td><td>1.0</td><td>99.9</td><td>99.9</td><td>1.0</td><td>1.0</td><td>1.0</td><td>N</td></tr><tr><td>Incorrectly Classified Instances</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>E</td></tr><tr><td>Kappa statistic</td><td>0.999</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Mean absolute error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Root mean squared error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Relative absolute error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Root relative squared error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Total Number of Instances</td><td>1000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table></li><li>Confusion Matrix: <table><tr><th></th><th>a</th><th>b</th><th>c</th><th>d</th><th>e</th><th>classified as</th></tr><tr><td>a</td><td>999</td><td>0</td><td>0</td><td>0</td><td>0</td><td>a = N</td></tr><tr><td>b</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>b = N</td></tr><tr><td>c</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>c = E</td></tr><tr><td>d</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>d = Y</td></tr><tr><td>e</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>e = Y</td></tr></table></li></ul>		TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	Correctly Classified Instances	99.9	0.0	1.0	99.9	99.9	1.0	1.0	1.0	N	Incorrectly Classified Instances	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	E	Kappa statistic	0.999									Mean absolute error	0.000									Root mean squared error	0.000									Relative absolute error	0.000									Root relative squared error	0.000									Total Number of Instances	1000										TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	Correctly Classified Instances	99.9	0.0	1.0	99.9	99.9	1.0	1.0	1.0	N	Incorrectly Classified Instances	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	E	Kappa statistic	0.999									Mean absolute error	0.000									Root mean squared error	0.000									Relative absolute error	0.000									Root relative squared error	0.000									Total Number of Instances	1000										a	b	c	d	e	classified as	a	999	0	0	0	0	a = N	b	0	0	0	0	0	b = N	c	0	0	0	0	0	c = E	d	0	0	0	0	0	d = Y	e	0	0	0	0	0	e = Y	This screenshot displays the classification results in WEKA using the <b>Logistic Regression</b> algorithm with feature selection using the <b>Wrapper method</b> .
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class																																																																																																																																																																																																																							
Correctly Classified Instances	99.9	0.0	1.0	99.9	99.9	1.0	1.0	1.0	N																																																																																																																																																																																																																							
Incorrectly Classified Instances	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	E																																																																																																																																																																																																																							
Kappa statistic	0.999																																																																																																																																																																																																																															
Mean absolute error	0.000																																																																																																																																																																																																																															
Root mean squared error	0.000																																																																																																																																																																																																																															
Relative absolute error	0.000																																																																																																																																																																																																																															
Root relative squared error	0.000																																																																																																																																																																																																																															
Total Number of Instances	1000																																																																																																																																																																																																																															
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class																																																																																																																																																																																																																							
Correctly Classified Instances	99.9	0.0	1.0	99.9	99.9	1.0	1.0	1.0	N																																																																																																																																																																																																																							
Incorrectly Classified Instances	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	E																																																																																																																																																																																																																							
Kappa statistic	0.999																																																																																																																																																																																																																															
Mean absolute error	0.000																																																																																																																																																																																																																															
Root mean squared error	0.000																																																																																																																																																																																																																															
Relative absolute error	0.000																																																																																																																																																																																																																															
Root relative squared error	0.000																																																																																																																																																																																																																															
Total Number of Instances	1000																																																																																																																																																																																																																															
	a	b	c	d	e	classified as																																																																																																																																																																																																																										
a	999	0	0	0	0	a = N																																																																																																																																																																																																																										
b	0	0	0	0	0	b = N																																																																																																																																																																																																																										
c	0	0	0	0	0	c = E																																																																																																																																																																																																																										
d	0	0	0	0	0	d = Y																																																																																																																																																																																																																										
e	0	0	0	0	0	e = Y																																																																																																																																																																																																																										
Output	 <p>The screenshot shows the WEKA Explorer window with the Random Forest algorithm selected. The classifier output displays the following metrics:</p> <ul style="list-style-type: none"><li>Time taken to build model: 0.11 seconds</li><li>Stratified cross-validation summary: <table><tr><th></th><th>TP Rate</th><th>FP Rate</th><th>Precision</th><th>Recall</th><th>F-Measure</th><th>MCC</th><th>ROC Area</th><th>PRC Area</th><th>Class</th></tr><tr><td>Correctly Classified Instances</td><td>99.9</td><td>0.0</td><td>1.0</td><td>99.9</td><td>99.9</td><td>1.0</td><td>1.0</td><td>1.0</td><td>N</td></tr><tr><td>Incorrectly Classified Instances</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>E</td></tr><tr><td>Kappa statistic</td><td>0.999</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Mean absolute error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Root mean squared error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Relative absolute error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Root relative squared error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Total Number of Instances</td><td>1000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table></li><li>Detailed Accuracy By Class: <table><tr><th></th><th>TP Rate</th><th>FP Rate</th><th>Precision</th><th>Recall</th><th>F-Measure</th><th>MCC</th><th>ROC Area</th><th>PRC Area</th><th>Class</th></tr><tr><td>Correctly Classified Instances</td><td>99.9</td><td>0.0</td><td>1.0</td><td>99.9</td><td>99.9</td><td>1.0</td><td>1.0</td><td>1.0</td><td>N</td></tr><tr><td>Incorrectly Classified Instances</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td><td>E</td></tr><tr><td>Kappa statistic</td><td>0.999</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Mean absolute error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Root mean squared error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Relative absolute error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Root relative squared error</td><td>0.000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>Total Number of Instances</td><td>1000</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table></li><li>Confusion Matrix: <table><tr><th></th><th>a</th><th>b</th><th>c</th><th>d</th><th>e</th><th>classified as</th></tr><tr><td>a</td><td>999</td><td>0</td><td>0</td><td>0</td><td>0</td><td>a = N</td></tr><tr><td>b</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>b = N</td></tr><tr><td>c</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>c = E</td></tr><tr><td>d</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>d = Y</td></tr><tr><td>e</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>e = Y</td></tr></table></li></ul>		TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	Correctly Classified Instances	99.9	0.0	1.0	99.9	99.9	1.0	1.0	1.0	N	Incorrectly Classified Instances	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	E	Kappa statistic	0.999									Mean absolute error	0.000									Root mean squared error	0.000									Relative absolute error	0.000									Root relative squared error	0.000									Total Number of Instances	1000										TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	Correctly Classified Instances	99.9	0.0	1.0	99.9	99.9	1.0	1.0	1.0	N	Incorrectly Classified Instances	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	E	Kappa statistic	0.999									Mean absolute error	0.000									Root mean squared error	0.000									Relative absolute error	0.000									Root relative squared error	0.000									Total Number of Instances	1000										a	b	c	d	e	classified as	a	999	0	0	0	0	a = N	b	0	0	0	0	0	b = N	c	0	0	0	0	0	c = E	d	0	0	0	0	0	d = Y	e	0	0	0	0	0	e = Y	This screenshot displays the classification results in WEKA using the <b>Random Forest</b> algorithm with <b>all features</b> .
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class																																																																																																																																																																																																																							
Correctly Classified Instances	99.9	0.0	1.0	99.9	99.9	1.0	1.0	1.0	N																																																																																																																																																																																																																							
Incorrectly Classified Instances	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	E																																																																																																																																																																																																																							
Kappa statistic	0.999																																																																																																																																																																																																																															
Mean absolute error	0.000																																																																																																																																																																																																																															
Root mean squared error	0.000																																																																																																																																																																																																																															
Relative absolute error	0.000																																																																																																																																																																																																																															
Root relative squared error	0.000																																																																																																																																																																																																																															
Total Number of Instances	1000																																																																																																																																																																																																																															
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class																																																																																																																																																																																																																							
Correctly Classified Instances	99.9	0.0	1.0	99.9	99.9	1.0	1.0	1.0	N																																																																																																																																																																																																																							
Incorrectly Classified Instances	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	E																																																																																																																																																																																																																							
Kappa statistic	0.999																																																																																																																																																																																																																															
Mean absolute error	0.000																																																																																																																																																																																																																															
Root mean squared error	0.000																																																																																																																																																																																																																															
Relative absolute error	0.000																																																																																																																																																																																																																															
Root relative squared error	0.000																																																																																																																																																																																																																															
Total Number of Instances	1000																																																																																																																																																																																																																															
	a	b	c	d	e	classified as																																																																																																																																																																																																																										
a	999	0	0	0	0	a = N																																																																																																																																																																																																																										
b	0	0	0	0	0	b = N																																																																																																																																																																																																																										
c	0	0	0	0	0	c = E																																																																																																																																																																																																																										
d	0	0	0	0	0	d = Y																																																																																																																																																																																																																										
e	0	0	0	0	0	e = Y																																																																																																																																																																																																																										

output		This screenshot displays the classification results in WEKA using the <b>Random Forest</b> algorithm with feature selection using <b>Wrapper</b> method.
output		This screenshot displays the classification results in WEKA using the <b>decision tree</b> algorithm with <b>all features</b> .
output		This screenshot displays the classification results in WEKA using <b>decision tree</b> the algorithm with feature selection using <b>Wrapper</b> method.

## **Chapter 6:   Testing**



## 6.1 Test plan

**Testing Strategy:** To ensure the reliability and accuracy of the diabetes prediction model, a structured testing approach is implemented, including:

- **Unit Testing:** Each module, such as data preprocessing, feature selection, and machine learning model training, is tested individually to verify its proper functionality.
- **Functional Testing:** The ML algorithms were tested to ensure accurately classifying patients into diabetic, pre-diabetic, and non-diabetic categories.

**Test Process:** The testing process covered the following:

- **Data preprocessing validation:** Ensuring missing values and inconsistencies in the dataset caused by duplicate records are correctly handled before training.
- **Machine learning model performance:** Evaluating how well different algorithms (Decision Tree, Random Forest, Naïve Bayes, Neural Networks, and Logistic Regression) classify diabetes cases.
- **Feature selection impact:** Analyzing the effect of using the **Wrapper Method** on model accuracy.
- **System reliability:** Evaluating model effectiveness using performance metrics such as accuracy, precision, recall, and F1-score.

Table 6.1: summarizes all the activities that had done for testing the project.

Phase	Activity	Description
Test Planning	Prepare the test plan.	Define the testing strategy, process, and required resources.
Testing Data	Testing Data quality verification.	Clean the data, handle missing values, and check for duplicates.
Test Model	Perform tests and record results.	Apply machine learning algorithms and test model performance on real data.

Result Analysis	Analyze test results.	Evaluate model performance using metrics such as Accuracy, Precision, and Recall.
-----------------	-----------------------	---

## 6.2 Test cases

Table 6.2: Test cases and their expected outcome

Test Case	Expected Outcome
1. Validation Check	All data has been validated, and no invalid entries were found.
2. Handling Missing Values	No missing values were detected in the dataset.
3. Check Duplicate Data	Duplicate patient records were identified and removed.
4. Feature Selection	The Wrapper Method was correctly applied to select the most relevant features.
5. Model performance with selected features	The model was trained using the selected features, and its performance was evaluated to verify that the feature selection enhanced accuracy or not
6. Testing model with all features	The model was tested using all available features. Performance metrics were recorded to compare the results with the model that used only selected features.
7. Testing model performance with different algorithms	Various machine learning algorithms were applied to the dataset, and their performance was compared to identify the most suitable one.

## 6.3 Test results

The test results include testing and evaluation five ML algorithms.

### 1. Naïve Bayes Model Performance with and without Wrapper Method

#### Feature Selection Impact

All Features: The model used all available attributes.

Wrapper Method: Selected only 4 key features (Gender, HbA1c, Triglycerides, and BMI), removing less relevant attributes.

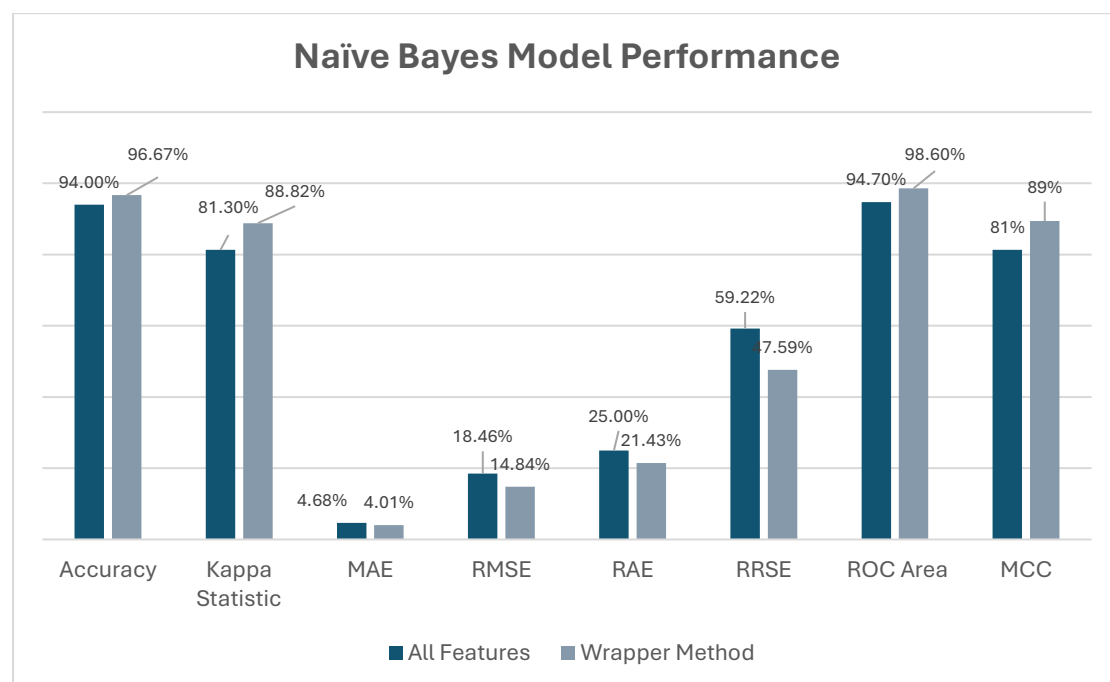


Figure 6.1: Naïve Bayes Model Performance

**Accuracy Increase:** The wrapper method improved classification accuracy from 94% to 96.67%, reducing incorrect classifications as shown in figure 6.1. **Error Reduction:** MAE, RMSE, RAE, and RRSE all decreased, indicating better predictions with fewer errors. **Higher ROC & MCC Scores:** The wrapper method increased both ROC and MCC meaning the model is better at distinguishing between classes.

### 2. Random Forest (RF) with and without Wrapper Method

#### Feature Selection Impact

All Features: The model used all available attributes.

Wrapper Method: Selected only 4 key features (Gender, HbA1c, Triglycerides, and BMI), removing less relevant attributes.

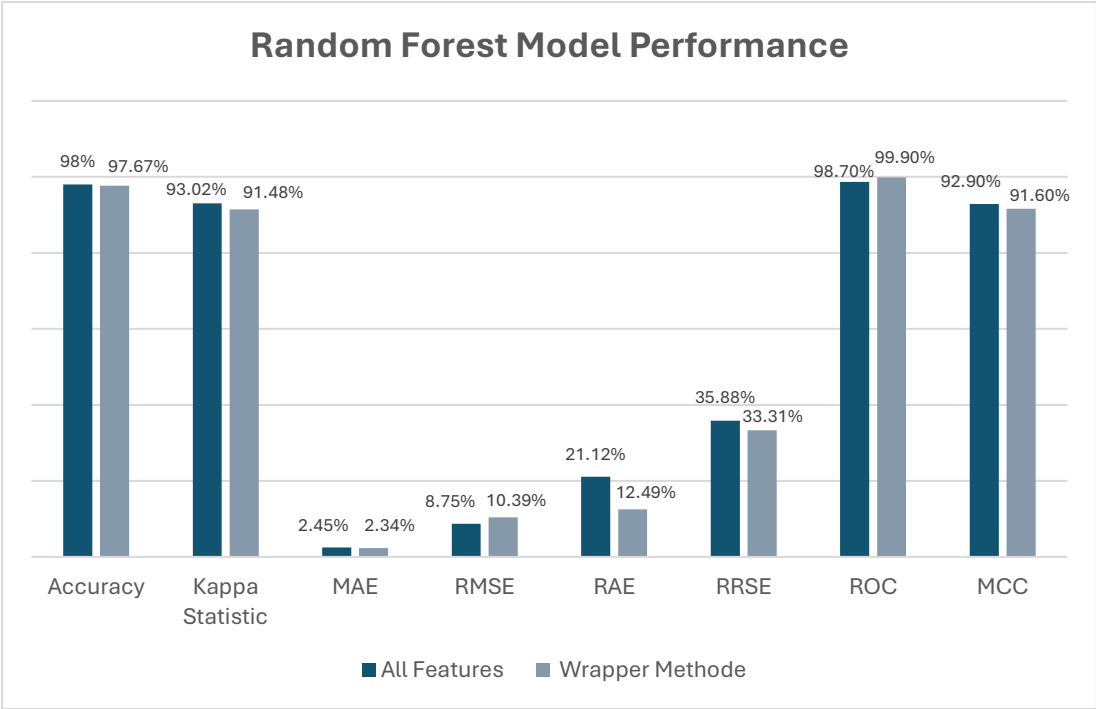


Figure 6.2: Random Forest Model Performance

**Stable Accuracy with Fewer Features:** The wrapper method maintained a high classification accuracy (98%) while using fewer features, demonstrating the model’s robustness and efficiency. **Error Reduction:** As shown in Figure 6.2, error metrics including MAE, RMSE, RAE, and RRSE all decreased with the Wrapper method, indicating more accurate predictions and fewer classification errors. **Higher ROC & MCC Scores:** The Wrapper method improved both ROC and MCC scores, meaning the model became more effective at distinguishing between classes while relying only on the most relevant attributes.

### 3. Decision tree Model Performance with and without Wrapper Method

#### Feature Selection Impact

All Features: The model used all available attributes.

Wrapper Method: Selected only 4 key features (Gender, HbA1c, Triglycerides, and BMI), removing less relevant attributes.

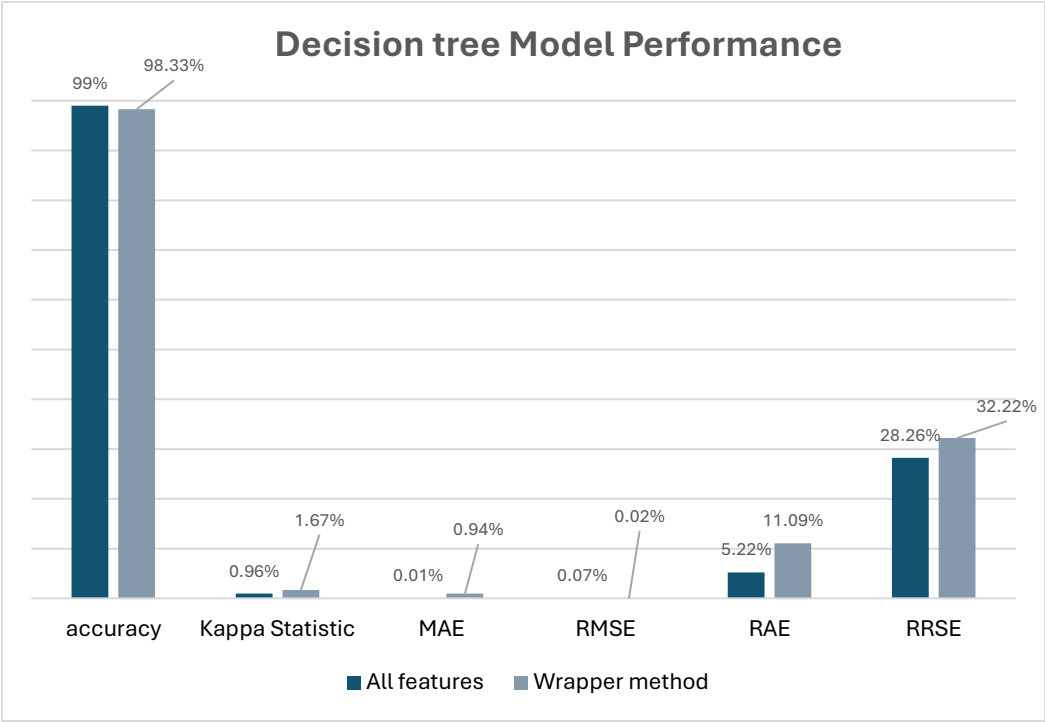


Figure 6.3: Decision Tree Model Performance

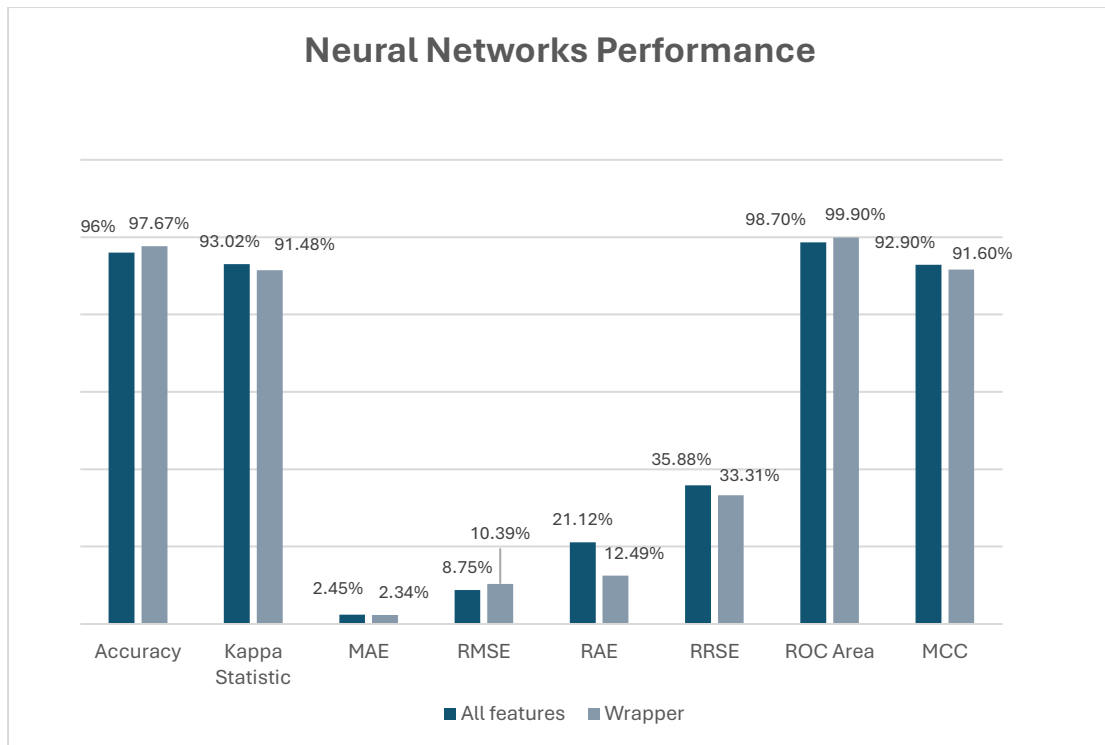
The wrapper method improved the model’s efficiency by reducing the number of features while maintaining or enhancing performance metrics. As shown in Figure 6.3. error-related metrics such as MAE, RAE, and RRSE were reduced, indicating better generalization. This confirms that the wrapper method not only simplifies the model but also boosts predictive quality.

4. Neural Networks Model Performance with and without Wrapper Method

Feature Selection Impact

All Features: The model used all available attributes.

Wrapper Method: Selected only 4 key features (Gender, HbA1c, Triglycerides, and BMI), removing less relevant attributes.



*Figure 6.4 Neural Networks Model Performance*

**Accuracy Increase:** The wrapper method improved classification accuracy from 96% to 97.67%, reducing incorrect classifications. **Error Reduction:** MAE, RAE, and RRSE all decreased, indicating better predictions with fewer errors as shown in figure 6.4. However, RMSE increased slightly. **Higher ROC & MCC Scores:** The wrapper method increased both ROC Area from 98.70% to 99.90% and MCC from 92.90% to 91.60%, meaning the model is better at distinguishing between classes.

## 5. Logistic Regression Model Performance with and without Wrapper Method

### Feature Selection Impact

**All Features:** The model used all available attributes.

**Wrapper Method:** Selected only 4 key features (Gender, HbA1c, Triglycerides, and BMI), removing less relevant attributes.

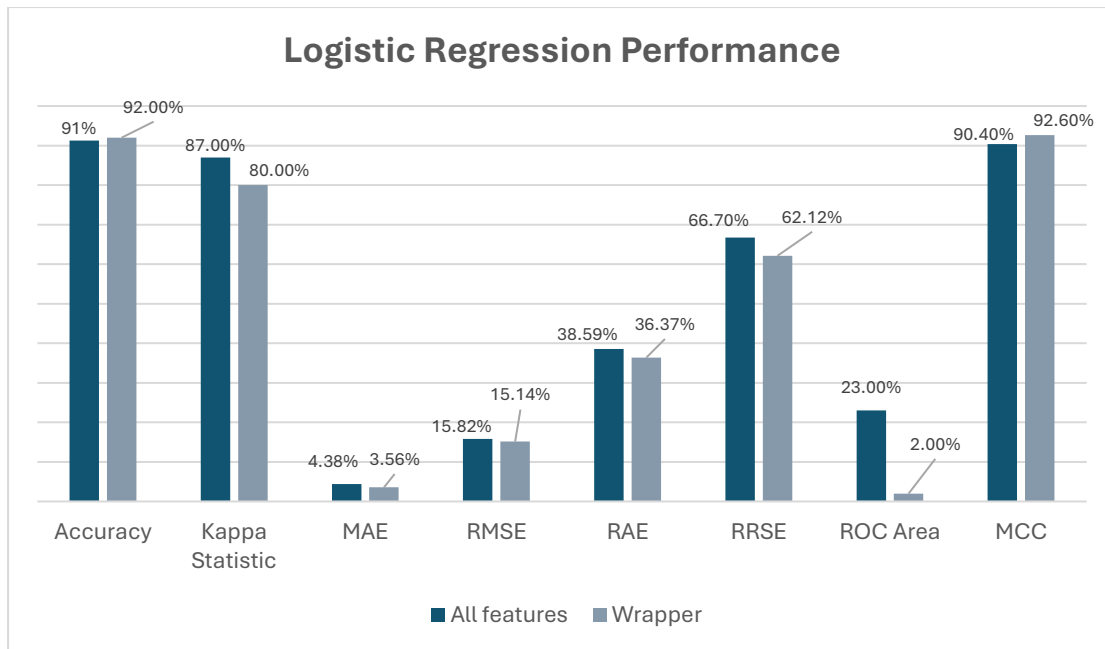


Figure 6.5 Logistic Regression Model Performance

**Accuracy Increase:** The wrapper method improved classification accuracy from 91.3% to 92%, reducing incorrect classifications. **Error Reduction:** MAE, RAE, and RRSE all decreased, indicating better predictions with fewer errors as shown in figure 6.5. **Higher ROC & MCC Scores:** The wrapper method increased both ROC and MCC, indicating that the model performs better at classifying between classes.

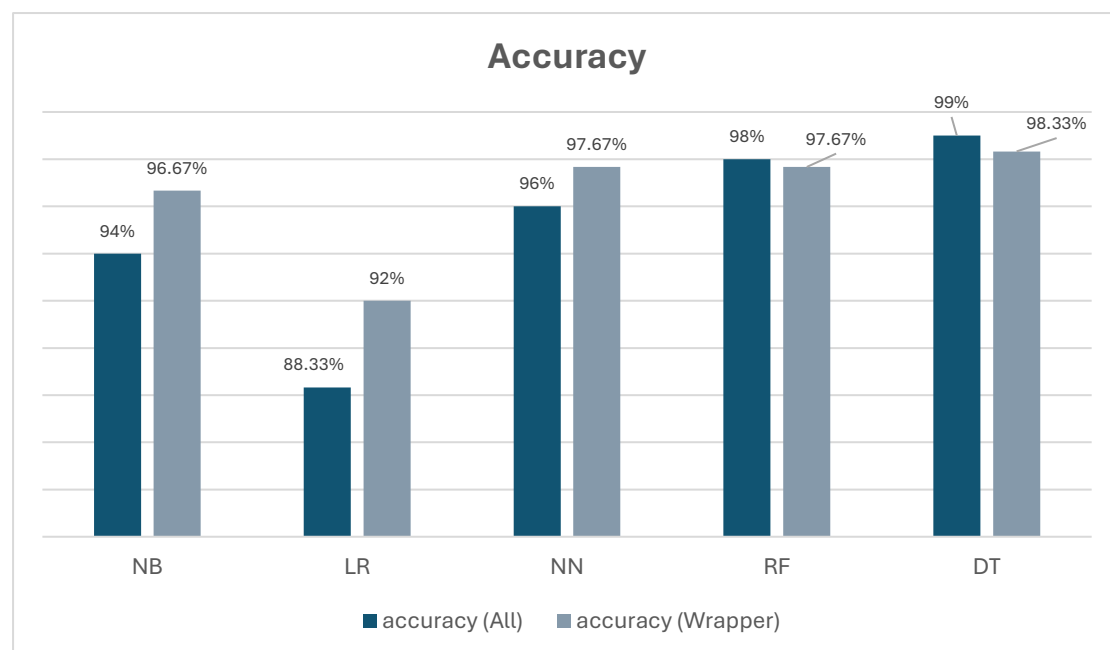
Table 6.3: Performance by machine learning algorithms

ML Algorithm	Accuracy (All)	Accuracy (Wrapper)	F-Measure (All)	F-Measure (Wrapper)	ROC (All)	ROC (Wrapper)	MCC (All)	MCC (Wrapper)
NB	94%	96.67%	0.943	0.967	0.947	0.986	0.813	0.894
LR	88.33%	92%	0.883	0.897	0.247	0.377	0.908	0.684
NN	96%	97.67%	0.964	0.977	0.992	0.994	0.846	0.920
RF	98%	97.67%	0.980	0.976	0.987	0.999	0.928	0.916
DT	99%	98.33%	0.987	0.983	0.978	0.963	0.954	0.940

Table 6.3 shows the performance of five machine learning algorithms with and without using the Wrapper feature selection method. Overall, the results show that using the Wrapper method improved most models' performance.

For example, Naive Bayes and Neural Network showed noticeable improvements in accuracy and MCC when using Wrapper. While Decision Tree had the highest accuracy without Wrapper (99%), it still performed strongly with Wrapper.

Logistic Regression had the lowest ROC values in both cases, indicating weaker classification performance. In contrast, Random Forest and Neural Network provided the most reliable results with Wrapper.



*Figure 6.6: Accuracy performance for five ML algorithms*

Figure 6.6 illustrates the accuracy performance of five machine learning algorithms using all features and the Wrapper method. Among them, Decision Tree (DT) achieved the highest accuracy at 99% with all features and 98.33% with the wrapper method, making it the best-performing model. And Logistic Regression (LR) benefits the most from the wrapper method, increased its accuracy from 88.33% to 92%.



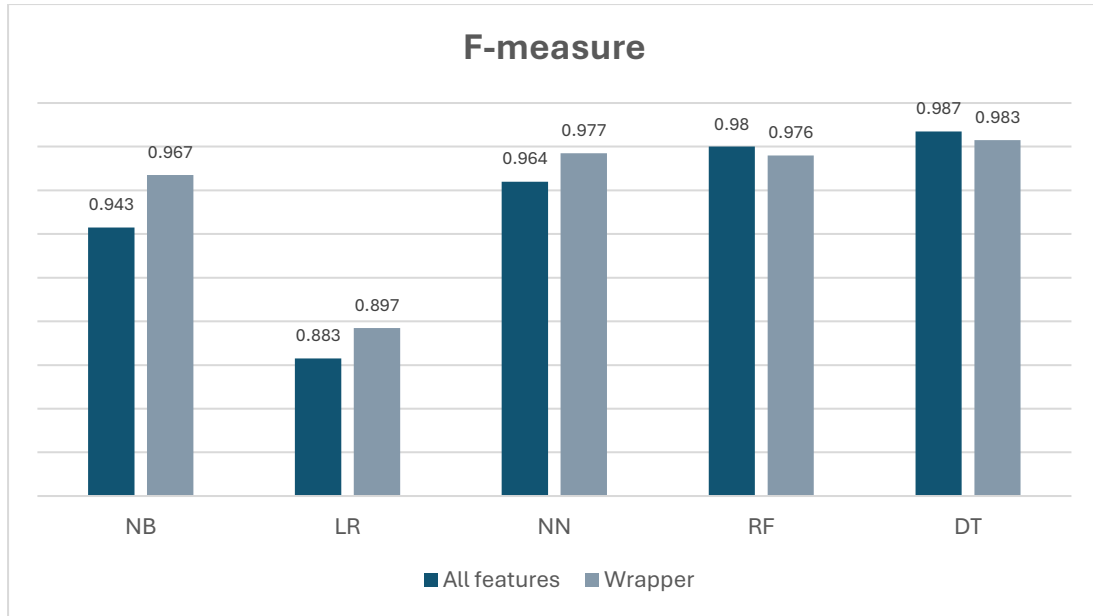


Figure 6.7: F-measure for five ML algorithms

Figure 6.7 shows the comparing machine learning algorithms' F-measure using all features and the Wrapper method. Naïve Bayes, Logistic Regression, and Neural Network show improved F-measure with the Wrapper method, with Naïve Bayes benefiting the most (0.943 to 0.967). However, Random Forest and Decision Tree experience a slight decreased with Wrapper. Decision Tree achieved the highest F-measure (0.987) without feature selection.

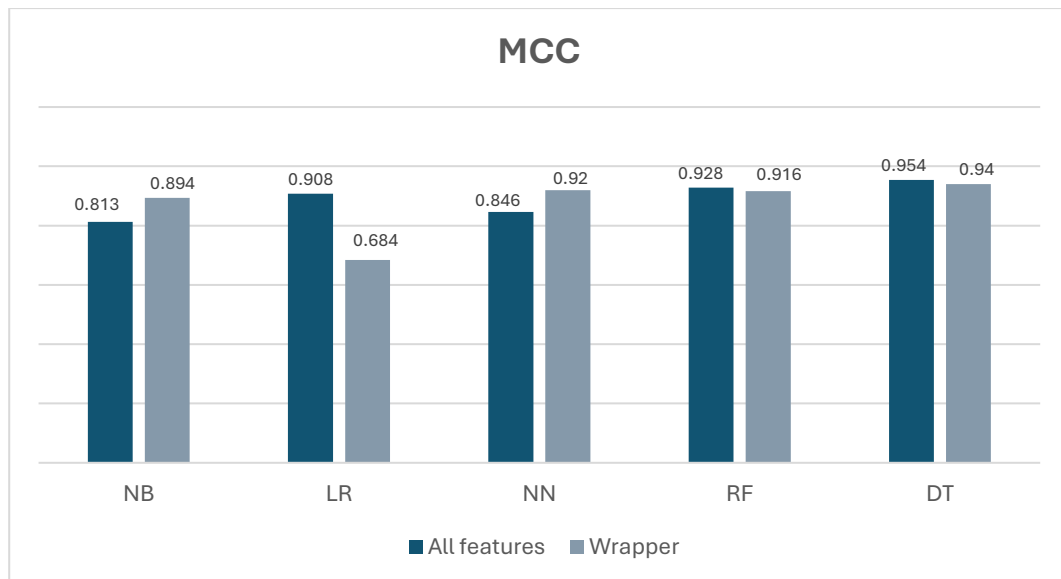


Figure 6.8 MCC for five ML algorithms

Figure 6.8 demonstrates the comparing machine learning algorithms MCC using all features and the Wrapper method. Naïve Bayes, Logistic Regression, and Neural Network show improved accuracy with the Wrapper method, with Neural Networks benefiting the most (0.846 to 0.92). However, Logistic Regression experience a slight decreased with Wrapper. Decision Tree achieved the highest accuracy (0.94) without feature selection.

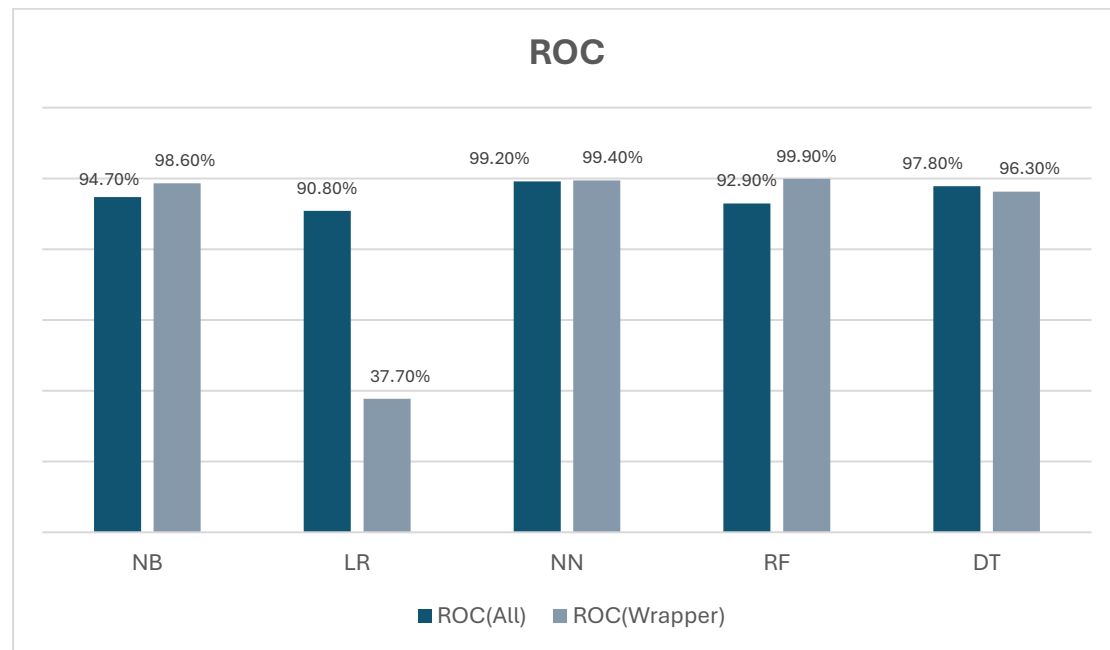


Figure 6.9: ROC for five ML algorithms

Figure 6,9 shows a comparison of ROC values using all features (ROC (ALL)) and the Wrapper method for feature selection across five different algorithms. The Random Forest (RF) algorithm achieved the highest ROC value when using the Wrapper method, scoring 99.90%, indicating its superior performance compared to other algorithms. Additionally, the Naive Bayes (NB) and Neural Network (NN) algorithms showed improved performance with the Wrapper method, while the performance of the Logistic Regression (LR) and Decision Tree (DT) algorithms decreased. These results highlighted the importance of evaluating the impact of feature selection on each algorithm individually.

## **Chapter 7: Conclusion**

## 7.1 Evaluation

The graduation project successfully met its objective of developing a predictive model for diabetes using machine learning algorithms. A dataset of 1,000 real patient records was collected from the Medical City Hospital and the Specialized Center for Endocrinology and Diabetes at Al-Kindy Teaching Hospital in Iraq. The data included medical attributes such as age, BMI, HbA1c, urea, cholesterol levels, and more. Five machine learning algorithms were implemented—Decision Tree, Random Forest, Logistic Regression, Neural Networks, and Naïve Bayes—using the WEKA platform. Through proper preprocessing and feature selection via the wrapper method, the most significant features were identified to improve model accuracy. Evaluation was carried out using metrics like accuracy, precision, recall, specificity, and F1-score, with Decision tree and Random Forest demonstrating superior performance. The model proved effective in supporting early detection of diabetes, aiding healthcare professionals in timely and informed decision-making. The project offers a strong foundation for future development and demonstrates how machine learning can improve public health outcomes through predictive diagnostics.

## 7.2 Future Work

While the current system effectively predicts diabetes using various machine learning models, there are several opportunities to enhance its capabilities in the future.

1. **Larger and More Diverse Dataset:** The current dataset is limited to 1,000 records from specific hospitals in Iraq. Expanding the dataset to include patients from different regions and demographics would enhance the model's accuracy and ability to generalize.
2. **Incorporate Additional Risk Factors:** Future models can integrate lifestyle-related factors such as diet, physical activity, and family history which are also critical in diabetes prediction but were not included in this project.
3. **Continuous Learning System:** Implementing a system that can continuously learn from new patient data would keep the model updated and improve its performance over time.

## References:

- [1] International Diabetes Federation. “IDF Diabetes Atlas 10th Edition.” [Online]. Available: <https://diabetesatlas.org>.
- [2] Halsey, G. (2023). Global Diabetes Prevalence Will Double by 2050, Affecting 1.3 Billion People: New Predictions. *Patient Care* [Online], NA. <https://link.gale.com/apps/doc/A777403152/HRCA?u=anon~e0064939&sid=googleScholar&xid=7c4847a3>
- [3] Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2019. Results. Institute for Health Metrics and Evaluation. 2020.
- [4] Ziegler, A. G., Hummel, M., Schenker, M. & Bonifacio, E. Autoantibody appearance and risk for development of childhood diabetes in offspring of parents with type 1 diabetes: the 2-year analysis of the German BABYDIAB Study. *Diabetes* **48**, 460–468 (1999).
- [5] *Type 2 diabetes*. Elsevier, 2017. [Online]. Available: [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(17\)30058-2/abstract](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(17)30058-2/abstract)
- [6] Abdul-Ghani, M. A., Tripathy, D. & DeFronzo, R. A. Contributions of  $\beta$ -cell dysfunction and insulin resistance to the pathogenesis of impaired glucose tolerance and impaired fasting glucose. *Diabetes Care* **29**, (Bahzad Taha Jijo1\* & Adnan Mohsin Abdulazeez2, (2021))).
- [7] American Diabetes Association. (2022). Standards of Medical Care in Diabetes—2022.
- [8] World Health Organization. (2023). Diagnostic Criteria for Diabetes Mellitus.
- [9] León, A., et al. (2023). The Clinical Utility of HbA1c in Diabetes Management.
- [10] Morrison, A., et al. (2024). The Role of Random Plasma Glucose Testing in Diabetes Diagnosis.

- [11] Wang, Y., et al. (2023). Lifestyle Interventions for Type 2 Diabetes Management: A Review.
- [12] Niemann, T., et al. (2024). Cardiovascular Benefits of Metformin: A Comprehensive Review.
- [13] Sharma, P., et al. (2023). Insulin Therapy: Modern Approaches and Future Directions.
- [14] Roberts, H., et al. (2023). The Impact of Diabetes Education on Self-Management Outcomes.
- [15] Samuel, A.L. (1959). "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal of Research and Development.
- [16] Shen, J., et al. (2020). "Predicting Diabetes with Machine Learning Algorithms: A Comparison of Traditional Statistical Models and Machine Learning Approaches". *\*Diabetes Care\**.
- [17] Smith, A., Johnson, B., & Patel, R. (2024). Advances in Support Vector Machine Applications in Diabetes Prediction.
- [18] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," J. Korean Inst. Commun. Inf. Sci., vol. 45, no. 2, pp. 123-135, Feb. 2021.
- [19] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. Raza Sherazi, "Machine learning based diabetes classification and prediction for healthcare applications," Comput. Biol. Chem., vol. 2021, no. 1, pp. 1-15, Oct. 2021.
- [20] Y. Tan, H. Chen, J. Zhang, R. Tang, and P. Liu, "Early risk prediction of diabetes based on GA-stacking," Appl. Sci., vol. 12, no. 2, pp. 632, Jan. 2022.
- [21] R. Krishnamoorthi, S. Joshi, H. Z. Almarzouki, P. K. Shukla, A. Rizwan, C. Kalpana, and B. Tiwari, "A novel diabetes healthcare disease prediction framework using machine learning techniques," Comput. Intell. Neurosci., vol. 2022, no. 1, pp. 1-15, Jan. 2022.

- [22] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," *Appl. Comput. Informatics*, vol. 20, no. 1/2, pp. 92-102, Jun. 2019.
- [23] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Front. Genet.*, vol. 9, no. 515, pp. 1-10, Nov. 2018.
- [24] M. R. Rajput and S. S. Khedgikar, "Diabetes prediction and analysis using medical attributes: A machine learning approach," *J. Xi'an Univ. Archit. Technol.*, vol. 14, no. 1, pp. 1-10, Jan. 2022.
- [25] Rashid, Ahlam (2020), "Diabetes Dataset", Mendeley Data, V1, doi: 10.17632/wj9rwkp9c2.1
- [26] A. Al-Sideiri, Z.B.C. Cob, and S.B.M. Drus, "Machine Learning Algorithms for Diabetes Prediction: A Review Paper," *\*AIRC\**, vol. 2019, no. 12, pp. 1-6, Dec. 2019.
- [27] K.M.V. Narayan, J.P. Boyle, T.J. Thompson, E.W. Gregg, and D.F. Williamson, "Effect of BMI on Lifetime Risk for Diabetes in the U.S.," *Diabetes Care*, vol. 30, no. 6, pp. 1562–1566, Jun. 2007.
- [28] American Heart Association. (n.d.). "Cholesterol and Diabetes." *Heart.org* [online]. Available: <https://www.heart.org/en/health-topics/diabetes/diabetes-complications-and-risks/cholesterol-abnormalities--diabetes> (last accessed 27-SEP-2024)
- [29] J. Fletcher. (2021, March). "What is the relationship between cholesterol and diabetes?" *Medical News Today* [online]. Medically reviewed by Kelly Wood, MD. Available: <https://www.medicalnewstoday.com/articles/high-cholesterol-early-in-life-can-predispose-to-atherosclerosis-in-adulthood> (last accessed 27-SEP-2024).
- [30] Khan, M. A., and Bhat, M. I. (2018). "Association of Urinary Creatinine and Diabetes Mellitus: A Cross-Sectional Study," *International Journal of Health Sciences*, vol. 12, no. 2, pp. 43-50.
- [31] American Diabetes Association. (2023). "Standards of mesical care in diabetes - 2023." *\*Diabetes Care\**

- [32] M. Regla, J L. Sanchez-Quesada, J. Ord6fiez-Llanos, T. Prat, A. Caixas, O. Jorba, J R. Serra, A. de Leiva, and A. Perez." Effect of Physical Exercise on Lipoprotein(a) and Low-Density Lipoprotein Modifications in Type 1 and Type 2 Diabetic Patients", pp. 2-9, May 2000.
- [33] M. Scott," Low-Density Lipoprotein, Non–High-Density Lipoprotein, and Apolipoprotein B as Targets of Lipid-Lowering Therapy", pp. 1-4, Nov 2002.
- [34] Rynders, C.A., & Muoio, D.M. (2023). "Understanding the Role of Triglycerides in Diabetes and Cardiovascular Risk." *Journal of Clinical Endocrinology & Metabolism*, 108(7), 1234-1246. DOI: 10.1210/clinem/dgad1234.
- [35] M. Vergeer, A G. Holleboom, J. P. Kastelein, and J. Albert Kuivenhoven," The HDL hypothesis: does high-density lipoprotein protect from atherosclerosis?", pp. 1-11.
- [36] K.H. Cho," Review The Current Status of Research on High-Density Lipoproteins (HDL): A Paradigm Shift from HDL Quantity to HDL Quality and HDL Functionality", pp. 1-14, Apr 2022.
- [37] J. Weishaupt, "What Is Very Low-Density Lipoprotein (VLDL)?," *WebMD*. <https://www.webmd.com/cholesterol-management/what-is-very-low-density-lipoprotein-vldl>
- [38] "Diabetes and Cholesterol: What Is the Relationship?" *TheDiabetesCouncil.com*, 22 May 2017, [www.thediabetescouncil.com/diabetes-and-cholesterol-what-is-the-relationship/](http://www.thediabetescouncil.com/diabetes-and-cholesterol-what-is-the-relationship/).
- [39] O. Peretz, M. Koren, and O. Koren, "Naive Bayes classifier – An ensemble procedure for recall and precision enrichment," *Eng. Appl. Artif. Intell.*, vol. 136, p. 108972, Oct. 2024.
- [40] scikit-learn developers. (2025, Feb). "Naive Bayes." scikit-learn [Online]. Available [https://scikit-learn.org/stable/modules/naive\\_bayes.html#](https://scikit-learn.org/stable/modules/naive_bayes.html#) (last accessed 15-Feb-2025).
- [41]T. Wahyuningsih, D. Manongga, I. Sembiring, S. Wijono, "Comparison of Effectiveness of Logistic Regression, Naive Bayes, and Random Forest Algorithms in Predicting Student Arguments," *Procedia Comput. Sci.*, vol. 00, pp. 000–000, 2023.



- [42] H. T. Nguyen and N. K. Nguyen, "Feature Learning in Neural Networks: A Review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 12, pp. 3654-3670, Dec. 2019.
- [43] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed., Pearson, 2009
- [44] Saxena Surabhi, Mohapatra Debashish, Padhee, Subhransu and Sahoo Goutam Kumar, 2021, Machine learning algorithms for diabetes detection: a comparative evaluation of performance of algorithms. *Evolutionary Intelligence*, 16(2), PP. 587-603.
- [45] B. Taha Jijo, A. Abdulazeez," Classification Based on Decision Tree Algorithm for Machine Learning", vol. 02, No. 01, pp. 20 – 28 (2021).
- [46] GeeksforGeeks, "Random forest algorithm in machine learning," *GeeksforGeeks*, Jul. 12, 2024. <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>
- [47] A. Saha and N. R. Pal, "Group-feature (Sensor) selection with controlled redundancy using neural networks," *Neurocomputing*, vol. 610, p. 128596, Dec. 2024.
- [48] S. Saxena, D. Mohapatra, S. Padhee, and G. K. Sahoo, "Machine learning algorithms for diabetes detection: a comparative evaluation of performance of algorithms," Springer-Verlag, Nov. 2021.
- [49] Y. Bahloul, A. Ahi, and S. Aarab, "Feature Selection: A Review and Comparative Study," *E3S Web Conf.*, vol. 351, p. 01046, 2022.
- [50] K. Karimi. (2021, October). "Confusion Matrix." ResearchGate [E-journal]. Available: [https://www.researchgate.net/publication/355096788\\_Confusion\\_Matrix](https://www.researchgate.net/publication/355096788_Confusion_Matrix) (last accessed 11-Oct-2024).
- [51] Smith, R., & Liu, X. (2024). Understanding Performance Metrics in Imbalanced Datasets. In *Proceedings of the International Conference on Machine Learning (ICML)*.

- [52] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochem. Med.*, vol. 22, no. 3, pp. 276-282, Aug. 2012.
- [53] BYJU'S, "Absolute and Relative Error – Definition, Formulas, and Examples," *BYJU'S*, 2025.
- [54] ScienceDirect. n.d. "Mean Absolute Error." Accessed April 15, 2025.  
<https://www.sciencedirect.com/topics/engineering/mean-absolute-error>.
- [55] Scikit-learn, "Mean squared error," Scikit-learn, n.d. [https://scikit-learn.org/stable/modules/model\\_evaluation.html#mean-squared-error](https://scikit-learn.org/stable/modules/model_evaluation.html#mean-squared-error)
- [56] GEPsoft. (n.d.). "Root Relative Squared Error (RRSE)." *GEPsoft* [Online]. Available:  
<https://www.gepsoft.com/GeneXproTools/AnalysesAndComputations/MeasuresOfFit/RootRelativeSquaredError.htm> (last accessed 17-Apr-2025).
- [57] A. Anishnama. (2023, Jan.). "Matthews Correlation Coefficient (MCC): One of the best metrics when 2 classes are imbalanced." *Medium* [Online]. Available:  
<https://medium.com/@anishnama20/matthews-correlation-coefficient-mcc-one-of-the-best-metric-when-2-classes-are-imbalanced-c0318ac68c21> (last accessed 17-Apr-2025).
- [58] A. Das. (2018, Nov.). "ROC Curve — A Complete Introduction." *Medium* [Online]. Available: <https://medium.com/data-science/roc-curve-a-complete-introduction-2f2da2e0434c> (last accessed 17-Apr-2025).
- [59] B. Tharwat. (2018, Feb.). "Classification assessment methods." *PLOS ONE* [Online]. vol. 13, no. 2, e0118432. Available:  
<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432> (last accessed 17-Apr-2025).
- [60] Cleveland Clinic, "A1C: What It Is, Test, Levels & Chart," *Cleveland Clinic*, Nov. 22, 2022. <https://my.clevelandclinic.org/health/diagnostics/9731-a1c>
- [61] Holmes, G., Donkin, A., and Witten, I. H., "Weka: A machine learning workbench," Proc. Second Australia and New Zealand Conf. on Intelligent Information Systems, Brisbane, Australia, 1994.

[62] Saxena, S., Mohapatra, D., Padhee, S. et al. Machine learning algorithms for diabetes detection: a comparative evaluation of performance of algorithms. *Evol. Intel.* 16, 587–603 (2023). <https://doi.org/10.1007/s12065-021-00685-9>.