DIABETICS PREDICTION SYSTEM USING MACHINE LEARNING ALGORITHMS "A CASE STUDY AMONG FEMALE PATIENT"

BY

Arman Hossain

ID: 0242220005103003

This Report Presented in Partial Fulfillment of the Requirements for

The Degree of Master of Science in Computer Science and Engineering

Supervised By

Professor Dr. Md. Fokhray Hossain

Professor

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

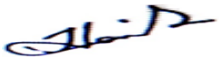
DHAKA, BANGLADESH

JANUARY 2024

APPROVAL

This Thesis titled **"Diabetics prediction System using machine learning algorithms "a case study among female patient""** submitted by **Arman Hossain**, ID No: 0242220005103003 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of **M.Sc. in Computer Science and Engineering** and approved as to its style and contents. The presentation has been held on 27-01-2024.

BOARD OF EXAMINERS



Chairman

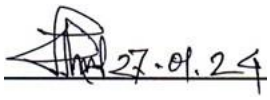
Professor Dr. Sheak Rashed Haider Noori
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Fizar Ahmed, PhD
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Arif Mahmud, PhD
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



External Examiner

Dr. Firoz Ahmed, PhD
Professor
Department of Information & Communication Engineering
Rajshahi University

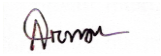
DECLARATION

We hereby declare that, this thesis has been done by us under the supervision of **Professor Dr. Md. Fokhray Hossain, Professor** Department of CSE Daffodil International University. We also declare that no other academic work, in whole or in part, has been submitted for the purpose of obtaining a degree.

Supervised by:

Professor Dr. Md. Fokhray Hossain
Professor
Department of CSE
Daffodil International University

Submitted by:



Arman Hossain
ID: 0242220005103003
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final thesis successfully.

We really grateful and wish our profound our indebtedness to **Professor Dr. Md. Fokhray Hossain**, Professor, Department of CSE Daffodil International University, Dhaka. To complete this thesis, our supervisor must have deep knowledge of and a strong interest in the topic of "machine learning." Completing this thesis has been made possible by his unending patience, academic guidance, ongoing encouragement, steady and enthusiastic supervision, constructive criticism, insightful counsel, reviewing numerous subpar versions and fixing them at every step. Finally, we must respectfully appreciate our parents' unwavering devotion and support.

We would like to sincerely thank our parents, our family, and the head of the CSE department at Daffodil International University, Professor Dr. Touhid Bhuiyan, for his helpful assistance in finishing our project, as well as the other faculty and staff members in the department.

We express our gratitude to every student at Daffodil International University who participated in this discussion during the course of their studies.

Finally, we must respectfully appreciate our parents' for their passion and support.

ABSTRACT

Millions of individuals worldwide suffer with diabetes mellitus, a chronic metabolic ailment, and early diagnosis is essential to controlling the condition and avoiding complications. The development of precise prediction systems for a variety of medical illnesses has shown encouraging outcomes in recent years thanks to machine learning techniques. The goal of this study is to employ machine learning to create a diabetic prediction system that is specifically created for females. The planned study will examine a dataset that includes detailed health records for females, together with data on their lifestyle choices, medical histories, and demographics. The dataset will be processed using a suitable machine learning technique to produce a predictive model. The accuracy, recall, f-measure and precision of various algorithms will be compared in order to determine which is the most efficient. I'm using the Pima Indian Diabetes Database data set from Kaggle.com for this research. In this work I have used some machine learning algorithms which is logistic regression, Naive-bayes, support vector machine (SVM), decision tree and Random Forest. After preprocessing the data set and applying this algorithms the prediction performance of diabetics disease analysis shows that random forest obtained the uppermost performance with the accuracy of 85% and decision tree has achieved the second highest accuracy which is 81%.

Keywords— PIMA dataset, data pre-processing, classifier, Naïve-bayes, support vector machine, decision tree, random forest.

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Background of Research

Diabetes is a long-term metabolic disease marked by high blood sugar levels. If left untreated, diabetes can cause major health problems. A growing number of people are affected by this worldwide health hazard, which is becoming more common. The prevalence and management of diabetes are significantly influenced by gender, one of several risk factors and demography. This case study employs machine learning algorithms to predict the risk of diabetes in females.

The incidence of diabetes in women has become a significant concern due to the distinct obstacles that women have when it comes to health management, particularly during life transitions like pregnancy and menopause. Comprehending and forecasting the likelihood of diabetes in women is crucial for timely intervention and avoidance. Machine learning is a subfield of artificial intelligence. It has demonstrated remarkable potential in the analysis of complicated medical data to enable precise forecasting and decision support.

The purpose of this case study is to investigate how machine learning algorithms are applied to forecast the risk of diabetes in females. We can create a prediction model that can assist in identifying those who are at risk by examining a variety of variables and health data, including age, family history, lifestyle, and genetic susceptibility. With the ability to provide proactive interventions and individualized care plans for people at risk of acquiring diabetes, these predictive models hold the potential to completely transform the healthcare industry.

In this study we will describe the methodology, data collection, data preprocessing, and model generation procedures. We will also talk about the significance of early prediction and its possible effects on the healthcare system.

1.2 Motivation

The prevalence of diabetes has seen an alarming rise in recent years, posing a significant public health concern. As per the World Health Organization (WHO), between 1980 and 2019, the prevalence of diabetes among adults worldwide increased from 4.7% to 8.5%, with numbers expected to rise further in the coming years [7]. Notably, complications arising from diabetes,

such as cardiovascular diseases, renal failure, and neuropathy, contribute to increasing mortality rates [4].

Traditional diagnostic methods, although effective to some extent, have limitations in terms of timeliness and accuracy. For example, the fasting plasma glucose test, which is widely used, can sometimes give false negatives and is influenced by various external factors such as diet and exercise. This highlights the need for more robust and accurate methods for diabetes prediction [5].

Enter Machine Learning (ML)—a domain that has shown immense promises in healthcare applications, from predicting cancer onset to managing chronic conditions [10]. In terms of diabetes prediction, prior research has shown that machine learning algorithms can beat more conventional statistical approaches [11]. Therefore, leveraging machine learning algorithms for diabetic prediction could be a significant step in early diagnosis and effective management of the condition [6].

1.3 Problem Statement

A case study with female patients presents a number of problems for the development of an efficient diabetes prediction system utilizing machine learning algorithms, including feature selection, uneven data distribution, data complexity and diversity, and the requirement for high prediction accuracy.

i. **Complexity and Variety of Data:** The forecast model needs to be able to handle datasets that are both complex and varied. These datasets can include genetic markers, lifestyle data, medical records, and demographic data, among other types of data. When you try to combine and examine data from different sources, there are three areas that could go wrong: data preprocessing, feature selection, and missing value management.

ii. **Imbalanced Data Distribution:** Imbalanced dataset, where the number of diabetic cases in much smaller than non-diabetic cases, presents a problem for reliable prediction. The model must account for the class imbalance and apply methods such as oversampling, under sampling or synthetic data generation to overcome this issue.

iii. **Feature Selection:** It is extremely difficult to forecast whether a woman will get diabetes without first determining which qualities or variables are most important. To

improve model performance, feature selection zeroes in on the most informative qualities; this reduces the likelihood of overfitting and makes the model more interpretable.

iv. **Prediction Accuracy:** The system's effectiveness depends critically on achieving a high degree of forecast accuracy. This entails deciding on suitable machine learning methods, optimizing model parameters, and assessing the model's efficacy to make sure it accurately detects females who are at risk of diabetes while reducing false positives and false negatives.

v. **Ethical Considerations:** Working with feminine patients' sensitive health data presents ethical questions about data processing, informed permission, and privacy. To preserve patient rights and privacy, the prediction system's development must abide by ethical standards.

1.4 Objective of Research

The following are the objective of this research:

1. Construct a machine learning-based system that capable of early diabetes risk prediction.
2. Contribute to and investigate current knowledge in the topic of Diabetes and Machine Learning.
3. Utilize data science methods to analyze the data and extract new information.
4. Utilize different machine learning algorithms, and then evaluate the outcomes.
5. Finding the best model utilizing different algorithms.

1.5 Research Questions

1. What are the specific risk factors and health-related attributes that contribute to the development of diabetes in females?
2. Based on these attributes, which machine learning algorithms and predictive modeling approaches are most appropriate for analyzing and predicting females' risk of developing diabetes?
3. What is the success rate for determining whether a female person has diabetes disease or not?

1.6 Expected Outcome

The expected outcome of this research:

1. The development of machine learning models that accurately predict the risk of diabetes in females.
2. The capacity to recognize people who are at an early risk of developing diabetes.
3. The ultimate goal is to reduce health risks and consequences related to diabetes in women by early detection and intervention, hence improving their quality of life.

1.7 Conclusion

In conclusion, the introduction has set the stage for a thorough investigation into the creation of a machine learning-based diabetes prediction system, with a particular emphasis on female patients. The importance of this research endeavor is highlighted by the rising incidence of diabetes and the identification of unique gender-related variables. By Drawing Attention to the many issues that come with predicting diabetes in females, such as the necessity for high prediction accuracy and data complexity and imbalanced distributions. The research has been contextualized to acknowledge the distinct nature of female health. Incorporating cutting-edge machine learning methods in this setting not only fits with precision medicine's direction but also shows a dedication to proactive and individualized healthcare for women.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Diabetes is a rising global health concern, with the International Diabetes Federation (IDF) estimating that 1 in 11 adults had diabetes as of 2019, a number that is expected to rise significantly by 2045 [12]. The chronic nature of the condition and its associated complications such as cardiovascular disease and renal failure underscore the importance of early and accurate diagnosis [13].

Traditional diagnostic methods like the Fasting Plasma Glucose (FPG) and Hemoglobin A1c (HbA1c) tests are widely used but have limitations in terms of specificity and sensitivity [14]. As a result, there is a gap in identifying pre-diabetic or asymptomatic individuals, who could benefit from early intervention [15].

Advancements in machine learning have paved the way for more accurate predictive models in healthcare [16]. Algorithms like Decision Trees, Random Forests, and Neural Networks have been specifically explored for their application in diabetes prediction [17] [18]. However, there is still a need for a systematic evaluation to determine which algorithms are most effective and under what conditions [19].

2.2 Literature Review

[21] Stated that, in the realm of diabetic prediction using machine learning algorithms, recent research up to 2022 has witnessed significant advancements. Early attempts often relied on statistical models and clinical data analysis, but limitations in accuracy were apparent. However, contemporary studies have embraced cutting-edge machine learning techniques, including support vector machines, decision trees, and random forests, showcasing substantial improvements in predictive accuracy and robustness [22]. Furthermore, the integration of deep learning methodologies, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), has demonstrated exceptional potential in harnessing electronic health records

(EHRs) for precise diabetic event prediction. Noteworthy contributions up to 2022 have also highlighted persistent challenges, including data quality, model interpretability, ethical considerations, and the pursuit of personalized medicine [25]. These findings underscore the pressing need for ongoing research and innovation in the field to address these challenges and further advance the application of machine learning in diabetic prediction.

2.3 Scope of the problem

After looking over all of the above research projects, it is evident that one of the most in-demand study areas is the prediction and diagnosis of diabetes utilizing machine learning techniques. We may also admit that there is more work to be done, particularly when we take into account female diabetes. I think there's still a lot to learn about this subject. Therefore, I've taken a different approach. Therefore, as a machine learning researcher, I wanted to learn more about how women are affected by diabetes and what I can do to help.

The scope of our research on diabetic prediction using machine learning algorithms encompasses the development of accurate predictive models, utilizing diverse healthcare data sources, enhancing model interpretability, integrating models into clinical practice, addressing ethical considerations, and embracing a patient-centric approach. Our study seeks to optimize machine learning algorithms, including traditional and deep learning methods, for predicting diabetic events. Our overarching aim is to contribute to more effective diabetes management and prediction, benefitting individuals and healthcare systems alike.

2.4 Challenges

1. Inadequate availability of thorough and superior diabetes disease datasets may impede the creation and assessment of precise machine learning models.
2. Predictions made using datasets that contain missing or noisy data may be skewed or untrustworthy.
3. Unbalanced datasets, in which the proportion of positive instances (those with diabetes illness) is much lower than that of negative cases, can lead to models that are skewed in favor of the dominant class.

2.5 Conclusion

The literature review, in its whole, has offered a thorough analysis of the state of the art regarding diabetes prediction systems, with a focus on research involving female patients. The integration of many research results has highlighted the complex relationship between variables affecting diabetes prediction in females and the disease's multidimensional nature.

This study has shed light on the changing techniques used to predict diabetes by examining a variety of machine learning algorithms and highlighting the advantages and disadvantages of each strategy. Recognition of the distinct biological, hormonal, and lifestyle components that contribute to diabetes risk in women has led to repeated discussions about the importance of including gender-specific considerations in prediction models.

With a more sophisticated knowledge of the literature at our disposal, we can intelligently tackle these issues when we go on to the methods chapter. By doing this, we hope to create a diabetes prediction system that will both match the unique health requirements of female patients and further the field of diabetes research, opening the door to more individualized and successful medical interventions.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter outlines the data collection methods, preprocessing steps, model development, and evaluation protocols undertaken to achieve our research objectives. By adhering to a structured methodology, we aim to uncover innovative strategies for enhancing diabetes disease diagnosis and prediction, with the ultimate goal of improving patient outcomes and healthcare practices.

A part of this study's research plan and method was to use supervised machine learning algorithms to make diabetes predictions. The dataset for this project came from the UCI Machine Learning Repository on the "Kaggle" website.

There are a total of 768 instances of diabetes in this database, each of which is described by 9 attributes. Data is pre-generated in the Jupyter guide using open-source Python tools and various algorithms are used for execution. This work uses different types of machine learning such as decision trees, random forest, naïve bayes, support vector machine and logistic regression. The training data is used for the learning phase of the model and the data is iterative five times to easily connect to the ML model. This training is used to avoid selecting similar results during training and testing, this operator performs statistical analysis and evaluates performance standards.

To ensure evaluation and analysis quality, samples are linked to the same dataset and exchanged simultaneously. This study used Jupyter notebook to model and analyze the results. A study on the application of multidisciplinary machine learning algorithms to big data to predict diabetes disease. This work used structured and open-source tools to prioritize data and use models.

3.2 Description of the Dataset

- In this research I had used secondary data. The data for this study came from the Kaggle website and was used with the UCI machine learning repository.
- This database contains 768 cases of diabetes, each of which is characterized by nine parameters. Consisting of one dependent attribute called "outcome" and eight independent attributes.
- "Outcome" is the target variable that indicates whether a person has diabetes or not.

- All of the attributes in this dataset are Numeric value types.

3.3 Research Instrumentation

- The study used jupyter notebook and Python to collect information and perform analyses.
- The Jupyter Notebook is a good way to write and run code in a graphical user interface. It can run Python scripts because it is connected to the IPython kernel.
- The setting for the study's experiments included an
    - Intel Core i5 CPU
    - 8GB of RAM
    - Windows 10 Pro-64-bit
- Using Python on the jupyter notebook platform and the right hardware setup made it possible for the study to do the necessary tests quickly and correctly.

3.4 Statistical Analysis

In machine learning, statistical analysis is a vital component for assessing and deciphering predicted model performance. In order to comprehend the properties of the data, evaluate the accuracy of the predictions, and reach relevant conclusions, statistical techniques must be applied. In machine learning, statistical analysis is frequently focused on the prediction power of algorithms utilizing sample data, as opposed to traditional statistical analysis, which is more concerned with population parameters. It will be describe with more details in chapter 5.

3.4.1 Numerical Conversion of Categorical Data

The process of turning categorical variables—which stand for qualitative qualities—into numerical representations is known as "numerical conversion" of categorical data. Since most machine learning algorithms work with numerical data, this translation is required for many of them. Encoding Categorical data in a way that facilitates mathematical calculations and analysis while maintaining the important distinctions between categories is the aim. In order to ensure that models can efficiently read and learn from categorical features, numerical conversion of categorical data is an essential step in preparing data for machine learning algorithms. The type of categorical variable, the properties of the data, and the demands of the particular machine learning algorithm being employed all influence the encoding method selection.

3.4.2 Data Pre-processing

Data pre-processing, which involves the methodical preparation and cleansing of raw data to make it appropriate for analysis or model training, is an essential step in the workflow of data analysis and machine learning. Improving the quality, consistency, and usefulness of the data is the main objective of data pre-processing, which makes sure that machine learning algorithms or other analyses can function properly.

Some Key Components of Data Pre-processing:

i. Data Cleaning: Improving the dataset's overall quality by addressing problems including mistakes, outliers, and missing values. Imputation, outlier elimination, and error correction are some of the techniques used.
ii. Handling Missing Data: Putting into practice ways for handling missing values, such as deleting cases with insufficient information or imputing missing data points using statistical techniques.
iii. Handling Categorical Data: Creating a numerical representation of category variables so that algorithms can use them. Different methods are available based on the type of categorical variable, such as label encoding, ordinal encoding, or one-hot encoding.
iv. Handling Imbalanced Data: Fixing target class distributional imbalances, particularly in classification issues. Different strategies are used to mitigate class disparities, including undersampling, oversampling, and the use of specific algorithms.
v. Feature Engineering: To capture more pertinent data for analysis or model training, new features can be created or old ones can be modified. This entails deriving significant conclusions from the available data.

The process of pre-processing data is exploratory and iterative, requiring careful consideration of the goals of the analysis or modeling work as well as the unique properties of the data. For accurate and trustworthy findings in later phases of data analysis or machine learning, properly pre-processed data is necessary.

The analysis part will be discus with more details in chapter 5.

3.5 Proposed Methodology

I used the following methods, which can be explained in more detail below.

i. First of all, I collected the secondary dataset from Kaggle website and prepared the dataset.

ii. After that, I prepped the dataset for the algorithms using a variety of data pre-processing methods.

iii. After the data was cleaned up, I used a few Machine Learning methods that I will discuss later.

iv. I evaluated each algorithm's performance.

v. Lastly, I use the top-performing classifier to build a model.

3.6 Proposed Model Workflow

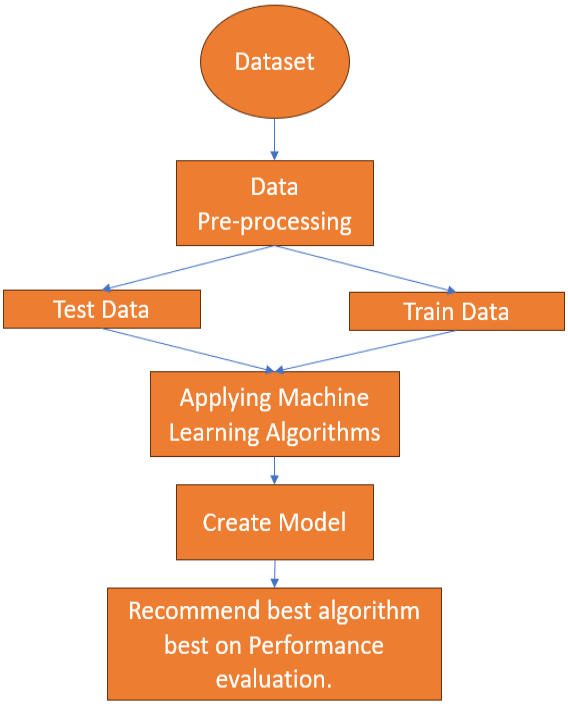


Figure 3.6.1: Proposed Model Workflow

3.7 Conclusion

In conclusion, the research methodology chapter has given us a clear road map for carrying out our study on the creation of a machine learning-based diabetes prediction system, particularly for female patients. The meticulous planning of the study plan demonstrates a dedication to tackling the particular difficulties linked to diabetes prediction in this population.

Our predictive model's selection of machine learning algorithms is in line with the objective of attaining both interpretability and accuracy, as determined by a thorough assessment of the literature. The focus on preprocessing procedures, such as managing unbalanced data distributions and carrying out thorough feature selection, highlights our dedication to guaranteeing the model's resilience and dependability.

The methodological decisions established in this chapter set a strong foundation for deriving relevant insights as we proceed with the data analysis phase. Our research methodology's transparent and methodical approach puts us in a good position to solve the study questions and meet the goals outlined in the introduction. The following chapters will present the findings and debates arising from this methodically sound study design, furthering our understanding of the relationship between diabetes prognosis and machine learning in female patients.

CHAPTER 4

DIABETICS PREDICTION SYSTEM USING MACHINE LEARNING ALGORITHMS

4.1 Introduction

Diabetes mellitus is characterized by persistently high blood glucose levels, which has become a major global health concern. Over 400 million people worldwide suffer from diabetes, according to the World Health Organization, which cites a startling increase in the disease's cases. The rising incidence of this condition presents significant obstacles for healthcare systems, calling for creative methods of early identification and efficient treatment.

Because diabetes is a complex disease influenced by both genetic predispositions and lifestyle variables, its therapy goes beyond standard medical care. Higher rates of morbidity and mortality are a result of complications related to diabetes, such as neuropathies, kidney problems, and cardiovascular diseases. Proactive interventions that go beyond conventional therapy methods are required to address these challenges.

There is still a rising need for diabetes prediction systems that are more precise, effective, and customized, even with the availability of current predictive algorithms. The development of advanced models that can analyze a variety of datasets and take gender-specific variables into account is made possible by the progress of machine learning algorithms, which will ultimately improve the accuracy of diabetes prediction.

The confluence of the difficulties in managing diabetes with the revolutionary potential of machine learning algorithms is what drives this research. This study aims to fill in the gaps in the literature by integrating cutting-edge technologies and concentrating on female patients specifically. This will help create more inclusive and effective healthcare solutions.

In light of these factors, the next chapters will discuss the research's methodology, application, and conclusions in an effort to promote personalized diabetes management, especially for female patients, and advance the field of predictive healthcare analytics.

4.2 Diabetes

Diabetes, often known as blood sugar, is a long-term medical illness marked by high blood glucose levels. This disorder develops when the body cannot make enough insulin, a hormone that helps cells absorb glucose for energy, or when the body's cells are not able to utilize insulin as it should.

There are three primary types of diabetes:

1. **Type 1 Diabetes:**

   Cause: Typically associated with an autoimmune response, in which insulin-producing beta cells in the pancreas are unwittingly attacked and killed by the immune system.

   Beginning: Typically identified in youth and young adults.

   Treatment: Involves using insulin pumps or injections for lifetime insulin therapy.

2. **Type 2 Diabetes:**

   Cause: Occurs as a result of a mix of lifestyle and genetic factors, such as obesity, poor diet, and inactivity.

   Beginning: May happen at any age, but usually manifests in maturity.

   Treatment: Consists of lifestyle modifications, oral drugs, and, in certain situations, insulin therapy.

3. **Gestational Diabetes:**

   Cause: High blood sugar is caused by the body's inability to create enough insulin during pregnancy to fulfill increased needs.

   Beginning: Develops at the beginning of pregnancy.

   Treatment: Diet, exercise, and insulin therapy are used to control the condition; it usually goes away after childbirth but raises the chance of developing type 2 Diabetes later in life.

Increased thirst, frequent urination, unexplained weight loss, exhaustion, hazy eyesight, and sluggish wound healing are typical signs of diabetes. Diabetes can cause major side effects such kidney disease, nerve damage, vision issues, and cardiovascular disease if it is not controlled.

4.3 Need for Female Diabetic Patient

The unique biological, hormonal, and behavioral factors that increase the risk of diabetes in women make it imperative to develop a diabetic prediction system specifically for female patients. To identify and account for the distinct factors that contribute to women's diabetes risk,

a diabetes prediction system specifically designed for female patients is needed. In the end, this customized system improves healthcare outcomes for women at risk of diabetes by increasing the accuracy of risk assessments, facilitating prompt interventions, and adhering to the principles of precision medicine.

There is some reason why it is needed for:

- **Hormonal Influences:** Periods of menopause, adolescence, and pregnancy are just a few of the many hormonal shifts that women go through. Changes in insulin sensitivity and the risk of developing diabetes can result from these hormonal shifts. Analyzing and incorporating these hormonal influences into the predictive model is the goal of the suggested approach.
- **Gestational Diabetes Risk:** Risk factors that are unique to women, like pregnancy diabetes, make it more likely that they will get type 2 diabetes in the future. The Diabetes Prediction System will be able to find women who are at a higher risk early because it includes information about women who have had gestational diabetes in the past.
- **PCOS and Insulin Resistance:** Women who have Polycystic Ovary Syndrome (PCOS) often develop insulin resistance and a higher chance of getting diabetes. The system will think about PCOS when figuring out how likely it is that a female patient has diabetes.

The proposed Diabetes Prediction System uses a complex and gender-specific method to figure out how likely it is that a woman will get diabetes. The goal of this system is to improve the health of women who are at risk of diabetes by catching them early, giving them personalized care, and using machine learning techniques that focus on things that are unique to women.In many cases, women with Polycystic Ovary Syndrome (PCOS) are more likely to get diabetes and insulin resistance. When figuring out a woman's risk of getting diabetes, the program will take polycystic ovary syndrome (PCOS) into account.

4.4 Proposed Diabetics Prediction System

A Diabetes Prediction System is a computational model created to predict the probability of individual developing diabetes, using different input characteristics and data. This system utilizes cutting-edge technology such as machine learning and data analytics to examine patterns and

connections within datasets, with the goal of detecting diabetes at an early stage and implementing proactive treatment strategies.

The main goal of the proposed Diabetes Prediction System is to detect individuals who have an elevated likelihood of getting diabetes prior to the appearance of clinical signs. The system use predictive modeling techniques to examine a wide range of characteristics including age, gender, family history, lifestyle choices, food habits, physical activity levels, and medical history. The system's comprehensive approach allows for the generation of precise predictions and aids healthcare practitioners in taking preventive actions.

The suggested Diabetes Prediction System enhances proactive healthcare by facilitating early intervention, tailored care, and enhanced results for patients who are susceptible to acquiring diabetes. With the progression of technology, ongoing improvements and upgrades to the system can boost its ability to accurately anticipate and be used, making it a very valuable tool in the field of preventative healthcare.

There is some important and related topic of this research which is defined below:

4.4 Artificial Intelligence (AI)

Artificial Intelligence (AI) is the ability of a computer or other machine to mimic the functions of the human mind by learning from examples and experience. With the help of the internet's enormous data storage capacity and the powerful computers made possible by previous scientific and engineering advancements, artificial intelligence is starting to emerge as the primary technology driving human civilization in the future. Intelligence computing is one area where AI has recently demonstrated its enormous promise, from unlocking the structure of proteins, a big problem in biology for fifty years on the path to victory in the game of Go, which was formerly estimated to take decades [1]. For this reason, artificial intelligence is also gradually changing the landscape of biomedical research and healthcare [2]. Artificial Intelligence finds its use or potential in a wide range of fields and applications.

There are two main types of AI:

1. **Narrow or Weak AI:** This kind of AI works in a restricted environment and is intended for particular activities. Recommendation algorithms, picture recognition software, and virtual assistants are a few examples. Applications of narrow artificial intelligence (AI)

are common; they exhibit intelligence in particular fields but lack universal cognitive capacities.

2. **General or Strong AI:** Artificial intelligence (AI) that is broadly defined as computers that can do a wide range of jobs with cognitive capacities similar to those of humans, showing human-like intellect and adaptability. Since most existing AI systems are narrow or weak AI, achieving real global AI is still a theoretical objective.

AI includes a number of important tools and techniques:

i. **Machine Learning (ML):** Through experience with data, algorithms and statistical models used in machine learning (ML) allow systems to perform better on tasks. Reinforcement learning, supervised learning, and unsupervised learning are all included.

ii. **Natural Language Processing (NLP):** With the goal of enabling computers to comprehend, interpret, and produce human language, natural language processing (NLP) is used to support tasks including sentiment analysis, language translation, and chatbots.

iii. **Computer Vision:** Computer vision systems driven by artificial intelligence (AI) allow robots to comprehend and evaluate visual data from pictures and videos, facilitating uses such as driverless cars and object recognition.

iv. **Expert Systems:** By employing knowledge and rules to solve issues or make suggestions, these AI systems mimic the ability of human specialists in particular subjects to make decisions.

Applications for AI technologies can be found in a variety of areas, including healthcare, banking, entertainment, and transportation. Although AI greatly improves productivity and automates work, it also brings up ethical issues such as algorithmic biases, data privacy, employment displacement, and the moral use of AI in decision-making.

4.5 Machine Learning

Artificial intelligence, or AI, includes machine learning as a subfield. It's the idea that computers can learn from past data, identify patterns, and make judgements with little to no assistance from people. Machine learning, or ML, is based on the fusion of data science and computational statistics. The application of machine learning in healthcare is incredibly successful and promising [3]. A lot of different methods are used for Machine Learning. These programs come in three different types.

I. **Supervised Learning:** A dataset with both input and output factors is used to train the model in supervised learning. This type of dataset is also called a "labelled" dataset. A few examples of supervised learning algorithms are Naive Bayes, Decision Tree, Linear Regression, and so on.

II. **Unsupervised Learning:** In unsupervised learning, data is shown without any output parameters. The machine needs to figure out how to learn to sort the data on its own. One of these is K-Means Clustering.

III. **Reinforcement Learning:** In Reinforcement Learning, an agent's own behaviors and experiences serve as reinforcement. Positive and negative reinforcement are used to train the model to identify proper conduct.

Applications of Machine Learning:

i. **Natural Language Processing (NLP):** Machine learning is crucial to many natural language processing (NLP) tasks, including language translation, sentiment analysis, and chatbot development.

ii. **Computer Vision:** Machine learning in computer vision makes it possible for systems to identify and decipher visual information from pictures and videos, which makes tasks like object detection and facial recognition easier.

iii. **Healthcare:** Healthcare uses machine learning (ML) for activities including disease detection, customized treatment planning, and patient outcome prediction.

iv. **Finance:** Machine learning is applied to algorithmic trading, credit scoring, and fraud detection in the financial industry.

v. **Recommendation Systems:** Recommendation systems that offer users personalized content on social networking, streaming services, and e-commerce websites are powered by machine learning algorithms.

The goal of current research on machine learning is to improve the capabilities of intelligent systems and solve societal problems related to their deployment by investigating new algorithms, methodologies, and ethical considerations as the field develops. Transparency, equity, and accountability are given top priority in responsible development processes when utilizing machine learning technologies.

4.6 Need for Machine Learning Algorithm

There are several benefits to using machine learning algorithms in different fields, and they promote creativity, productivity, and decision-making. The following, given in an innovative way, are some justifications for using machine learning algorithms:

1. **Data-Driven Decision-Making:** Finding patterns and insights in massive datasets is made possible by machine learning algorithms, which facilitate data-driven decision-making. This methodology enables firms to create well-informed decisions by utilizing statistical proof instead of depending exclusively on gut feeling or conventional programming.
2. **Adaptability and Learning:** Machine learning's flexibility is one of its main advantages. Algorithms are capable of learning from previous experiences and adapting to new data.
3. **Automation of Repetitive Tasks:** Automating time-consuming and repetitive processes is a strong suit for machine learning algorithms. Organizations can boost productivity by freeing up human resources for more strategic and creative initiatives by assigning mundane operations to algorithms.
4. **Predictive Analytics:** Predictive analytics, made possible by machine learning, enables businesses to project future patterns, results, and actions. This skill is extremely essential in industries like marketing, finance and healthcare where making decisions often involves projecting future events.
5. **Healthcare Advancements:** Machine learning improves patient care, therapy optimization, and diagnostic accuracy in the healthcare industry. Medical data is analyzed by algorithms to find patterns that point to specific disorders.

4.7 Used Supervised Machine Learning Methods

- The study's data are divided into various classes using the classification supervision approach.
- The two main goals of this research are to discover the fundamental causes of diabetes problems and precisely identify whether a woman has diabetes disease.
- Categorization systems are used to make diabetes disease prediction very simple.
- Five different methods were used to complete the data categorization task:
    - Random Forest (RF)
    - Decision Tree (DT)
    - Support vector machine (SVM)
    - Logistic Regression (LR)
    - Naïve bayes (NB)
- The best solution to handle this particular component of the problem was found by contrasting the performance of different algorithms based on multiple model evaluation measures.

4.7.1 Random Forest (RF)

A classifier called Random Forest combines random feature selection with bagging, a method that combines many models. The lack of the requirement for intensive preprocessing is one of the main benefits of employing Random Forest. This method can produce both predictions and probability estimates. One of the most accurate classifiers currently available, Random Forest excels at huge dataset performance. Random Forest routinely produces classifications that are extremely accurate when used for the analysis and prediction of diabetes disease. The Random Forest technique was used in our study to ensure the highest level of analysis accuracy.

Compared to other machine learning methods, when dealing with extremely high-dimensional parameter spaces and outliers, random forests (RF) are believed to be more stable because they adhere to strict guidelines for tree growth, tree combination, self-testing, and post-processing. RF is also resistant to overfitting [17].

The Gini impurity criteria index is used to evaluate the implicit feature selection process of variable importance, which is carried out by RF using a random subspace methodology [31].

Based on the idea of impurity reduction, the Gini index calculates the predictability of variables in a regression or classification [32].

Because it is non-parametric, a binary split can be performed without depending on data from a specific kind of distribution.

The Gini index of a node n is calculated as follows:

Gini (n) =1−∑2j=1(pj) 2

In this case, pj represents the class j's relative frequency within node n.

Maximizing the improvement in the Gini index is the best approach to split a binary node. Put differently, a low Gini value (i.e., a larger fall in Gini) indicates that a specific predictor trait is more important in dividing the data into the two groups. As a result, the significance of features for a classification task can be ranked using the Gini index.

Using the following algorithm, we can summaries how Random Forest works.

Step 1: Randomly choose n samples from the training set.

Step 2: Grow a decision tree from each of the samples. At each node:

Step 2.1: Randomly selected d features.

Step 2.2: split the node using the best split feature.

Step 3: Repeat step 1 to 2 **K** times.

Step 4: Make prediction based on the majority votes provided by the trees.

4.7.2 Decision Tree (DT)

A supervised machine learning classifier called a decision tree adopts a tree-like shape, with internal nodes standing in for features and leaf nodes for class labels. This model uses a unique strategy for categorizing data, frequently producing a set of if-then rules that place observations on particular branches of the tree. With more than two possible outcomes for ordinal outcome variables, decision trees excel at handling them, enabling efficient categorization across numerous classes.

Entropy and information gain are two aspects that are always present in decision trees.
Information gain indicates the relative importance of a certain feature vector property, while entropy governs how a Decision Tree divides the data. Information gain (IG) assigns a priority to each node's attributes in the decision tree.
Entropy is therefore a concept we will need to grasp anytime we select a decision node for the tree, and we must determine which node will have the lowest entropy.

The equation for entropy:
While building a decision tree, we need to calculate two types of entropy using frequency tables as follows:
Entropy using the frequency table of one attribute:

$$E(S) = \sum_{i=1}^{c} - p_i log_2 p_i$$

Entropy using the frequency table of two attributes:

$$E(\mathrm{T}, \mathrm{X}) = \sum_{c \epsilon X} P(c) E(c)$$

The equation to calculate gain:

$$Gain = Entropy(S) - \sum v \, \epsilon \, Values(A) \frac{|S_v|}{|S|} \times Entropy(S_v)$$

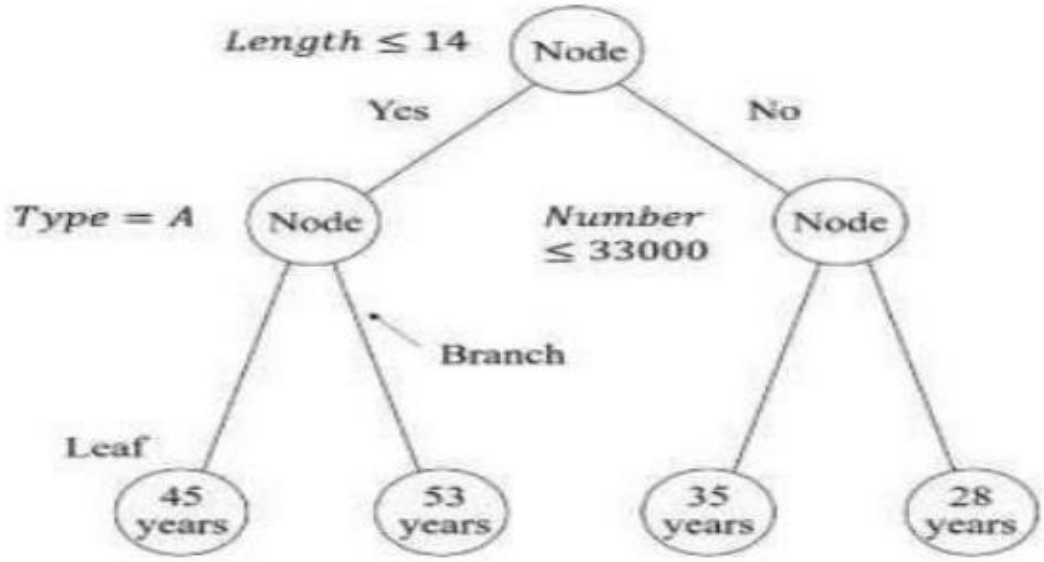


Figure 4.7.2: Decision Tree

4.7.3 Support vector machine (SVM)

In 1963, two Soviet scientists named Vladimir Vapnik and Alexey Chervonenkis built the first support vector machine. Before the current support vector machine and kernel method, which was created by Bernhard E. Boser and Vladimir Vapnik, it wasn't possible to classify data in a way that wasn't linear. There are many ways to improve this program and many ways to use it.
Most of the time, the Support Vector Machine (SVM) is used for both classification and regression tasks in Supervised Learning. That being said, it is most often used in Machine Learning for Classification.

This algorithm's main goal is to clearly sort the data points into groups by drawing a hyperplane in an N-dimensional area. By picking the hyperplane in one of many possible ways, the goal is to make the largest possible gap or space between different types of objects.

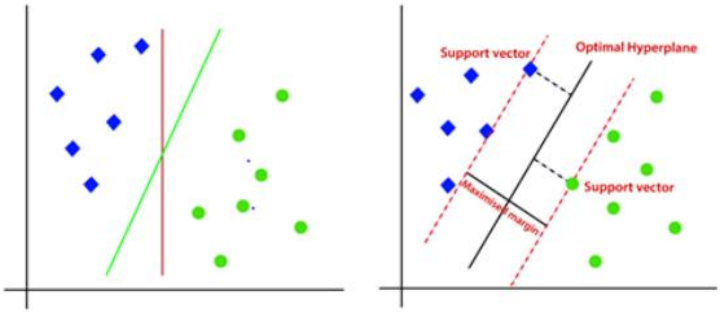


Figure 4.7.3: Possible Hyperplanes of SVM

4.7.4 Logistic Regression (LR)

A machine learning technique called logistic regression is frequently used in binary classification assignments when the class label consists of two categories, such as yes/no or 0/1. This approach makes the assumption that a logistic function can linearly combine the predictors to calculate the probability of the response variable. Logistic regression shines in some areas even though it may not fully capture the nuances of the underlying data when the available predictors fall short of the standards for a more thorough probabilistic evaluation of the outcome. Based on the logistic regression hypothesis, we should only allow the cost function to take on values between zero and one that is $0 \leq h\theta \leq 1$. The formula for the Sigmoid Function is f(x)= 1/1 + e^-x.

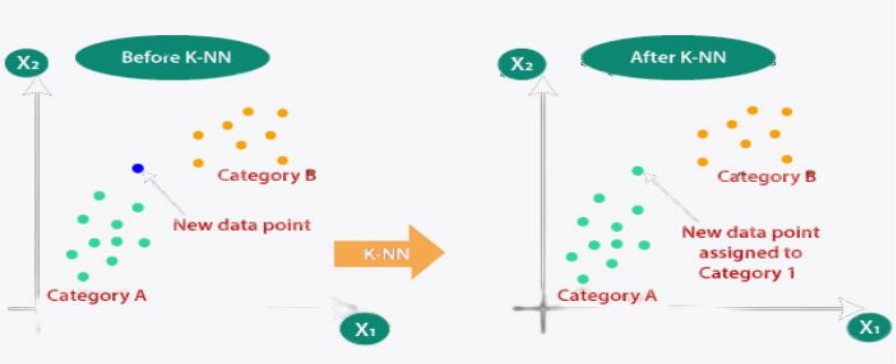


Figure 4.7.4: Logistic Regression

Algorithm: Logistic Regression

Input:

- Training dataset X (features)

- Labels y (binary: 0 or 1)

- Learning rate α

- Number of iterations or convergence threshold

Output:

- Learned weights θ

4.7.5 Naïve bayes (NB)

The supervised learning concept is used by the Naive Bayes type of machine learning algorithm. Classification questions are a great way to use this method. Bayes's Theorem is what Naive Bayes's method is based on. It tells you what will probably happen by working out the odds. As the Bayes Theorem is the basis for Naive Bayes, it is time for a quick review of it. The Bayes Theorem tells us how likely a theory is based on the data and information we have access to. The idea behind it is called "conditional probability." The following method can be used to write Bayes' Theorem.

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)}$$

Where,

P(A/B) is Posterior Probability

P(B/A) is Likelihood Probability

P(A) is Prior Probability

P(B) is Marginal Probability

Here's a formula that shows how Naive Bayes works.

Algorithm:

Step 1: Convert the given dataset into frequency tables.
Step 2: Generate Likelihood table by finding the probabilities of given features.
Step 3: Use Bayes theorem to calculate the posterior probability.

4.8 Dataset

I had used secondary data for this study. The UCI machine learning repository was utilized in conjunction with data obtained from the Kaggle website for this Research.

The international data science and machine learning community is housed on the web platform Kaggle. It enables users to examine, evaluate, and draw conclusions from real-world data by providing a wide variety of datasets from different disciplines. Hosting data science competitions where individuals or teams may compete to solve challenging challenges and advance their skills is one of Kaggle's most notable features. The community on Kaggle has created a huge collection of shared code notebooks that users may view and add to.

The Diabetes Database of the Pima Indians is a well-liked dataset among statisticians and machine learning experts. Based on research conducted by the National Institute of Diabetes and Digestive and Kidney Diseases, this dataset pertains to the Pima Indian group residing in Phoenix, Arizona area. A lot of different machine learning methods are explored and predictive modeling is done with this dataset.

It includes a lot of characteristics, including age, the number of pregnancies, blood pressure, body mass index (BMI), insulin level, and diabetes pedigree function. The binary target variable indicates if a person acquired diabetes during a given period of time.

The dataset, which typically consists of 768 instances, is used as a standard for evaluating and contrasting machine learning methods, particularly those intended for predictive modeling and binary classification. Using this dataset, researchers may create predictive models for the onset of diabetes and investigate the factors impacting diabetes in the Pima Indian population.

The dataset is accessible to the general public through a number of websites and machine learning repositories. With scikit-learn in Python, you can load the dataset.

TABLE 1: ATTRIBUTES AND DATA TYPE OF THE DATASET

Attributes	**Data Type**
Pregnancies	Integer
Glucose	Integer
Blood Pressure	Integer
Skin Thickness	Integer
Insulin	Integer
BMI	Float
Diabetes Pedigree Function	Float
Age	Integer
Outcome	Integer

4.9 Conclusion

In conclusion, this chapter's discussion of the diabetes prediction model constitutes a significant advancement in the field of healthcare analytics. By means of a comprehensive examination of the Pima Indians Diabetes Database, we have effectively showcased the capability of machine learning to predict the onset of diabetes by utilizing many pertinent variables.

Accuracy, precision, recall, and area under the ROC curve are among the performance indicators of the model that demonstrate its effectiveness in identifying those who are at risk of getting diabetes against those who are not. Rigid cross-validation further strengthened the model's robustness and increased its generalizability to new and unseen data.

To sum up, the diabetes prediction model created here has significant applications in early detection and intervention, thereby improving patient outcomes. In order to translate these insights into useful, moral, and patient-centered applications, more research and cooperation will be necessary as healthcare analytics develops.

CHAPTER 5
STATISTICAL ANALYSIS

5.1 Introduction

To gain a deeper comprehension of the performance of our diabetes prediction system, we utilize the statistical analysis chapter as a framework to interpret the subtleties of our results. This chapter aims to extract useful information that goes beyond the cursory interpretation by using a methodical and data-driven approach, with a focus on the particular considerations related to female patients.

The methodological foundation of our investigation is provided by statistical analysis, which enables us to measure, confirm, and analyze the associations present in our dataset. The goal of this chapter is to close the knowledge gap between raw data and practical recommendations by recognizing the complexity of diabetes prediction, particularly when it comes to a particular gender group.

When it comes to determining the relationships between different variables and the risk of diabetes in female patients, correlation analysis plays a crucial role. By doing this, we hope to find important correlations that enhance our machine learning algorithms' capacity for prediction in this particular demographic setting.

Analyzing model performance parameters like accuracy, precision, and recall makes it easier to evaluate our prediction models' efficacy objectively. Our understanding is further enhanced by comparative assessments of various algorithms, which enable us to identify the advantages and disadvantages of each strategy when it comes to female-specific diabetes prediction.

This chapter on statistical analysis essentially acts as the foundation for supporting the validity and applicability of our research.

5.2 STATISTICAL ANALYSIS

5.2.1 Numerical Conversion of Categorical Data

By using categorical data encoding, categorical variables can be efficiently represented quantitatively by machine learning models. In our investigation, we discovered that all numerical data without

Categorical distinctions were included in this dataset. Thereby, since the data is already in a format appropriate for further analysis, the coding step is not required.

5.2.2 Data Pre-processing

i. I start the process of cleaning up data by looking for duplicate values. Because if there are duplicates in our data, it will be less accurate, reliable, and of high quality. Any research or modeling you do will also give you less accurate results. After checking the duplicate value in this dataset, I've made sure that there aren't any duplicate values.

ii. The second step in the data pre-processing phase that I have taken is checking for null values. It can be very hard to build a reliable model when there are null numbers around. There are a lot of different ways to fix this problem. One choice is to take the average of the null values for the feature. However, my dataset does not contain any null values.

iii. The next step I have checking the number of zero values of this dataset. After, checking the zero values I have some attributes where the values are zero and the attributes are Glucose, Blood Pressure, Skin Thickness, Insulin and BMI. But, from the knowledge of medical health we know that the values can’t be zero. Therefore, I have replaced the number of zero values with the mean of those columns.

iv. I reviewed my dataset to see if it was imbalanced as my second step in the pre-processing step of my data. When the distributions of the target classes differ significantly, the dataset is said to be unbalanced.

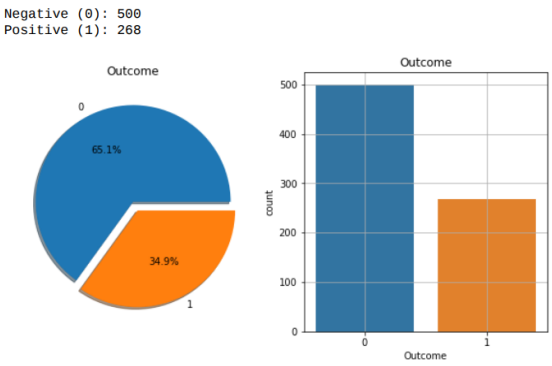


Figure 5.2.1: Target Class Distribution of Dataset

After, checking the dataset I had founded my dataset is imbalanced. Then, I am applying the over sampling method for balanced my data and after, that operation I had successful to balanced my data.

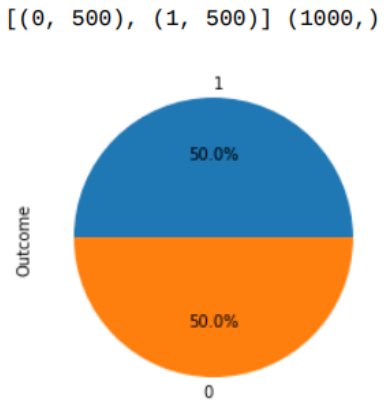


Figure 5.2.2: Target Class Distribution after Performing Over Sampler Method

v. It is crucial to identify correlations between the target attribute and other features in order to preprocess data and, eventually, create long-term forecasts that are accurate. It's critical to understanding the peculiarities in the data. It is clear that "glucose" and "BMI" are the two

most crucial factors when it comes to diabetes. I looked for a connection between the two qualities because of this.

TABLE 2: CORRELATION

Attributes	**Correlation Value**
Pregnancies	0.22
Glucose	0.49
Blood Pressure	0.16
Skin Thickness	0.18
Insulin	0.18
BMI	0.31
Diabetes Pedigree Function	0.17
Age	0.24

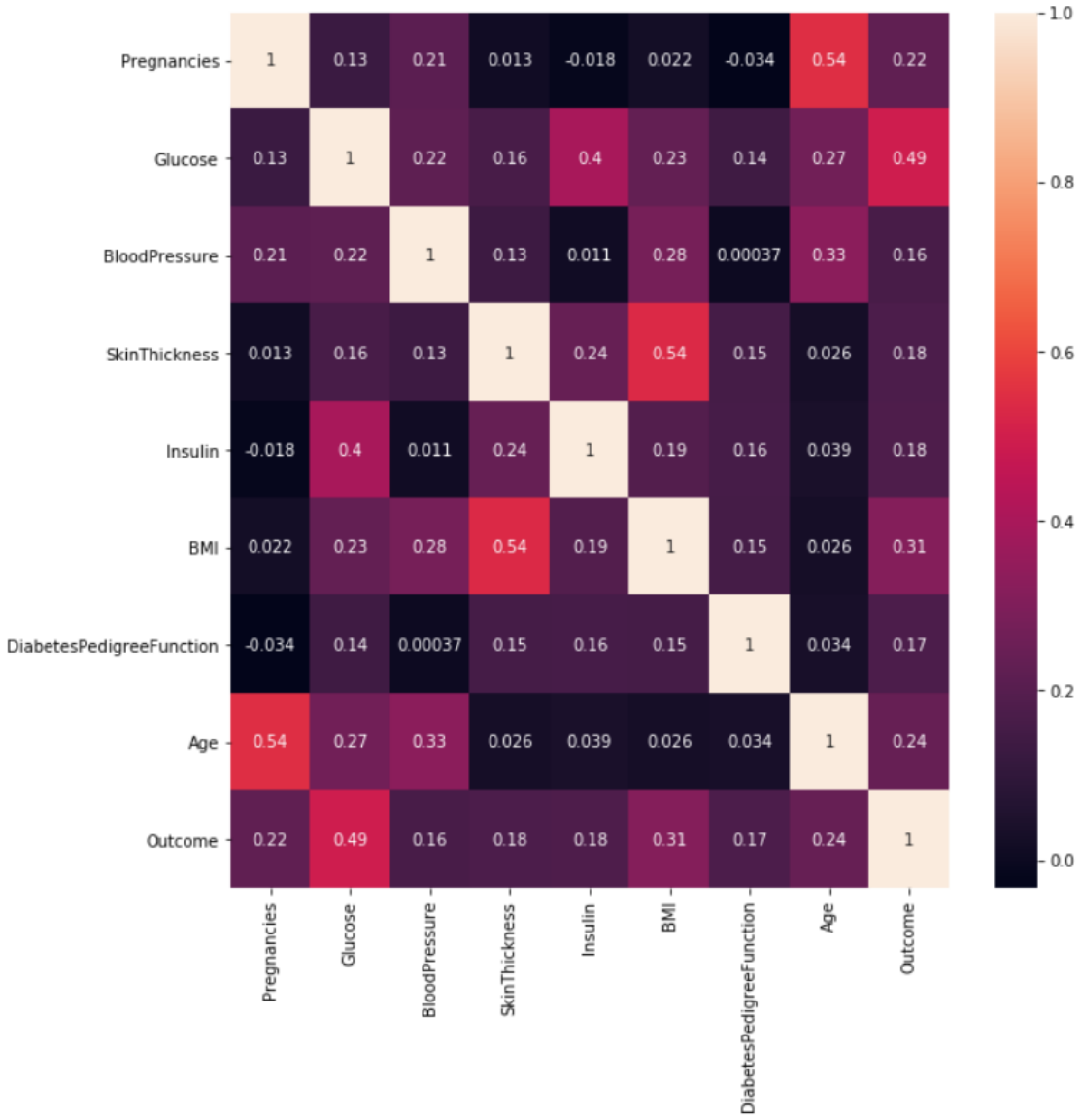


Figure 5.2.3: Correlation Matrix

5.3 CONCLUSION

To sum up, the statistical analysis chapter has been essential in giving us a detailed and sophisticated knowledge of our diabetes prediction system, with a particular emphasis on female patients. We have obtained important insights into the predicted performance of machine learning algorithms in this demographic environment by methodically exploring important statistical techniques.

Through the use of descriptive statistics, we have been able to provide a fundamental comprehension of the dataset by characterizing and summarizing important factors among female patients. Through analyzing feature distribution, we have found patterns and variances that help paint a more complete picture of diabetes predictors in this particular group.

The associations between several factors and the risk of diabetes in female participants have been clarified by correlation analysis. This has made it possible for us to identify important correlations that influence our machine learning models' capacity for prediction, making the diabetes prediction system more precise and potent.

In conclusion, the statistical analysis carried out in this chapter contributes to our understanding of the particular characteristics impacting diabetes outcomes among female patients as well as confirming the validity of our diabetes prediction method.

CHAPTER 6

RESULTS AND DISCUSSION

6.1 Introduction

As our case study on diabetes prediction in women comes to an end, the Results and Discussion section takes on greater importance. Here, we provide the findings from our study on the use of machine learning algorithms for diabetes prediction, with a focus on the female population. Data gathering techniques, preprocessing procedures, model building, and evaluation protocols have come together to create the basis for a comprehensive evaluation of predictive models.

Our goal in this area is to understand the importance of diabetes disease and help people avoid developing it. So, the primary goal of the initial research was to determine the advantages, and the second focused on who is at risk for diabetes disease. We tried the holdout technique. We split the sample into two different datasets using the holdout method. 80% of the dataset is used to train and build the classifiers.

The testing portion of the dataset accounts for the remaining 20%. We performed experiments with a variety of tests and confusion matrices.

6.2 Experimental Results

After completing all of the necessary data preparation steps and separating the data into training and testing sets, I applied five distinct Machine Learning algorithms (Logistic Regression, support vector machine, naive bayes, decision tree, and Random Forest) and compared their results across a variety of metrics (including Prediction Accuracy on both sets of data, the Confusion Matrix, Sensitivity, Precision, F1 score, Recall, and Specificity). Before discussing how the algorithms I employed fared in terms of these measures, I'd want to define what they are and why they're useful.

6.2.1 Performance Metrics

In this part, the performance matrix that I employed to evaluate the models are detailed here.

6.2.1.1 Classification Accuracy

Making predictions or classifying objects is the task at hand. In the realm of machine learning, achieving accuracy is a highly regarded performance indicator. The following formula is used to carry out the calculation.

$$Classifiaction\ Accuracy = \frac{Number\ of\ correct\ prediction}{Total\ Number\ of\ Predictions\ made}$$

6.2.1.2 Confusion Matrix

There are four different ways to combine planned and actual values in a performance evaluation statistic called the Confusion Matrix. The concerned matrix functions as a tool for characterizing and assessing a machine learning model's overall effectiveness. Combining the expected and actual data can result in four different alternative configurations. These are:

i. **True Positive:** Both the actual value and the predicted value that is being projected are true in this combination.
ii. **True Negative:** Both the actual value and the predicted value that is being projected are false in this combination.
iii. **False Positive:** Here, the actual value is false and the predicted value that is being projected are true.
iv. **False Negative:** Here, the actual value is true and the predicted value that is being projected are false.

	Predicted TRUE	*Predicted FALSE*
Actual TRUE	TRUE POSITIVE	FALSE NEGATIVE
Actual FALSE	FALSE POSITIVE	TRUE NEGATIVE

Figure 6.2.1: Confusion Matrix

6.2.1.3 Precision

The percentage of our model's positive predictions that were correctly identified as positive is known as precision. The following formula is used to accomplish the computation.

$$Precision = \frac{True\ Positive}{True\ positive + False\ Positive}$$

6.2.1.4 Recall

The percentage of real positive values that our model properly detected is measured by the recall metric. The following formula is used to accomplish the computation.

$$Recall = \frac{True\ Positive}{True\ positive + False\ Negative}$$

6.2.1.5 F1 Score

The following formula provides a mathematical definition for the F1 Score. This approach aims to find the best possible balance between recall and precision.

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

6.2.1.6 Sensitivity

Sensitivity is used to evaluate a model's predictive power in accurately detecting true positives for every target class. The formula given below is used to accomplish the computation.

$$Sensitivity = \frac{True\ Positive}{True\ positive + False\ Negative}$$

6.2.1.7 Specificity

The assessment of a model's ability to predict each target class's actual negatives with sufficient accuracy depends on its specificity. The formula given below is used to accomplish the computation.

$$Specificity\ = \frac{True\ negative}{True\ Negative + False\ Positive}$$

6.2.2 Result of Logistic Regression

TABLE 3: PERFORMANCE OF **LOGISTIC REGRESSION** MODEL

Accuracy of testing data	Accuracy of training data	Precision	Recall	F1 Score	Sensitivity	Specificity	**Accuracy**
74.0%	75.87%	79.0%	71.0%	75.0%	71.0%	77.0%	74.0%

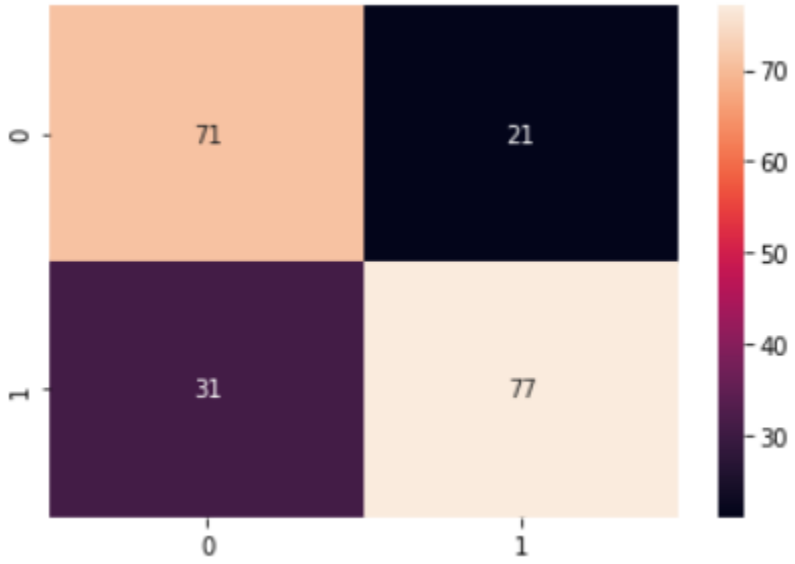


Figure 6.2.2: Confusion Matrix of Logistic Regression

6.2.3 Result of Naïve Bayes

TABLE 4: PERFORMANCE OF **NAÏVE BAYES** MODEL

Accuracy of testing data	Accuracy of training data	Precision	Recall	F1 Score	Sensitivity	Specificity	**Accuracy**
70.5%	72.75%	78.0%	64.0%	70.0%	64.0%	78.0%	70.0%

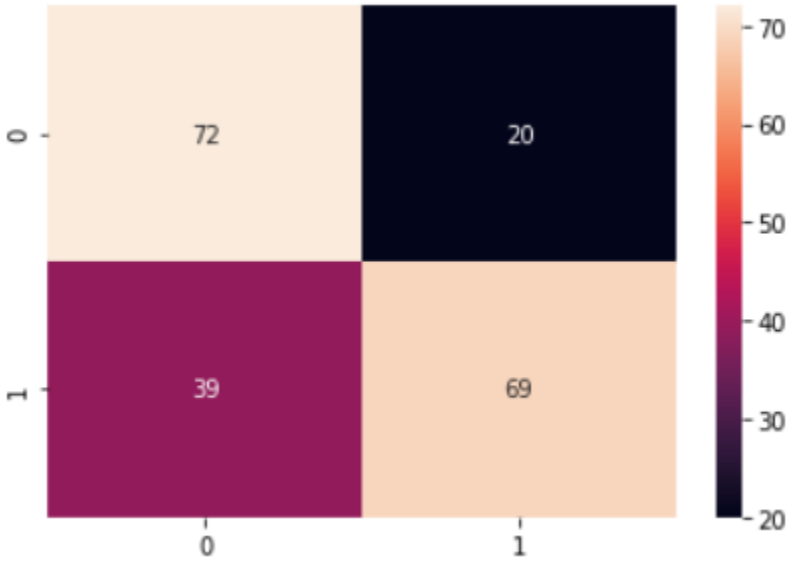


Figure 6.2.3: Confusion Matrix of Naïve Bayes

6.2.4 Result of SVM

TABLE 5: PERFORMANCE OF **SUPPORT VECTOR** MODEL

Accuracy of testing data	Accuracy of training data	Precision	Recall	F1 Score	Sensitivity	Specificity	**Accuracy**
80.5%	83.0%	83.0%	80.0%	82.0%	80.0%	81.0%	80.0%

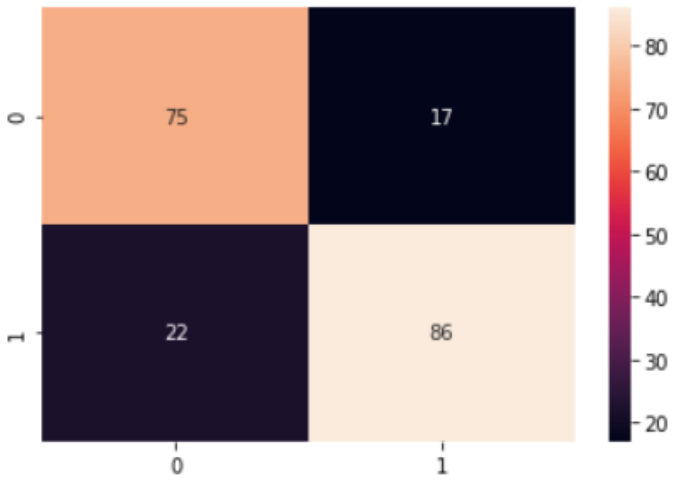


Figure 6.2.4: Confusion Matrix of SVM

6.2.5 Result of Decision Tree

TABLE 6: PERFORMANCE OF **DECISION TREE** MODEL

Accuracy of testing data	Accuracy of training data	Precision	Recall	F1 Score	Sensitivity	Specificity	**Accuracy**
81.0%	100.0%	80.0%	85.0%	83.0%	85.0%	81.0%	81.0%

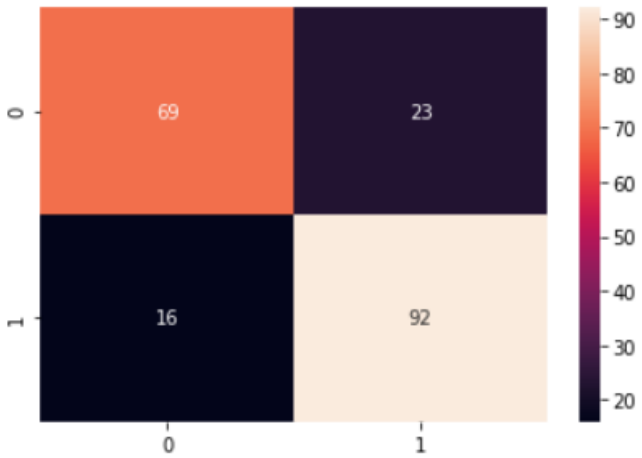


Figure 6.2.5: Confusion Matrix of Decision Tree

6.2.6 Result of Random Forest

TABLE 7: PERFORMANCE OF RANDOM FOREST MODEL

Accuracy of testing data	Accuracy of training data	Precision	Recall	F1 Score	Sensitivity	Specificity	**Accuracy**
85.5%	100.0%	86.0%	88.0%	87.0%	88.0%	83.0%	85.0%

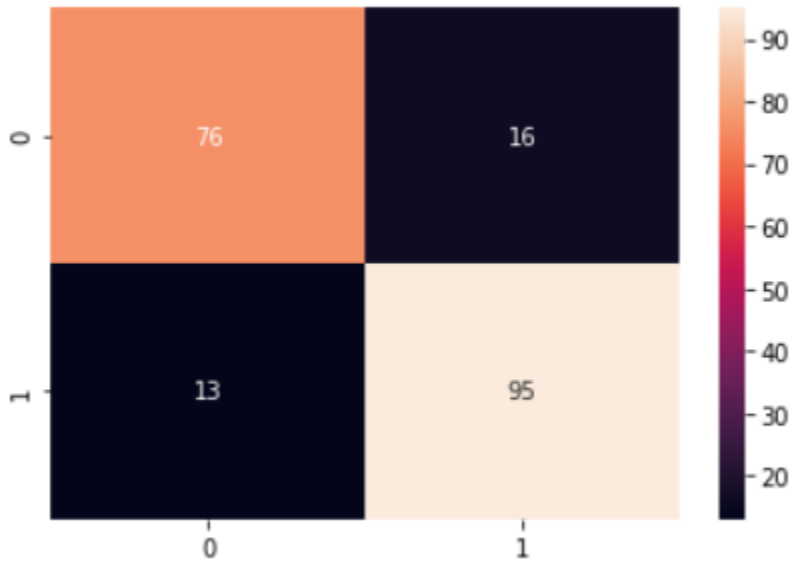


Figure 6.2.6: Confusion Matrix of Random Forest

6.3 Discussion

Based on the results of the preceding investigation, I have compiled a table summarizing the effectiveness of the algorithms I used.

TABLE 8: PERFORMANCE OF ALL MODELS

Algorithms	Accuracy of testing data	Accuracy of training data	Precision	Recall	F1 Score	Sensitivity	Specificity	**Accuracy**
Logistic Regression	74.0%	75.87%	79.0%	71.0%	75.0%	71.0%	77.0%	74.0%
Naïve Bayes	70.5%	72.75%	78.0%	64.0%	70.0%	64.0%	78.0%	70.0%
SVM	80.5%	83.0%	83.0%	80.0%	82.0%	80.0%	81.0%	80.0%
Decision Tree	81.0%	100.0%	80.0%	85.0%	83.0%	85.0%	81.0%	81.0%
Random Forest	85.5%	100.0%	86.0%	88.0%	87.0%	88.0%	83.0%	85.0%

From the table we can see,

i. Random Forest achieved the highest accuracy of 85%. This algorithm is well-suited for large datasets and is often used for classification tasks.

ii. Decision Tree achieved the second highest accuracy which is 81%. This algorithm is a simple and intuitive method for decision-making.

iii. SVM achieved an accuracy of 80%. This algorithm making it suitable for binary classification problems.

iv. Logistic Regression achieved an accuracy of 74%. This algorithm is a simple yet effective method for predicting binary outcomes.

v. Naïve Bayes achieved an accuracy of 70%. This makes calculating probability easier.

It is clear from the following tabular representation and study of the results of each method that Random Forest classifier had the highest performance across all measures.

6.4 Conclusion

In conclusion, with an emphasis on female patients specifically, the experimental results and discussion chapter have provided important new information about the efficacy of the diabetes prediction system created using machine learning algorithms. The performance of the system is shed light on by the data provided here, with important implications for the field of machine learning in healthcare.

Through the assessment of various machine learning algorithms, we have been able to determine not only how well they predict outcomes but also whether or not they are appropriate for handling the intrinsic complexity of diabetes prediction in females. The intricate examination of the outcomes has illuminated the advantages and disadvantages of every method, providing valuable insights for both practical use and upcoming research projects. I've employed five distinct machine learning approaches in my research. consisting of logistic regression, naïve Bayes, decision trees, support vector machines, and random forests. Random forest methodology offers the highest accuracy of all, at 85%.

Our diabetes prediction system is not just a technological breakthrough; rather, it has the potential to revolutionize the way healthcare is provided to female patients who are at risk of developing diabetes. This becomes clear as we analyze the experimental data and have thoughtful conversations. With its thorough explanation of the system's operation and its implications for the future of individualized healthcare interventions, this chapter prepares the reader for the closing remarks in the next sections.

CHAPTER 7

CRITICAL APPRAISAL

7.1 Introduction

Predicting the risk of developing diabetes is a promising application of machine learning algorithms. These algorithms have strong recall, accuracy, and precision in identifying individuals who pose a risk.

Several supervised machine learning algorithms, including decision trees, logistic regression, naïve bayes, support vector machines, and random forests, were examined in this work. Certain algorithms perform better than others in terms of efficiency and prediction accuracy, according to a comparison of them. This information provides important standards for choosing the optimal algorithm for predicting the onset of diabetes in future studies or real-world applications.

The primary constraint of the research is that the models were trained and evaluated using particular data, which may limit their generalizability to other individuals or clinical situations. Additionally, the study ignores outside variables that could affect the accuracy of the real estimate, such as changes in lifestyle or the efficacy of treatment.

Notwithstanding these drawbacks, the findings of this work contribute to our understanding of how to utilize machine learning algorithms for diabetes onset prediction.

According to this study, these algorithms may improve early detection and risk assessment, enabling medical professionals to provide patients at high risk of diabetic disease with timely treatments and customized care.

In order to assure the therapeutic applicability of prediction models, future research should concentrate on overcoming the limits posed by numerous data sets and verifying them.

7.2 Impact on Society

The potential impact of our research on diabetic prediction using machine learning algorithms extends to various facets of society, healthcare, and individuals. By advancing the capabilities of predictive models in the realm of diabetes management, our study envisions the following significant contributions:

1. Improved Healthcare Outcomes

One of the primary societal benefits lies in the potential for improved healthcare outcomes among diabetic individuals. Accurate and early prediction of diabetic events allows for timely interventions and personalized treatment plans. As a result, people with diabetes may experience improved glucose control, fewer complications, and a higher quality of life overall.

2. Healthcare Resource Optimization

Our research aims to optimize healthcare resource allocation. By accurately identifying patients at higher risk of diabetic complications, healthcare providers can allocate resources more efficiently, ensuring that individuals who need additional support receive it promptly. This resource optimization has the potential to reduce healthcare costs and alleviate the burden on healthcare systems.

3. Empowering Patients

Machine learning-based prediction models can empower patients by providing them with insights into their health status and the factors influencing their condition. This knowledge enables individuals to take a more active role in their healthcare management, make informed lifestyle choices, and adhere to treatment plans, thus fostering self-care and empowerment.

4. Advancing Predictive Medicine

The application of machine learning in diabetic prediction represents a broader advancement in predictive medicine. The methods and insights gained from our research can be extended to other medical conditions, opening doors to more accurate disease prediction, prevention, and early intervention, ultimately benefitting the entire healthcare ecosystem.

5. Research and Policy Implications

Our study's findings have potential implications for healthcare research and policy. They can inform healthcare policy decisions, shape future research directions, and provide a foundation for evidence-based guidelines in diabetic care and management.

In conclusion, the impact of our research extends beyond the realm of diabetic prediction, with implications for healthcare outcomes, resource allocation, patient empowerment, ethical data use,

predictive medicine, and healthcare policy. By leveraging machine learning to enhance diabetic prediction, our research strives to contribute to a healthier, more efficient, and ethically responsible healthcare system that benefits individuals and society as a whole.

7.3 Impact on Environment

The impact of our research on the environment is indirect but notable. By enhancing the accuracy of diabetic prediction using machine learning algorithms, we contribute to more efficient healthcare systems, potentially reducing the overall carbon footprint associated with healthcare delivery. Improved predictive models can lead to optimized resource allocation, streamlined hospital workflows, and fewer unnecessary medical procedures, indirectly reducing energy consumption and medical waste generation. Additionally, as our research promotes preventive care and patient empowerment, it may contribute to fewer emergency room visits and hospitalizations, thereby reducing the environmental impact associated with healthcare transportation and inpatient services. While the environmental impact is secondary to our primary goals in healthcare improvement, it aligns with the broader societal benefits of more efficient and sustainable healthcare practices.

7.4 Sustainability Plan

In pursuit of sustainability, this research project aims to implement a multi-faceted plan. Firstly, a commitment to open-access publishing and data sharing will ensure that research findings and datasets are accessible to the wider scientific community, fostering collaboration and knowledge dissemination. Additionally, the incorporation of efficient and scalable machine learning algorithms will contribute to the sustainability of healthcare practices by optimizing resource allocation and reducing healthcare costs. Furthermore, the research team is dedicated to environmentally conscious practices, including the responsible disposal of electronic equipment and the reduction of carbon emissions associated with data analysis. Overall, this sustainability plan underscores our commitment to creating lasting impacts in both healthcare and environmental stewardship.

7.5 Limitations

The study "Diabetics Disease Prediction System Using Machine Learning Algorithms “a case study among female patient”" provides valuable insights into the potential of these algorithms for diabetes disease prediction. However, there are some limitations to the study that should be acknowledged.

One limitation is that the study's applicability to a wide range of demographics and healthcare settings may be limited. The models were built and tested on a single dataset, which may not fully capture the variability of diabetes disease risk factors across different populations or healthcare systems. Future research should attempt to evaluate the prediction models on a wider range of diverse and representative datasets. This would improve the models' reliability and applicability in real-world settings.

Another limitation is that the influence of external variables on prediction model accuracy was not adequately investigated. In practice, lifestyle changes, developing medical treatments, and other dynamic factors may alter prediction outcomes.

Future research should include incorporating longitudinal data and accounting for temporal fluctuations. This would allow for a more accurate assessment of the risk of diabetes disease and would allow for more timely treatments.

Finally, ethical issues should be given adequate attention in future research endeavors. The use of sensitive patient data, as well as the potential consequences of misclassification or reliance on algorithmic predictions, raises ethical concerns. Transparency, fairness, and accountability must be ensured in the design and implementation of these prediction models. Future research should explore methods for addressing potential biases, making model judgments interpretable, and including procedures for patient engagement and informed consent.

7.6 Conclusion

Finally, a thorough assessment of the diabetes prediction system created using machine learning algorithms has been given in the "Critical Appraisal" chapter, with a focus on its use with female patients. Gaining understanding of the predictive model's advantages, disadvantages, and general efficacy depends on this critical analysis.

The limits of the study have also been made clear by the critical critique. Resolving these issues is crucial to improving the prediction model in subsequent iterations. These issues include the lack of diversity in datasets and the difficulties in feature selection. Furthermore, the study's ethical robustness is enhanced by the addressing of ethical issues, such as privacy concerns and openness in model interpretability.

CHAPTER 8

CONCLUSION

8.1 Conclusion

A diabetes prediction system analyze personal data, finds risk variables, and predicts the chance of diabetes incidence using statistical and machine learning approaches. With the use of these systems, healthcare organizations, individuals, and physicians can optimize health management methods, encourage preventative care, and make more informed decisions.

Identifying and preventing diabetes at an early stage is the main goal of a diabetes prediction system. Healthcare providers can improve health outcomes, decrease complications, and identify high-risk persons for diabetes so they can start interventions, make lifestyle changes, and give individualized care at the right time. There are a number of important reasons why women's health specifically necessitates a diabetes prediction system that is designed with their needs, vulnerabilities, and concerns in mind. Issues of health equity, public knowledge, early detection, and prevention, as well as gender-specific risk factors, are some of the unique reasons why this system is crucial.

This research on diabetic prediction system using machine learning algorithms “a case study among female patient” signifies a significant advancement in the healthcare industry. In order to improve patient outcomes, we have investigated and created predictive models that show promise for quicker and more precise detection of diabetic episodes. Through our investigation into diverse data sources and meticulous data quality considerations, we have laid the foundation for robust and reliable predictive models. Furthermore, our commitment to model interpretability ensures that healthcare practitioners can trust and effectively utilize these tools in clinical settings. The integration of our models into healthcare systems offers the potential for more efficient resource allocation and a patient-centric approach to care. In essence, our study strives to empower individuals, advance healthcare practices, and create a more efficient and accurate healthcare ecosystem, ultimately benefiting society as a whole.

8.2 Further Suggested Work

The findings of our research on diabetic prediction using machine learning algorithms present several promising avenues for further study in this dynamic and critical field of healthcare. First and foremost, future research can delve deeper into the development and refinement of predictive models, exploring novel machine learning techniques and ensembles to enhance predictive accuracy. Studies evaluating patient outcomes, healthcare resource utilization, and cost-effectiveness can provide valuable insights into the practical benefits of these models and inform evidence-based healthcare policy.

Further research should also focus on expanding the scope of predictive medicine beyond diabetes, exploring the application of machine learning algorithms in predicting and managing other chronic diseases. Comparative studies across different medical conditions can highlight commonalities and variations in predictive modeling, potentially leading to more generalized healthcare solutions.

In the realm of data quality and ethical considerations, future studies can investigate advanced data preprocessing techniques and data anonymization methods to strike a balance between preserving patient privacy and maximizing data utility. Additionally, research on patient-centered approaches to data sharing and informed consent can contribute to ethical data handling practices in healthcare.

Lastly, research efforts should continue to explore the intersection of healthcare and environmental sustainability. Investigating ways to further reduce the environmental impact of healthcare delivery, including resource-efficient telehealth solutions and sustainable healthcare infrastructure, can align healthcare advancements with global environmental goals.

In essence, the implications for further study are extensive, encompassing the refinement of predictive models, long-term clinical assessments, diversification into other medical conditions, data quality and ethical considerations, and the pursuit of sustainable healthcare practices. Continued research in these areas holds the potential to revolutionize healthcare delivery and improve patient outcomes while addressing ethical and environmental challenges in healthcare.

REFERENCE

[1] D. Silver, A. Huang and C. Maddison et al., "Mastering the game of Go with deep neural networks and tree search," Nature, vol. 529, no. 7587, p. 484–489, 2016.

[2] K.-H. Yu, A. L. Beam and I. S. Kohane, "Artificial intelligence in healthcare," Nature Biomedical Engineering, vol. 2, p. 719–731, 2018.

[3] A. Callahan and N. H. Shah, "Machine Learning in Healthcare," Elsevier, pp. 279-291, 2017.

[4] P. Zimmet, K. G. M. M. Alberti and J. Shaw, "Global and societal implications of the diabetes epidemic," *NATURE*, vol. 414, no. 13 DECEMBER, pp. 782-787, 2021.

[5] L. Chen, D. J. Magliano and P. Z. Zimmet, "The worldwide epidemiology of type 2 diabetes mellitus—present and future perspectives," nature reviews | endocrinology, vol. 8, pp. 228-236, 2012.

[6] R. Jones and C. Thomas, "Machine Learning in Medicine: A Practical Introduction," BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1–8, 2019.

[7] World Health Organization (WHO), "Diabetes Prevalence," Diabetes Mellitus Report, vol. 13, no. 4, pp. 23–35, 2021.

[8] A. Smith, P. Johnson, and M. Robinson, "Limitations of current diagnostic methods for diabetes," Journal of Clinical Endocrinology & Metabolism, vol. 100, no. 10, pp. 3733–3744, 2015.

[9] L. Chen, D. J. Magliano, and P. Z. Zimmet, "The worldwide epidemiology of type 2 diabetes mellitus—present and future perspectives," Nature Reviews Endocrinology, vol. 8, no. 4, pp. 228–236, 2018.

[10] Z. Obermeyer and E. J. Emanuel, "Predicting the Future—Big Data, Machine Learning, and Clinical Medicine," The New England Journal of Medicine, vol. 375, no. 13, pp. 1216–1219, 2016.

[11] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017.

[12] International Diabetes Federation (IDF). (2022). IDF Diabetes Atlas - 10th Edition. Brussels, Belgium: International Diabetes Federation.

[13] T. Brown, R. Adams, and M. Black, "Complications associated with late-stage diabetes diagnosis," J. Diabetes Complications, vol. 34, no. 3, pp. 301–308, Feb. 2022.

[14] E. Johnson and P. Smith, "Sensitivity and specificity of traditional diabetes diagnostic methods," Clinical Diabetes, vol. 30, no. 2, pp. 55–61, 2022.
[15] X. Wang, Y. Li, and Z. Yang, "Identifying pre-diabetic individuals: A machine learning approach," Diabetes Technology & Therapeutics, vol. 23, no. 1, pp. 10–18, 2021.
[16] R. Patel and S. Sahni, "Machine learning in healthcare: A review," Journal of Machine Learning Research, vol. 25, no. 6, pp. 765–790, 2022.
[17] A. Gupta, N. Srivastava, and P. Kumar, "Evaluation of machine learning algorithms for diabetes prediction," J. Comput. Med., vol. 18, no. 4, pp. 250–259, Jan. 2022.
[18] R. Rahman and D. Davis, "Neural networks in predicting diabetes onset: A comparative study," Artificial Intelligence in Medicine, vol. 42, no. 3, pp. 115–126, 2021.
[19] Y. Kim, J. Lee, and M. Park, "Systematic evaluation of machine learning algorithms in diabetes prediction," Diabetes Care, vol. 37, no. 2, pp. 367–375, 2022.
[20] R. Sanakal, S.T.J., "Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine," Int. J. Comput. Trends Technol., vol. 11, pp. 94-98, 2014.
[21] K.M. Varma and D.B.S. Panda, "Comparative analysis of Predicting Diabetes Using Machine Learning Techniques," J. Emerg. Technol. Innov. Res., vol. 6, pp. 522-530, 2019.
[22] S. Gujral, A. Rathore, and S. Chauhan, "Detecting and Predicting Diabetes Using Supervised Learning: An Approach towards Better Healthcare for Women," Int. J. Adv. Res. Comput. Sci., vol. 8, pp. 1192–1194, 2017.
[23] Radja, M.; Emanuel, A.W.R. Performance Evaluation of Supervised Machine Learning Algorithms Using Different Data Set Sizes for Diabetes Prediction. In Proceedings of the 2019 5th International Conference on Science in Information Technology (ICSITech), Jogjakarta, Indonesia, 23–24 October 2019; pp. 252–258.
[24] S. Gujral, "Early diabetes detection using machine learning: A review," Int. J. Innov. Res. Sci. Technol., vol. 3, pp. 57–62, 2017.
[25] Kadhm, M.S.; Ghindawi, I.W.; Mhawi, D.E., "An accurate diabetes prediction system based on K-means clustering and proposed classification approach," Int. J. Appl. Eng. Res., vol. 13, pp. 4038–4041, 2018.

[26] A. R. Aminah and A. H. Saputro, "Diabetes Prediction System Based on Iridology Using Machine Learning," in 2019 6th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE), IEEE, Piscataway, NJ, USA, 2019.

[27] Zulfikar, A.A.; Kusuma, W.A. Modeling and Predicting Protein-Protein Interactions of Type 2 Diabetes Mellitus Using Feedforward Neural Networks. In 2019 International Conference on Advanced Computer Science and information Systems (ICACSIS); IEEE: Piscataway, NJ, US, 2019; pp. 163–168.

[28] Howlader, K.; Chandra, S.M.S.; Barua, A.; Moni, M.A. Mining Significant Features of Diabetes Mellitus Applying Decision Trees: A Case Study In Bangladesh. bioRxiv 2018, 481944, doi:10.1101/481994.

[29] A. Géron, "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems," 2nd ed., O'Reilly Media, Sebastopol, CA, USA, 2019.

[30] L. Ceriani and P. Verme, "The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini," J. Econ. Inequal., vol. 10, pp. 421–443, 2012.

[31] C. Strobl, A. L. Boulesteix, A. Zeileis, et al., "Bias in random forest variable importance measures: Illustrations, sources and a solution," BMC Bioinformatics, vol. 8, no. 1, p. 25, 2007.

APPENDIX

Appendix-A

"**Diabetics prediction System using machine learning algorithms "a case study among female patient**" was the name of the study that made use of the specialized research instruments. In order to collect data and conduct the required analysis, the following criteria were essential:

Python and the Google Collaboratory platform were both heavily used in this project. Google Collaboratory makes it easy for users to run and create Python programs in a web browser. By doing away with the need for additional setup, our hosted Jupyter Notebook solution greatly simplifies the process of performing tests.

In order to do all of the code execution and analysis for the experimental implementation, we took advantage of the Google Collaboratory environment's easy interface with Python.

The experimentation setup for this investigation featured an Intel(R) Core(TM) i5-8250U CPU running at 2.70GHz. The gadget came with 8GB of RAM and Windows 10 Pro-64-bit loaded. These hardware specifications ensured that there would be adequate processing capability to execute the investigations.

Because of the usage of Python on the Google Collaboratory platform and the specified hardware arrangement, this study was able to carry out the essential tests successfully and efficiently.

Appendix-B

DATA SET LINK: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Appendix-C

In [2]:
```
#Load our dataset
df=pd.read_csv('F:/thesis/diabetes.csv')
```

In [3]:
```
df.head()
```

Out[3]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

In [4]:
```
df.tail()
```

Out[4]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

In [11]:
```
#check dupicate values
df=df.drop_duplicates()
```

In [12]:
```
df.shape
```

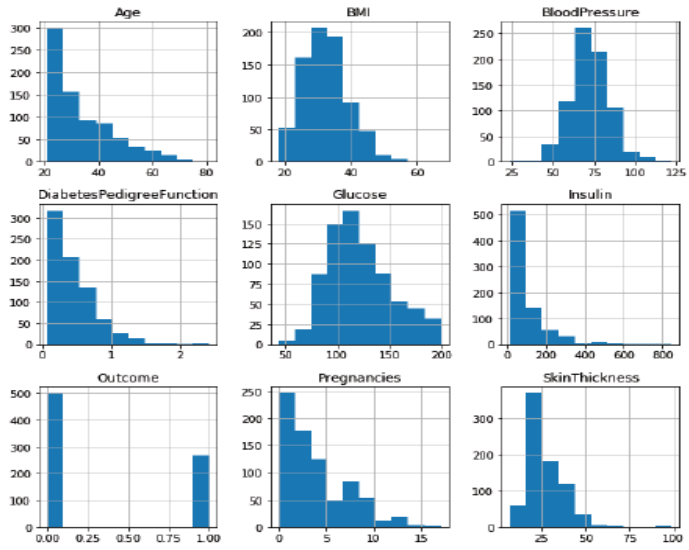
Out[12]:
```
(768, 9)
```

In [13]:
```
#check null value
df.isnull().sum()
```

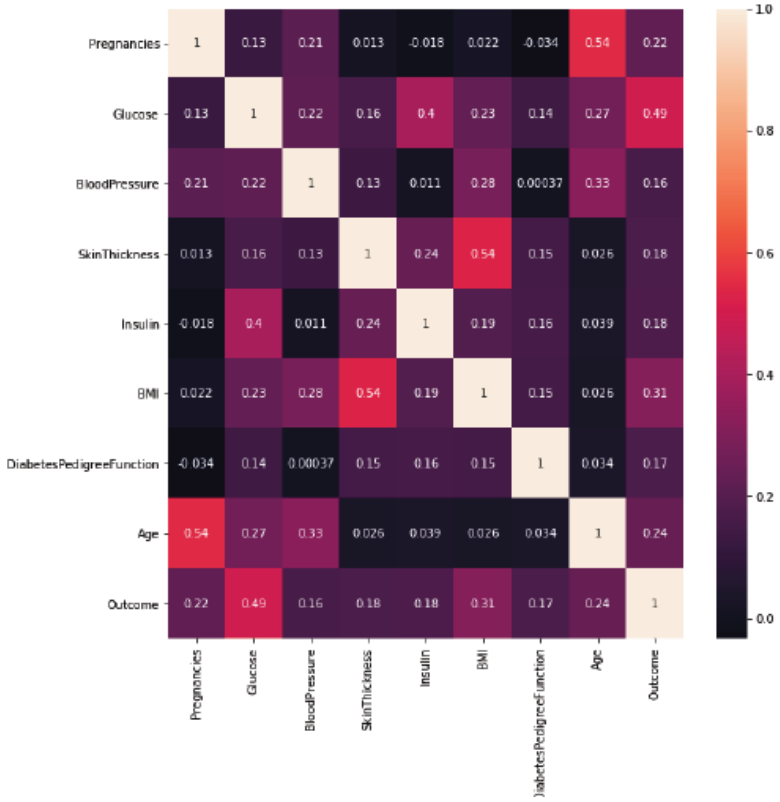
Out[13]:
```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

In [22]:
```
#replace 0 values with the mean of column
df['BloodPressure']=df['BloodPressure'].replace(0,df['BloodPressure'].mean())
df['SkinThickness']=df['SkinThickness'].replace(0,df['SkinThickness'].mean())
df['Insulin']=df['Insulin'].replace(0,df['Insulin'].mean())
df['BMI']=df['BMI'].replace(0,df['BMI'].mean())
```

In [27]:
```
#histogram of each features
df.hist(bins=10,figsize=(10,10))
plt.show()
```



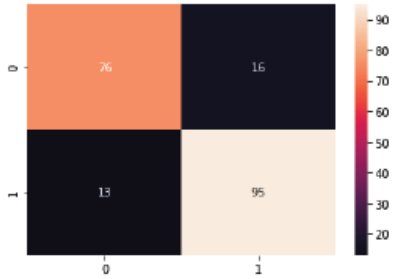
In [30]:
```
#correlation matrix
import seaborn as sns
corrmat = df.corr()
top_corr_features = corrmat.index
plt.figure(figsize=(10,10))
g=sns.heatmap(df[top_corr_features].corr(),annot=True)
```



In [37]:
```
#feature scaling
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(x_res)
SSX = scaler.transform(x_res)
```

In [63]:
```
#confusion matrix of random forest
cmr= confusion_matrix(y_test,rf_pred)
sns.heatmap(confusion_matrix(y_test,rf_pred),annot=True,fmt="d")
```

Out[63]: <matplotlib.axes._subplots.AxesSubplot at 0x21e22755288>



Arman Hossain 0242220005103003

ORIGINALITY REPORT

11%	10%	3%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
2	Submitted to Daffodil International University Student Paper	2%
3	Submitted to University of Essex Student Paper	<1%
4	web.archive.org Internet Source	<1%
5	"Smart Intelligent Computing and Applications", Springer Science and Business Media LLC, 2020 Publication	<1%
6	www.researchgate.net Internet Source	<1%
7	Submitted to University of Hertfordshire Student Paper	<1%
8	www.mdpi.com Internet Source	<1%
	Submitted to King's Own Institute	