EARLY PREDICTION AND DETECTION OF DIABETES MELLITUS USING VARIOUS MACHINE LEARNING APPROACHES

BY

Nafis Rayat Shopnil
ID: 201-15-3236

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Ms. Taslima Ferdaus Shuva
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

Tanvirul Islam
Lecturer
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

July 2024

APPROVAL

This Project titled "Early Prediction and Detection of Diabetes Mellitus Using Various Machine Learning Approaches", submitted by Nafis Rayat Shopnil to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 14-07-2024.

BOARD OF EXAMINERS

Dr. S.M Aminul Haque — **Board Chairman**
Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Md. Abbas Ali Khan — **Internal Examiner 1**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Md. Aynul Hasan Nahid — **Internal Examiner 2**
Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Abu Sayed Md. Mostafizur Rahaman — **External Examiner**
Professor
Department of CSE
Jahangirnagar University

©Daffodil International University, Bangladesh

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Ms. Taslima Ferdaus Shuva, Assistant Professor, Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

15/07/24

Name: Ms. Taslima Ferdaus Shuva
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Name: Tanvirul Islam
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Nafis Rayat Shopnil
ID: 201-15-3236
Department of CSE
Daffodil International University

©Daffodil International University, Bangladesh

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty for His divine blessing making it possible for us to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Ms Taslima Ferdaus Shuva**, **Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to the **Dr. Sheak Rashed Haider Noori** (Head of the Department of CSE), for his kind help in finishing our project and also to other faculty members and the staff of the Department of CSE, Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Excess blood glucose levels are indicative of diabetes mellitus (DM), a chronic metabolic disease. Improving patient outcomes and minimizing complications from diabetes need early detection and care of the condition. In this work, we suggest a dataset-based machine learning method for the early identification of diabetes. The dataset gets divided into training and testing sets, missing value management, and feature scaling are among the preparatory procedures that it goes through. After then, each algorithm is trained on the data that has been processed, and cross-validation methods are used to evaluate its performance. We investigate the effectiveness of various machine learning strategies algorithms perform in categorizing people as either diabetes or non-diabetic depending on their clinical and demographic characteristics using Random Forest (RF) and Extreme Gradient Boosting (XGB), Logistic Regression (LR) and Gradient Boosting (GB). Using an ensemble approach called Random Forest, many decision trees are combined to decrease overfitting and increase forecast accuracy. Another ensemble approach, gradient boosting, improves model performance by building trees one after the other to fix mistakes in the prior ones. The statistical model known as logistic regression is useful for classification jobs because it calculates the likelihood of a binary result. The Support Vector Classifier builds hyperplanes to divide various classes and is well-known for its efficiency in high-dimensional domains. Random Forest method performed best with an 85% accuracy and f1 score 0.86. The suggested machine learning approach exhibits encouraging outcomes for diabetes early detection, which may help medical professionals identify those who are at risk.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Introduction

Diabetes mellitus (DM) is an ongoing metabolic condition characterised by high blood glucose levels that are the result of deficiencies in the secretion or actions of insulin. In order to improve patient outcomes and minimize consequences including cardiovascular disease, renal failure, and blindness, premature diabetes diagnosis and treatment are essential. The use of methods for machine learning (ML) has demonstrated potential in the research of big datasets and the discovery of intricate patterns that conventional statistical methods could miss. Machine learning algorithms don't need to be explicitly coded; they can learn from past data to generate predictions or choices. Machine learning has been more prevalent in the sector of healthcare in the last decade, with functions, such as threat evaluation, prediction, and illness diagnosis. In this work, we suggest a dataset-based machine learning method for the early identification of diabetes. Our intention is creating a predictive model that, using clinical and demographic data, can reliably identify people as either diabetics or non-diabetics. We assess the efficacy of different machine learning techniques, such as Random Forest (RF) and Extreme Gradient Boosting (XGB), Logistic Regression (LR) and Support Vector Classifier (SVC) in detecting diabetes. The outcomes of this research possibly help healthcare experts to recognise patients who are at risk and enable prompt interventions to delay the emergence of issues connected to diabetes. They may also offer enlightening details on the use of methods for machine learning for earlier diabetes comprehension.

1.2 Motivation

Utilizing the described predictive model utilising machine learning for determining diabetes, discuss its advantages in terms of early diagnosis, timely treatment, and preventing the consequences of the disease. Effective diabetes management and the avoidance of complications depend on early detection. Algorithms utilizing machine learning have the capacity to detect people who are diabetics at an early stage, which would allow for prompt interventions and lifestyle adjustments. Conventional techniques for diagnosing diabetes might not be readily available or economically viable, especially in environments with low resources. The goal of the project is to raise the affordability and reachability of diabetes screening by creating a machine learning-based method for diabetes detection. Machine learning algorithms can look at a lot of data and find trends in clinical and demographic data that may not be evident using traditional statistical methods. By leveraging these algorithms, the research intends to develop a more correct and trustworthy method for diabetes detection. The fine- tuning and adjustment of the assessment and intervention methods to the identified risk in light of new information that may be obtained from the patient data can be done due to the learn-ability of the machine learning algorithms. Scope of the project is to contribute to the progression of personalized healthcare approaches for diabetes prevention and management. The investigation advances the current state of affairs of technology in healthcare by leveraging machine learning algorithms for detection and diagnosis of diseases. By integrating cutting-edge technology into clinical practice, the research aims to make things better for patients and healthcare.

1.3 Rational of the Study

Large-scale medical record collections can be analyzed by Machine Learning algorithms, which can spot minute patterns in things like blood sugar, body composition, and even genetic markers. It could be challenging for conventional techniques to identify these patterns. It may be possible to use ML models, enabling early detection and focused screening. By doing this, problems might be avoided and long-term health results could be enhanced. In the future, machine learning models may investigate non-invasive data sources, even though blood tests are still vital for diagnosis. Analyzing vocal patterns, retinal scans, or wearable sensor data may be necessary for this. Although ML has great potential for diabetes detection, a logical approach recognizes both its advantages and disadvantages. Through emphasis on data quality, interpretability, and cooperation, scientists can leverage machine learning capabilities to enhance diabetes diagnosis and treatment. Recall that ML is a tool to supplement current practices, not to replace them.

1.4 Research Questions

- What is Machine Learning?
- Which algorithms are applied in this study?
- Which algorithm performs best?
- What are the types of Diabetes?

1.5 Expected Output

The purpose of this work is to show how machine learning in conjunction with Random Forest (RF), Extreme Gradient Boost (XGB), Logistic Regression (LR), Gradient Boosting (GB) can accurately and successfully detect diabetes. Studies may report on the overall accuracy of the model, or they may break it down into sensitivity (how well the model identifies people with diabetes) and specificity (how well the model identifies people without diabetes). This quantifies the frequency with which the model accurately forecasts. if diabetes is present or not. Accuracy is usually expressed as a percentage. Papers may also report on recall and F1 score in addition to accuracy. These measurements are able to offer us a more detailed picture of the model's effectiveness. The machine learning model's efficiency and conventional diagnostic techniques, including blood testing, may be compared in the study. This makes it easier to determine whether the machine learning strategy is a good substitute.

1.6 Report layout

The suggested report has a thorough format that walks readers through the methodology, conclusions, and ramifications of the study. Every chapter has a distinct function and adds to the document's overall coherence and depth.

Chapter 1: Introduction

The research is put in motion by the introduction chapter, which gives background information on the importance of diabetes diagnosis, applying machine learning, and the requirement for a comparative study of detection and prediction techniques. It presents an overview of the study's goals, research questions, and general framework**.**

Chapter 2: Background

A thorough analysis of pertinent literature and earlier research is provided in the background chapter. A thorough grasp of the theoretical underpinnings, technological developments, and machine learning techniques pertaining to diabetes diagnosis is provided in this part. It lays out the background for the current investigation and identifies knowledge gaps.

Chapter 3: Research Methodology

The strategy used to carry out the study is described in the research methodology chapter. It sheds light on the model architecture, data gathering techniques, research design, and the reasoning behind particular methodology selections. The ethical issues and potential constraints on the investigation are also covered in this chapter.

Chapter 4: Experimental Result

The findings of experiments from the use of machine learning in diabetes detection are presented in this chapter. Comprehensive evaluations of the effectiveness of various detection and segmentation techniques are offered, accompanied by graphical depictions as well as parameters for statistics to corroborate the results.

Chapter 5: Impact on Society

The chapter on social influence delves into the wider consequences that the findings of the study hold for healthcare and diagnostics in medicine. It talks about how better diabetes detection techniques can benefit treatment outcomes, patient care, and public health in general. Future technological developments and their effects on society are taken into account.

Chapter 6: Summary, Conclusion and Implications for Future Research

The whole report is summarized in this last chapter. It provides recommendations for real-world applications and more research in addition to summarizing the main results and restating the research's conclusions. A path for academics interested in building upon the current study is provided in the discussion of the implications for future research.

CHAPTER 2
Background

2.1 Preliminaries

The initial phases of a research investigation are the first steps and preparations needed before starting. A comprehensive foundation encompassing several aspects of machine learning is needed in order to get ready for the research on "Diabetes Detection using Machine Learning." First and foremost, one needs to be fully informed about diabetes, including its types, epidemiology, risk factors, pathophysiology, and contemporary diagnosis methods. Understanding this is critical to realizing the need for improved prediction techniques. It's also necessary to have a basic understanding of machine learning, including its core concepts, use in the healthcare sector, and methodology. We made use of machine-learning techniques to create the process of building models easier. Additionally, I assessed the models' speed on various platforms. In the section that follows, I did some comparative analysis with other works. Correct definition of a literature review is also essential for a thorough understanding of this research.

2.2 Related Works

Mithesh et al. shows the comparisons of performance of machine learning methods for the analysis of diabetes such as Naïve Bayes, Decision Tree, Support Vector Machine and K-Nearest Neighbours. Naïve Bayes shows the best accuracy of 72% [1]

Mitushi et al. gives an extensive review of performance of machine learning algorithms for detecting diabetes such as SVM, Logistic Regressions, Decision Tree, KNN and Random Forest and Gradient Boosting. From this study we see that Random Forest shows the highest accuracy of 77%. [2]

Priyanka et al. talks about the analysis comparing the use of two machine learning methods for diabetes and cardiovascular disease classification: artificial neural networks and Bayesian networks. Artificial Neural Network shows the best accuracy. The training accuracy is 89.56% and testing accuracy is 81.49%. [3]

Muhammed et al. compares the execution of machine learning algorithms like KNN and Naïve Bayes. From these two algorithms Naïve Bayes Performed best with an accuracy of 76.07%. [4]

Olta et al. assesses techniques for data mining and their effectiveness that can be applied to the analysis of the diabetes data that has been gathered. Decision Tree shows higher performance with the accuracy of 79% [5]

Shiva et al. describes about the comparison of Machine Learning Algorithms like Logistic Regression, Decision Tree Classifier, Random Forest, Ada Boost and Gradient Boosting for Diabetes diagnosis. Logistic Regression (LR) shows highest accuracy of 65.3%. [6]

Orlando et al. In this study author compares the algorithms that are used in the early detection of diabetes. In this study different machine learning techniques are used likely KNN, Decision Tree, Logistic Regression, BNB and SVM. And also SMOTE technique was used in this study. From this study we can see that KNN (SMOTE) performs well with accuracy of 79.6% [7]

Deepti et al. The author demonstrates how the three Machine Learning Classifier algorithms were probably Naïve Bayes, Support Vector Machine and Decision Tree when used in the prediction of diabetes. From all these mentioned algorithms Naïve Bayes outperformed with an accuracy of 76.30%. [8]

Peri at al. This work aims at comparing the various selectivity and sensitivity models to better determine potential diabetic patients based on patient's characteristics and laboratory data got from emergency hospital visits. This paper compares the performance of machine Learning algorithms like SVM, Decision Tree, KNN etc. The average accuracy obtained by these algorithms was 68-74%. [9]

Monalisa et al. The aim of the study is to show us how various ML Algorithms performs in forecasting diabetes. In this study Gradient Boost performs with the accuracy of 81% [10]

Jitnarayan et al. In this study the author shows that how Machine Learning classification models such as Logistic Regression, Naïve Bayes, KNN, Decision Tree, Random Forest and SVM perform in predicting diabetes. In this paper Logistic Regression performs best with an accuracy of 79.17%. [11]

Victor et al. In this respect, the present research setting proposes an e-diagnostic system, which is based on the IoMT and incorporating ML algorithms for the diagnosis of diabetes mellitus. In this paper three machine learning models are used likely Random Forest, Decision Tree and Naïve Bayes. Random Forest outperformed with the accuracy of 79.57%. [12]

2.3 Comparative Analysis and Summary

The publication " Early Prediction and Detection of Diabetes Mellitus Using Various Machine Learning Approaches " represents a significant advancement in the application of AI in healthcare. In order to detect diabetes I have used Machine Learning algorithms including Random Forest Classifier (RF) and Extreme Gradient Boost (XGB). These algorithms detect diabetes through database based on patient's pregnancies, glucose, blood pressure, skin thickness, insulin, BMI and age. To improve the precision and effectiveness of Diabetes detection, scientists are searching for trends and risk variables that conventional statistical analysis misses. The practical use of these machine learning models in clinical settings is one of the primary focus areas, with a focus on assisting medical practitioners in early diagnosis and customized treatment planning to improve patient outcomes. The important subject of the detection of diabetes is much enhanced by this work. It provides a thorough examination of how machine learning can revolutionize the diagnosis and management of Diabetes.

2.4 Scope of the Problem

1. **Current Diagnostic Challenges:** The current procedures, which require blood samples, are intrusive and comprise the oral glucose tolerance test (OGTT) and the fasting blood sugar (FBS) test. These tests are not always available and can be costly, particularly in environments with limited resources. Diabetes might have mild symptoms that appear gradually, delaying diagnosis and treatment. Because of the complexity of the relationship between diabetes and many risk variables (age, BMI, genetics, lifestyle), reliable identification of the condition requires the use of sophisticated analytical tools.

2. **Role of Machine Learning:** Using machine learning algorithms, large amounts of data may be examined to identify patterns that are difficult to see with conventional statistical techniques. determining who is most likely to acquire diabetes before they have any clinical symptoms. enhancing the accuracy of diabetes prediction models to lower false negative and positive results. customizing forecasts according to unique patient information, increasing the applicability of diagnostic understandings.

3. **Machine Learning Methods in Focus:** Based on labeled training data, algorithms like logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks are used to anticipate diabetes. Methods such as clustering may be implemeted to

find patterns in the data that are not immediately obvious but are related to the risk of diabetes.

4. **Data Sources and Features:** Thorough patient histories that include clinical notes, prescription records, and test findings. data from ongoing monitoring, including blood sugar levels, exercise, and sleep habits. genome-wide association studies have shown genetic predispositions (GWAS). Age, gender, socioeconomic level, food, exercise routines, and BMI.

2.5 Evaluation Metrics: Proportion of authentic outcomes in the entire population, that comprises true positives and true negatives. Precision can be expressed as a combination of all precisely predicted favorable results to all anticipated positives, while recall is the relation of all actual positives to all correctly predicted positive observations. Offers a balance between precision and recall through the harmonic mean.

Challenges

Biased models may result from imbalanced datasets, in which the proportion of cases with diabetes is considerably smaller than that of cases without the disease. It can be difficult to select pertinent variables from a vast set of potential predictors. It can be difficult to select pertinent variables from a vast set of potential predictors. When a model learns noise from training data, it becomes overfitted and performs poorly when applied to new data. Interpretability may be lacking in complex machine learning models, such deep learning algorithms, making it challenging to determine the elements influencing the predictions. It's possible that models developed for one population won't translate well to others with distinct clinical traits or demographics. Imputation methods include using algorithms that can handle missing variables or imputation approaches like mean and median imputation.

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

Research Subject:

"Early Prediction and Detection of Diabetes Mellitus Using Various Machine Learning Approaches" is a research project that focuses on the use of machine learning methods for medical diagnostics, particularly for the identification of diabetes. This paper aims at screening diabetes, which is a long-term metabolic disorder, and establish the efficiency of various artificial intelligence algorithms. The paper investigates the limitations of current diabetes diagnosis methods and considers potential solutions involving machine learning algorithms. The research's comparative analysis component examines how different machine learning techniques perform differently from one another. The process of detection entails determining whether diabetes is present based on patient dataset variables such as blood pressure, glucose, thickness of the skin, insulin, and BMI. Depending on the outcome attribute, the patient may or may not have diabetes. The goal of the research is to provide a thorough understanding of their individual advantages and disadvantages in order to direct the creation of more reliable and efficient diagnostic instruments.

Instrumentation:

At first google colab was used for implementation. Then pickle library was used for saving and loading trained model. xgboost was used for building xgboost model. Then plot tree was used for visualizing decision tree. Then numpy and pandas libraries were used to manipulate numerical data. SimpleInputer was used for handling missing values. Seaborn library was used to create statistical data visualizations. RandomOverSampler and RandomUnderSampler methods were used for addressing imbalanced datasets. Then popular ensemble learning method called RandomForestClassifier was used for classification. XGBClassifier was used to classify XGBoost implementation. Others models such as Logistic Regression for linear classification problems, SVC for SVM classification, Gradient Boosting, Linear Regression etc. were used for this study. confusion_matrix, accuracy_score, recall_score and f1_score were used for evaluation metrics. Sklearn.metrics was used to evaluate model performance on classification tasks. Matplotlib.pyplot library was used to visualize data.

3.2 Data Description

There are important features of patients including pregnancies, glucose, blood pressure, skin thickness, insulin, bmi, diabetes pedigree function and age. The patients have diabetes or not is outcome variable.

Features_Name	**Description**	**Type**
Pregnancies	No. of times pregnant	Continuous
Glucose	Plasma Glucose concentration at 2 hours	Continuous
BloodPressure	Diastolic Blood Pressure	Continuous
SkinThickness	Triceps Skin Fold Thickness	Continuous
Insulin	2 hour serum insulin level	Continuous
BMI	Body Mask Index	Continuous
DiabetesPedegreeFunction	Probability of diabetes based on family history	Nominal
Age	Patient age	Nominal

Table 3.2: Features description of the dataset

3.3 Data Source

This study uses the data to identify diabetes and compares the results with those of previous investigations. A sizable quantity of patient data, including medical records, are included. The dataset was collected via Kaggle. This dataset has 9 features and the shape instance is 768.

3.4 Data Preprocessing

Data preprocessing is necessary for creating a machine learning model which is more accurate. The process of cleaning data is called data pre-processing. Here I used SimpleImputer from sklearn.impute tool for handling missing values. Then I used RandomOverSampler and RandomUnderSampler from imblearn techniques in order to address imbalanced datasets that means when there are substantially fewer samples in one class. And I also fill this missing value with the mean of each feature. In order to determine the statistics (mean) for each feature, the imputer is fitted using the training data (X_train).

3.4.1 Feature Engineering: The accuracy and performance of machine learning models are influenced by feature engineering. It increases machine learning's capacity for prediction and assists in revealing hidden patterns in the data. Here I used binary classification to the target variable (Outcome) to represent the patient having diabetes or not. Then I separated the data into two dataframes called max and min based on the values of outcome feature.

3.4.2 Feature Selection: Here I selected the features based on correlation and random forest feature importance. I converted these selected features from dataframe into numpy array. I can simply input features and the target variable into training and prediction algorithms by dividing them into separate numpy arrays.

3.4.3 Data Splitting: The dataset is partitioned into two sections: one for training and the other for testing. Here 80% data was used for training model and 20% data was used for testing purpose. Stratify the claim that each class is equally represented in the training and testing sets.

3.5 Data Analysis and Visualization

The following figure shows the correlation matrix for the features of data. In order to create heatmap visualization matplotlib library and seaborn library are used.

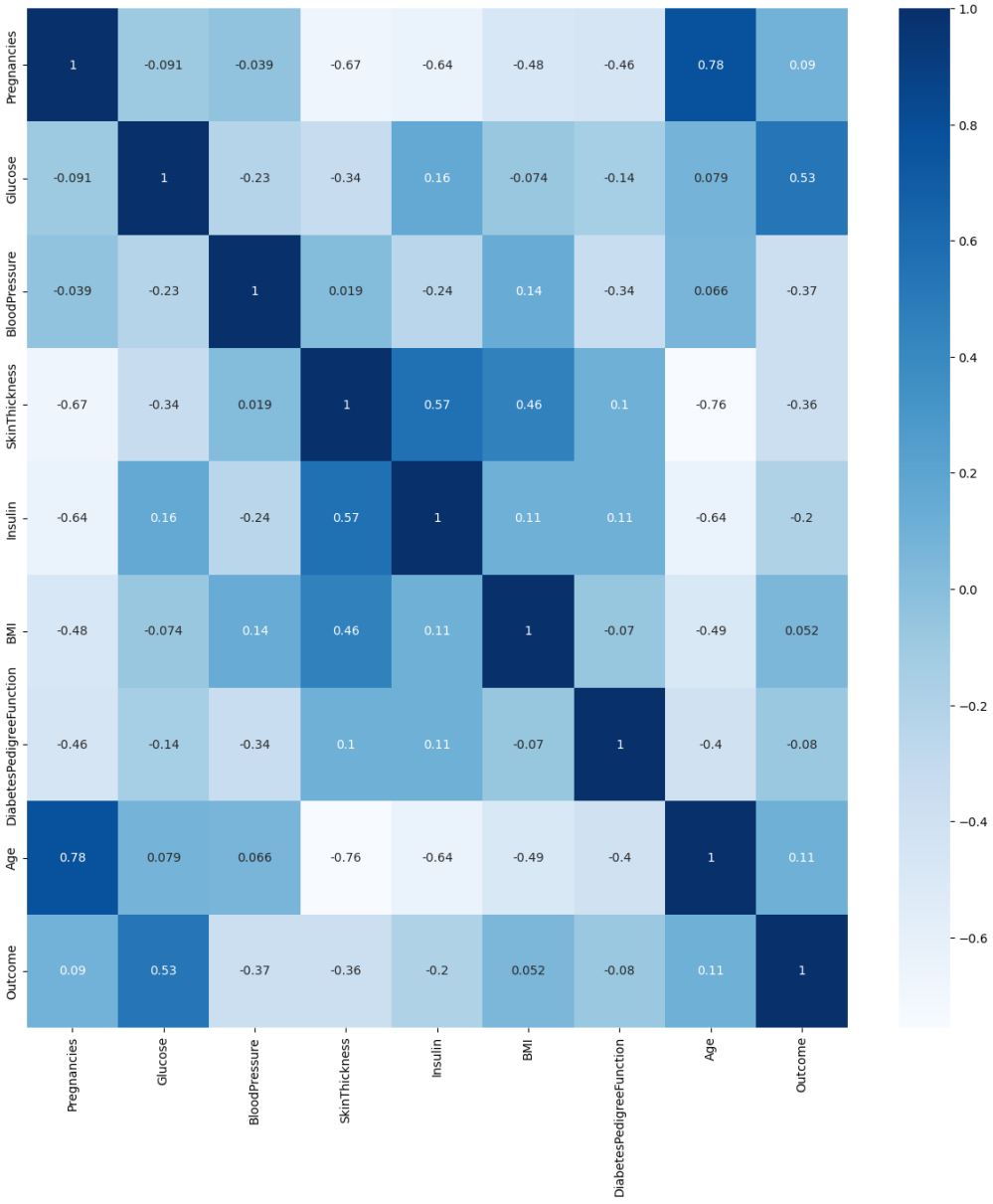

Figure 3.5.1 Data Correlation

The following graph shows the visual summary of distribution of categories of outcome feature. With regard to the visualization of categorical variable frequency, the countplot function is particularly helpful. The categories of the 'Outcome' variable are represented by the x-axis. The x-axis will have two categories, for example, if 'Outcome' contains values of 0 (no diabetes) and 1 (diabetes). The number of observations for each category is shown on the y-axis. A colored bar representing each category is displayed, with the height of the bar corresponding to the total number of observations in that category.

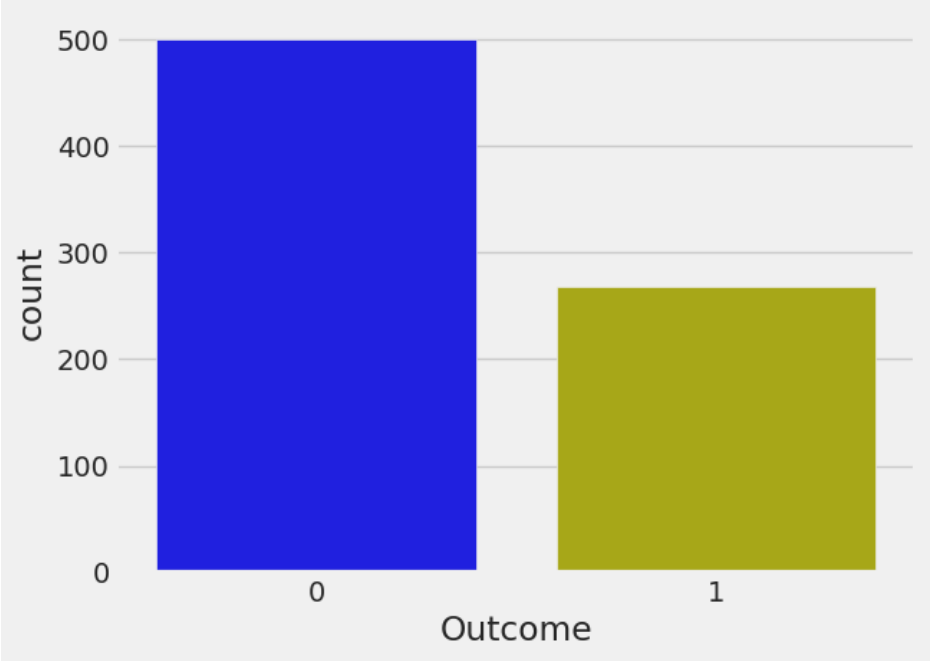

Figure 3.5.2 Count Plot

The following figure is the density graph of all variables. Here y axis represents density and x axis represents attributes. Understanding the distribution of various traits is made easier with the use of density graphs. Choosing the right models and preprocessing methods requires a grasp of this. Outliers that may not be visible in summary statistics can be identified with the use of density graphs. Many machine learning methods may be adversely affected by outliers. A certain algorithm's sensitivity to the feature distribution is higher.

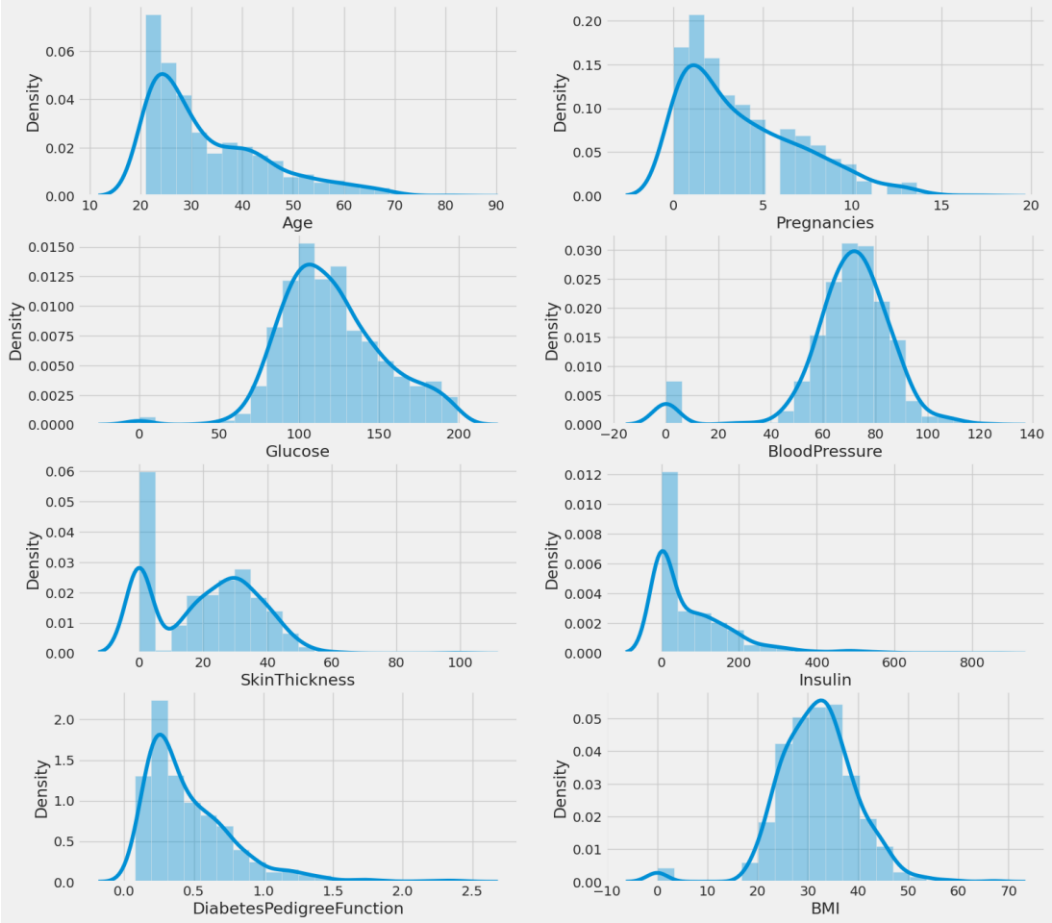

Figure 3.5.4: Density Graph

3.6 Machine Learning Models:

3.6.1 Random Forest Classifier: Random Forest is an ensemble learning technique that is frequently employed in machine learning applications, including classification and regression programs. It is well known for managing complex datasets with exceptional accuracy, resilience, and handling. Random Forest builds multiple decision trees and aggregates their forecasts to generate predictions that are more trustworthy and accurate. It combines projections from several models to produce a final forecast. The projections from many decision trees are combined by the Random Forest method. In this Study I found the algorithm has accuracy of 85%. The following figure is a confusion matrix of Random Forest Classifier. The predicted categories are represented on the X-axis, while the genuine levels are represented on the Y-axis.

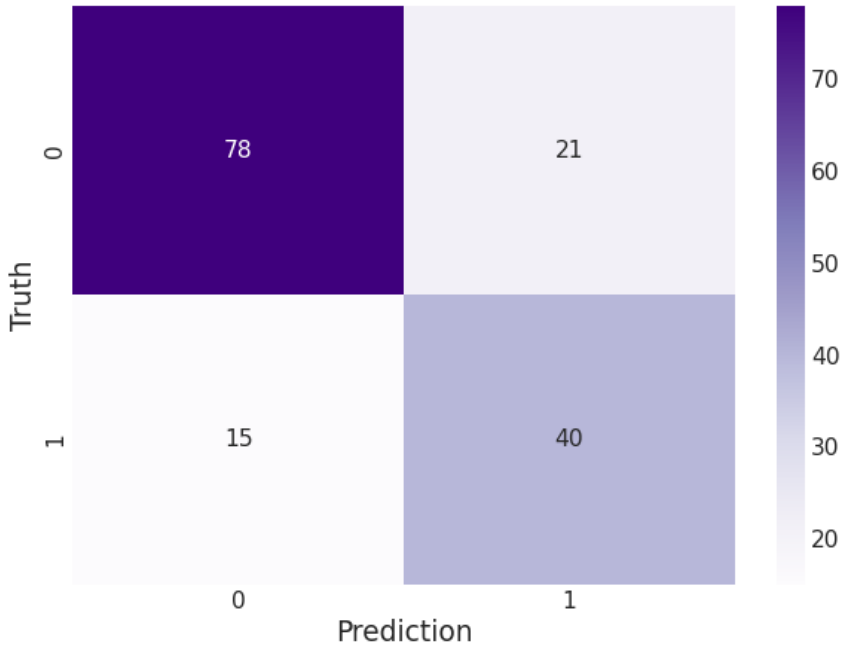

Figure 3.6.1 Confusion Matrix of Random Forest Classifier

3.6.2 Extreme Gradient Boost Classifier: A supervised learning technique called XGBoost (Extreme Gradient Boosting) is quite good at certain machine learning tasks, such as classification issues**.** Gradient boosting, which iteratively trains models by concentrating on the mistakes made by prior models, is the foundation upon which XGBoost develops. Missing values and categorical characteristics are only two of the complicated data structures that XGBoost can handle with ease. It is hence adaptable to real-world datasets. From this study we discovered the algorithm has accuracy of 74.68%. The following figure is decision tree. Typically, a tree diagram represents the decision tree structure in the visualization produced by plot_tree.

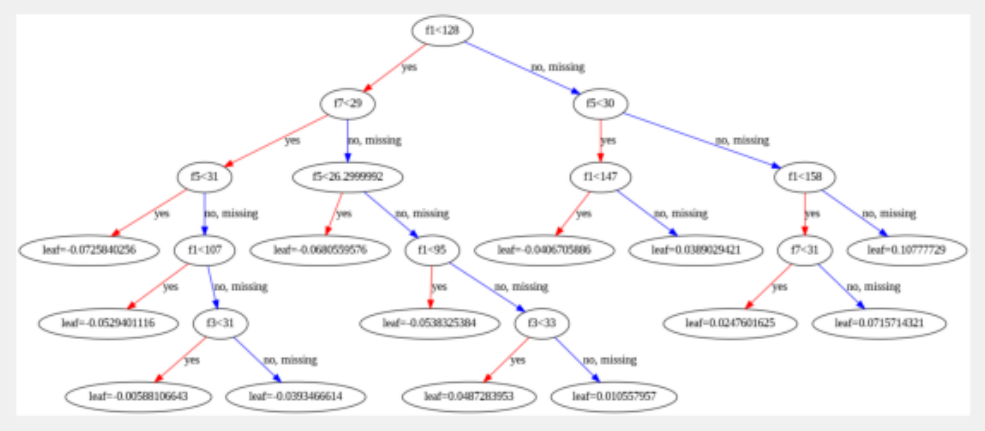

Figure 3.6.2: Plot Tree of XGB Classifier

3.6.3 Logistic Regression: A basic statistical technique that is frequently applied in machine learning for categorization issues is logistic regression. Because it calculates probability that something will happen, it's an effective tool for forecasting binary outcomes (yes/no, 0/1). The interpretability of logistic regression is one of its main benefits. Each feature's coefficient gives information about how much of an impact it has on the model's predictions. Understanding how the model determines its classifications depends on this. We discovered the algorithm has accuracy of 78%.

3.6.4 Gradient Boosting: One machine learning method that works well for both regression and classification applications is gradient boosting. It builds a series of models iteratively, with each one aimed at enhancing the predictions of the one before it. This ensemble method can handle complex data structures and frequently results in large accuracy gains. Models are gradually built via gradient boosting. Every new model is trained to reduce the errors caused by the preceding model (s), which is known as a weak learner and is often a basic decision tree. The ensemble can learn from its mistakes and enhance its performance as a whole thanks to this sequential technique. We found the accuracy was 78%

3.6.5 Support Vector Classifier: One potent machine learning technique that is frequently used for classification problems is the Support Vector Classifier (SVC). It is particularly good at finding the best hyperplane—a high-dimensional decision boundary—to divide the various data classes. SVC, also known as support vector classification, is a classifier that finds a hyperplane that optimizes the margin between the closest data points of various classes, in contrast to those classifiers that represent the class labels directly. These support vectors are essential for establishing the boundaries of the categorization. We discovered the accuracy of this model was 75%.

Figure 3.7: Methodology

CHAPTER 4

Experimental Result

4.1 Introduction

In the previous subsection, we outlined the dataset on which this work was carried out, the methodologies of handling an Electric Power Industry dataset, and the kind of machine learning required for this study. Here, it is planned to provide a description of the evaluation of the entire procedure and the results of the models that utilize the processed data. Since it was important to determine, which of the algorithms, will perform with the highest accuracy, several were implemented. They are at the moment within the process of analysis.

4.2 Performance Evaluation

In this work, a range of classifiers are tested during training on the testing dataset in order to identify the best classification algorithm. Classifiers are evaluated using a variety of performance assessment metrics. Success The process of ascertaining the accuracy, effectiveness, and efficiency of a model is referred to as evaluation in machine learning. To compare the model's predictions with the actual results, specified metrics are usually applied. The crucial tasks for classification metrics are accuracy, recall and f1 score. Understanding the model's advantages and disadvantages can impact future efforts to improve performance on untested data. The confusion matrix of each classifier contains the true positives, false negatives, true negatives, and false positives parameters, which are the source of these measurements.

4.2.1 Accuracy: Accuracy is one measure applied to assess classification model performance. The percentage of all occurrences—true positives and true negatives—that were correctly anticipated is computed. The accuracy of the classifier determines its overall efficacy.

In mathematics accuracy is defined as,

Accuracy= Total number of predictions/Number of correct predictions.

On the contrary for binary classification

Accuracy = (TP+TN) / (TP+TN+FP+FN)

4.2.2 Recall: Indeed the recall also known as sensitivity or true positive rate is a statistic that is used in measuring efficiency of a classification model, especially when there are imbalances in the classes. It quantifies the percentage of real positive examples that the model accurately recognizes. When there is a significant penalty associated with missing a positive instance, recall becomes even more crucial.

Recall is calculated by following formula,

Recall = TP / (TP+FN)

4.2.3 F1 Score: An indication of the success of the used model in classification tasks is F1 score. It solves a drawback of merely accuracy, particularly when dealing with unbalanced datasets. The harmonic mean of memory and precision determines the F1 score. Models with extremely low precision or recall levels are penalized as a result. F1 score can be calculated by,

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

4.3 Feature Importance Analysis: Feature significance in machine learning helps determine how each piece of input data influences a model's prediction. It is necessary to understand the model, simplify it by removing unnecessary components, improve performance, and get further insight into the data. The assessment of feature importance can be accomplished by examining coefficients in linear models and by using tree-based methods like Random Forest. The figure of features impotence in our model is shown below.

Features	**Score**
Pregnancies	0.07289
Glucose	0.25689
Blood Pressure	0.08124
Skin Thickness	0.07160
Insulin	0.08840
BMI	0.17126
Diabetes Pedigree Function	0.11887
Age	0.13883

Table 4.3: Random Forest Feature Importance

4.4 Result Analysis: Based on experiments using a variety of machine learning classification techniques to automatically identify diabetes using the diabetes dataset. Random forest model performs best with the accuracy of 85%. On the otherhand the lowest performance of the model was Support Vector Classifier with accuracy of 75%.

Model	Accuracy
Random Forest	85%
Gradient Boosting	78%
Logistic Regression	78%
Support Vector Classifier	75%

Table 4.4 Accuracy of all models

4.5 Performance Measurement: The evaluation of a model's accuracy, efficacy, and efficiency is a component of performance measurement in machine learning. It entails evaluation of how the model is expected to perform when the test data is fed to it through the use of measures like the classification accuracy. However, confusion matrix is very useful when dealing with imbalanced datasets to evaluate the prediction accuracy and obtaining other important metrics, as precision, recall and F1-score. The application of these evaluation methods is contingent upon the specifics of the problem, the data available, and the model in question.

Algorithms	**Accuracy**	**Recall**	**F1 Score**
Random Forest	85%	0.88	0.86
Gradient Boosting	78%	0.75	0.78
Logistic Regression	78%	0.77	0.79
Support Vector Classifier	75%	0.78	0.76

Table 4.5: Performance measurement of algorithms

4.6 Comparison of evaluation metrics: The execution of four machine learning models on a classification test is displayed in the following graph. The y-axis displays the values of three distinct metrics: accuracy, recall, and F1 score, while the x-axis displays the various models that were assessed. This is a bar chart that contrasts four machine learning models' output on a categorization task. It accomplishes this by assessing every model with test data that hasn't been seen before and computing three distinct metrics: accuracy, recall, and F1 score. These metrics are visualized in the resulting chart, which makes it simple to compare the advantages and disadvantages of each model.

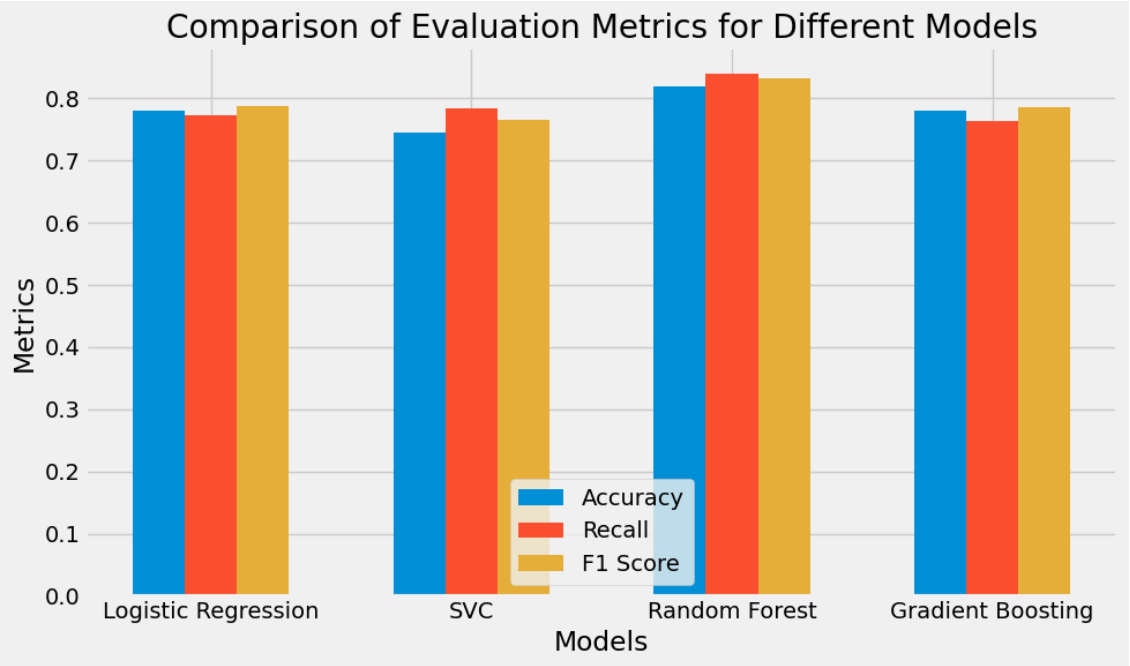

Figure 4.6: Comparison of Evaluation Metrics

CHAPTER 5

Impact on Society

5.1 Impact on Society

Large-scale medical record collections can be analyzed by ML algorithms, which can spot minute patterns that indicate when diabetes may start. In contrast to more conventional techniques, this enables early detection. Individual risk factors such as blood indicators, lifestyle choices, and genetics can be evaluated by ML models. This makes it possible to implement high-risk population-focused targeted screening programs that result in earlier intervention and better health outcomes. Machine learning algorithms are capable of analyzing patient data to forecast a patient's potential response to various treatment options. This enables medical professionals to tailor treatment regimens, resulting in improved oversight and a decrease in problems. Based on lifestyle decisions, machine learning models are able to forecast the risk of developing diabetes. This enables people to take action before the disease manifests, such as altering their diet or increasing their physical activity. By averting costly consequences, early detection and better management result in significant cost savings for healthcare systems.

5.2 Impact on Environment

Complex machine learning algorithms demand a lot of processing power to train and run, which increases data centers' energy usage. This means that if the energy source is not renewable, there will be an increase in greenhouse gas emissions. The creation, production, and utilization of specialized hardware (such as GPUs) for machine learning applications leads to resource depletion and electronic waste. More research is being done to create hardware designs and machine learning algorithms that use less energy. By doing this, the environmental effect of training and using these models can be greatly decreased. ML operations can drastically run their data centers on alternative energy sources like solar or wind power, they can lessen their carbon footprint.

5.3 Ethical Aspects

A possible approach to enhancing diabetes detection is machine learning (ML). But along with its obvious advantages come moral questions that need to be carefully thought through. This study investigates these moral issues and offers recommendations for the appropriate design and deployment of machine learning-based diabetes detection systems. Strict laws are required to guarantee data gathering with informed consent, the anonymization of private medical information, and strong data security protocols to avert breaches. Regarding data ownership and sharing for commercial or research objectives, transparency is essential. If ML systems are trained on unbalanced datasets, they may reinforce preexisting biases in the healthcare industry. To prevent prejudiced results, varied and representative datasets must be guaranteed. The applicability of ML-based detection systems may be restricted by differences in access to resources and technology. Accessible and affordable healthcare strategies are essential to achieving equity in healthcare. Machine learning need to supplement, not replace, the knowledge of medical professionals. For an appropriate diagnosis and treatment strategy, human judgment is still essential.

5.4 Sustainability Plan

When discussing the possible uses of machine learning to detect cardiovascular diabetes, sustainability refers to the substantial effects on people. This made me think about the sustainability plan for our study. At first I will make it efficiency of energy through algorithmic optimization, hardware optimization, renewable energy sources and data center efficiency. Then I will manage hardware lifecycle through extending lifespan and recycling programs. Then I will make it data sustainable through minimization, anonymization, security and governance. Then I will make it operational sustainable through model maintenance, scalability and efficiency. We can guarantee the long-term advantages of machine learning (ML)-based diabetes detection for both the environment and society by putting this complete sustainability plan into action.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

The study compares various machine learning methods for diabetes detection by evaluating the performance of various models. Three metrics are computed by the code: F1 score, accuracy, and recall. These measures are used to evaluate how well the models classified people as either non-diabetic or diabetic. The fact that the code makes reference to training and testing data suggests the researchers divided their diabetes data into two groupings. The machine learning models were trained on one piece of data (training data), and their performance on unseen data was assessed on a second set of data (testing data). This code also computes heatmap that reveals the relationship of these features. Through the identification of highly correlated features, researchers are able to eliminate redundant features from the dataset. This code also create a countplot that serves several purposes to understand the dataset likely visualizing diagnosis distribution, exploring feature distribution etc. A confusion matrix is computed by this code. The rows of this table represent the patients are with diabetes or non-diabetes. And the columns represent the prediction of the models.

6.2 Conclusions

In this work, we investigated the use of different biological characteristics in conjunction with machine learning algorithms to diagnose diabetes. Based on patient data, our findings show how well machine learning algorithms can identify whether a patient has diabetes or not. Additionally, the predictive performance of the model is highly influenced by [mention crucial features], among other features, according to our feature importance study. This knowledge may help physicians concentrate on important biomarkers when making diagnoses and treating patients. Even though our findings are encouraging, there is yet room for development. In order to improve the model's prediction power, future studies may investigate the incorporation of further clinical data, such as patient demographics and lifestyle variables. To sum up, there is a lot of potential for bettering early diagnosis and treatment outcomes through the use of machine learning approaches for diabetes detection. These models have the potential to be useful tools for healthcare professionals in the battle against diabetes with additional development and validation.

6.3 Implication for Further Study

This subject has been the subject of numerous studies, and I endeavored to predict which algorithms will yield the most efficient results. I intend to employ these models in a real-world scenario in the future and develop a more precise model that can accurately detect diabetes in patients in clinics, hospitals, and other healthcare settings before it deteriorates. With an emphasis on explainable AI (XAI) for transparency and integration with wearable devices for continuous health monitoring, this field of research is fast progressing. Through tackling the obstacles and emphasizing sustainability, machine learning (ML) holds the potential to transform diabetes detection and usher in a healthier future for everybody.

REFERENCES

[1] Warke, M., Kumar, V., Tarale, S., Galgat, P., & Chaudhari, D. J. (2019). Diabetes diagnosis using machine learning algorithms. *Diabetes*, *6*(03), 1470-1476.

[2] Soni, M., & Varma, S. (2020). Diabetes prediction using machine learning techniques. International Journal of Engineering Research & Technology (IJERT), 9(9), 921-925.

[3] Indoria, P., & Rathore, Y. K. (2018). A survey: detection and prediction of diabetes using machine learning techniques. International Journal of Engineering Research & Technology (IJERT), 7(3), 287-291.

[4] Febrian, M. E., Ferdinan, F. X., Sendani, G. P., Suryanigrum, K. M., & Yunanda, R. (2023). Diabetes prediction using supervised machine learning. Procedia Computer Science, 216, 21-30.

[5] Llaha, O., & Rista, A. (2021, May). Prediction and Detection of Diabetes using Machine Learning. In RTA-CSIT (pp. 94-102).

[6] Reddy, S. S., Sethi, N., & Rajender, R. (2020). A Comprehensive analysis of machine learning techniques for incessant prediction of Diabetes mellitus. International Journal of Grid and Distributed Computing, 13(1), 1-22.

[7] Iparraguirre-Villanueva, O., Espinola-Linares, K., Flores Castañeda, R. O., & Cabanillas-Carbonell, M. (2023). Application of machine learning models for early detection and accurate classification of type 2 diabetes. Diagnostics, 13(14), 2383.

[8] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585. [9] Arjun, P., & Verma, J. (2020). Methods for detection of diabetes mellitus using machine learning techniques. Methods, 7(11).

[9] Panda, M., Mishra, D. P., Patro, S. M., & Salkuti, S. R. (2022). Prediction of diabetes disease using machine learning algorithms. IAES International Journal of Artificial Intelligence, 11(1), 284.

[10] Sahoo, J., Dash, M., & Pati, A. (2020). Diabetes prediction using machine learning classification algorithms. International Research Journal of Engineering and Technology, 7(8).

[11] Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2023). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications, 35(22), 16157-16173.

[12] Sharma, T., & Shah, M. (2021). A comprehensive review of machine learning techniques on

diabetes detection. Visual Computing for Industry, Biomedicine, and Art, 4(1), 30.

[13] Ghosh, P., Azam, S., Karim, A., Hassan, M., Roy, K., & Jonkman, M. (2021). A comparative study of different machine learning tools in detecting diabetes. Procedia Computer Science, 192, 467-477.

[14] Priya, R., & Aruna, P. (2013). Diagnosis of diabetic retinopathy using machine learning techniques. ICTACT Journal on soft computing, 3(4), 563-575.

[15] He, B., Shu, K. I., & Zhang, H. (2019, April). Machine learning and data mining in diabetes diagnosis and treatment. In IOP conference series: materials science and engineering (Vol. 490, No. 4, p. 042049). IOP Publishing.

[16] Jelinek, H. F., Cornforth, D. J., & Kelarev, A. V. (2016). Machine learning methods for automated detection of severe diabetic neuropathy. J. Diabet. Complicat. Med, 1(02), 1-7.

[17] Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, 292-299.

[18] Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018, September). Prediction of diabetes using machine learning algorithms in healthcare. In 2018 24th international conference on automation and computing (ICAC) (pp. 1-6). IEEE.

[19] Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. Ict Express, 7(4), 432-439.

[20] Sonar, P., & JayaMalini, K. (2019, March). Diabetes prediction using different machine learning approaches. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 367-371). IEEE.

[21] Saru, S., & Subashree, S. (2019). Analysis and prediction of diabetes using machine learning. International journal of emerging technology and innovative engineering, 5(4).

[22] Khaleel, F. A., & Al-Bakry, A. M. (2023). Diagnosis of diabetes using machine learning algorithms. Materials Today: Proceedings, 80, 3200-3203.

[23] Panda, M., Mishra, D. P., Patro, S. M., & Salkuti, S. R. (2022). Prediction of diabetes disease using machine learning algorithms. IAES International Journal of Artificial Intelligence, 11(1), 284.

[24] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585.

[25] Lyngdoh, A. C., Choudhury, N. A., & Moulik, S. (2021, March). Diabetes disease prediction

using machine learning algorithms. In 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES) (pp. 517-521). IEEE.

[26] 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.

[27] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.

[28] SVKR Rajeswari, V. (2021). Prediction of diabetes mellitus using machine learning algorithm. Annals of the Romanian society for cell biology, 5655-5662.

[29] Al-Sideiri, A., Cob, Z. B. C., & Drus, S. B. M. (2019, December). Machine learning algorithms for diabetes prediction: A review paper. In Proceedings of the 2019 International Conference on Artificial Intelligence, Robotics and Control (pp. 27-32).

[30] Mir, A., & Dhage, S. N. (2018, August). Diabetes disease prediction using machine learning on big data of healthcare. In 2018 fourth international conference on computing communication control and automation (ICCUBEA) (pp. 1-6). IEEE.

306

ORIGINALITY REPORT

21%	16%	12%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
2	Submitted to Daffodil International University Student Paper	2%
3	Submitted to George Bush High School Student Paper	2%
4	doctorpenguin.com Internet Source	1%
5	"Advanced Computing", Springer Science and Business Media LLC, 2024 Publication	1%
6	Submitted to University of Technology, Sydney Student Paper	1%
7	www.mdpi.com Internet Source	1%
8	www.coursehero.com Internet Source	1%
9	export.arxiv.org	

	Internet Source	1%
10	"Proceedings of the Fifth International Conference on Trends in Computational and Cognitive Engineering", Springer Science and Business Media LLC, 2024 Publication	<1%
11	Submitted to Sunway Education Group Student Paper	<1%
12	Submitted to Liverpool John Moores University Student Paper	<1%
13	ijmi.ir Internet Source	<1%
14	Submitted to University of Hertfordshire Student Paper	<1%
15	medium.com Internet Source	<1%
16	link.springer.com Internet Source	<1%
17	Reddy Shiva Shankar, P. Neelima, V. Priyadarshini, K. V. S. S. R. Murthy. "Chapter 33 Comprehensive Analysis to Predict Hepatic Disease by Using Machine Learning Models", Springer Science and Business Media LLC, 2022	<1%