VAASAN AMMATTIKORKEAKOULU
UNIVERSITY OF APPLIED SCIENCES

Bishal Gyawali

DIABETES MELLITUS PREDICTION WITH CLASSIFICATION ALGORITHMS USING WEKA

School of Technology

2024

VAASAN AMMATTIKORKEAKOULU

UNIVERSITY OF APPLIED SCIENCES

Cloud-Based Software Engineering

ABSTRACT

Author	Bishal Gyawali
Title	Diabetes Mellitus Prediction with Classification Algorithms Using Weka
Year	2024
Language	English
Pages	43 + 6 Appendix
Name of Supervisor	Rayko Toshev

Early diabetes identification is crucial for controlling chronic illnesses. This study uses WEKA to compare the performance of four classification algorithms (Multilayer Perceptron, Logistic Regression, Random Forest, and Extra Trees) for diabetes prediction.

Accuracy, precision, recall, and f-measure were evaluated across various train-test splits. The multilayer perceptron regularly outperformed others, indicating its usefulness in diabetes prediction. Logistic regression and random forest both produced encouraging results. Extra trees have regularly underperformed.

These findings emphasize the potential of classification algorithms for early diabetes diagnosis, which can help healthcare practitioners make more informed decisions. Future research might investigate sophisticated algorithms, combine many data sources, and assess therapeutic impact in real- world scenarios.

Keywords Diabetes prediction, classification algorithms, multilayer perceptron, logistic regression, random forest, extra tree

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

1 INTRODUCTION

1.1 Background of the Research

My academic interest in machine learning, particularly its applications in healthcare, combined with personal experiences witnessing family members battle diabetes, motivated this thesis. Seeing family members battle with diabetes has motivated me to utilize statistics to forecast and treat the disease. Focusing on diabetes prediction with WEKA, a user-friendly platform with powerful medical data algorithms, aligns with my objective of creating a predictive model for early diabetes identification. This research not only improves my machine learning abilities but also aims to make a significant contribution to healthcare by potentially saving lives and lowering the worldwide diabetes burden.

According to Kangra and Singh (2023, p. 1728): "Today, the world is facing a lot of chronic diseases such as heart disease, cancer, diabetes, and tuberculosis. The early detection of these illnesses is crucial. The patient must endure these diseases for a very long time. Numerous studies are being done to control these diseases. But these diseases are becoming more prevalent day by day" Therefore, there is an urgent need for further research and interventions to effectively control these chronic diseases. The World Health Organization supports this assumption, stating that noncommunicable diseases (NCDs) account for 71% of all deaths worldwide, with early identification being a major method in lowering this burden (WHO, 2018).

The healthcare sector is experiencing rapid evolution, marked by the continuous accumulation of data due to the generation and storage of vast amounts of information, including electronic medical records, reports, and diagnostic findings. Data mining plays a pivotal role in harnessing this wealth of healthcare data, facilitating the extraction of novel and valuable insights from extensive datasets. Within healthcare, data mining serves as a predictive tool for various diseases and aids physicians in the diagnostic process (Koh & Tan, 2011; Raghupathi & Raghupathi, 2014; Alpan & Ilgi, 2020).

As per WHO (2019), “Diabetes is a chronic metabolic disorder characterized by elevated blood glucose (or blood sugar) levels, leading to progressive damage in the heart, blood vessels, eyes, kidneys, and nerves. About 422 million people worldwide have diabetes, with the majority living in low- and middle-income countries.” Diabetes cases are rising globally, posing a major health threat with 1.5 million deaths annually (WHO, 2019). This rise in diabetes prevalence is also noted by the International Diabetes Federation (American Diabetes Association, 2019), which states that diabetes affects 463 million people worldwide, and this number is expected to reach 700 million by 2045. Hence Saeedi et al. (2019) provides information that diabetes has increased by nearly double in the last twenty years which calls for increased awareness and control measures.

1.2 Problem Statement

The detection of diabetes in the starting phase is important to manage the new patient without problems. It is difficult to diagnose diabetes in individuals until serious health issues appear (American Diabetes Association, 2018). It is impractical for everyone to identify the early symptoms of diabetes and try to find medical observations. They may not even know their risk status or cannot always afford proper check-ups, especially in low- and middle-income countries (American Diabetes Association, 2019).

This research demonstrates how early detection of diabetes using classification algorithms can increase prediction accuracy. It also shows how these algorithms can speed up medical response. This research uses machine learning methods such as Logistic Regression, Multilayer Perceptron, Random Forest, and Extra Tree algorithms to create strong models that enable early detection of diabetes among people with different aspects related to their health (Kangra & Singh, 2023; Saeedi et al., 2019).

1.3 Fields of Science for Thesis

This research talks about how various scientific and technical fields combine. It aims to find the best classification algorithms for predicting diabetes. The focus of this work is in the domain of computer sciences, particularly dealing with machine learning. The research involves different machine learning methods such as Logistic Regression, Multilayer Perceptron, Random Forest, or Extra Trees to create a diabetes prediction model (Hastie, Tibshirani, & Friedman, 2009). These models are developed and assessed utilizing the complimentary platform WEKA (Hall et al., 2009). The dataset, diabetes health indicators, is available on Kaggle (Teboul, 2021). It emphasizes the significance of information technology for data management and access (Grolemund & Wickham, 2017). To understand the outcomes, the results must be interpreted by employing the statistical and probability theories (Casella & Berger, 2002). The models can, however, be evaluated by use of other measures like precision, recall as well as the F-measure.

The diabetic health indicators dataset used in the study can be found on Kaggle (Teboul, 2021). The major idea is to illustrate how important it is to manage and obtain data using information technology (Grolemund & Wickham, 2017). Statistical and probability theories are necessary for interpreting the results (Casella & Berger, 2002). The study employs indicators such as accuracy. It also uses accurate Recall measures and F-measures to check the performances of the models. Understanding the statistical basis of these variables makes the study possible. This allows for more accurate judgments regarding each method's performance (James et al., 2013). Furthermore, the results are consistent with the wider subject of data science. Data science entails data gathering, cleaning, preprocessing, analysis, and model building (Provost & Fawcett, 2013). This study utilizes several data science methods. It analyzes the diabetes health indicator dataset and creates categorization models.

1.4 Research Philosophy and Time Horizon

This study on diabetes prediction is consistent with a positivist worldview. Positivism stresses the scientific method and objective inquiry to learn about the world (Creswell, 2009). It is considered that there is an outside world. Observation and experimentation can quantify this reality, which is independent of the observer (Mertens, 2010; Phillips & Burbules, 2000).

This study takes a positivist approach for various reasons. First, it is based on well-defined machine-learning algorithms. These are objective approaches to data analysis (Russell & Norvig, 2019). Second, the study makes use of a publicly available dataset that includes measurable diabetic health markers (Teboul, 2021). The assessment measures (accuracy, precision, recall, and F-measure) are all objective and statistically valid (Sokolova & Lapalme, 2009). Finally, WEKA is a popular platform. It is used for machine-learning tasks. WEKA offers a consistent environment for model development and validation (Hall et al., 2009).

This study utilized a secondary dataset from the Teboul (2021). This is coherent with the cross-sectional approach. The data only represents one point in time. The secondary dataset eliminates the need for data collection. This study during the timeline focused on data analysis and also focused on building a classifier. Experimentation and result interpretation occurred over four months.

1.5 Research Questions and Objective

The main research questions are:

- How do Logistic Regression (LR), Multilayer Perceptron (MLP), Random Forest (RF), and Extra Tree (ET) classifiers compare in terms of evaluation metrics (accuracy, precision, recall, and F-measure) for predicting diabetes?

- How does varying the percentage split (70%, 80%, and 90%) of dataset impact the performance of LR, MLP, RF, and ET classifiers for diabetes prediction?

The objective of the research is to compare the performance of Logistic Regression (LR), Multilayer Perceptron (MLP), Random Forest (RF), and Extra Tree (ET) classifiers using WEKA for diabetes detection.

Research on Diabetes depends on being able to detect it early because this way it is possible to treat the condition promptly before it leads to the problems associated with it (American Diabetes Association, 2018). Precise classification algorithms assist in detecting vulnerable individuals, thereby making detection strategies more easily available and faster (Zheng et al., 2017). These algorithms also aid in risk classification, customizing treatment strategies, and prioritizing high-risk patients (He et al., 2019). Identification is crucial in following the progress of diseases and in making adjustments to the treatment schedules. Research on these algorithms not only increases diagnosis accuracy but also enhances diabetes research by encouraging innovation and promoting evidence-based healthcare practices (Hinton, 2018).

1.6 Research Significance

This study contributes to the advancement of machine learning for early diabetes diagnosis by:

- Early diagnosis through effective prediction algorithms can lead to timely intervention and potentially better patient outcomes.
- By identifying high-performing algorithms, healthcare organizations can prioritize implementation of the most effective models, optimizing resource allocation for diabetes prediction.

- Findings can inform the development of improved medical decision support systems, aiding healthcare professionals in more accurate and efficient diabetes risk assessments.

1.7 Structure of the Thesis

This thesis is structured in five major chapters, references, and appendices to systematically explore the potential of machine learning for diabetes prediction.

Chapter 1: The introduction establishes the basis. It delves into the consequences of diabetes and the difficulties associated with its diagnosis. The potential of machine learning for prediction (using WEKA) is then discussed, followed by the study challenge, questions, and objectives. Finally, the chapter discusses the study's significance.

Chapter 2: The literature review covers existing research on machine learning for diabetes prediction, with an emphasis on algorithms utilized in WEKA. It addresses previous research algorithms, datasets, and assessment techniques.

Chapter 3: Methodology provides the reader with the tools. This chapter describes WEKA and the methods used, followed by information on the diabetes dataset and any data pre-processing approaches.

Chapter 4: Experiment and Analysis focuses on practical application. It describes the data splitting method, assessment measures, and outcomes for each algorithm. Finally, it assesses the findings, emphasizing their strengths and flaws.

Chapter 5: Conclusion and Future. Work concludes the thesis. It presents major discoveries (including algorithm performance), recognizes limits, and suggests future research directions to enhance diabetes prediction with machine learning.

2 LITERATURE REVIEW

2.1 Machine Learning

According to Fuentes (2023), "Machine Learning (ML) focuses on the creation of systems or models that can learn from data and improve their performance in specific tasks without the need to be explicitly programmed, making them learn from past experiences or examples to make decisions on new data." ML has two main types: Supervised learning and Unsupervised learning.

Supervised learning, the most common type of machine learning, uses labeled datasets to predict categorical or continuous values by mapping input features to labels (Crabtree, 2023). Algorithms like linear regression, logistic regression, decision trees, and support vector machines are commonly used (Crabtree, 2023). It is essential for tasks such as image recognition (Crabtree, 2023). In contrast, **unsupervised learning** uses unlabeled datasets to identify patterns and relationships, employing algorithms like k-means and PCA for clustering and dimensionality reduction (Crabtree, 2023). This method is useful in marketing for customer segmentation without pre-existing labels (Crabtree, 2023). Both approaches are crucial for enabling intelligent machine learning (Goodfellow et al., 2016).

2.2 Classification

According to Han et al. (2012), Building models in data analysis for categorizing points of data is referred to as classification. The model's build is called classifiers. It identifies patterns that distinguish classes which helps them to predict class labels accurately. Applications of classification are widespread and include image recognition, spam email detection, medical diagnosis, customer segmentation, and sentiment analysis (Han et al., 2012; James et al., 2013).

To ensure the effectiveness of the classifier in real-world applications, evaluation matrixes are used (Sokolova & Lapalme, 2009). These metrics assess the performance of the classifiers, ensuring they can accurately and reliably predict class labels across various datasets and conditions. Evaluation metrics are critical for validating the practical utility of classifiers, guiding improvements, and ensuring their robustness in diverse applications.

2.3 Diabetes Mellitus

As per the WHO (2019), “Diabetes Mellitus (DM) is a chronic metabolic disorder characterized by elevated blood glucose (or blood sugar) levels, leading to progressive damage in the heart, blood vessels, eyes, kidneys, and nerves.” The global burden of diabetes is substantial and continues to grow, affecting millions worldwide (Saeedi et al., 2019). Kangra and Singh (2023) describe diabetes as having four main classifications: type-1, type-2, pre-diabetes, and gestational diabetes. Type-1 diabetes is a chronic condition where the immune system attacks and destroys beta cells in the pancreas that release insulin (Kangra & Singh, 2023; Atkinson et al., 2014). Type-2 diabetes causes low insulin secretion and excessive blood sugar levels (Kangra & Singh, 2023; DeFronzo et al., 2015). Pre-diabetes is characterized by raised blood sugar levels that are not high enough to be identified as type 2 diabetes (Kangra & Singh, 2023; Tabák et al., 2012). Pregnant women with high blood sugar are diagnosed with gestational diabetes (Kangra & Singh, 2023; McIntyre et al., 2019).

2.4 Related Works

Kangra and Singh (2023) compared the ML algorithms DT, KNN, LR, NB, RF, and SVM for predicting diabetes using WEKA. They found SVM best for the Pima Indians Diabetes dataset (74% accuracy), while KNN and RF excelled on the Germany dataset (98.7% accuracy) (Kangra & Singh, 2023).

Anusha (2023) researched the application of machine learning algorithms to predict diabetes. The study compared the LR, RF, and KNN algorithms. Logistic regression had the highest accuracy, at 82.7% (Anusha, 2023).

Bhat et al. (2023) suggested an ML technique for early diabetes prediction (PIDD dataset). The authors tested many algorithms (SVM, LR, KNN, and RF), with random forest attaining the greatest accuracy (92.85%). K-fold cross-validation was utilized for verification.

Özsezer and Mermer (2022) predicted diabetes risk using machine learning models on a diabetes health indicators dataset (253,680 entries, 21 variables from CDC). KNN achieved 0.74 accuracy; LR 0.72; DT 0.84; RF 0.84; NB 0.84. Splitting data 80:20 (training: testing) resulted in comparable accuracy for Random Forest (84.4%) and Decision Tree (84.7%), making them top performers for risk estimation (Özsezer & Mermer, 2022).

In their study, Chang et al. (2022) utilized machine learning algorithms to predict diabetes diagnosis on a CDC health indicators dataset (253,680 entries, 21 variables). They compared various algorithms (DT, RF, KNN, LR, and NB) based on accuracy, precision, recall, and F1 score. Random Forest has the best accuracy rate (82.26%) (Chang et al., 2022).

Alpan and Ilgi (2020) evaluated data mining techniques for diabetes classification on a UCI dataset (520 instances, 17 attributes). Classification algorithms (Bayes Network, NB, J48, RT, RF, KNN, and SVM) were evaluated. KNN achieved the highest accuracy (98.07%).

Sisodia and Sisodia (2018) evaluated the utility of classification algorithms in predicting diabetes. They discovered that Naive Bayes outperformed Decision Trees and Support Vector Machines (SVM), achieving an accuracy of 76.30%. Their findings demonstrate the potential of machine learning to improve early detection and prognosis of diabetes, thereby potentially reducing the burden on healthcare systems.

Research conducted by Alehegn, Joshi, and Mulay (2018) aimed at establishing how effective machine learning algorithms are in predicting the arrival of Diabetes Mellitus (DM). The Pima Indian Diabetes Data Set (PIDD) yields a high accuracy of 90.36% when using the Ensemble Method (PEM). Single algorithms, such as Decision Trees, have lower accuracy, ranging from 85% to 80%. They conclude that PEM outranks the Decision tree.

Hassan Malaserene and Leema (2020) used the Pima Indian Diabetes Dataset to predict Diabetes Mellitus (DM) with classification techniques. They tested different algorithms, including Decision Trees, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM). SVM performed best with an accuracy of 90.23%. This shows its value in classifying medical data.

Patil and Tamane (2018) published a paper comparing different algorithms to predict whether someone will be diagnosed with diabetes mellitus (DM). They evaluated Logistic Regression, KNN, SVM, and Decision Trees using the Pima Indian Diabetes dataset. According to their findings, Logistic Regression and Gradient Boost were the best algorithms. They achieved an accuracy rate of 79%, making them useful for early DM diagnosis.

Khanam and Foo (2021) studied how well neural networks and machine learning algorithms can predict diabetes using data from the Pima Indian Diabetes dataset. They tested Support Vector Machine and Logistic Regression. Both algorithms had a 70% accuracy rate. However, they discovered that a neural network model performed even better, with an accuracy of 88.6%. This outperformed all other methods they tried. They concluded that neural networks and machine learning methods are effective for predicting diabetes.

3 METHODOLOGY

3.1 Background

This section outlines the study process, following the common steps employed by most researchers (see Figure 1). Steps are elaborated in the following subsection:

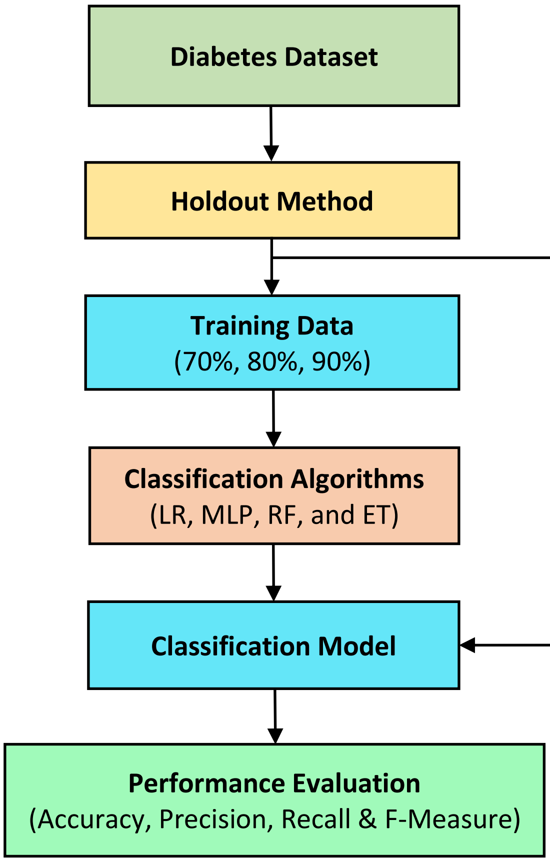

Figure 1. Block diagram of research methodology

3.3 Holdout Method

To evaluate the classifier, the holdout method is used. The holdout method is a data-splitting technique that divides the original dataset into two parts: training and testing. The training set is used to construct and train the data mining model, and the test set is used to assess the model's ability to generalize to new data. In this research, different train-test splits (70:30, 80:20, and 90:10) are used.

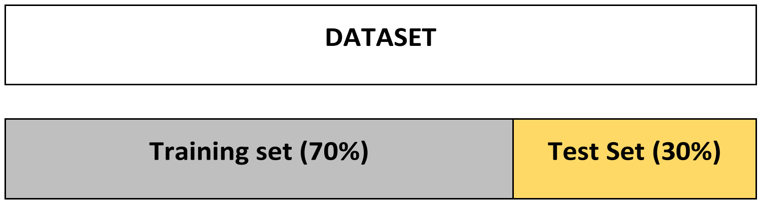

Figure 2. An example of a holdout method for split ratio of 70:30

3.4 Algorithms

In this study, four classification algorithms are implemented to train and test the classifiers.

Random Forest is a supervised learning technique used for both classification and regression tasks, although it is more commonly employed for classification problems (Anusha, 2023). The strength of a random forest increases with the number of trees it contains (Anusha, 2023). The random forest algorithm constructs decision trees based on data samples, gathers predictions from each tree, and then selects the most frequent prediction as the final output (Breiman, 2001). This ensemble technique prevents overfitting by averaging the results from multiple trees, making it generally more robust than single decision trees (Liaw & Wiener, 2002). Figure 3 illustrates the workflow of the Random Forest Algorithm (Anusha, 2023).

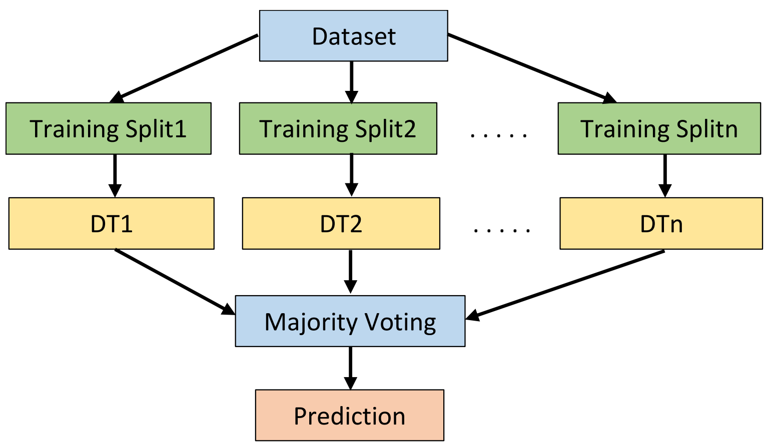

Figure 3. Working of Random Forest algorithm (Adapted from Anusha, 2023)

Logistic Regression (LR) is a statistical model that utilizes a logical function to create a binary-dependent variable, estimating the relationship between dependent and independent variables based on probabilities, and accommodating categorical dependent variables (Bhat et al., 2023). Maalouf (2011) emphasizes the enduring relevance of LR in data mining, especially for binary classification tasks.

According to Nosratabadi et al. (2021), "Multilayer perceptron (MLP) is a type of neural network that has a supervised learning technique using the back-propagation method" (p. 408). As illustrated in Figure 4, MLP utilizes a layered structure typically consisting of an input layer, one or more hidden layers, and an output layer. Each neuron within a layer connects to all neurons in the subsequent layer, enabling the network to learn complex, non-linear relationships (Nosratabadi et al., 2021).

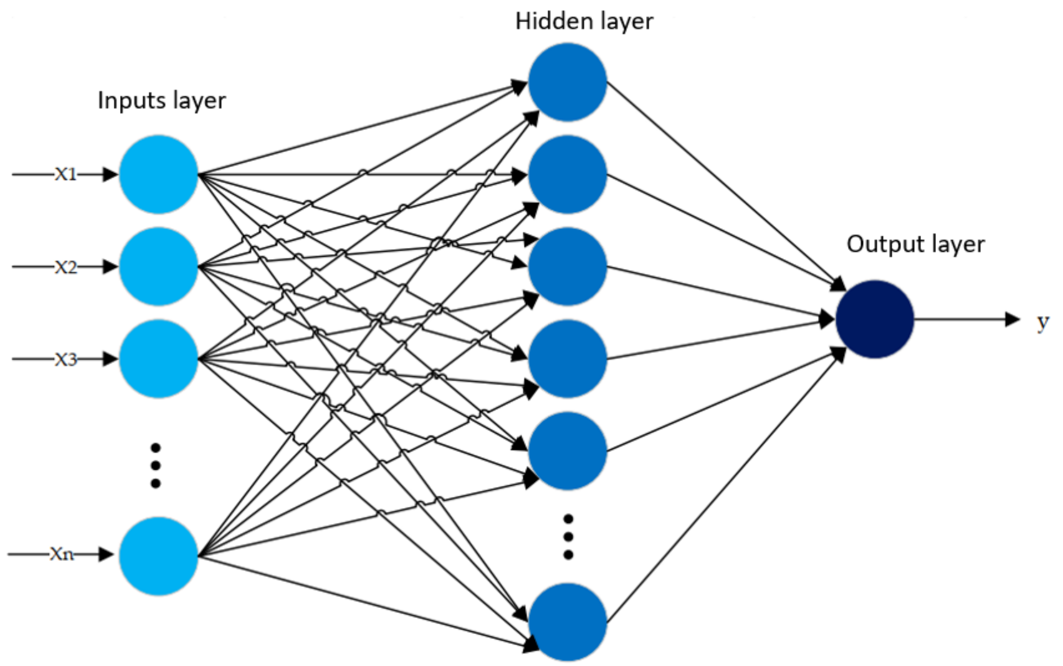

Figure 4. Architecture of the multilayer perceptron neural networks. (Adopted from Nosratabadi et al., 2021)

Geurts et al. (2006) introduce a novel tree-based ensemble method, named Extra-Trees, for supervised classification and regression tasks. This algorithm randomizes both attribute and cut-point selection strongly during tree node splitting, potentially leading to the construction of totally randomized trees independent of the output values. The degree of randomization can be adjusted through a parameter choice tailored to specific problems. Besides its computational efficiency, the algorithm's main strength lies in its robustness and accuracy (Geurts et al., 2006).

According to Saeed (2023), the Extra-Trees classifier is a bagging machine learning algorithm that utilizes random trees built from training data samples to achieve high classification accuracy and reduce overfitting.

3.5 Performance Evaluation

A confusion matrix is a table used to evaluate the performance of a classifier in a classification task (Han & Kamber, 2012).

	Actual Positive (YES)	**Actual Negative)** (NO)
Predicted Positive (YES)	True Positive (TP)	False Positive (FP)
Predicted Negative (NO)	False Negative (FN)	True Negative (TN)

Figure 5. Components of confusion matrix

The commonly used metrics to evaluate the performance of classification algorithms are accuracy, precision, recall, and F-measure (Sokolova et al., 2006). Accuracy reflects the proportion of correctly classified instances across all classes (Equation 1). It provides a general overview of the model's performance but can be misleading in imbalanced datasets (Sokolova et al., 2006). Precision (Equation 2) measures the proportion of true positives among all instances predicted as positive. It reflects the model's ability to avoid false positives. Recall (Equation 3) focuses on the proportion of true positives identified by the model from all actual positive cases. It highlights the model's ability to avoid false negatives. F-measure (Equation 4), also known as the F1-score, combines precision and recall into a single metric, providing a more balanced view of the model's performance.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

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

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$F - \text{Measure} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (4)$$

4 EXPERIMENTAL SETUP AND RESULTS ANALYSIS

4.1 Background

This chapter gives information about the implementation and performance analysis of our used algorithms LR, MLP, RF, and ET for diabetes prediction. The whole experiments were performed on Windows 10 64-bit operating system. The system specifications used for this experiment are an Intel® Core™ i5-3230M CPU with 8 GB RAM. WEKA version 3.8.6 was also installed in this system.

The study utilized the Waikato Environment for Knowledge Analysis (WEKA) to achieve the research objective (Witten, Frank, Hall, & Clark, 2011) which is a widely accepted suite of data mining tools. It is popular for its broad collection of machine-learning algorithms designed for data mining tasks and related analyses (Witten et al., 2011). It provides a strong platform for various analyses with tools for data preparation, clustering, classification, association, regression, and visualization (Hall et al., 2009). Weka is written in Java and distributed under the General Public License (Hall et al., 2009). It was developed at the University of Waikato, New Zealand in an attempt to ease the process of studying by promoting openness and the ease of accessing data for research (Hall et al., 2009).

4.2 Dataset Description

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's main framework of health-related phone surveys that collect state information around U.S. inhabitants with respect to their health-related chance behaviors, constant well-being conditions, and utilization of preventive administrations (*Behavioral Risk Factor Surveillance System*, 2017). BRFSS presently collects information in all 50 states as well as the area of Columbia and three U.S. domains since 1984. BRFSS completes more than 400,000 grown-up interviews each year, making it the biggest ceaselessly conducted well-being study framework in the world (*Behavioral Risk Factor Surveillance System*, 2017). In this study, a CSV of the dataset accessible on Kaggle for the year 2015 was utilized (Teboul, 2021). This dataset contains the responses of 253,680 respondents and 21 features with two classes on the target variable ("NO" means no diabetes or "YES" means prediabetes or diabetes). These features are either questions straightforwardly inquired of members, or calculated factors based on personal member response (Teboul, 2021).

Table 1. Data features description (Adapted from Özsezer & Mermer, 2022)

Features Name	**Description**
High Blood Pressure (BP)	0 = No High BP, 1 = High BP
High Cholesterol (Chol)	0 = No High Cholesterol, 1 = High Cholesterol
Cholesterol Check	0 = No, 1 = Yes
Body Mass Index (BMI)	Numerical value representing weight adjusted for height
Smoker	0 = No (has not smoked 100+ cigarettes in lifetime), 1 = Yes
Stroke	0 = No, 1 = Yes (has had a stroke)
Heart Disease or Attack	0 = No (no coronary heart disease or myocardial infarction), 1 = Yes
Physical Activity	0 = No physical activity in past 30 days (excluding work), 1 = Yes
Fruit Consumption	0 = No, 1 = Consumes fruits 1 or more times per day
Vegetable Consumption	0 = No, 1 = Consumes vegetables 1 or more times per day
Heavy Alcohol Consumption	0 = No, 1 = Meets criteria for heavy drinking
Health Insurance Coverage	0 = No, 1 = Has any health insurance/prepaid plan
Difficulty Affording Healthcare	0 = No, 1 = Faced difficulty affording healthcare in the past year
General Health Perception	Scale 1-5 (1 = Excellent, 5 = Poor)
Mental Health Days	Number of days in the past 30 with poor mental health (scale 1-30)
Physical Health Days	Number of days in the past 30 with poor physical health (scale 1-30)
Difficulty Walking	0 = No, 1 = Has serious difficulty walking/climbing stairs
Sex	0 = Female, 1 = Male
Age Category	13-level category (18-29 to 80+)

Education Level	Scale 1-6 (Never attended school to college graduate+)
Income Category	Scale 1-8 (less than $10,000 to $75,000 or more)
Diabetes Status	"NO" = No Diabetes, "YES" = Diabetes

4.3 Experimental Setup and Parameters

WEKA version 3.8.6 was used for the experiments. It is a popular open-source machine-learning software tool. Table 2 provides an overview of the experimental parameters used in the investigation.

Table 2. Experimental Parameters

Classifer1: "weka.classifiers.functions.Logistic" **Classifer2:** "weka.classifiers.functions.MultilayerPerceptron" **Classifer3:** "weka.classifiers.trees.RandomForest" **Classifer4:** "weka.classifiers.trees.ExtraTree"
Relation: "diabetes_binary_health_indicators_BRFSS2015.csv"
Instances: "253680"
Test options: "Percentage Split (70:30, 80:20 and 90:10)"

The Extra-Trees classifier's run information is displayed in Figure 6 and includes information about the scheme, relation, number of instances, attributes, and test mode. Similar trials employing training-testing splits of 70:30, 80:20, and 90:10 was carried out for Random Forest, Multilayer Perceptron, and Logistic Regression analysis.

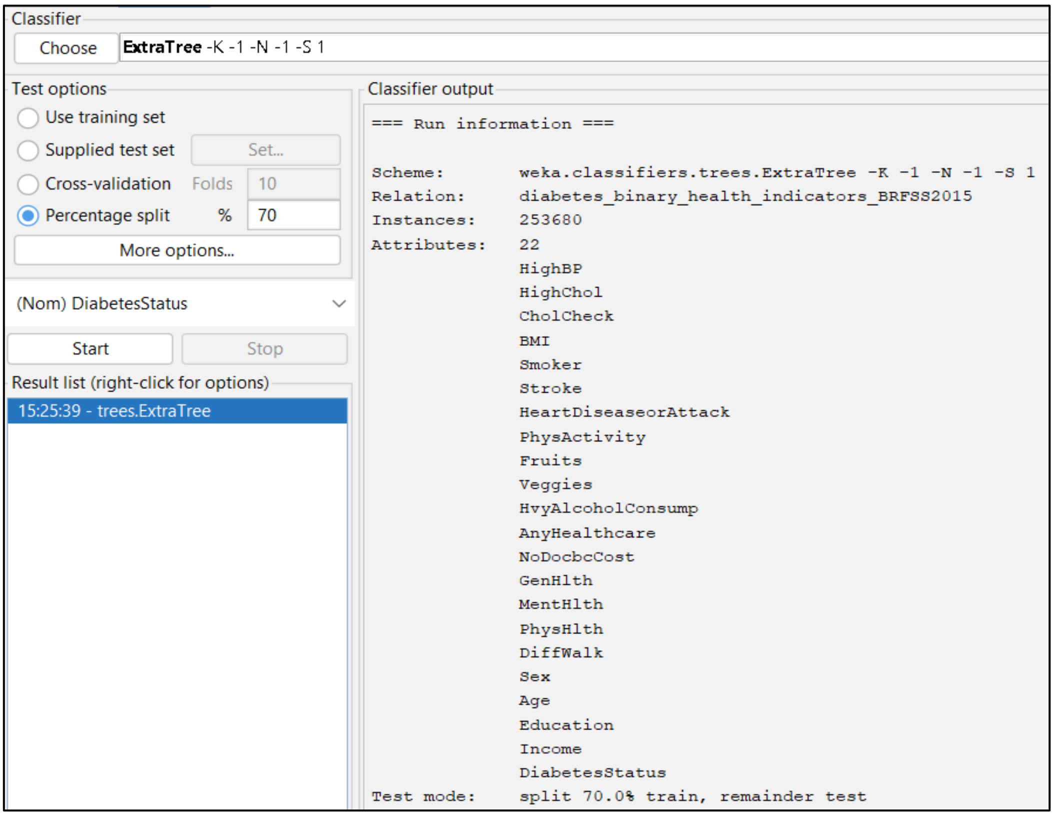

Figure 6. Output of Extra Tree algorithm showing run information in WEKA

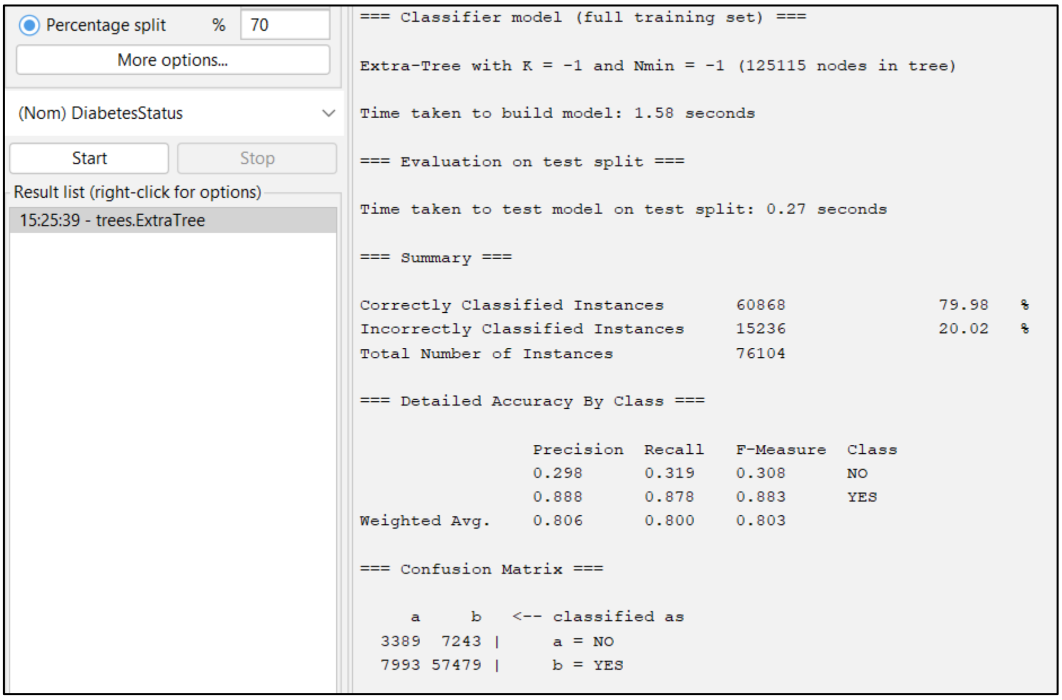

Figure 7. Classified instances by Extra Tree algorithm

Figure 7 presents the WEKA output, where the Extra-Trees algorithm achieved a classification accuracy of 79.98%. The study evaluated other algorithms (Logistic Regression, Multilayer Perceptron, Random Forest) using training/testing splits of 70:30, 80:20, and 90:10.

4.5 Results Analysis

This section presents and analyzes the performance of four machine learning algorithms (MLP, LR, RF, and ET). This research evaluates these algorithms using various train-test splits (70:30, 80:20, and 90:10) and assesses their efficacy through key performance metrics: accuracy, precision, recall, and f-measure.

Table 3. Classification performance of algorithms (70:30 split)

	70:30 Split			
Algorithm	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F-Measure (%)**
Multilayer Perceptron	86.59	83.5	86.6	82.8
Logistic Regression	86.43	83.2	86.4	83
Random Forest	85.82	82.4	85.8	83
Extra Tree	79.98	80.6	80	80.3

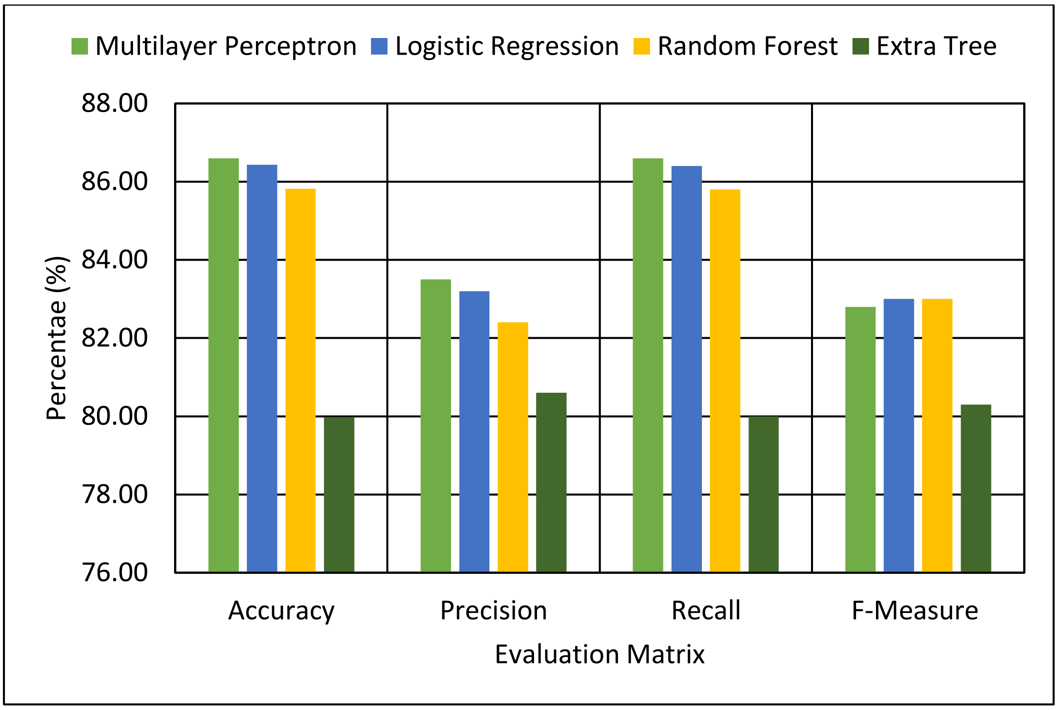

Figure 8. Classification performance of algorithms (70:30 split)

For the 70:30 split (see Table 3 and Figure 8), MLP marginally surpasses LR and RF in accuracy and recall, while ET exhibits the lowest performance across all metrics.

Table 4. Classification performance of algorithms (80:20 split)

	80:20 Split			
Algorithm	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F-Measure (%)**
Multilayer Perceptron	86.57	83.6	86.6	83
Logistic Regression	86.35	83.1	86.4	82.8
Random Forest	85.78	82.3	85.8	82.9
Extra Tree	79.90	80.5	79.9	80.2

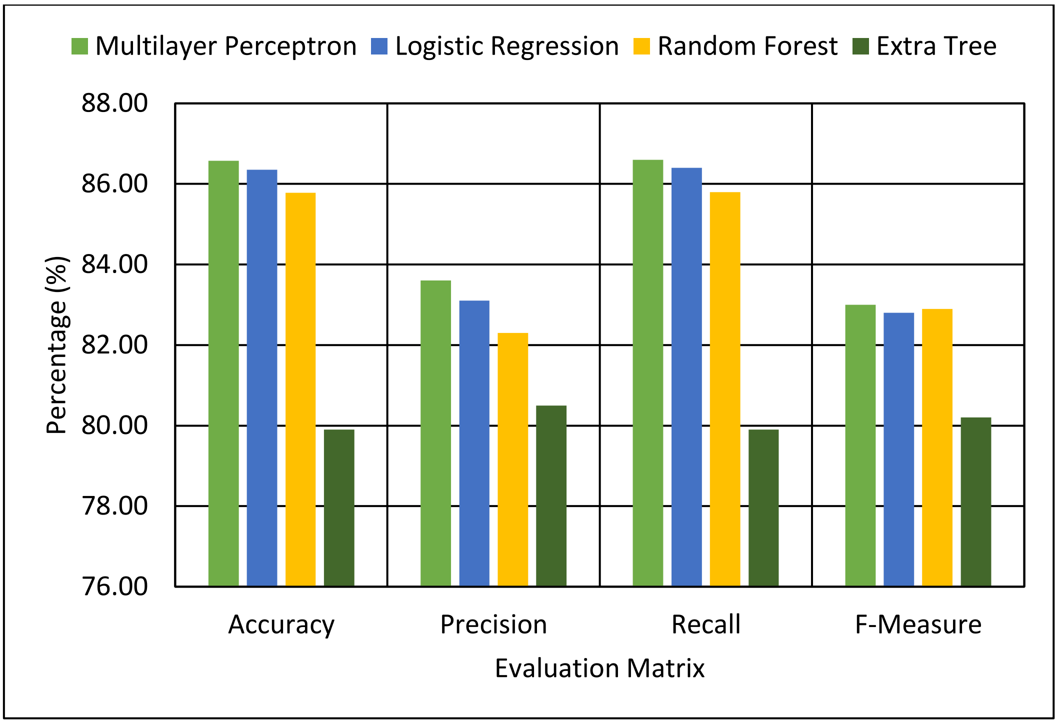

Figure 9. Classification performance of algorithms (80:20 split)

The trend persists in the 80:20 split (see Table 4 and Figure 9), with MLP consistently leading in accuracy, precision, recall, and f-measure, followed closely by LR and RF, while ET consistently performs the poorest.

Table 5. Classification performance of algorithms (90:10 split)

	90:10 Split			
Algorithm	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F-Measure (%)**
Multilayer Perceptron	86.35	83.1	86.3	82.9
Logistic Regression	86.26	82.9	86.3	82.7
Random Forest	85.66	82.2	85.7	82.9
Extra Tree	79.81	80.4	79.8	80.1

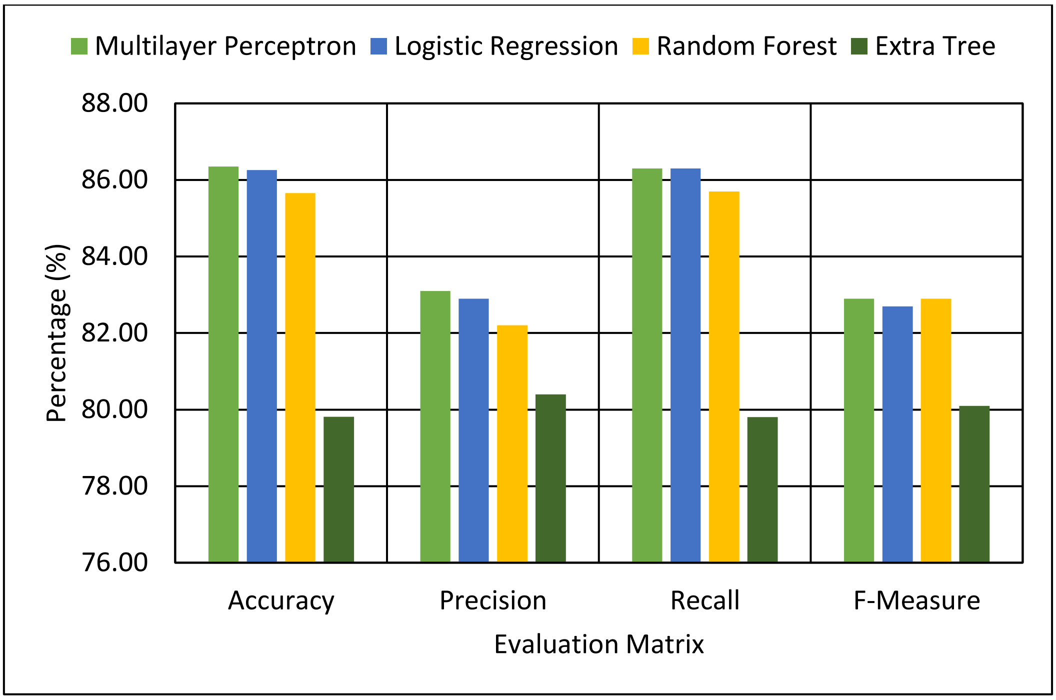

Figure 10. Classification performance of algorithms (90:10 split)

In the 90:10 split (Table 5 and Figure 10), performance metrics slightly decrease across the board due to the smaller test set; however, the relative performance of algorithms remains consistent. MLP maintains a slight edge in accuracy, recall, and f-measure, with LR and RF closely trailing behind.

Throughout all train-test splits, MLP consistently outshines with the highest accuracy, precision, recall, and f-measure, establishing itself as the most robust algorithm for diabetes mellitus prediction in this study. LR and RF also perform commendably, with LR exhibiting slightly superior results. Conversely, the consistently inferior performance of the ET al algorithm underscores its ineffectiveness for this dataset and task.

Table 6. Accuracy across different train-test split ratios

Algorithm	70:30 Split	80:20 Split	90:10 Split
Multilayer Perceptron	86.59	86.57	86.35
Logistic Regression	86.43	86.35	86.26
Random Forest	85.82	85.78	85.66
Extra Tree	79.98	79.90	79.81

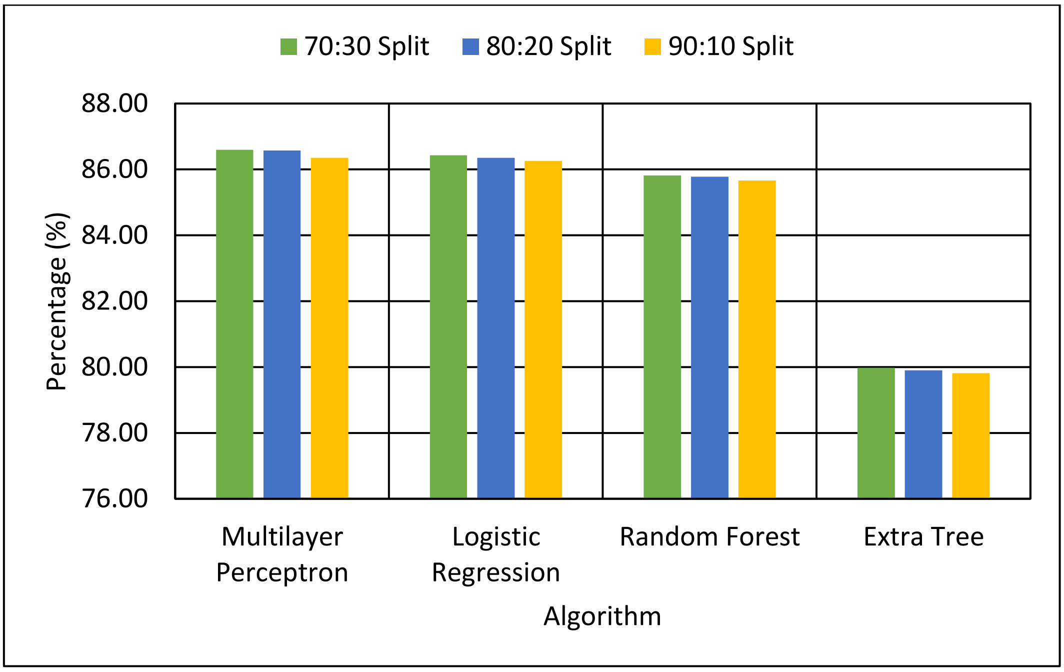

Figure 11. Accuracy across different train-test split ratios.

For accuracy (see Table 6 and Figure 11), the MLP shows stable performance across all split ratios, with minor decreases as the test set size reduces (86.59% for 70:30, 86.57% for 80:20, and 86.35% for 90:10). LR and RF exhibit similar trends, with minor variations in accuracy (LR: 86.43% to 86.26%, RF: 85.82% to 85.66%). The ET classifier consistently shows the lowest accuracy across all splits, with minor fluctuations around 80%. This suggests that larger test sets may provide more reliable performance estimates for these models.

Table 7. Precision across different train-test split ratios

Algorithm	70:30 Split	80:20 Split	90:10 Split
Multilayer Perceptron	83.5	83.6	83.1
Logistic Regression	83.2	83.1	82.9
Random Forest	82.4	82.3	82.2
Extra Tree	80.6	80.5	80.4

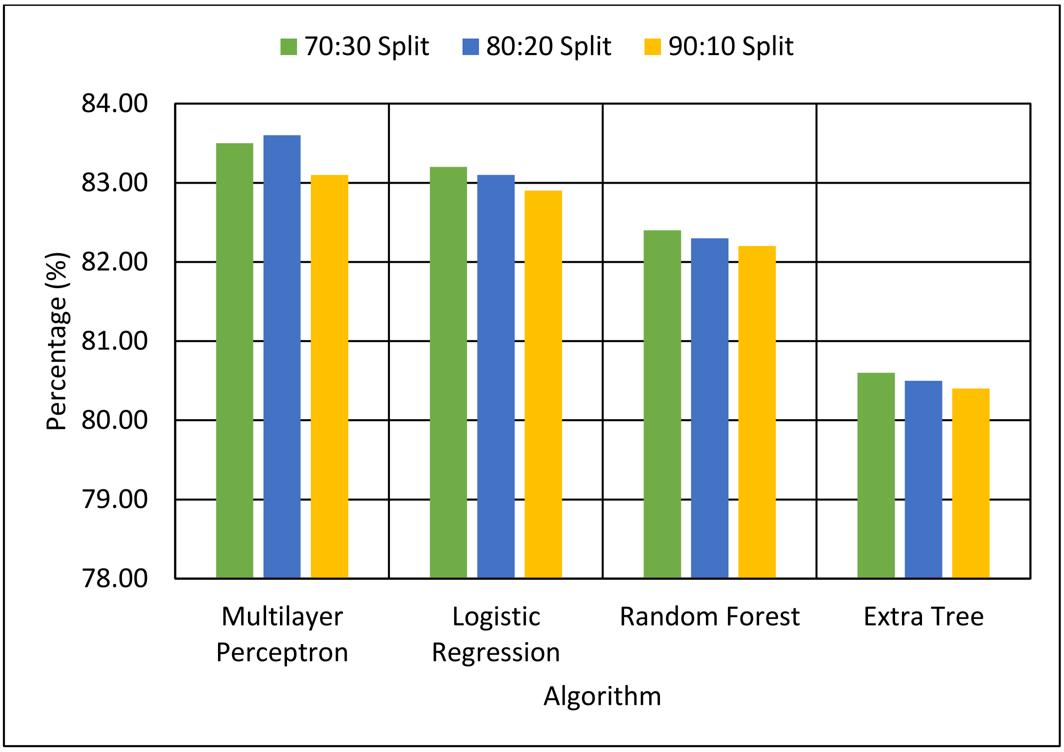

Figure 12. Precision across different train-test split ratios

Regarding Precision (see Table 7 and Figure 12), both MLP and LR maintain stable precision across different splits, with only slight variations (MLP: 83.5% to 83.1%, LR: 83.2% to 82.9%). RF also shows stable precision, with values ranging from 82.4% to 82.3%. ET's precision remains around 80.6% to 80.4%, the lowest among the models. This indicates that precision is relatively unaffected by the split ratio, suggesting that the models' ability to correctly identify positive cases (diabetes) is stable regardless of the test set size.

Table 8. Recall across different train-test split ratios

Algorithm	70:30 Split	80:20 Split	90:10 Split
Multilayer Perceptron	86.6	86.6	86.3
Logistic Regression	86.4	86.4	86.3
Random Forest	85.8	85.8	85.7
Extra Tree	80	79.9	79.8

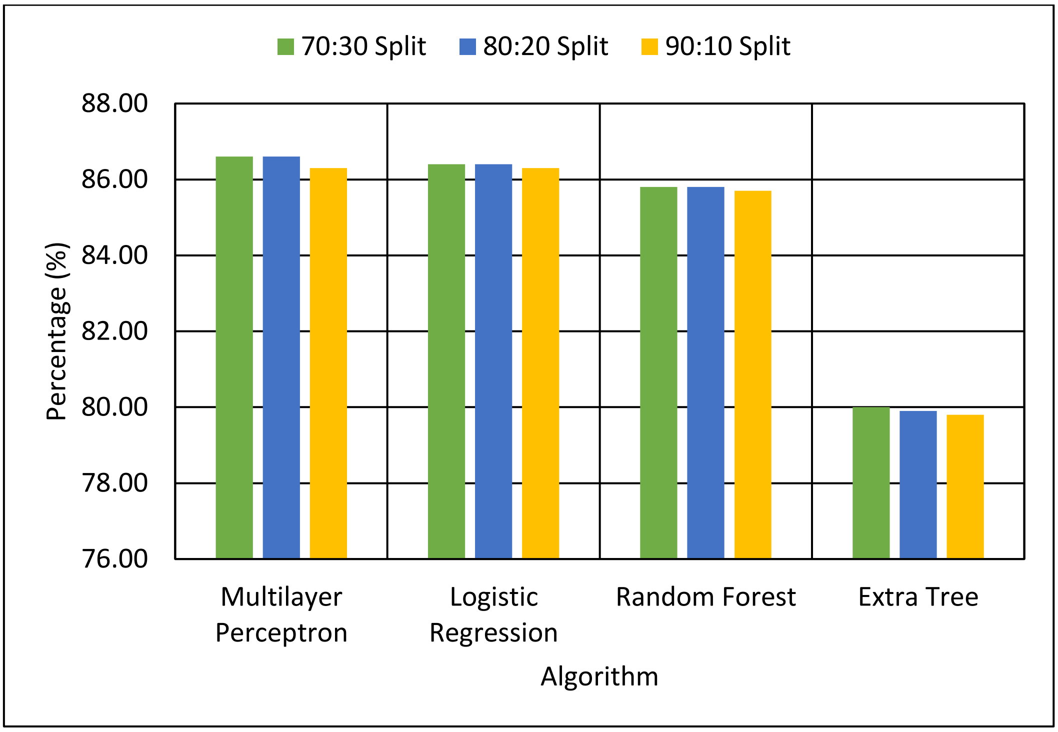

Figure 13. Recall across different train-test split ratios

In terms of Recall (see Table 8 and Figure 13), MLP and LR exhibit high recall values, with minimal changes across splits (MLP: 86.6% to 86.3%, LR: 86.4% to 86.3%). RF shows a slight increase in recall with a smaller test set (85.8% to 85.7%), while ET has the lowest recall, ranging from 80% to 79.8%. The consistent recall values across split ratios suggest that the model's ability to identify true positive cases is relatively unaffected by the size of the test set.

Table 9. F-measure across different train-test split ratios

Algorithm	70:30 Split	80:20 Split	90:10 Split
Multilayer Perceptron	82.8	83	82.9
Logistic Regression	83	82.8	82.7
Random Forest	83	82.9	82.9
Extra Tree	80.3	80.2	80.1

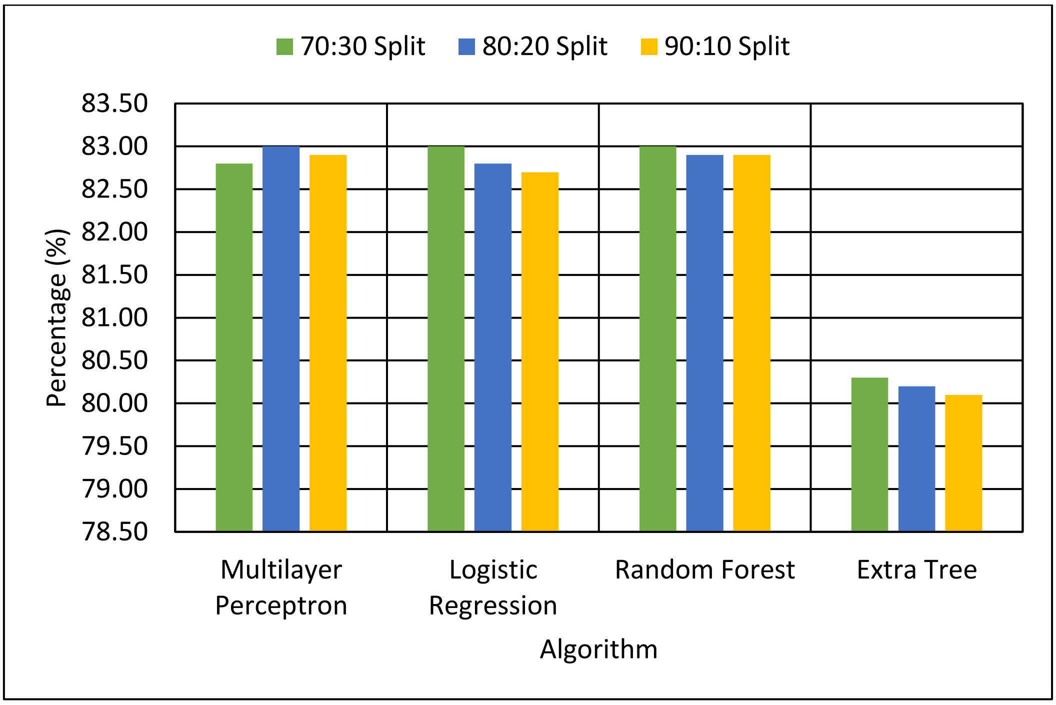

Figure 14. F-measure across different train-test split ratios

For f-measure (see Table 9 and Figure 14), which balances precision and recall, MLP and LR display stable values with slight increases in consistency (MLP: 82.8% to 82.9%, LR: 83% to 82.7%). RF also maintains a stable F-measure across splits (83% to 82.8%), while ET's F-measure is consistently lower, around 80.3% to 80.1%. The consistent F-measure values across split ratios suggest that the models' overall performance, balanced between precision and recall, remains relatively stable regardless of the test set size.

5 CONCLUSION AND FUTURE WORK

5.1 Conclusion

The fight against chronic diseases like diabetes is becoming increasingly important as more and more people are affected worldwide. Early detection is crucial for effective management. This study investigated how well different machine learning algorithms could be used to predict diabetes using WEKA. This study compared four algorithms: multilayer perceptron, logistic regression, random forest, and extra trees. The study tested the algorithms using different portions of the data for training and testing (70:30, 80:20, and 90:10).

Research Question: How do LR, MLP, RF, and ET classifiers compare in terms of evaluation metrics (accuracy, precision, recall, and F-measure) for predicting diabetes?

Answer: The analysis of classification algorithms for predicting diabetes revealed distinct performance differences. The analysis of classification algorithms for predicting diabetes showed different results. MLP had the highest accuracy with rates of 86.59%, 86.57%, and 86.35% for the 70:30, 80:20, and 90:10 splits. Due to this MLP is the best classifier as it has good precision, recall, and F-measure. LR followed with accuracy rates of 86.43%, 86.35%, and 86.26%, and stable precision and recall. So, it indicates reliable performance. RF performed slightly lower but was still competent and it had good recall and F-measures. However, its accuracy was a bit lower. ET had the worst performance, with accuracy around 79.98%, 79.90%, and 79.81%, and lower precision, recall, and F-measures as compared to the others. Overall, MLP was the best for classifying diabetes, followed by LR and RF. ET is the least effective algorithms as compared to others. This shows the importance of choosing the right machine learning models for accurate diabetes predictions.

Research Question: How does varying the percentage split (70%, 80%, and 90%) of dataset impact the performance of LR, MLP, RF, and ET classifiers for diabetes prediction?

Answer: Varying the percentage split (70%, 80%, and 90%) of the dataset impacts the performance of LR, MLP, RF, and ET classifiers for diabetes prediction. MLP maintains high and stable accuracy across all splits. Its accuracy slightly decreases from 86.59% at 70:30 to 86.35% at 90:10. Its precision, recall, and F-measure also remain consistent. This shows MLP performs well regardless of the train-test split. LR also shows similar stability. Its accuracy decreases slightly from 86.43% to 86.26%. Precision and recall values stay around 83.1% and 86.4%. The F-measure scores remain stable. This indicates LR is reliable across different splits. RF experiences a small decline in accuracy from 85.82% to 85.66%. Precision and recall stay consistent around 82.3% and 85.8%. This suggests RF performs well with slight variation across splits. ET consistently has the lowest performance. Its accuracy ranges from 79.98% to 79.81%. Precision, recall, and F-measure values are lower. This suggests ET is less effective regardless of the data split. Overall, MLP and LR are the most reliable classifiers for diabetes prediction. They show minimal impact from varying data splits. RF performs well but shows more variation. ET is the least effective across all metrics.

These results help explore machine learning for early diabetes diagnosis. Healthcare professionals can use this knowledge to choose the right algorithms for risk assessment. They should consider factors like model accuracy and interpretability. The study highlights the importance of checking model performance across different data splits. This is crucial for real-world applications.

In conclusion, insights have been provided by this research on how machine learning algorithms may be used in forecasting diabetes. Future research can explore advanced algorithms and feature engineering techniques. These efforts aim to further improve model performance and generalizability.

5.2 Future Work

Some areas for future research include:

- Considering using advanced models like deep learning and ensemble methods to improve diabetes prediction.
- Using feature engineering to create informative and scalable features to help building predictive models that improve performance.
- Developing features selection methodologies to detect the features that are most critical so that in the process, the model is easier to comprehend (interpretative) and performs well on unseen data (generalizability).
- Incorporation with Clinical Practice in Real-Life Situation.

Future diabetes prediction research should focus on more robust models such as Deep Learning and Ensembles. These new methods can work much better than older ones, especially in the health sector. For instance, various neural networks variations can identify complex patterns. Ensemble forecasting requires combining many models to enhance precision. Using these technologies could lead to more accurate and quicker diabetes diagnoses, significantly improving patient outcomes. To find features that improve model performance, we need to use feature engineering methods. These methods create features full of useful information. Machine learning algorithms make accurate predictions by using data closely related to real-world information. This works best when we use extra knowledge about specific domains to create new important features. This might include finding new ways to update information from medical records and combining indicators to create a clearer picture of diabetes risk factors.

Furthermore, when we aim to discover key traits and create easy-to-understand models applicable to various datasets, the crucial step is employing feature selection methods. Streamlining models by concentrating on essential parts enhances clarity, which is crucial in healthcare since professionals depend on accurate, clear predictions. Models with fewer but significant factors enhance clinical decision-making by increasing accuracy and clarity, aiding practitioners in making better-informed choices. Furthermore, it is vital to integrate these models into actual medical practice. This involves designing user-friendly interfaces, such as dashboards and visual tools, to assist medical professionals in quickly understanding model estimates and assumptions. To keep healthcare applications fair and trustworthy, we must address ethics and privacy concerns. It is crucial to handle these issues carefully, especially when working together on research involving data.

5.3 Implication of thesis for the Medical Sector

This research shares important discoveries that could greatly benefit the medical community, especially in predictive analytics and managing diabetes. Advanced machine learning models such as Multilayer Perceptron (MLP), Logistic Regression (LR), and Random Forest (RF) greatly improve the accuracy of diagnosing diabetes. They enable early detection and intervention, leading to better management of the condition. This initial evaluation could decrease fatal case incidences and cut down the spread of the disease improving patient outcomes in the end.

Incorporating predictive models into clinical processes allows medical personnel to improve clinical judgment. This enhances treatment accuracy. Medical personnel gain access to sophisticated decision-support tools. These tools enhance their decision-making capabilities. Predictive analytics helps healthcare by finding patients with a high risk of health issues. This ensures resources are used effectively, improves overall care, and reduces unnecessary spending.

Furthermore, using machine learning to create individualized treatment plans addresses specific patient characteristics. Ultimately, this way helps in more effective and patient-focused treatment. The findings provide a basis for future investigation. They stimulate the discovery of new algorithms and methodologies.

These advancements enhance predictive capabilities and expand applications beyond diabetes. This fosters creativity in medical research and healthcare delivery.

REFERENCES

Alehegn, M., Joshi, R., & Mulay, P. (2018). Analysis and prediction of diabetes mellitus using machine learning algorithm. *International Journal of Pure and Applied Mathematics, 118*(9), 871-878.

Alpan, K., & Ilgi, G. S. (2020). Classification of Diabetes Dataset with Data Mining Techniques by Using WEKA Approach. *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. https://doi.org/10.1109/ismsit50672.2020.9254720

American Diabetes Association. (March 22, 2018). *Economic costs of diabetes in the U.S. in 2017*. Diabetes Care, 41(5), 917– 928. https://doi.org/10.2337/dci18-0007

American Diabetes Association. (2019). IDF Diabetes Atlas, 9th edition.

Anusha, C. (2023). A machine learning approach for prediction of diabetes mellitus. *International Journal of Emerging Trends in Engineering Research, 11*(6), 207–213. https://doi.org/10.30534/ijeter/2023/031162023

Atkinson, M. A., Eisenbarth, G. S., & Michels, A. W. (2014). Type 1 diabetes. *Lancet, 383*(9911), 69–82. https://doi.org/10.1016/s0140-6736(13)60591-7

Behavioral Risk Factor Surveillance System. (2017, August 24). Kaggle. https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system?select=2015.csv

Bhat, S. S., Selvam, V., & Ansari, G. A. (2023). Predicting Life Style of Early Diabetes Mellitus using Machine Learning Technique. *International Journal of Computing*, 345–351. https://doi.org/10.47839/ijc.22.3.3230

Breiman, L. (2001). Random Forests. Machine Learning, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Casella, G., & Berger, R. L. (2002*). Statistical inference*. Brooks/Cole.

Chang, V., Ganatra, M. A., Hall, K., Golightly, L., & Xu, Q. (2022). An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. *Healthcare Analytics*, 2, 100118. https://doi.org/10.1016/j.health.2022.100118

Crabtree, M. (2023, July). *What is machine learning? Definition, types, tools & more | DataCamp.* Datacamp. Retrieved March 20, 2024, from https://www.datacamp.com/blog/what-is-machine-learning

Creswell, J. (2009). Research design: Qualitative, quantitative, and mixed methods approach. Sage publications.

DeFronzo, R. A., Ferrannini, E., Groop, L., Henry, R. R., Herman, W. H., Holst, J. J., Hu, F. B., Kahn, C. R., Raz, I., Shulman, G. I., Simonson, D. C., Testa, M. A., & Weiss, R. (2015). Type 2 diabetes mellitus. *Nature Reviews. Disease Primers, 1*(1). https://doi.org/10.1038/nrdp.2015.19

Fuentes, E. (2023, June 26). *Introduction to artificial intelligence and machine learning.* Community.aws. Retrieved March 20, 2024, from https://community.aws/content/2drbbXokwrIXivItJ8ZeCk3gT5F/introduction-to-artificial-intelligence-and-machine-learning?lang=en

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press.

Grolemund, G., & Wickham, H. (2017). *R for Data science.*

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software. SIGKDD Explorations, 11(1), 10–18. https://doi.org/10.1145/1656274.1656278

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining. Concepts and Techniques* (3rd ed.). Morgan Kaufmann Publishers.

Hassan, A. S., Malaserene, I., & Leema, A. A. (2020). Diabetes Mellitus Prediction using Classification Techniques. *International Journal of Innovative Technology and Exploring Engineering, 9*(5), 2080–2084. https://doi.org/10.35940/ijitee.e2692.039520

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data Mining, Inference, and Prediction, Second Edition.* Springer Science & Business Media.

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine, 25*(1), 30–36. https://doi.org/10.1038/s41591-018-0307-0

Hinton, G. (2018). Deep Learning—A technology with the potential to transform health care. *JAMA, 320*(11), 1101. https://doi.org/10.1001/jama.2018.11100

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. In *Springer texts in statistics.* https://doi.org/10.1007/978-1-4614-7138-7

Kangra, K., & Singh, J. (2023). Comparative analysis of predictive machine learning algorithms for diabetes mellitus. *Bulletin of Electrical Engineering and Informatics, 12*(3), 1728–1737. https://doi.org/10.11591/eei.v12i3.4412

Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express, 7*(4), 432–439. https://doi.org/10.1016/j.icte.2021.02.004

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of Healthcare Information Management, 19*(2), 65-72.

Liaw, A., & Wiener, M. (2002). Classification and Regression by RandomForest. *R News, 2*(3), 18–22. https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf

Maalouf, M. (2011). Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies, 3*(3), 281. https://doi.org/10.1504/ijdats.2011.041335

McIntyre, H. D., Catalano, P., Zhang, C., Desoye, G., Mathiesen, E. R., & Damm, P. (2019). Gestational diabetes mellitus. *Nature Reviews. Disease Primers, 5*(1). https://doi.org/10.1038/s41572-019-0098-8

Mertens, D. (2010). Research and evaluation in education and psychology: integrating diversity with quantitative, qualitative, and mixed methods (3rd ed.). Sage publications.

Nosratabadi, S., Ardabili, S., Lakner, Z., Makó, C., & Mosavi, A. (2021). Prediction of food production using machine learning algorithms of multilayer perceptron and ANFIS. Agriculture, 11(5), 408. https://doi.org/10.3390/agriculture11050408

Özsezer, G., & Mermer, G. (2022). Diabetes Risk Prediction with Machine Learning Models. *Artificial Intelligence Theory and Applications, 2*(2), 1–9.

Patil, R., & Tamane, S. (2018). A comparative analysis on the evaluation of classification algorithms in the prediction of diabetes. *International Journal of Electrical and Computer Engineering, 8*(5), 3966. https://doi.org/10.11591/ijece.v8i5.pp3966-3975

Phillips, D. C., & Burbules, N. C. (2000). *Postpositivism and educational research.*

Provost, F., & Fawcett, T. (2013). *Data science for business.* Oreilly & Associates Incorporated.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems, 2*(1), 3.

Russell, S., & Norvig, P. (2019). *Artificial intelligence: A Modern Approach.* Pearson Higher Education.

Saeed, M. H. (2023). Diabetes type 2 classification using machine learning algorithms with up-sampling technique. *Journal of Electrical Systems and Information Technology*, *10*(1). https://doi.org/10.1186/s43067-023-00074-5

Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, A. A., Ogurtsova, K., Shaw, J. E., Bright, D., & Williams, R. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Research and Clinical Practice, 157,* 107843. https://doi.org/10.1016/j.diabres.2019.107843

Sisodia, D., & Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science, 132,* 1578–1585. https://doi.org/10.1016/j.procs.2018.05.122

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-Score and ROC: a family of discriminant measures for performance evaluation. *In Lecture notes in computer science* (pp. 1015–1021). https://doi.org/10.1007/11941439_114

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Tabák, A. G., Herder, C., Rathmann, W., Brunner, E. J., & Kivimäki, M. (2012). Prediabetes: a high-risk state for diabetes development. *Lancet, 379*(9833), 2279–2290. https://doi.org/10.1016/s0140-6736(12)60283-9

Teboul, Alex. (2021, November 8) *Diabetes Health Indicators Dataset.* Kaggle. Retrieved March 19, 2024, from https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: practical machine learning tools and techniques*. Morgan Kaufmann.

World Health Organization. (2018). Noncommunicable diseases country profiles 2018.

World Health Organization: WHO. (2019, May 13). *Diabetes*. Retrieved March 6, 2024, from https://www.who.int/health-topics/diabetes#tab=tab_1

Zheng, Y., Ley, S. H., & Hu, F. B. (2017). Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nature Reviews. Endocrinology, 14(2), 88–98. https://doi.org/10.1038/nrendo.2017.151

Appendices

APPENDIX 1. SUMMARY RESULT OF MULTILAYER PERCEPTRON

```
=== Summary ===

Correctly Classified Instances       65902               86.5947 %
Incorrectly Classified Instances     10202               13.4053 %
Total Number of Instances            76104

=== Detailed Accuracy By Class ===

                 Precision  Recall   F-Measure  Class
                 0.586      0.138    0.224      NO
                 0.876      0.984    0.927      YES
Weighted Avg.    0.835      0.866    0.828

=== Confusion Matrix ===

     a     b   <-- classified as
  1471  9161 |     a = NO
  1041 64431 |     b = YES
```

Figure 15. Summary result of MLP across 70:30 split

```
=== Summary ===

Correctly Classified Instances       43924               86.5736 %
Incorrectly Classified Instances      6812               13.4264 %
Total Number of Instances            50736

=== Detailed Accuracy By Class ===

                 Precision  Recall   F-Measure  Class
                 0.587      0.155    0.245      NO
                 0.876      0.982    0.926      YES
Weighted Avg.    0.836      0.866    0.830

=== Confusion Matrix ===

     a     b   <-- classified as
  1104  6034 |     a = NO
   778 42820 |     b = YES
```

Figure 16. Summary result of MLP across 80:20 split

```
=== Summary ===

Correctly Classified Instances       21905               86.3489 %
Incorrectly Classified Instances      3463               13.6511 %
Total Number of Instances            25368

=== Detailed Accuracy By Class ===

                 Precision  Recall   F-Measure  Class
                 0.561      0.156    0.244      NO
                 0.876      0.980    0.925      YES
Weighted Avg.    0.831      0.863    0.829

=== Confusion Matrix ===

     a     b   <-- classified as
   558  3027 |     a = NO
   436 21347 |     b = YES
```

Figure 17. Summary result of MLP across 90:10 split

APPENDIX 2. SUMMARY RESULT OF LOGISTIC REGRESSION

```
=== Summary ===

Correctly Classified Instances       65779               86.433  %
Incorrectly Classified Instances     10325               13.567  %
Total Number of Instances            76104

=== Detailed Accuracy By Class ===

                 Precision  Recall   F-Measure  Class
                 0.552      0.153    0.240      NO
                 0.877      0.980    0.926      YES
Weighted Avg.    0.832      0.864    0.830

=== Confusion Matrix ===

     a     b   <-- classified as
  1629  9003 |     a = NO
  1322 64150 |     b = YES
```

Figure 18. Summary result of Logistic Regression across 70:30 split

```
=== Summary ===

Correctly Classified Instances       43812               86.3529 %
Incorrectly Classified Instances      6924               13.6471 %
Total Number of Instances            50736

=== Detailed Accuracy By Class ===

                 Precision  Recall   F-Measure  Class
                 0.555      0.150    0.236      NO
                 0.876      0.980    0.925      YES
Weighted Avg.    0.831      0.864    0.828

=== Confusion Matrix ===

     a     b   <-- classified as
  1072  6066 |     a = NO
   858 42740 |     b = YES
```

Figure 19. Summary result of Logistic Regression across 80:20 split

```
=== Summary ===

Correctly Classified Instances       21882               86.2583 %
Incorrectly Classified Instances      3486               13.7417 %
Total Number of Instances            25368

=== Detailed Accuracy By Class ===

                 Precision  Recall   F-Measure  Class
                 0.551      0.150    0.236      NO
                 0.875      0.980    0.925      YES
Weighted Avg.    0.829      0.863    0.827

=== Confusion Matrix ===

     a     b   <-- classified as
   538  3047 |     a = NO
   439 21344 |     b = YES
```

Figure 20. Summary result of Logistic Regression across 90:10 split

APPENDIX 3. SUMMARY RESULT OF RANDOM FOREST

```
=== Summary ===

Correctly Classified Instances       65309               85.8155 %
Incorrectly Classified Instances     10795               14.1845 %
Total Number of Instances            76104

=== Detailed Accuracy By Class ===

                 Precision  Recall   F-Measure  Class
                 0.480      0.184    0.266      NO
                 0.880      0.968    0.921      YES
Weighted Avg.    0.824      0.858    0.830

=== Confusion Matrix ===

    a     b   <-- classified as
  1956  8676 |     a = NO
  2119 63353 |     b = YES
```

Figure 21. Summary result of Random Forest across 70:30 split

```
=== Summary ===

Correctly Classified Instances       43522               85.7813 %
Incorrectly Classified Instances      7214               14.2187 %
Total Number of Instances            50736

=== Detailed Accuracy By Class ===

                 Precision  Recall   F-Measure  Class
                 0.486      0.183    0.265      NO
                 0.879      0.968    0.921      YES
Weighted Avg.    0.823      0.858    0.829

=== Confusion Matrix ===

    a     b   <-- classified as
  1303  5835 |     a = NO
  1379 42219 |     b = YES
```

Figure 22. Summary result of Random Forest across 80:20 split

```
=== Summary ===

Correctly Classified Instances       21730               85.6591 %
Incorrectly Classified Instances      3638               14.3409 %
Total Number of Instances            25368

=== Detailed Accuracy By Class ===

                 Precision  Recall   F-Measure  Class
                 0.481      0.189    0.272      NO
                 0.879      0.966    0.920      YES
Weighted Avg.    0.822      0.857    0.829

=== Confusion Matrix ===

     a     b   <-- classified as
   678  2907 |     a = NO
   731 21052 |     b = YES
```

Figure 23. Summary result of Random Forest across 90:10 split

APPENDIX 4. SUMMARY RESULT OF EXTRA TREE

```
=== Summary ===

Correctly Classified Instances       60868               79.98   %
Incorrectly Classified Instances     15236               20.02   %
Total Number of Instances            76104

=== Detailed Accuracy By Class ===

                 Precision  Recall   F-Measure  Class
                 0.298      0.319    0.308      NO
                 0.888      0.878    0.883      YES
Weighted Avg.    0.806      0.800    0.803

=== Confusion Matrix ===

     a     b   <-- classified as
  3389  7243 |     a = NO
  7993 57479 |     b = YES
```

Figure 24. Summary result of Extra Tree across 70:30 split

```
=== Summary ===

Correctly Classified Instances       40540               79.9038 %
Incorrectly Classified Instances     10196               20.0962 %
Total Number of Instances            50736

=== Detailed Accuracy By Class ===

                 Precision  Recall   F-Measure  Class
                 0.299      0.319    0.309      NO
                 0.887      0.878    0.882      YES
Weighted Avg.    0.805      0.799    0.802

=== Confusion Matrix ===

     a     b   <-- classified as
  2280  4858 |     a = NO
  5338 38260 |     b = YES
```

Figure 25. Summary result of Extra Tree across 80:20 split

```
=== Summary ===

Correctly Classified Instances       20247               79.8132 %
Incorrectly Classified Instances      5121               20.1868 %
Total Number of Instances            25368

=== Detailed Accuracy By Class ===

                 Precision  Recall   F-Measure  Class
                 0.301      0.324    0.312      NO
                 0.887      0.876    0.882      YES
Weighted Avg.    0.804      0.798    0.801

=== Confusion Matrix ===

     a     b   <-- classified as
  1161  2424 |     a = NO
  2697 19086 |     b = YES
```

Figure 26. Summary result of Extra Tree across 90:10 split