**CSE 4000:** Thesis/ Project

**Project/Thesis No.:**

**A Study On Effectiveness Of Machine**

**Learning Models For Diabetes Prediction**

**By**

**Simon Walter Tudu**

**Roll: 1807121**



**Department of Computer Science and Engineering**

**Khulna University of Engineering & Technology**

**Khulna 9203, Bangladesh**

**February, 2024**

# **A Study on Effectiveness of Machine Learning Models for Diabetes Prediction.**

By

**Simon Walter Tudu**

Roll: 1807121

A thesis submitted in partial fulfillment of the requirements for the degree of

“Bachelor of Science in Computer Science & Engineering”

**Supervisor:**

**Dr. Sheikh Imran Hossain**

Assistant Professor

Department of Computer Science and Engineering Signature

Khulna University of Engineering & Technology

Department of Computer Science and Engineering

Khulna University of Engineering & Technology

Khulna 9203, Bangladesh

December, 2023

# **Acknowledgment**

I am grateful to God for guiding and blessing me during my thesis. I want to express my sincere appreciation and gratitude to Dr. Sheikh Imran Hossain, my respected teacher and supervisor, for providing valuable suggestions and continuous support throughout my research journey. Furthermore, I would like to extend my heartfelt thanks to my dear parents and family members for their unwavering encouragement and support.

I want to express my gratitude to the anonymous reviewers and scholars who helped me refine my research. Their feedback and insights were critical in achieving academic excellence. I'm humbled and thankful to everyone who contributed to my thesis journey, and their support reaffirmed my belief in the power of collaboration in scholarly pursuits.

**Authors**

# **Abstract**

In this thesis, we will examine the utilization of machine learning methods to identify the initial stages of diabetes. We know that Diabetes is a disease that is spreading quickly throughout the world. It can lead to many serious complications, such as cardiovascular disease, kidney failure, diabetic retinopathy, and neuropathy. These complications can increase the chances of illness and death. Therefore, early detection and treatment could spare many lives. This study aims to determine the effectiveness of machine learning models in predicting diabetes using a variety of datasets.

The research uses five distinct datasets, including numerical and textual data, to assess the performance of eight different machine-learning models. The main objective of the study is to compare the accuracy and complexity of these models in predicting diabetes. By conducting thorough experimentation and analysis, the study aims to identify the most effective model for diabetes prediction across different datasets. The findings of this research provide valuable insights into the suitability of different machine learning approaches for diabetes prediction, which can help in making informed decisions in healthcare applications.

Keywords: Machine Learning, Deep Learning, Diabetes disease

# **Contents**

**PAGE**

**Table of Contents**

[**A Study on Effectiveness of Machine Learning Models for Diabetes Prediction.** ii](#_Toc159112986)

[**Acknowledgment** ii](#_Toc159112987)

[**Abstract** iii](#_Toc159112988)

[**Contents** iv](#_Toc159112989)

[**List of Table** vi](#_Toc159112990)

[**List of Figure** vii](#_Toc159112991)

[**CHAPTER I** 1](#_Toc159112992)

[**Introduction** 1](#_Toc159112993)

[**1.1**  **Introduction** 1](#_Toc159112994)

[**1.2**  **Problem Statement** 1](#_Toc159112995)

[**1.3**  **Objectives** 2](#_Toc159112996)

[**1.4** **Unfamiliarity of the Problem** 3](#_Toc159112997)

[**1.5** **Project Planning** 3](#_Toc159112998)

[1.5.1 Project Timeline 3](#_Toc159112999)

[1.5.2 Legal and Ethical Aspects 4](#_Toc159113000)

[**1.6** **Applications of the Work** 5](#_Toc159113001)

[**1.7** **Organization of the Report** 6](#_Toc159113002)

[**CHAPTER II** 7](#_Toc159113003)

[**Literature Review** 7](#_Toc159113004)

[**2.1**  **Introduction** 7](#_Toc159113005)

[**2.1** **Literature Review** 8](#_Toc159113006)

[**2.2** **Discussion of research gap solution** 9](#_Toc159113007)

[Methodology 11](#_Toc159113008)

[**3.1** **KNN algorithm (K-Nearest Neighbors)** 11](#_Toc159113009)

[**3.2** **Logistic regression algorithm** 12](#_Toc159113010)

[**3.3** **Naive Bayes** 12](#_Toc159113011)

[**3.4** **Decision tree algorithm** 13](#_Toc159113012)

[**3.5** **Random Forest algorithm** 13](#_Toc159113013)

[**3.6** **SVM algorithms** 13](#_Toc159113014)

[**3.7** **Multilayer perceptron (MLP)** 13](#_Toc159113015)

[**3.8** **Ada-boost** 13](#_Toc159113016)

[**3.9** **Data Splitting** 14](#_Toc159113017)

[**3.11** **Hyper parameter Model** 16](#_Toc159113018)

[**3.12**  **Conclusion** 16](#_Toc159113019)

[**CHAPTER IV** 17](#_Toc159113020)

[**Implementation, Results and Discussions** 17](#_Toc159113021)

[**4.1** **Introduction** 17](#_Toc159113022)

[**4.2** **Experimental Setup** 17](#_Toc159113023)

[**4.3**  **Evaluation Metrics** 19](#_Toc159113024)

[**4.4**  **Dataset** 19](#_Toc159113025)

[**4.5** **Implementation and Results** 22](#_Toc159113026)

[4.5.1 Quantitative results 22](#_Toc159113027)

[4.5.2 Qualitative results 23](#_Toc159113028)

[4.5.3 Analysis of the results: 23](#_Toc159113029)

[**4.6** **Objectives Achieved** 24](#_Toc159113030)

[4.3 **Morality or Ethical Issues** 25](#_Toc159113031)

[4.4 **Socio-Economic Impact and Sustainability** 25](#_Toc159113032)

[4.5 **Financial Analyses and Budget** 25](#_Toc159113033)

[**CHAPTER V** 27](#_Toc159113034)

[Conclusions 27](#_Toc159113035)

[5.1 **Conclusion and Challenges Faced** 27](#_Toc159113036)

[**5.2** **Future Works** 27](#_Toc159113037)

[**References** 27](#_Toc159113038)

[31](#_Toc159113039)

# **List of Table**

[**Table 2. 1: Research gaps of different models** 9](#_Toc159113867)

[**Table 2. 2: Abbreviations** 9](#_Toc159113868)

[**Table 4. 1: Data table** 19](#_Toc159113897)

[**Table 4. 2: Accuracy Score** 22](#_Toc159113898)

[**Table 4. 3: F-1 Score** 22](#_Toc159113899)

# **List of Figure**

[Figure 1. 1: Gantt Chart of the Project Plan 4](#_Toc159113923)

[**Figure 3. 1: KNN** 12](#_Toc159113938)

[Figure 3. 2: Train-Validation 14](#_Toc159113939)

[**Figure 3. 3: Stratified 5-fold Validation** 15](#_Toc159113940)

[**Figure 3. 4: Proposed Methodology** 15](#_Toc159113941)

[**Figure 3. 5: Hyper parameter Model Tuning** 16](#_Toc159113942)

[**Figure 4. 1: Critical difference** 23](file:///C:\Users\ASUS\Downloads\CSE%204000.docx#_Toc159113955)

[**Figure 4. 2: Bubble Chart Accuracy vs Complexity** 23](file:///C:\Users\ASUS\Downloads\CSE%204000.docx#_Toc159113956)

# **CHAPTER I**

# **Introduction**

## **1.1 Introduction**

Diabetes is a result of the pancreas not producing enough insulin or the body not being able to use the insulin it produces properly. Insulin is a hormone that controls blood sugar levels. An imbalance in blood sugar control can lead to hyperglycemia or high blood glucose [4].

There are three types of diabetes. Type 1, also known as IDDM or Insulin-Dependent Diabetes Mellitus, requires patients to receive insulin injections. This is because their body is unable to produce enough insulin on its own. Type 2, or Non-Insulin-Dependent Diabetes Mellitus (NIDDM), occurs when body cells cannot properly use insulin. Type 3, Gestational Diabetes, happens when pregnant women experience an increase in blood sugar levels and have not yet been diagnosed with diabetes. [3].

According to the International Diabetes Federation, there are an estimated 382 million individuals worldwide living with diabetes. This number is expected to rise to 592 million by the year 2035 [5].

## **1.2 Problem Statement**

Diabetes is a dangerous disease that cannot be cured. Once it affects you, it will be a lifelong condition. High levels of glucose in your blood can lead to various health issues such as kidney disease, heart disease, stroke, eye problems, dental disease, foot problems, and nerve damage. Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. [1].

Machine learning techniques play a crucial role in predicting diabetes and identifying individuals who are at risk or in need of monitoring. By using textual datasets, we can develop prediction models that enable early intervention and better patient care.

The growing incidence of diabetes is a major concern for public health. Machine learning algorithms have shown promising results in predicting the risk of diabetes. The aim of this study is to examine how well machine learning models can predict diabetes using various datasets. The study assesses different machine learning models and their performance parameters, including sensitivity, specificity, and predicted accuracy. The findings can help healthcare professionals choose the most effective machine learning techniques for diabetes risk assessment, leading to improved patient care and better management plans for the disease.

## **1.3 Objectives**

Our aim is to create precise machine learning models for diabetes diagnosis and classification using five datasets. To achieve this, we'll focus on preparing the data and selecting informative features for improved accuracy.

1. **Model Analysis:** Evaluate the performance of eight different types of machine learning models in predicting the risk of diabetes. The models include support vector machines, logistic regression, random forests, Naïve Bayes, k-nearest neighbor, decision tree, Ada-boost, and deep learning methods.
2. **Data Examination:** Analyze five distinct datasets containing important clinical and demographic information related to diabetes prediction. Use techniques like data cleaning, exploration, and feature manipulation to extract useful insights from the data.
3. **Model Training and Assessment:** Train each machine learning model using appropriate methods. Evaluate their performance using techniques such as cross-validation and train-test splits.
4. **Comparative Analysis:** Compare the prediction accuracy, sensitivity, specificity, and AUC-ROC of different machine learning models. Identify which models are most effective in predicting diabetes risk.
5. **Identifying Best Models:** Determine the most optimal machine learning models for predicting diabetes based on their performance measures and experimental results.
6. **Consideration of Influential Factors:** Explore how factors like feature selection, dataset size, and class distribution impact the accuracy of machine learning models for diabetes prediction.
7. **Interpretability Assessment:** Analyze to understand why certain features are important for predicting diabetes. Assess the relevance of identified features in real-world scenarios.
8. **Recommendations and Future Directions:** Provide recommendations for using suitable machine learning models in clinical settings. Suggest areas for future research to further improve diabetes prediction techniques.

## **Unfamiliarity of the Problem**

I am new to diabetes prediction and lack experience with model selection and dataset analysis. There are multiple aspects of the problem that are unfamiliar:

1. **Analyze Models:** Examine several machine learning models to find out how well they function with various diabetes prediction datasets. This entails assessing their suitability for the job as well as their strengths and limitations.
2. **Dataset Exploration**: Examine different datasets, including clinical and demographic data that include vital information for diabetes prediction. Recognize the unique characteristics of each dataset and how they affect how well various models function.
3. **Optimal Model Selection:** The best machine learning model for anticipating the onset of diabetes should be chosen after careful investigation and analysis. To decide which option is best, take accuracy, computational efficiency, and simplicity of implementation into account.

## **1.5 Project Planning**

In the project planning process, there are some crucial milestones that we need to bear in mind. To ensure that we complete the project within the given timeframe, we have created a timeline. Additionally, we have carefully considered the moral and legal implications of our thesis topic. This meticulous planning ensures that we comply with all legal and ethical requirements while remaining true to our study objectives.

### 1.5.1 Project Timeline

There are two stages in the process of thesis research, which are both aimed at achieving our primary objective. The first stage is expected to be completed during the first semester, while the second stage is planned to be finished during the second semester. Our goal is to complete the thesis in its entirety by February 2024.

The expected duration of these phases is explained in the Gantt chart in Figure 1.2 below:

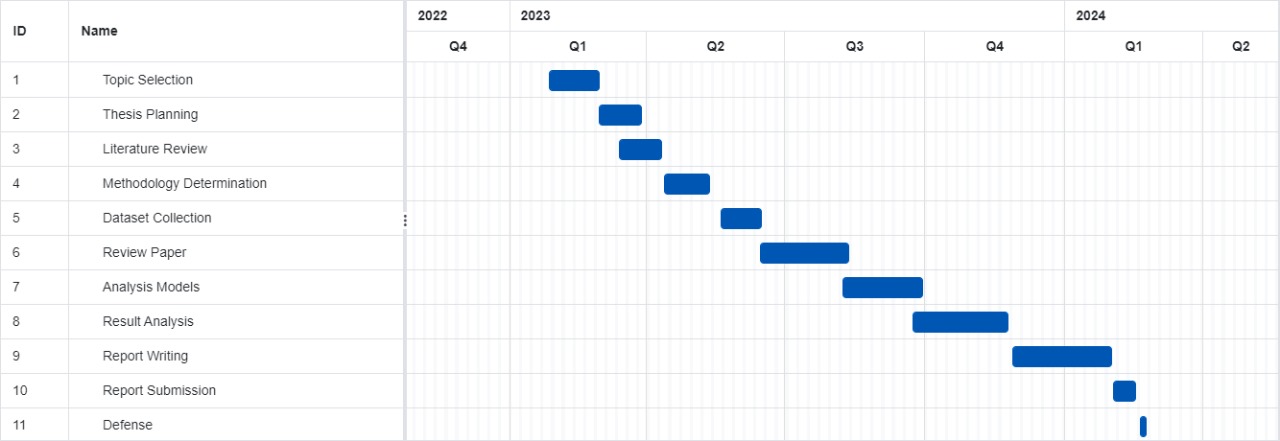


Figure 1. 1: Gantt Chart of the Project Plan

### 1.5.2 Legal and Ethical Aspects

A crucial aspect of machine prediction of diabetes is its impact on culture, society, and health as well as its technological aspects.

1. **Societal Health Impact**: The project aims to improve public health outcomes and reduce the frequency of diabetes-related complications by developing precise machine-learning models that forecast the onset of diabetes. Early detection and intervention can be facilitated through these models.
2. **Safety Considerations:** Safety is the top priority of the project. Strict validation and assessment processes are followed to ensure the accuracy and reliability of the prediction models. By reducing the risk of misdiagnosis or delayed treatment, this focus on safety enhances patient safety and well-being.
3. **Legal Compliance:** The project ensures that it abides by all relevant laws and regulations, including patient data protection laws like HIPAA, as it recognizes the importance of legal compliance in healthcare research. This ensures that the law is upheld, patient data is treated ethically, and public trust in the research process is established.
4. **Cultural Sensitivity:** The project follows a culturally sensitive approach in its methods and interventions, recognizing the diversity of cultures among the populations that suffer from diabetes. By considering cultural nuances and preferences, the project aims to enhance the accessibility and inclusivity of healthcare practices, leading to more equitable health outcomes for diverse populations.

## **1.6 Applications of the Work**

The implementation of machine learning techniques in creating a diabetes prediction model has the potential to significantly impact patient care and healthcare. Here are some ways in which this model can be put to use:

1. **Early Detection and Prevention:** Predictive algorithms can identify individuals at high risk of developing diabetes before any clinical symptoms occur, enabling the timely implementation of lifestyle changes, nutrition, and exercise programs that can delay or prevent the onset of diabetes.
2. **Personalized Treatment Plans:** Healthcare professionals can create customized treatment plans for patients using predictive models. This involves monitoring blood sugar levels and recommending specific interventions to improve glycemic control based on the individual patient's characteristics.
3. **Reducing Complications:** Early detection of diabetes risk leads to better management, reduced complications, and improved blood sugar levels, reducing the risk of neuropathy and cardiovascular disease.
4. **Remote Monitoring:** Machine learning-based prediction models can help create remote monitoring systems that empower patients to monitor their risk factors and receive timely warnings for follow-up appointments or actions. This helps enhance patient adherence to treatment regimens and involvement.
5. **Public Health Planning:** Aggregated data from predictive models can be highly beneficial to public health authorities. This can help them plan interventions, allocate resources, and create awareness campaigns that focus on high-risk populations.
6. **Clinical Decision Support:** Healthcare professionals can use the prediction model as a decision-support tool to make clinical decisions. With its accurate predictions, the model helps prioritize patients, optimize clinic workflow, and ultimately improve patient outcomes.
7. **Diabetes Research:** Predictive models can aid in identifying potential risk factors and biomarkers for diabetes, guiding further research on the development of this disease.
8. **Health Insurance and Wellness Programs:** Health insurance providers and corporate wellness programs can use predictive models to identify individuals who may be at risk and offer them targeted wellness initiatives. These initiatives may involve health education courses, wellness competitions, and programs that provide incentives to encourage participation.
9. **Patient Education and Empowerment:** The use of predictive models can give patients more control over their health, help them follow medical advice more closely, and make better choices for a healthier lifestyle.
10. **Telemedicine and Telehealth:** Healthcare professionals can remotely consult and monitor patients' health using predictive models in telemedicine and telehealth.

## **1.7 Organization of the Report**

To assist readers in navigating the document, this section provides an overview of the thesis structure. It briefly describes the chapters and their contents.

**Chapter I:** This chapter covers the history of the machine learning model for diabetes prediction and emphasizes its importance. It also includes a discussion of the social and cultural aspects of the machine learning model for diabetes prediction research.

**Chapter II:** This section reviews current error-correcting techniques and presents a comparison of several strategies. The chapter concludes by highlighting the advantages and disadvantages of different approaches.Overview of the Organization

**Chapter III:** This chapter provides a complete description of the research methodology and design utilized in the study, along with a rationale for the chosen strategy. To ensure transparency and accuracy, any limitations or restrictions that are inherent in the research process are also discussed.

**Chapter IV:** The research findings are presented and carefully analyzed in this chapter. The methodology of the study is reviewed, and the collected data is analyzed to clarify its significance and implications.

**Chapter V:** In this chapter, the broader socioeconomic, health, environmental, safety, ethical, legal, and cultural implications of the research findings are examined. The implementation of the research findings is associated with potential risks and ethical considerations, which are addressed.

**Chapter VI:** The chapter focuses on the complex engineering issues that were tackled during the research. The chapter highlights any new or innovative ideas or solutions that were developed, and stresses the importance of the research in advancing the field of engineering.

**Chapter VII**: The main conclusions of the research are presented in the conclusion chapter. This chapter also revisits the goals of the study and discusses the broader implications of the findings. Finally, the chapter provides recommendations for future directions that can be pursued to bring the thesis to a close.

# **CHAPTER II**

# **Literature Review**

## **2.1 Introduction**

The use of machine learning (ML) models for diabetes prediction has gained much attention due to its potential to improve early diagnosis and treatment of the condition. To get insights into the efficacy of various machine learning models and their suitability for different patient groups, it is essential to study the literature in this area. By looking at previous studies, researchers can learn about the datasets, assessment methods, and machine-learning approaches currently in use for diabetes prediction. This review aims to provide a comprehensive overview of the state-of-the-art in machine learning-based diabetes prediction, highlighting developments, challenges, and future research directions. The review will guide the selection of relevant ML models, datasets, and assessment methods to investigate the effectiveness of ML models in predicting diabetes.

## **2.1 Literature Review**

Results from related research that included analysis and predictions on diverse healthcare datasets utilizing a variety of techniques and methods are provided.

To predict diabetes, a variety of machine learning techniques were utilized such as Support

Vector Classifier, Random Forest, Decision Tree, Extra Tree, AdaBoost, Perceptron, Linear

Discriminant Analysis, Logistic Regression, K-Nearest Neighbor, Gaussian Nave Bayes, Bagging, Gradient Boost Classifier. Of these, Logistic Regression achieved a high accuracy rate of 96%. However, the application of an AdaBoost the pipeline resulted in the top model with an accuracy rate of 98.8% [3].

Five different models to identify diabetes using support vector machine (SVM) techniques, including linear kernel, radial basis kernel, SVM-RBF, k-NN, ANN, and MDR. By employing the Boruta wrapper algorithm to select relevant features from the dataset, they ensured an unbiased selection process. Among all the models, the SVM-linear model demonstrated the highest accuracy and precision in predicting diabetes, at 0.89 and 0.88, respectively. Ultimately, they were able to improve accuracy and precision while reducing the number of parameters by implementing the Boruta feature selection technique [2].

The suggested possible approach for analyzing the PIMA Indian Diabetes Dataset involves utilizing several classifiers such as SVM, KNN, Random Forest, Decision Tree, Logistic Regression, and Gradient Boosting. This method resulted in achieving a classification accuracy rate of 77%. [4].

Utilized the Bangladesh Diabetes Classification Dataset (DDC) to predict early diabetes using an ML-based ensemble model. They focused on preprocessing and feature selection to improve the dataset quality, which required extensive examination of ablative processes. Despite including only four features such as BMI, age, systolic pressure, and occupation, the model produced more accurate estimates [1].

## **2.2 Discussion of research gap solution**

This section discusses various related models, their advantages and disadvantages. The overall birds-eye view of these restrictions is presented in Table 2.1.

**Table 2. 1: Research gaps of different models**

|  |  |  |
| --- | --- | --- |
| DATASET | ALGORITHM | ACCURACY |
| Dataset contains 800 records and 10 attributes | DT, GNB, LDA, SVC, RF,  Extra Trees, AdaBoost, Perceptron,  LR, GBC, Bagging, KNN | DT=86%,  GNB=93%,  LDA= 94%,  SVC=60%,  RF=91%,  Extra Trees=91%,  AdaBoost=93%,  Perceptron=76%,  LR=96%,  GBC=93%,  Bagging=90%,  KNN=90% |
| PIMA | Linear Kernel SVM, Radial Basis Kernel SVM, k-NN,  ANN,  MDR | Linear Kernel SVM=0.89, Radial Basis Kernel SVM=0.84, k-NN=0.88, ANN=0.86,  MDR=0.83  (Boruta wrapper algorithm is used for feature selection) (Pre-processing involved removal of outliers and k-NN imputation to predict the missing values) |
| Patients with age  18-90 old in Canada | LR,RF,GBM | (GBM-85.1%)  (RF-85.5%)  (LR-84.6%) |

**Table 2. 2: Abbreviations**

|  |  |
| --- | --- |
| DT | Decision Tree |
| GNB | Gaussian NB |
| LDA | Linear Discriminant Analysis |
| SVC | Support Vector Machine |
| RF | Random Forest |
| LR | Logistic Regression |
| GBC | Gradient Boost Classifier |
| KNN | K-Nearest-Neighbor |
| ANN | Artificial Neural Network |
| MDR | Multidimensional Reduction |
| GBM | Gradient Boosting Algorithm |

Research on machine learning models for diabetes prediction is lacking. There are no comparative studies on different algorithms and datasets. Our study aims to address this gap by testing eight machine learning models using five datasets. We will evaluate model performance to determine the advantages and disadvantages of various algorithms for diabetes prediction. Our study emphasizes the need for standardized evaluation processes in machine learning research to bridge the research gap. We recorded our experimental processes to create a repeatable foundation for further research, enhance understanding, guide future endeavors, and analyze machine learning models' effectiveness in predicting diabetes.

Our study emphasizes the need for standardized evaluation processes in machine learning research to bridge the research gap. We recorded our experimental processes to create a repeatable foundation for further research, enhance understanding, guide future endeavors, and analyze machine learning models' effectiveness in predicting diabetes.

CHAPTER III

# Methodology

During the process of assessing the accuracy of machine learning models for predicting diabetes, we selected eight different models including Support Vector Machine (SVM), Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbours (KNN), Naive Bayes (Gaussian), Multi-Layer Perceptron (MLP), and AdaBoost.

To train and evaluate each model, we used the Stratified K-Fold cross-validation methodology, which involved variations such as 5-fold train-test splits and 10-fold train-validation splits. This approach helped to eliminate the risk of overfitting and ensured that each model was evaluated thoroughly across multiple datasets.

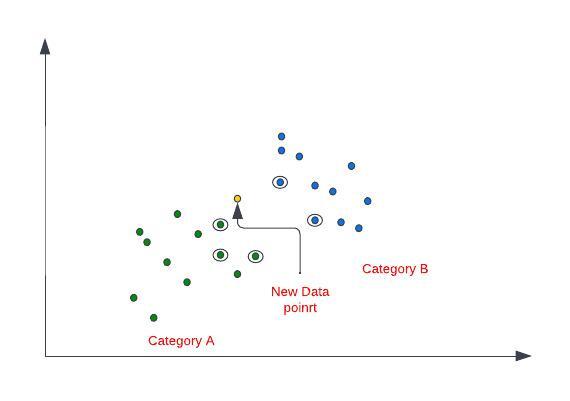
We conducted experiments with different hyperparameters for each model to maximize their performance. We iteratively trained and evaluated the models, computing the accuracy and F1-score for each fold as performance measures. We used the best average performance over multiple folds to determine each model's ideal hyperparameters.

Finally, we examined the data to compare the effectiveness of several models and determine which ones are most useful for diabetes prediction. Thanks to this thorough methodology, we were able to compare and assess the performance of different machine learning models in a systematic manner. This also provided us with important insights into the models' suitability for this crucial healthcare use.

## **3.1 KNN algorithm (K-Nearest Neighbors)**

The KNN algorithm classifies data points based on the labels of their nearest neighbors. It's straightforward and intuitive, making it easy to implement. However, it can become computationally expensive, especially for large datasets.

KNN algorithm predicted diabetes using parameter tuning and stratified k-fold cross-validation. Accuracy was evaluated on a testing set. Comparative analysis with other models was done to assess its efficacy.



**Figure 3. 1: KNN**

## **3.2 Logistic regression algorithm**

The logistic regression algorithm is used to solve issues related to binary categorization. It applies a logistic function to represent the relationship between the attributes of a data point and the class labels. This helps estimate the probability of a given data point belonging to a particular class.

Logistic regression was used to predict diabetes. The model was trained with stratified k-fold cross-validation and parameter optimization. The accuracy and performance of the model were evaluated on a separate testing set. Its performance was compared to other machine learning models.

## **3.3 Naive Bayes**

Naive Bayes is a simple and efficient probabilistic classifier that assumes feature independence. However, it may not perform well on complex datasets where feature independence doesn't hold.The distinctive feature of this particular technique is probabilistic processing.

Naive Bayes algorithm was used to predict diabetes. The model was trained using parameter tuning and cross-validation techniques. Test data was used to evaluate performance. A comparative study was carried out to assess the prediction efficacy of alternative models.

## **3.4 Decision tree algorithm**

Decision trees classify data points into classes by dividing them based on features using a tree-like structure. Despite being straightforward to comprehend, decision trees often suffer from overfitting.

## **3.5 Random Forest algorithm**

The Random Forest Algorithm is a technique that combines multiple decision trees in an ensemble manner. It is used for regression and classification applications and is less prone to overfitting. Each tree is trained on a distinct portion of the data, and the final forecast is obtained by a majority vote.

## **3.6 SVM algorithms**

Support vector machines (SVM) are algorithms that identify the hyperplane with the largest margin between data points of different classes. They are suitable for both regression and classification tasks, particularly with high-dimensional data.

## **3.7 Multilayer perceptron (MLP)**

I implemented Multi-Layer Perceptron (MLP) algorithm with different architectures to capture patterns in data. I used logistic and ReLU activation functions to enhance model flexibility. Stratified k-fold cross-validation was used for consistent validation. Accuracy and F1 score were used to assess model efficacy on unseen test data. I compared MLP's predictive capabilities with other models for diabetes prediction.

## **3.8 Ada-boost**

I used AdaBoost algorithm to create a group of weak learners that captured unique features of the data by building different models with varying numbers of decision trees. I assigned greater weights to misclassified samples and trained each weak learner in turn to improve overall prediction accuracy. By combining the forecasts of all weak learners through weighted majority voting, I created a strong learner with increased predictive power. I evaluated the performance of the AdaBoost algorithm using cross-validation and compared its effectiveness in predicting diabetes with other machine learning algorithms.

## **3.9 Data Splitting**

Two common methods for splitting data in machine learning are:

1. **Stratified 5-fold Train-Test Split:** The dataset is divided into two subsets for training and testing, typically using 80/20 split.
2. **Stratified 10-fold Cross-Validation:** The dataset is divided into k equal-sized folds. Each fold is used as the testing set once, while the remaining folds are used for training. This process is repeated k times, and the average performance measures are obtained.3.8

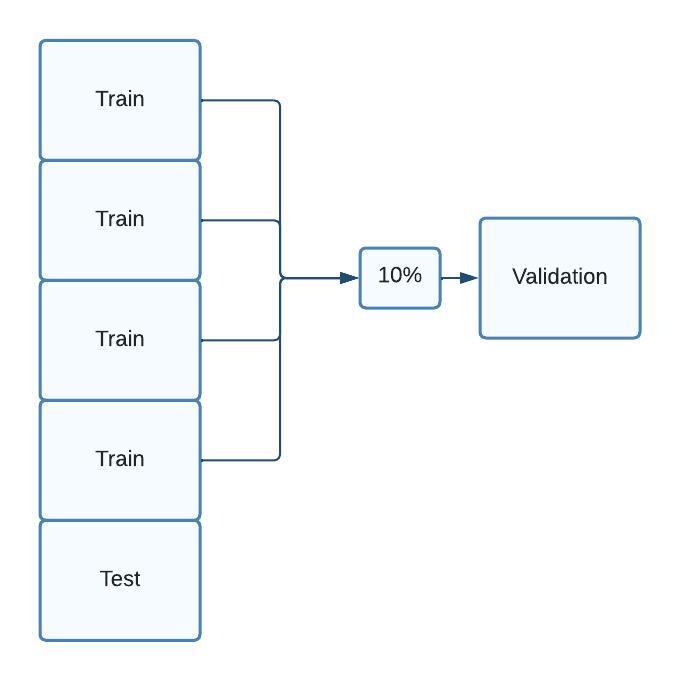
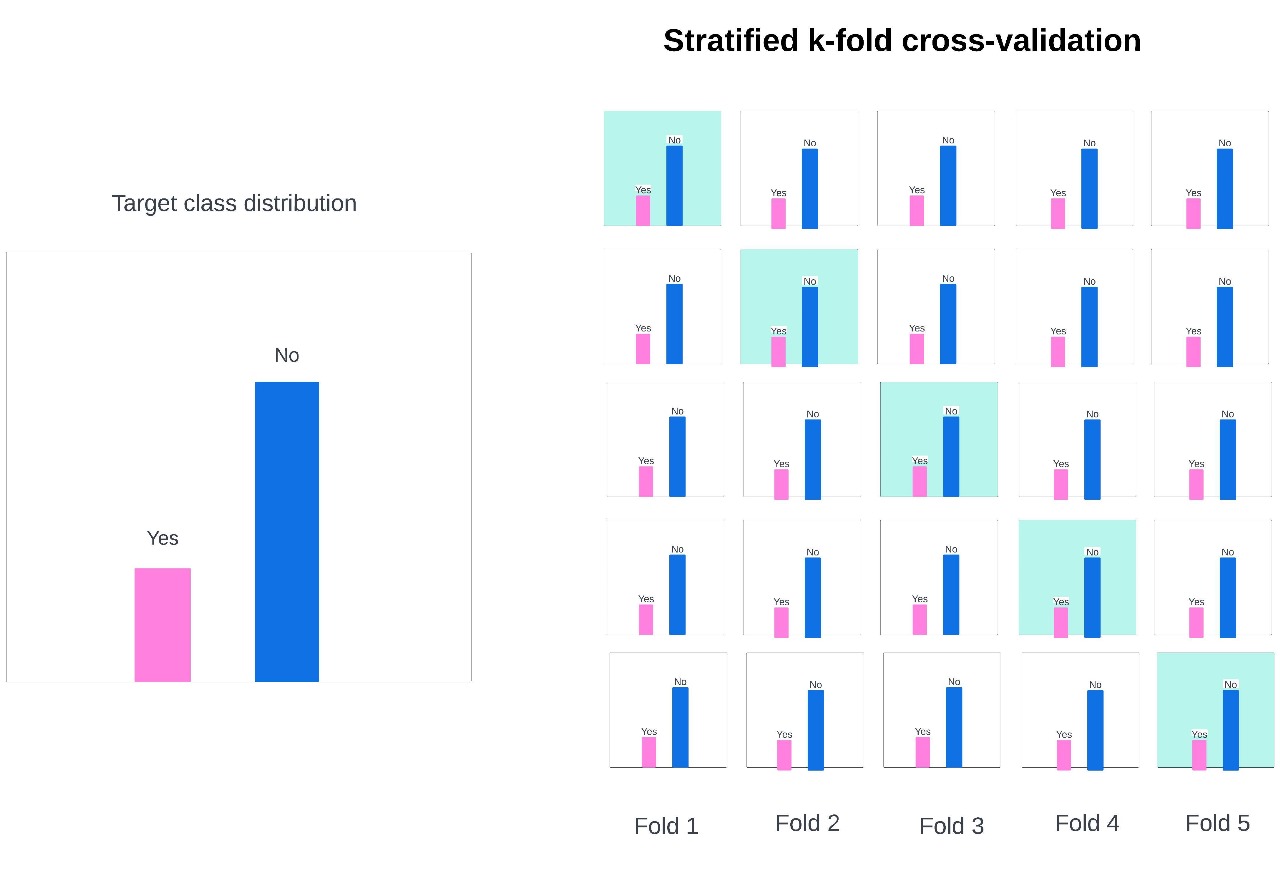
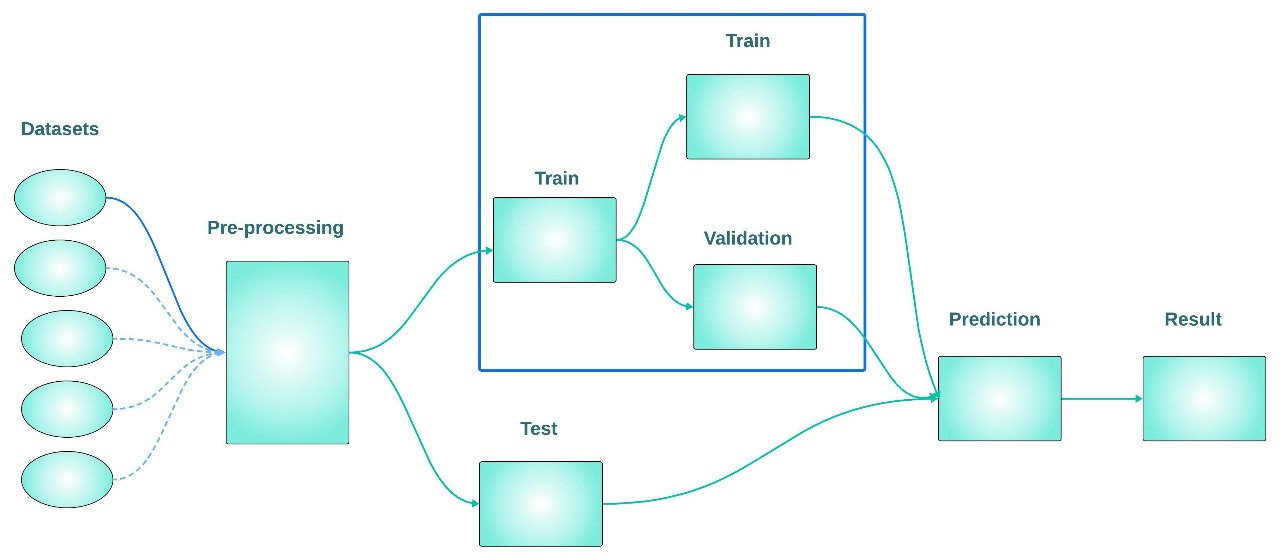


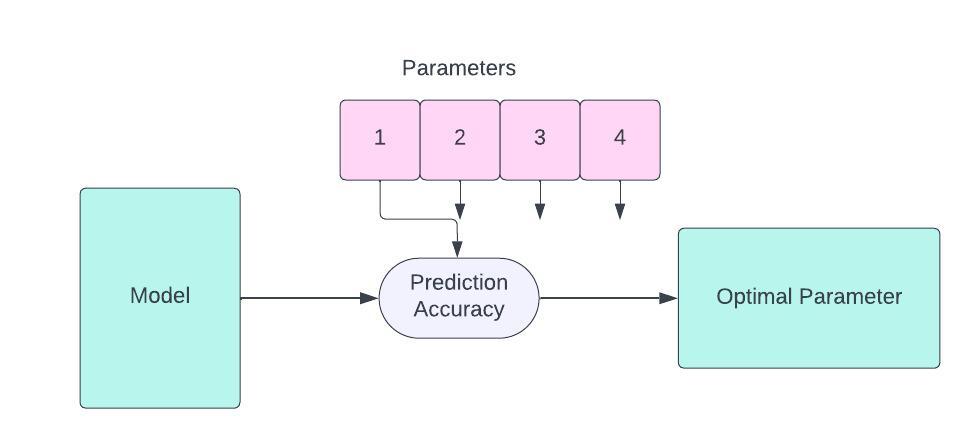
Figure 3. 2: Train-Validation



**Figure 3. 3: Stratified 5-fold Validation**

**3.10 Proposed Methodology:** 

**Figure 3. 4: Proposed Methodology**

**3.11 Hyper parameter Model**

**Figure 3. 5: Hyper parameter Model Tuning**

## **3.12 Conclusion**

The following thesis delves into the effectiveness of various machine learning models for predicting diabetes. A large dataset is used along with sophisticated evaluation methods. The study evaluates eight different models including SVM, Logistic Regression, Random Forest, Decision Tree, KNN, Naive Bayes, MLP, and AdaBoost on five datasets. The goal is to provide a comprehensive review of predictive performance. The results demonstrate the varying degrees of efficacy and accuracy of the models, which are influenced by feature selection techniques and dataset properties. The importance of clinical relevance and prediction accuracy in model selection is emphasized by interpretability analysis. The algorithms that prioritize diabetes prediction offer important new insights into the field by balancing interpretability and accuracy for real-world implementation in healthcare settings.

# **CHAPTER IV**

## **Implementation, Results and Discussions**

## **4.1 Introduction**

We tried to predict the onset of diabetes using various machine learning models such as SVM, Logistic Regression, Random Forest, Decision Tree, KNN, Naive Bayes, MLP, and AdaBoost. To reduce the impact of resource limitations, we used a stratified k-fold cross-validation strategy, dividing datasets into 5 and 10 folds for training and validation, respectively. Each model was thoroughly reviewed based on accuracy and F1 score to ensure robustness and generalizability. After modeling, we conducted comprehensive comparative studies to determine the prediction efficacy among datasets.datasets.

## **4.2 Experimental Setup**

To conduct an experiment using machine learning, it is crucial to have a hardware and software configuration that can handle the task efficiently.

**Hardware Configuration:**

1. Computing Device: It is highly recommended to use a computer with high-performance capabilities or a cloud computing service that has a GPU to speed up the model training.
2. GPU: It is advisable to use NVIDIA GPUs, as they are highly optimized and work well with popular deep learning frameworks.
3. Memory: Ensure that the GPU has adequate memory (VRAM) to handle the computational requirements of training deep learning models, especially for larger architectures and datasets.
4. Storage: Make sure that your system has enough storage space to hold intermediate findings from your experiments, model checkpoints, and datasets.

**Software Configuration:**

1. Operating System: Choose a well-known Linux distribution for its strong performance and compatibility with deep learning frameworks, such as Ubuntu.
2. Python: Install Python (3.6 or higher), the main programming language used to create machine learning frameworks and algorithms.
3. CUDA Toolkit: Install the CUDA toolkit to enable GPU acceleration and maximize computational speed while using NVIDIA GPUs.
4. cuDNN Library: Installing the cuDNN library will improve training efficiency by enabling faster deep learning operations on NVIDIA GPUs.
5. Anaconda: Use Anaconda to simplify the setup and administration of experimental environments by efficiently managing Python environments and package dependencies.
6. Deep Learning Frameworks: Depending on your needs and preferences, choose a deep learning framework like PyTorch or TensorFlow, which have a large community and a lot of functionality.

**Configuring Software:**

1. Operating System: Choose a Linux distribution that is well-known for its strong performance and compatibility with deep learning frameworks, like Ubuntu.
2. Install Python (3.6 or higher), the main programming language used to create machine learning frameworks and algorithms.
3. CUDA Toolkit: To enable GPU acceleration and maximise computational speed while using NVIDIA GPUs, install the CUDA toolkit.
4. Installing the cuDNN library will improve training efficiency by enabling faster deep learning operations on NVIDIA GPUs.
5. Anaconda: Use Anaconda to simplify the setup and administration of experimental environments by efficiently managing Python environments and package dependencies.
6. Deep Learning Frameworks: Based on your needs and preferences, pick one of the many features and robust community support offered by TensorFlow or PyTorch, two popular deep learning frameworks.

## **4.3 Evaluation Metrics**

Various evaluation metrics are utilized to assess the efficacy of machine learning models in predicting the onset of diabetes:

1. **Accuracy:** The parameter Accuracy measures the percentage of correctly identified cases in relation to the total number of occurrences. It provides a general indication of the model's ability to predict both positive and negative cases.

Accuracy = (1)

1. **F1 Score:** A measure of accuracy that balances recall and precision by taking the harmonic mean of both. Useful for datasets with class imbalance as it considers false positives and negatives.

(2)

## **4.4 Dataset**

**Table 4. 1: Data table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Column Name | Samples | Yes | No |
| Dataset-1 (Pima Indians Diabetes Database) | 1. Pregnancies 2. Glucose 3. BloodPressure 4. SkinThickness 5. Insulin BMI 6. DiabetesPedigre 7. Function 8. Age 9. Outcome | 768 | 268 | 500 |
| Dataset-2(A Comprehensive Dataset for Predicting Diabetes with Medical & Demographic Data) | 1. gender    1. Male    2. Female 2. Age 3. Hypertension 4. Heart\_disease 5. Smoking\_history    1. No Info,    2. never,    3. former,    4. current,    5. not current,    6. ever 6. Bmi 7. HbA1c\_level 8. Blood\_glucose\_level 9. diabetes | 100000 | 82 | 918 |
| Dataset-3(Predict a Model to detect Person has Diabetes or Not) | 1. Pregnancies 2. Glucose 3. BloodPressure 4. SkinThickness 5. Insulin 6. BMI 7. DiabetesPedigreeFunction 8. Age 9. Outcome | 2000 | 1316 | 684 |
| Dataset-4(Healthcare Diabetes Dataset-  A Comprehensive Dataset for Diabetes Risk Assessment) | 1. Id 2. Pregnancies 3. Glucose 4. BloodPressure 5. SkinThickness 6. Insulin 7. BMI 8. DiabetesPedigreeFunction 9. Age 10. Outcome | 2768 | 1816 | 952 |
| Dataset-5  (70,692 survey responses from cleaned BRFSS 2015) | * Diabetes, * Hypertension and * Stroke Prediction * This is a clean dataset of over 70,692 survey responses to the CDC's BRFSS2015. * There are 18 feature variables and 3 target variables,   + Diabetes,   + Hypertension, and   + Stroke,   each target variable having 2 classes. This dataset is balanced   1. Age 2. Sex 3. HighChol 4. CholCheck 5. BMI 6. Smoker 7. HeartDiseaseorAttack 8. PhysActivity 9. Fruits 10. Veggies 11. HvyAlcoholConsump 12. GenHlth 13. MentHlth 14. PhysHlth 15. DiffWalk 16. Stroke 17. HighBP 18. Diabetes | 70,692 | 35346 | 35346 |

## **4.5 Implementation and Results**

### 4.5.1 Quantitative results

**Table 4. 2: Accuracy Score**

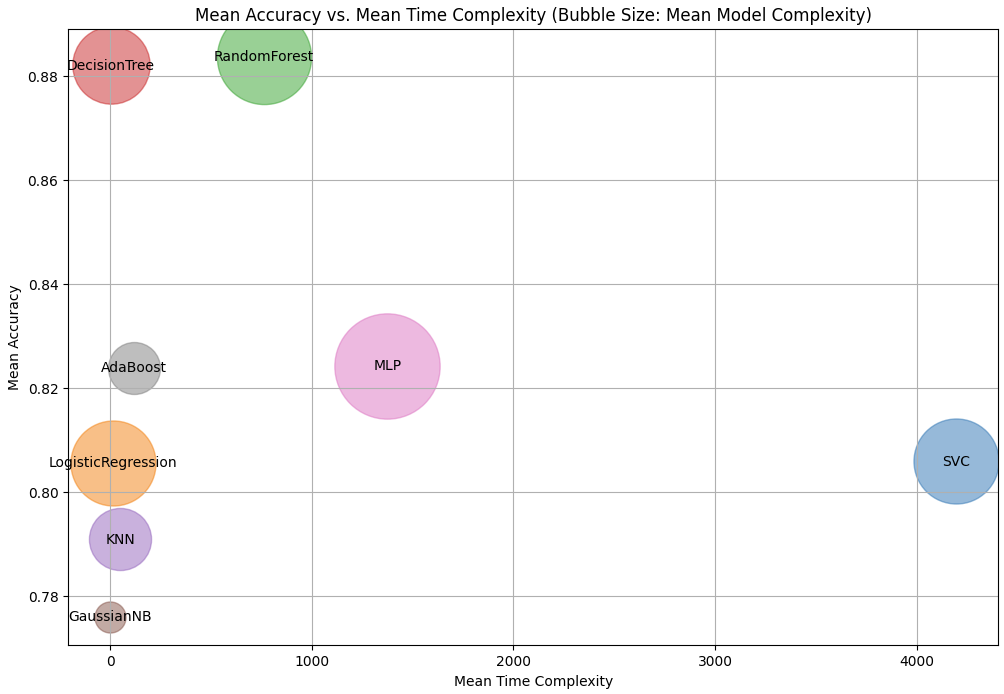
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Test Accuracy Score** | | | | | |
| **Model** | **DATASET-1** | **DATASET-2** | **DATASET-3** | **DATASET-4** | **DATASET-5** |
| SVC | 0.774713522 | 0.96046 | 0.775 | 0.77167338 | 0.747779042 |
| LogisticRegression | 0.770783465 | 0.96023 | 0.773 | 0.776371743 | 0.747496148 |
| RandomForest | 0.762948816 | 0.96981 | 0.969 | 0.989163147 | 0.727168553 |
| DecisionTree | 0.748637637 | 0.97187 | 0.962 | 0.989164453 | 0.738343823 |
| KNN | 0.69526356 | 0.96302 | 0.7875 | 0.82984378 | 0.678704791 |
| GaussianNB | 0.744733045 | 0.90347 | 0.751 | 0.757233599 | 0.723589677 |
| MLP | 0.747381377 | 0.97189 | 0.803 | 0.847913906 | 0.75107505 |
| AdaBoost | 0.77206519 | 0.97195 | 0.8145 | 0.811414601 | 0.74953312 |

**Table 4. 3: F-1 Score**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Test (F-1) Score** | | | | | |
| **Model** | **DATASET-1** | **DATASET-2** | **DATASET-3** | **DATASET-4** | **DATASET-5** |
| SVC | 0.638692383 | 0.714274011 | 0.627140027 | 0.628355269 | 0.758243682 |
| LogisticRegression | 0.63508249 | 0.725716142 | 0.623869043 | 0.633119938 | 0.752432688 |
| RandomForest | 0.642049455 | 0.79488311 | 0.954291113 | 0.984314382 | 0.736455142 |
| DecisionTree | 0.573577911 | 0.801697752 | 0.944986434 | 0.98432148 | 0.748851576 |
| KNN | 0.543407521 | 0.74506928 | 0.67580264 | 0.752996911 | 0.634221998 |
| GaussianNB | 0.622040894 | 0.532228298 | 0.626722566 | 0.629596523 | 0.721919095 |
| MLP | 0.632113337 | 0.804710987 | 0.702475091 | 0.757706116 | 0.764969475 |
| AdaBoost | 0.644404256 | 0.806368309 | 0.712213438 | 0.711110312 | 0.755184141 |

### Qualitative results

**Figure 4. 1: Critical difference**



**Figure 4. 2: Bubble Chart Accuracy vs Complexity**

### Analysis of the results:

The bubble chart output displays a visual representation of the average accuracy and time complexity of various machine learning models. Each bubble on the chart corresponds to a model, with the mean accuracy on the y-axis and mean time complexity on the x-axis. The size of each bubble represents the average model complexity, with larger bubbles indicating a higher level of complexity. The chart shows the balance between the complexity of a model, the time it takes to execute, and its accuracy.

If you see larger bubbles in the upper-right area of the plot, it means that more accurate models typically have higher time and model complexity. On the other hand, smaller bubbles in the lower-left corner indicate that less accurate models may have lower time and model complexity.

To choose the best machine learning model, consider accuracy, time, and complexity trade-offs. Visualize model performance to compare accuracy and computing complexity.

## **4.6 Objectives Achieved**

We successfully completed two major tasks in our project to predict diabetes:

Gathering and Preparing Data: We carefully collected a large dataset containing relevant information for the prediction of diabetes. Ensuring the accuracy of the data was crucial, so we carefully handled any missing information and organized the data in a useful way, making it ready for analysis.

Building and Testing Models: We created eight different computer models to predict diabetes, using various techniques to analyze the data. Each model was tested to evaluate its accuracy in predicting diabetes.

Analysing Model Performance: We evaluated the performance of different models to predict diabetes using various metrics such as accuracy and F1-score. This helped us determine the best-performing models.

Verifying Our approaches: To confirm the reliability and consistency of our findings, we used stratified k-fold cross-validation method to verify the accuracy of our approaches.

Learning from Our Experiences: Comparing the performances of different models under different conditions helped us identify the most effective models for predicting diabetes. This gave us valuable insights that will guide our future studies and progress in this field.

## 4.3 **Morality or Ethical Issues**

The use of machine learning techniques to predict diabetes raises moral and ethical issues, including privacy, data bias, ownership and control of the data, healthcare professional involvement, transparency, misdiagnosis, regulation, accountability, algorithmic medical decision-making, beneficence and non-maleficence, equity and access. Important considerations include privacy, informed consent, data bias, data ownership, engagement of healthcare professionals, interpret-ability, misdiagnosis, long-term effects, regulation, accountability, algorithmic medical decision-making, beneficence and non-maleficence, equi-ty, and access. To manage these problems responsibly, cooperation between healthcare practitioners, data scientists, and legislators is required. Machine learning can maximize the promise of diabetes prediction while protecting patient safety, privacy, and equity by taking into account these moral and ethical issues.

## 4.4 **Socio-Economic Impact and Sustainability**

Machine learning techniques for predicting diabetes can have major socioeconomic effects, such as lower healthcare costs, better patient outcomes, better resource allocation, patient empowerment, and lower health insurance premiums. These advantages support preventative care, data-driven decision-making, long-term health management, a smaller carbon footprint, healthcare access and equity, and a positive impact on global health. They are in line with sustainable healthcare practices. But there are issues that must be resolved, including data privacy, technology accessibility, potential biases, and continual monitoring and updating of predictive models. To maximize the socioeconomic advantages and sustainability of machine learning for diabetes prediction, cooperative efforts involving healthcare providers, policymakers, technology developers, and communities are essential.

## 4.5 **Financial Analyses and Budget**

This thesis project requires careful budget management for efficient allocation of resources for data collection, experimentation, and computational assistance. The funding plan used for the project is outlined in this section.

We have taken a lot of care in optimising our spending by making use of institutional resources and open-source solutions, given the limitations we face. We have found that the use of open-source Integrated Development Environments (IDEs), such as Google Colab, substantially reduces costs without compromising the quality of the output. Additionally, we prefer using publicly available datasets wherever possible, as this reduces acquisition costs while still giving proper recognition to the creators of the datasets.

Provisions have been made in the budget to ensure the project's smooth progress despite any unforeseen expenses, as detailed in Table 4.1.

|  |  |
| --- | --- |
| Budget | Estimated Cost (BDT) |
| Research Tools and Technology | 16,000 - 22,000 |
| Data Procurement and Management | 8,000 - 10,000 |
| Travel Expenses and Professional Development | 4,000 - 6,000 |
| Publication Costs and Dissemination | 3,000 - 4,000 |
| Incidental and Administrative Expenses | 2,000 - 3,000 |
| Total | 33,000 - 4500 |

The thesis will fulfill its goals and maintain financial responsibility by sticking to a flexible budget that accounts for unforeseen circumstances.

### 4.9 Conclusion

We used a heterogeneous dataset consisting of five independent datasets to evaluate the predictive power of eight machine learning models for diabetes. Our study showed that machine learning accurately predicts diabetes consequences. We also provided insights into the advantages and disadvantages of each model for different datasets. Our findings highlight the potential of machine learning to revolutionize medical procedures, but further research and optimization may be required in the future to enhance its ability to function effectively across various datasets and real-world scenarios.

**CHAPTER V**

**Societal, Health, Environment, Safety, Ethical, Legal and Cultural Issues**

### 5.1 Intellectual Property Considerations

## During our thesis, we conducted an investigation into machine learning models for predicting diabetes. However, we encountered significant issues related to intellectual property. Our intellectual property is not limited to the development of innovative approaches, but also includes the data preparation procedures, model architectures, and optimization strategies that we have customized for diabetes prediction. To ensure the protection of these resources, we have maintained strict documentation. We acknowledge the importance of the datasets we have chosen, which along with our model approaches, constitute crucial intellectual property. It is crucial to carefully manage intellectual property when developing software tools and scripts for training and testing models. This includes controlling access and version control. Our interdisciplinary study involves collaborating with specialists from diverse fields to gain fresh perspectives and reach joint agreements regarding intellectual property ownership or licensing. Our primary focus is to follow ethical guidelines for intellectual property management and maximize the value and impact of our research. Therefore, we remain vigilant in addressing and handling intellectual property issues throughout our investigation.

### 5.2 Ethical Considerations

As we work on our thesis investigating the effectiveness of machine learning models for diabetes prediction, we prioritize ethical considerations. It is important to us to handle data ethically, through informed consent, confidentiality, and privacy protection laws. This means that we first obtain permission before using any sensitive or identifiable information, and we anonymize data where necessary to protect privacy. Moreover, we strive to minimize potential biases in our datasets and algorithms to ensure fairness and justice in our research results. Our study adheres to ethical guidelines and emphasizes openness, honesty, and integrity in all aspects of the inquiry. We also promote justice and decency in our interactions with each other. we prioritize the health and safety of all individuals involved in our research, including participants and partners, by implementing appropriate safety protocols and ethical guidelines. Simply put, we aim to conduct our research in a socially responsible and ethical manner, upholding the values of fairness, honesty, and decency in all aspects of our work.

### 5.3 Safety Considerations

Our thesis focuses on how accurately machine learning models can predict diabetes, with safety being a top priority. Data privacy and confidentiality are crucial aspects of our research, and we take strict measures to protect patient data from misuse or illegal access. Our team uses robust data anonymization techniques to protect people's privacy and adhere to ethical principles. To ensure the security of study data, we also take precautions with our computational resources. Moreover, we prioritize the health and safety of all participants, collaborators, and researchers involved in the study process. In order to conduct our research effectively, we must properly manage data and adhere to institutional policies and procedures. Our safety measures are designed to minimize potential privacy and data security risks while upholding the integrity and ethics of our study.

### 5.4 Legal Considerations

Legal issues play a crucial role in ensuring compliance with relevant laws and regulations governing healthcare data, as illustrated in our thesis on the effectiveness of machine learning models for predicting diabetes. It is vital to adhere to data protection regulations such as GDPR (General Data Protection Regulation) in Europe or HIPAA (Health Insurance Portability and Accountability Act) in the United States, given the sensitive nature of the medical data used in our study. This requires implementing robust data security measures to protect patient privacy and confidentiality at every stage of the study, from data collection to analysis and dissemination of results.To ensure the protection of patient rights and minimize legal risks, it is important to adhere to ethical considerations related to informed consent and transparent data handling practices. We prioritize ethical norms and ensure the legitimacy and acceptance of our research in the academic and professional fields by addressing these legal issues. To avoid potential legal problems and contribute to the ethical development of machine learning in healthcare, it is essential to incorporate legal compliance into the planning and implementation of our projects.

### 5.5 Impact of the Project on Societal, Health, and Cultural Issues

Our thesis has a significant impact on societal, health, and cultural issues, improving various aspects of these fields.  
  
5.5.1. Societal Impact:

* Our initiative utilizes machine learning algorithms to predict diabetes, aiming to improve healthcare accessibility and equity.
* Our research aims to enhance societal well-being by accurately predicting and addressing healthcare delivery and access gaps, especially in marginalized communities.

5.5.2. Health Impact:

* The use of machine learning algorithms for predicting diabetes can have significant implications for health outcomes.
* Early identification of those at risk can reduce diabetes prevalence.
* Our research improves public health by bettering disease management and prevention efforts.

5.5.3. Cultural Impact:

* Although our research primarily focuses on healthcare, it has significant implications for culture.
* The integration of cutting-edge technology in healthcare may influence how various cultures perceive health and innovation in healthcare systems.
* We can promote a culture of innovation and growth in healthcare delivery by embracing technological advancements and implementing cultural shifts in healthcare practices.

### 5.6 Impact of Project on the Environment and Sustainability

The significance of a study or thesis on sustainability and environment lies in its ability to address sustainability issues and promote sustainable practices.

5.6.1. Environmental Impact:

* Our research on predicting diabetes through machine learning models provides direct support for environmental conservation initiatives.
* We optimize computer resource usage to reduce energy consumption during model training and evaluation.
* Moreover, our study advocates for the implementation of digital healthcare technologies to reduce environmental impact and minimize the need for physical infrastructure.

5.6.2 Sustainability Impact:

* Our primary focus is healthcare, but our research indirectly promotes sustainability through technological innovation.
* Our project improves healthcare sustainability by preventing and managing illnesses.
* Our aim is to enhance sustainability in healthcare by utilising machine learning to create efficient solutions that reduce waste.

5.6.3. Waste Reduction:

* Our project aims to indirectly reduce waste by optimizing healthcare services, even though it is not directly related to waste management.
* Our project aims to indirectly reduce waste by optimizing healthcare services, even though it is not directly related to waste management.
* Our study promotes preventative healthcare, which can reduce healthcare industry waste generation over time.

**CHAPTER VI**

**Addressing Complex Engineering Problems and Activities**

## 6.1 Complex engineering problems associated with the current thesis

1. Preprocessing tasks, including data cleaning, normalization, and feature engineering, must be carefully considered when managing and preparing five different datasets for diabetes prediction due to differences in quality, format, and distribution of data.
2. Selecting the most suitable machine learning models for diabetes prediction and optimizing their hyperparameters can be a challenging task. Out of the eight models considered, it is crucial to conduct thorough testing and validation to ensure optimal performance.
3. To ensure accurate evaluation and comparison of models across datasets, careful planning and effective resource utilization is required for the 5-fold and 10-fold variations of the stratified k-fold cross-validation method.
4. Resolving class imbalance in diabetes prediction datasets can be difficult. Ensemble approaches, class weighting, or resampling can help reduce bias and enhance predictive performance during model evaluation and training.
5. Managing computational resources effectively is crucial when conducting multiple experiments with large datasets and complex models. In order to reduce processing time and resource consumption, optimization techniques such as parallel processing, distributed computing, or model compression may be employed.
6. To draw solid conclusions on the efficacy of machine learning models for diabetes prediction, analyzing and interpreting results across multiple datasets and models using advanced statistical analysis and visualization techniques is necessary.

Complex engineering activities associated with the current project/thesis

1. We used a thorough investigation and experimentation process to choose eight appropriate machine learning models for predicting diabetes, taking into account factors like algorithm complexity, performance metrics, and computational efficiency
2. To ensure efficient model training and evaluation, strong preprocessing techniques must be used on all five datasets to address inconsistencies, missing values, and outliers. This will guarantee consistency and quality.
3. We can use stratified k-fold approach with 5-fold and 10-fold variations to validate model performance on different datasets, but it requires careful preparation and adequate computational resources for reliable evaluation.
4. This process involves iteratively experimenting and validating optimal settings for each dataset in order to fine-tune the model hyperparameters for each chosen algorithm to maximize performance and generalization.
5. To pick the ideal diabetes prediction model, cross-validation trials are performed, and performance outcomes are analyzed using various metrics and statistical analyses.

**CHAPTER VII**

**Conclusions**

# **References**

1. Aishwariya Dutta, Md Kamrul Hasan, Mohiuddin Ahmad, Md Abdul Awal, Md Akhtarul Islam, Mehedi Masud, and Hossam Meshref. Early prediction of diabetes using an ensemble of machine learning models. *International Journal of Environmental Research and Public Health*, 19(19):12378, 2022.
2. Harleen Kaur and Vinita Kumari. Predictive modelling and analytics for diabetes using a machine learning approach. *Applied computing and informatics*, 18(1/2):90–100, 2022.
3. Aishwarya Mujumdar and V Vaidehi. Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, 165:292–299, 2019.
4. Mitushi Soni and Sunita Varma. Diabetes prediction using machine learning techniques. *International journal of engineering research and technology*, 9, 2020.
5. Fikirte Girma Woldemichael and Sumitra Menaria. Prediction of diabetes using data mining techniques. In *2018 2nd International Conference on Trends in Electronics and*

*Informatics (ICOEI)*, pages 414–418, 2018.

**REFERENCES**

Use IEEE formatting style

[ *2 blank line of size 12, 1.5 line*]

**Please follow the link below for example**

Chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://ieeeauthorcenter.ieee.org/wp-content/uploads/IEEE-Reference-Guide.pdf

[ *For References*

*Times Roman, normal, justified, 12 size, 1.2 line space,*

*Space before 6 pt, after 6 pt*

*Bullet is left aligned,*

*Paragraph-> left indentation=0, hanging =0.31”]*

**N.B.** This is the preferable format of progress report for Thesis writing. The sections written in italic forms are fixed. However, the Supervisor can extend the chapters/points of report for Thesis writing (If necessary).