**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.** i](#_Toc159087231)

[**Acknowledgment** ii](#_Toc159087232)

[**Abstract** iii](#_Toc159087233)

[**Contents** iv](#_Toc159087234)

[**List of Table** vi](#_Toc159087235)

[**List of Figure** vii](#_Toc159087236)

[**CHAPTER I** 1](#_Toc159087237)

[**Introduction** 1](#_Toc159087238)

[**1.1**  **Introduction** 1](#_Toc159087239)

[**1.2**  **Problem Statement** 1](#_Toc159087240)

[**1.3**  **Objectives** 2](#_Toc159087241)

[**1.4** **Unfamiliarity of the Problem** 3](#_Toc159087242)

[**1.5** **Project Planning** 3](#_Toc159087243)

[1.5.1 Project Timeline 3](#_Toc159087244)

[1.5.2 Legal and Ethical Aspects 4](#_Toc159087245)

[**1.6** **Applications of the Work** 5](#_Toc159087246)

[**1.7** **Organization of the Report** 6](#_Toc159087247)

[**CHAPTER II** 7](#_Toc159087248)

[**Literature Review** 7](#_Toc159087249)

[**2.1**  **Introduction** 7](#_Toc159087250)

[**2.1** **Literature Review** 8](#_Toc159087251)

[**2.2** **Discussion of research gap solution** 9](#_Toc159087252)

[**CHAPTER III** 11](#_Toc159087253)

[**Methodology** 11](#_Toc159087254)

[**3.1** **KNN algorithm** 11](#_Toc159087255)

[**3.2** **Logistic regression algorithm** 12](#_Toc159087256)

[**3.3** **Na¨ıve Bayes** 12](#_Toc159087257)

[**3.4** **Decision tree algorithm** 12](#_Toc159087258)

[**3.5** **Random Forest algorithm** 12](#_Toc159087259)

[**3.6** **SVM algorithms** 13](#_Toc159087260)

[**3.7** **Data Splitting** 13](#_Toc159087261)

[**3.8** **Data Preprocessing** 13](#_Toc159087262)

[**3.9** **Feature Selection** 14](#_Toc159087263)

[**3.10** **Ensemble Methods** 14](#_Toc159087264)

[**CHAPTER IV** 15](#_Toc159087265)

[**Results and Discussions** 15](#_Toc159087266)

[**4.1** **Results** 15](#_Toc159087267)

[**4.2** **Objectives Achieved** 15](#_Toc159087268)

[**4.3** **Morality or Ethical Issues** 15](#_Toc159087269)

[**4.4** **Socio-Economic Impact and Sustainability** 16](#_Toc159087270)

[**4.5** **Financial Analyses and Budget** 16](#_Toc159087271)

[**CHAPTER V** 17](#_Toc159087272)

[**Conclusions** 17](#_Toc159087273)

[**5.1** **Conclusion and Challenges Faced** 17](#_Toc159087274)

[**5.2** **Future Works** 17](#_Toc159087275)

[**References** 17](#_Toc159087276)

[20](#_Toc159087277)

# **List of Table**

[Table 3. 1 : Research gaps of different models 8](#_Toc159020893)

[Table 3. 2 8](#_Toc159020894)

# **List of Figure**

[Figure 1. 1: Essential Learning process to develop a predictive model 2](#_Toc159020722)

[Figure 1. 2: Gantt Chart of the Project Plan 3](#_Toc159020723)

# **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:

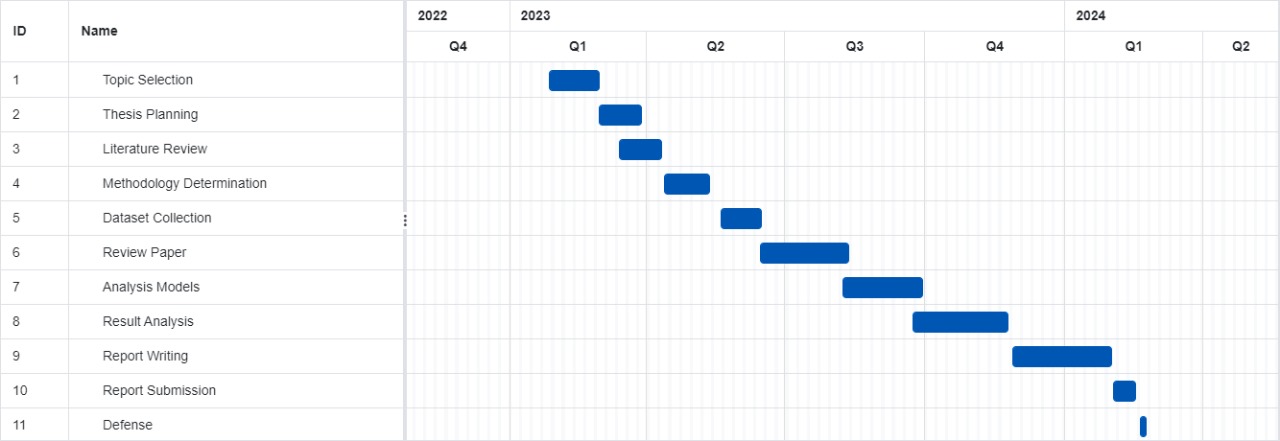


Figure 1. 1: Gantt Chart of the Project Plan

### 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**

The preprocessing techniques ensure the precision and dependability of the data. It utilizes various machine-learning strategies like K-Nearest Neighbors, Random Forest, SVM, ANN, and Decision Tree. The textual (PIDD) dataset being used is extensive and contains variables such as age, blood pressure, skin thickness, glucose levels, insulin, and BMI. Diabetes is defined by a blood pressure reading that exceeds 80 and a glucose reading that surpasses 105. This study contributes to our knowledge of diabetes classification and prediction by utilizing sophisticated machine-learning techniques.

## **3.1 KNN algorithm**

This study uses data preprocessing to ensure data quality and reliability, followed by machine learning techniques like K-Nearest Neighbors, Random Forest, SVM, ANN, and Decision Tree. The comprehensive dataset includes parameters like pregnancy, glucose levels, blood pressure, skin thickness, insulin, BMI, ancestral background, and age. This study contributes to understanding diabetes prediction and classification using sophisticated machine learning methodologies. The three distance indices are applicable only to continuous variables, and the Hamming distance is suitable for predetermined categories. Normalizing numerical variables between 0 and 1 is crucial in datasets with both numerical and category variables.

## **3.2 Logistic regression algorithm**

Logistic regression is an extended linear model used to predict diabetes in a binary study with two possible outcomes. It uses the sigmoid function to transform its output into a range between 0 and 1, representing probabilities.

## **3.3 Na¨ıve Bayes**

A method for automatic learning that makes use of the probability and statistics’ Bayes theorem. It is stupid because it thinks that all features are equal, to put it frankly. When doing classification tasks that need linear, supervised learning, this method is frequently used. It has already been utilized with success for content classification and text processing.

The distinctive feature of this particular technique is probabilistic processing.

## **3.4 Decision tree algorithm**

Decision trees are a powerful guided learning method for making future predictions in various contexts. They use algorithms to find natural breaks in datasets, resulting in a prototype that accurately forecasts target variable values using simplified rules. The leaf Node gives the result and the Internal nodes are the labels of the data we are using.

## **3.5 Random Forest algorithm**

Frequently Used Machine Learning Algorithm One of the more well methods of machine learning is supervised learning, of which Random Forest is a subset. Both classification and regression can be performed using ML approaches. It uses classifiers, which combine the output from several classifiers to tackle challenging problems and improve overall model performance. Random Forest is an algorithm that averages multiple decision trees trained on random samples to improve predictive accuracy. It takes forecasts from all trees and makes final predictions based on the majority of projections. The more trees, the less likely overfitting increases. This technique is useful in medicine and land usage analysis.

## **3.6 SVM algorithms**

Support Vector Machine (SVM) is a supervised machine learning technique that is used for classification issues. It creates an n-dimensional hyperplane with n features, splitting preprocessed data into training and testing sets. SVM models the training set and the testing set for diabetes predictions. Kernel functions generate outputs, with polynomial kernels employed. Overfitting and misclassification can be reduced by increasing C and adjusting regularization parameters.

## **3.7 Data Splitting**

K-Fold Cross-Validation is a widely used technique in machine learning and statistics to assess predictive model performance and mitigate overfitting or bias issues. It involves dividing a dataset into K subsets or ”folds” of equal size, training and validating the model K times, and evaluating the model K times. The average performance is then averaged after K iterations, providing a more reliable indicator of the model’s performance on unseen data. K-Fold Cross-Validation offers several advantages, including a more robust estimate of generalization performance, mitigation of random variations in train-test splits, maximizing data utilization. Common values for K include 5 and 10, but other values can be used depending on the dataset size and computational resources.

## **3.8 Data Preprocessing**

Trainable automated classification decision-making frameworks rely on datasets with missing values, such as NaNs, null, blanks, or undefined placeholders. To develop a robust and effective classification model, statistical and machine learning approaches like median and

KNN-based imputation techniques are employed [1].

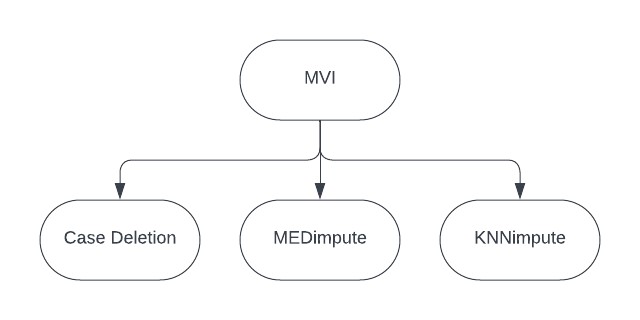


Figure 3.1: Missing Value Imputation

## **3.9 Feature Selection**

One technique for feature selection (FS) is called RF, which uses a tree-based approach to rank features based on how much they reduce impurities at each node. This technique tends to have the most significant reductions at the beginning of the tree and slight reductions towards the end. By trimming the tree below a certain node, we can obtain a subset of relevant features [1].

## **3.10 Ensemble Methods**

Ensemble methods can significantly enhance model performance by reducing overfitting, improving stability, and capturing complex patterns in the data. However, they might increase computational requirements and model complexity. The choice of ensemble method depends on the problem at hand, the diversity of models available, the trade-off between performance gains and computational costs.

# **CHAPTER IV**

# **Results and Discussions**

## **4.1 Results**

There are many datasets(textual) available on diabetes, so dataset collection is not a big concern. I want to work with the PIMA Indian Diabetes Dataset based on the textual dataset, but the noticeable point is that there is much work in the same dataset like PIDD in which different machine learning algorithm is used. Some term that enhances model performance is correct data splitting (k-fold validation), missing value imputation, and feature selection to avoid unnecessary features in the ensemble because they can form a complex pattern to predict diabetes.

## **4.2 Objectives Achieved**

In order to work on this topic, I made it a point to understand the severity of diabetes disease and familiarize myself with various machine-learning terms. Machine learning techniques are crucial in the field of Biomedical Engineering as they help predict diseases such as diabetes at an early stage. By utilizing these methods, individuals can take necessary actions to address their health concerns.

## **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**

A machine learning-based diabetes prediction system’s financial analysis and budgeting must take into account a variety of expenses and potential returns on investment. Data gathering, preparation, infrastructure, model development and training, validation and testing, deployment and integration, maintenance and upgrades, and regulatory compliance are important financial considerations. To determine healthcare cost savings, productivity increases, revenue creation, and insurance premium reduction, returns on investment (ROI) calculations should be made. The project’s budget allocation and return on investment should be calculated for each phase. Risk analyses should be carried out to account for various scenarios and identify potential uncertainties. Long-term budgeting should take continuing costs for upkeep, updates, and compliance into account. Stakeholders should be given a concise value proposition that outlines the financial advantages and improved patient outcomes.

# **CHAPTER V**

# **Conclusions**

## **5.1 Conclusion and Challenges Faced**

A machine learning method is required for the implementation of the diabetes prediction model, but it is very difficult to build a proper model that performs errorless in every textual data set. Our dataset is textual data (PIDD), We know the machine learning model to train and test data but there is a question if it gives perfect results or not. So, at this time I am studying it and my target is to write a review paper on this topic. After getting more knowledge on it I will try to contribute by building a good model to predict diabetes, which can give us expected output.

## **5.2 Future Works**

Although there are many models that perform with high accuracy on the textual dataset by an ensemble of machine learning models. I wanted to build a model that gives low error on any dataset and gives expected output with high accuracy. It can expand to create a complete space for both doctors and patients with proper communication. It can be expanded to create an online solution.

# **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).