

**EARLY PREDICTION OF DIABETES USING MACHINE LEARNING
CLASSIFIERS**

BY

Mithun Mondal
ID: 211-25-954

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

Supervised By

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

Co-Supervised By

Dr. Md. Ismail Jabiullah
Professor
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

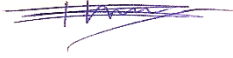
DHAKA, BANGLADESH

JANUARY 2022

APPROVAL

This Research titled “**Early Prediction of Diabetes using Machine Learning Classifiers**”, submitted by Mithun Mondal, ID No: 211-25-954 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 22-01-2022.

BOARD OF EXAMINERS



Chairman

Dr. Touhid Bhuiyan

Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

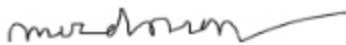


Internal Examiner

Abdus Sattar

Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Internal Examiner

Md. Riazur Rahman (RR)

Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



External Examiner

Dr. Mohammad Shorif Uddin

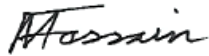
Professor

Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

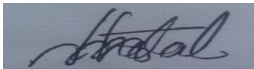
I hereby declare that this work has been done by me under the supervision of **Dr. Md. Fokhray Hossain, Associate Dean & Professor, Department of CSE**, Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



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

Submitted by:



Mithun Mondal
ID: 211-25-954
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible to complete the last semester research successfully.

I am really grateful and wish for profound indebtedness to **Dr. Md. Fokhray Hossain, Associate Dean & Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of “*Machine Learning*” to carry out this work. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

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

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Diabetes Mellitus is one of the most vastly dispersed, lethal and life threatening ailments not only around the globe but also in Bangladesh. It deteriorates the health condition gradually when the human body can not manufacture adequate insulin or could not acknowledge it in a decent fashion, which results in anomalously increased blood sugar levels. Countless complexities including high mortality, damages of numerous organs occur if the patients continue to live without medical treatment. So, identification of this illness in the premature phase and timely medical therapy can retain more humankind from serious injuries. The astonishing advancements in health sciences have contributed to a noteworthy volume of data. Machine learning algorithms have extensively gained popularity in medical science to diagnose and predict the likelihood of this sickness using these tons of raw data. The intention of this research work is to make a side by side analysis of multiple machine learning classifiers and their results of prognosis to this deadly disease beforehand. Decision Tree, Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbours and Naive Bayes have been applied in supervised circumstances to predict the possibility of the disease. The fresh dataset at hand is imbalanced and has been accumulated from UCI repository and having sixteen dimensions and one outcome class. That's why pre-processing tasks like missing or null value replacement, label encoding, importance feature selection, SMOTE resampling methodology to balance class variables, have been conducted on the data. Scikit Learn, a python free module has been used for analysing and visualizing the experiments. Lastly, outcomes of the algorithms have been compared to put a verdict that the Random Forest classifier outperforms others with 98.38% accuracy level.

TABLE OF CONTENTS

CONTENTS	PAGE
Approval	i
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v

CHAPTERS

CHAPTER 1: Introduction	1-3
1.1 Introduction	1
1.2 Motivation	1
1.3 Objective	2
1.5 Report Layout	3
CHAPTER 2: Literature Review	4-6
2.1 Introduction	4
2.2 Related Works	4
2.3 Conclusion	6

CHAPTER 3: Diabetes Mellitus 7-13

3.1 Introduction	7
3.2 Common Types	7
3.3 Complications	9
3.4 Global Scenario	10
3.5 Bangladesh Perspective	11
3.6 Treatment Cost	12
3.7 Conclusion	13

CHAPTER 4: Machine Learning & Knowledge Discovery 14-19

4.1 Introduction	14
4.2 Machine Learning Categories	14
4.2.1 Supervised Learning	15
4.2.2 Unsupervised Learning	16
4.2.3 Reinforcement Learning	17
4.3 Knowledge Extraction	17
4.3.1 Classification	18
4.3.2 Clustering	18
4.3.3 Reduction of Dimensionality	19
4.3.4 Collaborative Filtering	19
4.4 Conclusion	19

CHAPTER 5: Methodology & Result Analysis	20-50
5.1 Introduction	20
5.2 Dataset and Tool Description	22
5.3 Preprocess Dataset	27
5.3.1 Label Encoding	27
5.3.2 Feature Selection	28
5.3.3 SMOTE Resampling Technique	30
5.3.4 Normalization	32
5.4 Dataset Train and Test Splitting Techniques	33
5.5 Applied Classifiers	34
5.5.1 Decision Tree	34
5.5.2 Random Forest	35
5.5.3 K-Nearest Neighbours	36
5.5.4 Support Vector Machine	37
5.5.5 Naive Bayes	39
5.5.6 Logistic Regression	40
5.6 Comparative Analysis	41
5.7 Conclusion	50
CHAPTER 6: CRITICAL APPRAISAL	51-53
6.1 Introduction	51
6.2 SWOT Analysis	51

6.2.1 Strengths	52
6.2.2 Weakness	52
6.2.3 Opportunities	52
6.2.4 Threats	53
6.3 Conclusion	53
CHAPTER 7: Conclusion	54-55
7.1 Conclusion	54
7.2 Further Suggested Works	55
REFERENCES	57-60

LIST OF FIGURES

FIGURES	PAGE NO
Figure 1: Overall Process Diagram	20
Figure 2: Target Class Distribution	23
Figure 3: Count of Instance in various Age Groups	24
Figure 4: Categorical Features Distribution	25
Figure 5: Nominal Scale Values to Encoded Values	27
Figure 6: Correlation Matrix	28
Figure 7: Normal Vs Resampling	30
Figure 8: 3-Nearest Neighbours	36
Figure 9: Support Vectors, Hyperplane, Margin	37
Figure 10: Classifiers' Performance using Cross Validation	43
Figure 11: Classifiers; Performance using Split	44
Figure 12: ROC curves of Six Algorithms	46
Figure 13: Confusion Matrix using Split Technique	48
Figure 14: Confusion Matrix using Cross Validation Method	49

LIST OF TABLES

TABLES	PAGE NO
Table 1: Dataset Description	21
Table 2: Details of the Dataset	22
Table 3: Selected Features	29
Table 4: Performance Matrix	40
Table 5: Processed Dataset	41
Table 6: Comparison of Correct & Incorrect Instances using Cross Validation and Splitting	42
Table 7: Comparison of Performance Metrics	45

CHAPTER 1

INTRODUCTION

1.1 Introduction

Diabetes Mellitus is a collection of metabolic illnesses that has a substantial impact on the human body around the globe. That's one of the deadliest chronic diseases since it elevates blood sugar levels [15]. DM afflicted roughly 422 million people in 2014, with 1.5 million deaths reported in 2012 [24]. Diabetes impacted 8% (12.88 million) of this country's total population in 2016, and 3% of total fatalities of the total age group are caused by diabetes according to WHO report [6]. As a result, one of the most pressing demands of our generation is the development of a cutting-edge expert system that can detect diabetes at an early stage with minimal complexity and in an expedited manner. Machine learning techniques are used to extract and find new and interesting aspects from enormous amounts of data. Immense research into all areas of diabetes generates massive amounts of data. This huge volume of raw data can be used to predict and analyze diabetes.

1.2 Motivation

Diabetic individuals are at risk for significant problems like blindness, nerve damage, heart attack, renal failure and stroke. This disease must be diagnosed as soon as possible in order to receive proper treatment. A patient's appointment at a medical center and consultation with a clinician is the time-consuming identification process. The patient's condition may deteriorate in the meantime due to the frustrating process. One variation type-1 is a life-and-death, ever lasting illness that is commonly visible in young and grownup persons. Another variation type-2 diabetes is a non-insulin-dependent illness

that affects adults on a regular basis. In the initial phases of Type 2 diabetes, patients experience few to no noticeable signs of the disease. 61.50 % of diabetics were ignorant that they had the disease, and just 35.20 % were receiving treatment on a regular basis [7].

1.3 Objective

In Bangladesh, the cost of diabetes care is quite high, owing mostly to the high cost of medicine and hospitalization. It greatly magnifies medicare demand as well as spending. It is also projected to have a significant financial implication on Bangladesh's medical management systems [8]. Most of the low income people can not bear the excessive cost. This research work will help them by predicting the likeness of diabetes. It will provide a comprehensive analysis of the hurdles of predicting diabetes. Finally, the goal of this study is to deliver an easily accessible tool for the user which detects diabetes in an early stage. The general people of Bangladesh may benefit from this research based tool.

1.4 Report Layout

The layout of this report consists of a total seven chapters to fully present the work. Those include: Introduction, Literature Review, Diabetes Mellitus, Machine Learning & Knowledge Discovery, Methodology & Result Analysis, Critical Appraisal and Conclusion.

Chapter 1: The chapter illustrates an overall snapshot of the work. It has Introduction, Motivation, Objective and Report Layout.

Chapter 2: In this part, previous works related to this topic have been discussed. It includes Introduction, Related Works and Conclusion.

Chapter 3: This chapter broadly describes Diabetes Mellitus. It consists of Introduction, Common Types, Complications, Global Scenario, Bangladesh Perspective, Treatment Cost and Conclusion.

Chapter 4: This section highlights different categories of machine learning and ways of extracting knowledge from raw dataset. Introduction, Machine Learning Categories, Knowledge Extraction and Conclusion are the main parts of this section. Clustering, Reduction of Dimensionality and Collaborative Filtering are sub parts of Knowledge Discovery.

Chapter 5: This is the most important chapter of this report. This section narrates the dataset, its characteristics, pre-processing tasks, dataset split techniques. Applied classifiers and comparative result analysis are the crucial parts here. Lastly, this chapter ends with the conclusion part.

Chapter 6: Critical appraisal has been done in this part. Those include Introduction, SWOT Analysis and Conclusion.

Chapter 7: This is the last chapter of this report. It contains the summary and future extension plan.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The rising number of diabetic patients around the globe has prompted experts to do extensive study in exploring hidden patterns in clinical statistics. The previous computational works on pattern identification in diabetic disease are summarized in this section. Not only are numerous methodologies discussed, but also multiple diabetic datasets are examined in order to provide a fair comparison. Finally, in order to reveal unknown diabetic illness trends, all available classification algorithms are applied to a single dataset and their accuracy is compared, which is the study's major goal. Several researchers had used the machine learning techniques to predict diabetes using Pima Indian diabetes dataset (PIDD). This dataset has 9 attributes, 768 records and all patients are particularly female.

2.2 Related Works

Using the ANN approach on PIDD, Alam, T.M.et al. [1] demonstrated 75.70 % accuracy. On PIDD, Sisodia et al. [2] discovered that the NB classifier outperforms the SVM, NB, and DT machine learning algorithms, with an accuracy of 76.30 %. Tigga et al. used logistic regression on PIDD to predict diabetes [3]. Among all the parameters in PIDD, they discovered that the number of pregnancies, BMI, and glucose level are the most relevant variables for diabetes prediction. The PID dataset is used for analysis, and the results are processed and shown using RStudio. With a forecast accuracy of 75.32 %, their model is doing fairly well. According to Amour Diwaniet 's research [4], then

Naive Bayes using cross validations and decision tree algorithms are used to train and test all of the patient's information. The performance of the algorithm was then assessed, analyzed, and compared to that of other classification algorithms using WEKA. With a 76.3021 percent accuracy Naive Bayes performs best according to the result. After applying Principal Component Analysis and Minimum Redundancy Maximum Relevance (mRMR) methods for feature reduction, Zou et al. used Random Forest, Decision Tree, and ANN for classification algorithms on PIDD [5]. They discovered that the best accuracy for Pima Indians is 77.21 percent, which they derived from the random forest using the mRMR.

Sajida et al. [11] address the importance of Adaboost and Bagging ensemble machine learning algorithms [9] in classifying diabetic patients, depending on risk variables, employing the J48 decision tree as the basis. Orabi et al. established a diabetes prediction system in [10], with the main goal of predicting whether or not a diabetes candidate will get diabetes at a given age.

Lee et al. [15] applied Naive Bayes, Logistic Regression and 10-fold cross validation in their study. The goal of the work is to investigate the relationship of the hyper triglyceridemic waist phenotype and diabetes type 2, as well as to evaluate the predictive power of various phenotypes based on individual anthropometric measurements and triglyceride levels. The association of hyper triglyceridemic waist phenotypic with type 2 diabetes is weaker than WC or triglyceride levels. It's simple to use to figure out which type 2 diabetes predictor is best in different regions. Both Naïve Bayes and logistic regression are machine learning techniques that are compared by Lee et al [17]. As a result, the Naïve Bayes method outperforms logistic regression. Also Women perform better than men in terms of prediction. One of the primary differences between the former and this current study is the amount of parameters that are measured.

Hina et al. [12] compared different algorithms with a regression model to predict diabetes patients using Pima. Prema et al. [13] used ensemble voting classifications for the Pima dataset, and compared their method with different algorithms in terms of accuracy

performances. Nilasi et al. [15] in their proposed system, they used clustering, removing noise and then classifying patients. The Self Organizing Map (SOM) method for clustering, Principal Component Analysis method (PCA) for removing the noise and Neural Networks (NA) for classification. They claimed that their new developed method has significantly improved the accuracy of diabetes prediction.

2.3 Conclusion

This literature survey revealed that the research on analyzing and predicting diabetes of Pima dataset was focused only on using data mining techniques and comparing the accuracy of each algorithm used. However, an imbalanced class occurs when one of the classes has a smaller amount, named minority class, than the majority class. The issue of imbalanced data usually leads to misclassification problems where the minority class tends to be misclassified as compared to the majority class. Therefore, it is rather challenging to use the raw data without normalizing and balancing to get higher accuracy.

CHAPTER 3

DIABETES MELLITUS

3.1 Introduction

Diabetes Mellitus is a bunch of metabolic disorders caused mostly by inadequate insulin secretion [18]. Insulin deficiency causes hyperglycemia (high blood sugar) and poor carbohydrate, lipid, and protein metabolism. Diabetes arises when the human body's blood glucose/blood sugar level is extremely high. This disease develops when a human body gland named pancreas, is incompetent to create enough insulin, and the insulin that is produced is unable to be utilised by the body's cells, according to medical experts. Diabetes mellitus is the most familiar endocrine illness, affecting about 200 million individuals globally. Diabetes' progression is intricately tied to a variety of problems, the most prevalent of which is persistent hyperglycemia. Usually diabetes mellitus comprises a vast scope of pathophysiological diseases.

3.2 Common Types

Diabetes is commonly assumed to be divided into two categories, however gestational diabetes is also very prevalent. Because your body has difficulties manufacturing insulin, a hormone that moves and stores sugar, all types of diabetes cause high blood sugar.

- **Type 1:** This type of diabetes is marked by the loss of beta cells as a result of an autoimmune response, which usually results in complete insulin insufficiency. The onset is usually sudden, lasting anywhere between days and weeks. Around

95% of type 1 diabetes patients already have the disease before the age of 25, with the disease occurring equally in both sexes and with a greater proportion in the white community. A small number of patients, typically of African or Asian origin, lack antibodies but present with a similar clinical presentation. As a result, they are included in this classification, and their diabetes is referred to as the "idiopathic form" of type 1 diabetes [23].

- **Type 2:** This is the most prevalent type of diabetes, and it's closely associated to having a family history of diabetes, being older, being obese, and not getting enough exercise. Women, especially those who have had gestational diabetes before, as well as blacks, Hispanics, and Native Americans, are more prone to develop the disease. Insulin resistance and hyperinsulinemia cause poor glucose tolerance over time. Glucose intolerance and hyperglycemia are fueled further by the exhaustion of defective beta cells. Type 2 diabetes mellitus has a complex etiology that is most likely genetically based, but it also includes substantial behavioral components.
- **Gestational:** This form of diabetes develops as a result of hormonal changes that occur during pregnancy. The hormones produced by the placenta can reduce your body's insulin sensitivity. This could lead to elevated blood sugar levels during pregnancy. Most women with gestational diabetes mellitus, on the other hand, maintain good glucose homeostasis during the first half of their pregnancy and then experience a relative insulin shortage in the second half, resulting in hyperglycemia.

3.3 Complications

- **Vision Loss:** The retina, optic nerve, and focal point are all affected by retinopathy. Swelling of the retinal region can arise as a result of late-night vision impairment concerns, diminishing mental interaction. A diabetic's eye vision should be treated with a few tests and medications at first. The treatment includes visual sharpness testing, tonometry, student enlargement, and optic intelligibility tomography.
- **Kidney Neuropathy:** An excessive blood sugar level destroys the kidney's arteries, leading in chronic kidney infection or diabetic neuropathy. The kidney's job is to excrete waste and extra water from the bloodstream. The kidneys should work extra hard to filter the blood due to hypertension and high blood sugar levels. Those can culminate in renal impairment or the necessity for dialysis. Treatment options include kidney replacement therapy, as well as kidney and pancreas transplants.
- **Liver Problems:** Gluconeogenesis and glycogenolysis processes in the liver help to regulate blood glucose levels in the bloodstream. The formation of the liver tumor is facilitated by a fatty liver. Renal impairment, disturbed digestion, insulin sensitivity and hyperglycemia, as well as malnutrition, are all concerns. Impacted people must take numerous antitoxin drugs, and liver administration also includes lifestyle changes, pharmacological treatment, TZDs, and weight loss.
- **Heart Problems:** According to the American Heart Association, 68% of the total population has heart problems that cause mortality. Heart attacks, strokes, atherosclerosis (hardening of the supply pathways), stress, and burden on the heart can all lead to mortality. Because of the high viscosity of blood due to high sugar levels, it sticks to veins, putting more demand on supply pathways and veins to continue further. It continually harms the veins and nerves, causing people to experience circulatory system or organ failure.

3.4 Global Scenario

Diabetes is a serious public health concern in most nations, both locally and internationally. 465 million (9.3%) persons worldwide have diabetes in 2019, and this number is expected to climb to 700 million (10.9 percent) by 2045 as per the International Diabetes Federation [25]. In the same way, pre-diabetes can be spread out in adults, was reckoned to be 374 million (7.5%) in 2019 and is expected to rise to 548 million (8.6%) by 2045. Type 2 diabetes mellitus patients have a 10-year reduced life expectancy and 80 percent of patients with type 2 die from cardiovascular problems [26]. Additionally, it is predicted that between 2010 and 2030, developing nations would have 69 percent more adults with diabetes and developed countries will have 20 percent more [27]. Around 79 percent of diabetics live in low and mid level income countries, with asian countries accounting for more than 60% [26]. Diabetes and pre-diabetes prevalence has risen steadily in both urban and rural parts of South Asia, owing mostly to lifestyle changes and the shift to urbanisation and industrialisation [28, 29, 30]. Diabetes and its accompanying health consequences are on the rise in developing countries, threatening to undo economic achievements.

The diabetes prevalence is greater than the IDF's predicted general age-adjusted diabetes prevalence of 11.3 percent in Southeast Asia in 2019 [31]. China registered 116 million cases and India with 77 million cases are the nations in the area with the highest number of adults with diabetes aged 20–79 years in 2019 according to IDF [31]. In 2019, the International Diabetes Federation ranked Bangladesh 10th among countries with the largest number of adults of 20–79 years with diabetes (8.4 million cases) and it is anticipated to be ranked 9th in 2030 and 2045 [31]. In Bangladesh, more than one out of every ten persons (18+) has diabetes, resulting in a population of 14 million people by 2020. This increasing case in our country points out that the country has one of the

highest diabetes burdens in the Southeast Asian area, emphasizing the urgent need for policies that promote the implementation of diabetes prevention programs.

3.5 Bangladesh Perspective

Bangladesh has a somewhat greater prevalence of pre-diabetes than diabetes. One possible explanation is that as the Bangladeshi workforce has transitioned away from agriculture and toward manufacturing services and industries, people's energy usage has reduced considerably. Obesity and insulin sensitivity are raised consequently from the combination of increased caloric intake and decreased energy production caused by sedentary lifestyles, which escalates the likelihood of pre-diabetes. In Bangladesh, according to the International Centre for Diarrhoeal Condition Research, 7.1 million individuals had diabetes in 2015, with 3.7 million instances going untreated and 129 000 fatalities related to the disease.

Bangladesh is still urbanizing, with the rate of urbanization increasing from 28.97 percent in 2008 to 36.63 percent in 2018. Diabetes prevalence is substantially greater in urban people (11.5 percent) than in rural populations (6.2 percent). Random eating habits, lack of physical activities, frequent smoking and alcohol consumption are probable reasons for obesity and diabetes that are linked to urbanization. In addition, our findings revealed, overall prevalence of diabetes was slightly greater in males than in females (7.34 percent compared with 6.70 percent). This finding is in line with earlier research. On the contrary, there exists a subtle difference in pre-diabetes prevalence between male and female in the pooled data. Over the last two decades, the pervasiveness of diabetes has been boosted from 4.0 percent in 1995–2000 to 10.4 percent in 2010–2019.

3.6 Treatment Cost

This is costly for both affected people and their families, as well as governments. Diabetes medication, for example, represents a considerable portion of a diabetic patient's family income, 5–25 percent in India and 5–10 percent in the United States [38]. Diabetes has a major national influence on direct health-care expenses in all countries. A diabetic person costs the health-care system 2.5 times more than someone who does not have the ailment. Diabetes' estimated total yearly direct health-care costs in developed countries ranged from 0.54 billion US dollars in Denmark to 60 billion US dollars in the United States in 1998 [39].

Low and lower middle-income people have limited access to health insurance and available medical services when compared to citizens in high-income nations. As a result, they have to bear a larger part of the pocket for health-care costs. Moreover, in some low income countries, diabetes patients have to face nearly all costs associated with diabetes care [33]. Diabetes prevalence has risen faster in Southeast Asia than in any other big region of the world [34]. In 2011, the estimated prevalence of diabetes among adults in Bangladesh was 9.7%, with a projected increase to 13.7 million by 2045 [34]. According to the Bangladesh National Health Accounts, Bangladesh spent US\$2.3 billion on health in 2010 (or US\$16.20 per person per year), with out-of-pocket payments accounting for 64% of the cost [35]. Bangladesh, on the other hand, spent US\$88 per person per year on health in 2014, according to the WHO. It was found that on average, a household spent 7.5 percent of its total income on health care, with the lowest 20% of households spending around 13.5 percent [36]. As a result, out-of-pocket health-care costs imposed a significant financial burden on the Bangladeshi people.

3.7 Conclusion

Diabetes mellitus is a condition that is affected by a variety of factors and can change over time and place. As a result, it must be screened on a regular basis. Because of its pandemic character, it has become a serious public issue all across the world. Bangladesh is no different. Age, gender, educational achievement, wealth, obesity, hypertension, and physical task level were all major forecasters of diabetes mellitus prevalence in Bangladeshi adults. Diabetes epidemics are linked to physical activity. When the amount of these activities becomes lower, this raises the incidence of obesity. These activities lessen the probability of having diabetes, as evidenced by various research. Diabetes problems can be reduced by engaging in physical activity [19]. It's also crucial for diabetic management [20]. Diabetic patients should engage in regular physical activity [21].

CHAPTER 4

MACHINE LEARNING & KNOWLEDGE EXTRACTION

4.1 Introduction

Artificial Intelligence is a broad field that encompasses a wide range of topics, ranging from logic to text tonality analysis. Machine learning is one of artificial intelligence's most successful directions (ML). It's the branch of science that studies how machines learn from their experiences. Many different types of problems have been successfully solved using machine learning. In many aspects, machine learning is now a well-established discipline. Choice of valid dataset and pre-processing, election of appropriate algorithms and solution standard evaluation are all part of the ML application methodology. The quest for optimal use of the collected potential of massive data, exploration for rapid learning methods and inspection of application features, based on the application property, are all part of the development of this vast domain.

4.2 Machine Learning Categories

Typically, machine learning tasks are divided into three groups.

4.2.1 Supervised Learning

The system must learn a target function, which is an expression of a data model, inductively, in supervised learning. Instances refer to the collection of possible input values for a function, also known as the function's domain. A set of attributes is used to describe each case. A subset of all situations for which the output variable value is known is referred to as training data or instances. In order to infer the best target function from a

training set, the learning system evaluates alternative functions, termed hypotheses and denoted by h . Classification and regression are two forms of supervised learning algorithms. Classification models try to forecast specific classes, such as blood types; on the contrary, regression models foresee numerical values. Most common popular classifiers are Decision Tree (DT), Instance Based Learning (IBL), such as k-Nearest Neighbors (k-NN), Genetic Algorithms (GA), Artificial Neural Networks (ANN) and Support Vector Machines (SVM).

4.2.2 Unsupervised Learning

The system seeks to find the hidden structure of data or correlations between variables in unsupervised learning. Rule of Association Mining is a considerably later development than machine learning, and it is influenced more by database research. Cluster analysis, often referred as clustering, is the problem of grouping items so that objects in the same group (called a cluster) are more comparable to one another than objects in other groups. It is a standard technique for statistical data analysis used in many domains, that includes machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics. This can be considered as the main task of data mining.

- **Association Rule Learning:** Since its introduction as a market basket analysis tool, association rule mining has evolved into one of the most effective methods for undertaking unsupervised exploratory data analysis across a vast scope of technical and commercial fields, including biology and genetics. Biological sequence analysis, gene expression data analysis and other applications in human science are among the most prominent.
- **Clustering:** Clusters are patterns that emerge from clustering, which divides a huge raw dataset into small groups of data so that records in the same group are as

similar as feasible and samples that reside in other groups differ as little as possible [40].

4.2.3 Reinforcement Learning

Reinforcement Learning is a group of strategies in which the model tries to learn itself by interacting directly with the circumstance in order to produce some kind of cumulative reward. It's worth noting that the system has no prior knowledge of the environment's behavior, and the only method to learn is through trial and error. Because of its independence from its surroundings, reinforcement learning is typically used in autonomous systems.

4.3 Knowledge Extraction

The discipline of knowledge discovery encompasses ideas, methodologies and approaches aimed at making proper use of data and extracting meaningful information from it. Data mining, which exemplifies the use of machine learning algorithms in analysis of data, is the crucial step in the full process. Knowledge extraction collects valid data from structured and unstructured sources to create a meaningful database that can be used to find meaningful and helpful patterns in huge, semantically ambiguous datasets. Sets with a degree of membership are called fuzzy datasets. A membership function with a value between 0 and 1 defines the degree of membership. Knowledge discovery entails a systematic search of huge amounts of raw dataset for patterns that can be classified as knowledge. This gained information has been offered in the form of models, which may be queried as needed. To handle user-oriented questions and difficulties, knowledge discovery combines concepts from computer science and machine learning with those from statistics. Knowledge can be expressed in a variety of ways, including actor classes, attribute association models, and dependencies. Core machine techniques for

classification, clustering, dimensionality reduction, and collaborative filtering as well as scalable distributed platforms, are used in big data knowledge discovery.

4.3.1 Classification

The development of predictive analytics which is capable of emulating human decision-making requires classification. These classifiers are useful for issues with well-defined limits, such as those with inputs that follow a predefined set of attributes and category outputs. Usually the classification process creates an experience archive by evaluating new inputs and comparing them to previously recognized patterns. This input variable is connected with the established predictive behavioral pattern if a pattern can be matched. When a pattern can't be matched, it's quarantined for additional analysis to see if that is a previously unknown valid pattern or an uncommon pattern. Machine-based classification algorithms employ supervised-learning techniques, in which the algorithms learn from training sets of exact decision-making made with precisely built variables. A learning algorithm can be used to synthesize a model and this model will be employed to categorize new data; Those two are the primary phases in classification.

4.3.2 Clustering

Clustering is a method of knowledge discovery in which data points from a collection are grouped together based on comparable qualities (or characteristics). Those who belong to the similar group have homogeneous features to those who belong to heterogeneous groups. Clustering is often accomplished by an iterative trial and error method. The goal is to create a function that uses a numerical number to represent the degree of similarity between two items or data points. Clustering parameters like the clustering technique, the

distance function, the density threshold and the number of clusters vary depending on the applications and dataset.

4.3.3 Reduction of Dimensionality

Reduction of dimensionality is the process of minimizing random variables by using selection and extraction of features. Reduced dimensionality provides for faster training times, better generalization, and less overfitting. The process of synthesizing a group of the original variables for model creation by deleting extra or irrelevant characteristics is known as feature selection. Feature extraction, on the other hand, is the process of merging qualities to change a space of high dimensional elements into a smaller dimension space.

4.3.4 Collaborative Filtering

Collaborative filtering is the searching process for information or trends across numerous sources of data using a holistic approach. This method investigates a subject of research by collecting preferences from a large number of individuals who share similar interests and makes recommendations based on their choices. Despite very sparse data, rising numbers of users and objects, synonymy, data noise, and privacy concerns, these filtering algorithms are projected to generate adequate recommendations.

4.4 Conclusion

The efficiency of the machine learning algorithms are calculated from the ability of corresponding methods to extract information, patterns and generate models from data.

These approaches use established properties gained from training data to do predictive analysis. By comparing fresh information with the past one, in the form of patterns, machine learning facilitates the exploration of relevant or previously undiscovered knowledge. New information or patterns are filtered out using these patterns. This new information is integrated into the current knowledge database once it has been evaluated against a set of associated behavioral patterns.

CHAPTER 5

METHODOLOGY & RESULT ANALYSIS

5.1 Introduction

In many real life scenarios, one of the most significant decision-making approaches is classification. This study's main purpose is to improve classification efficiency and determine whether data is diabetes positive or negative. The bigger the number of samples for various classification problems, the lower the classification accuracy. The method performs well in terms of speed in many situations, however the quality of the classification process is low. This study looked into a variety of classification techniques for diabetes positive and negative data. As a result, classifiers such as Decision Tree, Random Forest, K-Nearest Neighbours, Support Vector Machine, Logistic Regression and Naive Bayes are found to be the best fit for constructing the Diabetes prediction system.

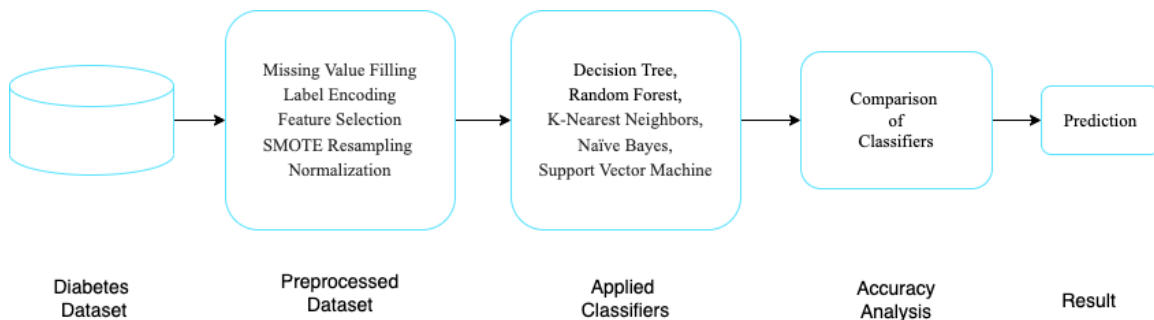


Figure 1: Overall Process Diagram

5.2 Dataset and Tool Description

In this research, a benched-mark UCI repository dataset [41] has been used to detect diabetes at earlier stages. This data was gathered using a direct questionnaire delivered to 520 patients at the Sylhet Diabetes Hospital in Bangladesh who had recently been diagnosed with diabetes or were experiencing diabetes-related symptoms. This dataset contains 17 attributes; Among them 16 attributes - Age, Gender, Polyuria, Polydipsia, Sudden Weight Loss, Weakness, Polyphagia, Genital Thrush, Visual Blurring, Itching, Irritability, Delayed Healing, Partial Paresis, Muscle Stiffness, Alopecia, Obesity are considered as independent variables and the remaining Class attribute is regarded as dependent variable. Table 1 describes the properties considered in this dataset. Figure 2 depicts the target class distribution. Figure 3 and Figure 4 draw the distributions of the numeric feature (Age) and categorical dimensions respectively.

Table 1: Dataset Dimension

	No of Attributes	No of Records
Dataset	17	520

Table 2: Details of the Dataset

Attributes	Descriptions	Encodings/Values
Age	Age in years	20 - 65
Gender	Patient's Sex	Male / Female
Polyuria	Patient needs to frequently urinate, particularly at night	Yes / No
Polydipsia	Patient's thirst is increased & need for fluids	
Sudden Weight Loss	Noticeable weight loss	
Weakness	Patients' physical condition	
Polyphagia	It signals patients' increased appetite	
Genital Thrush	It mainly affects the vagina, though may affect the penis	
Visual Blurring	Swelling in the macula	
Itching	It's a condition that develops when diabetes leads to nerve damage	
Irritability	Patients' mood swings if s/he frequently has high and low blood glucose	
Delayed Healing	It happens due to poor blood circulation	
Partial Paresis	It denotes a condition in which muscle movement has become impaired	
Muscle Stiffness	Patients experience muscle tightness and find it difficult to move their muscles.	
Alopecia	Hair falls out in little places as a result	
Obesity	It is caused by the underlying pathology of accumulation of excess body fat	

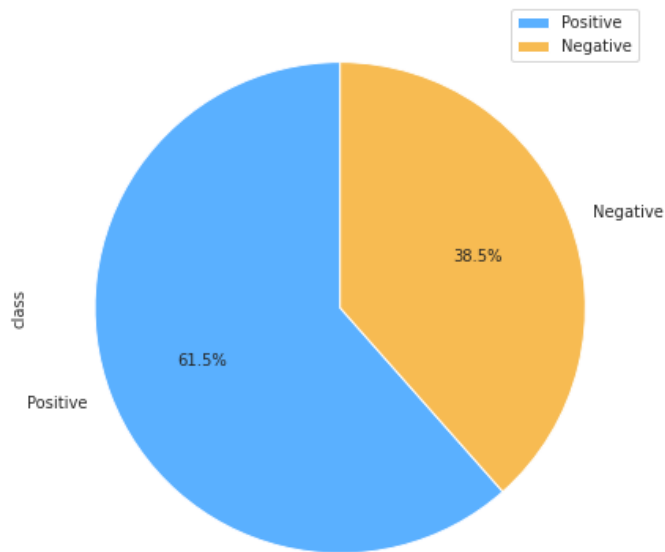
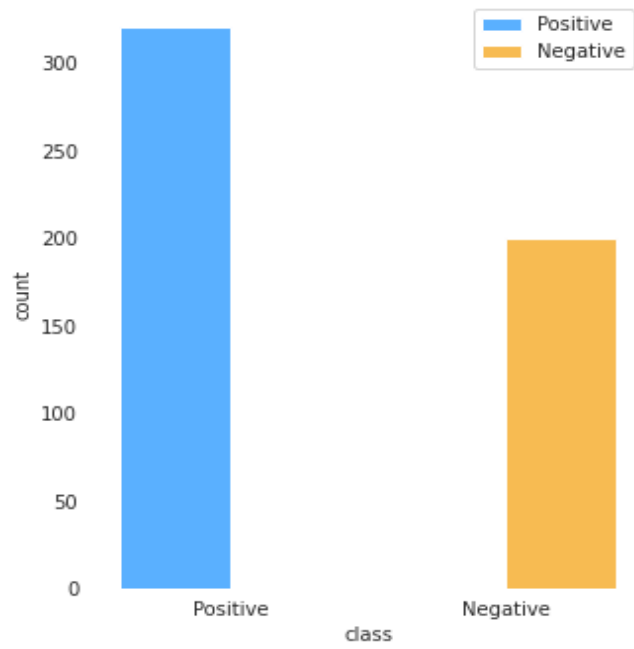


Figure 2: Target Class Distribution

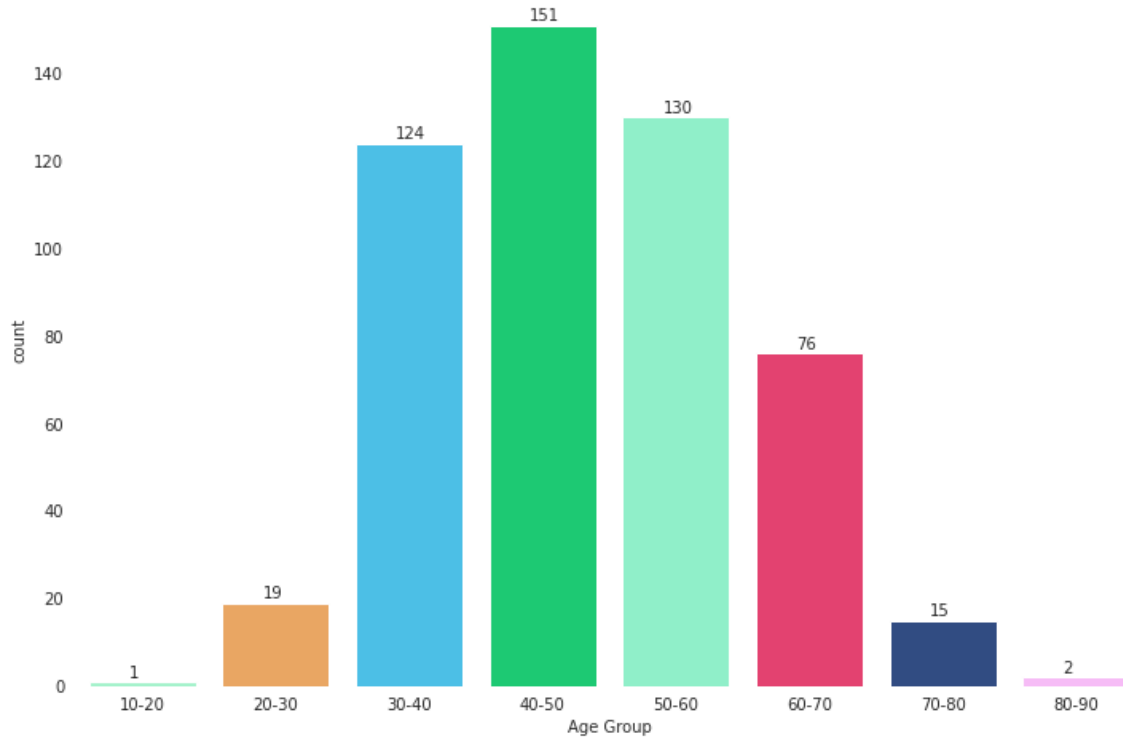


Figure 3: Count of Instance in various Age Groups

In this research work, Scikit-learn (formerly scikits learn and also known as sklearn) has been used to pre-process, visualize and analyze which is the most usable and robust machine learning library. It's a Python-based machine learning package that's available for free. It provides a suite of efficient machine learning and statistical modeling methods like classification, regression, clustering, and dimensionality reduction through a Python consistency interface.

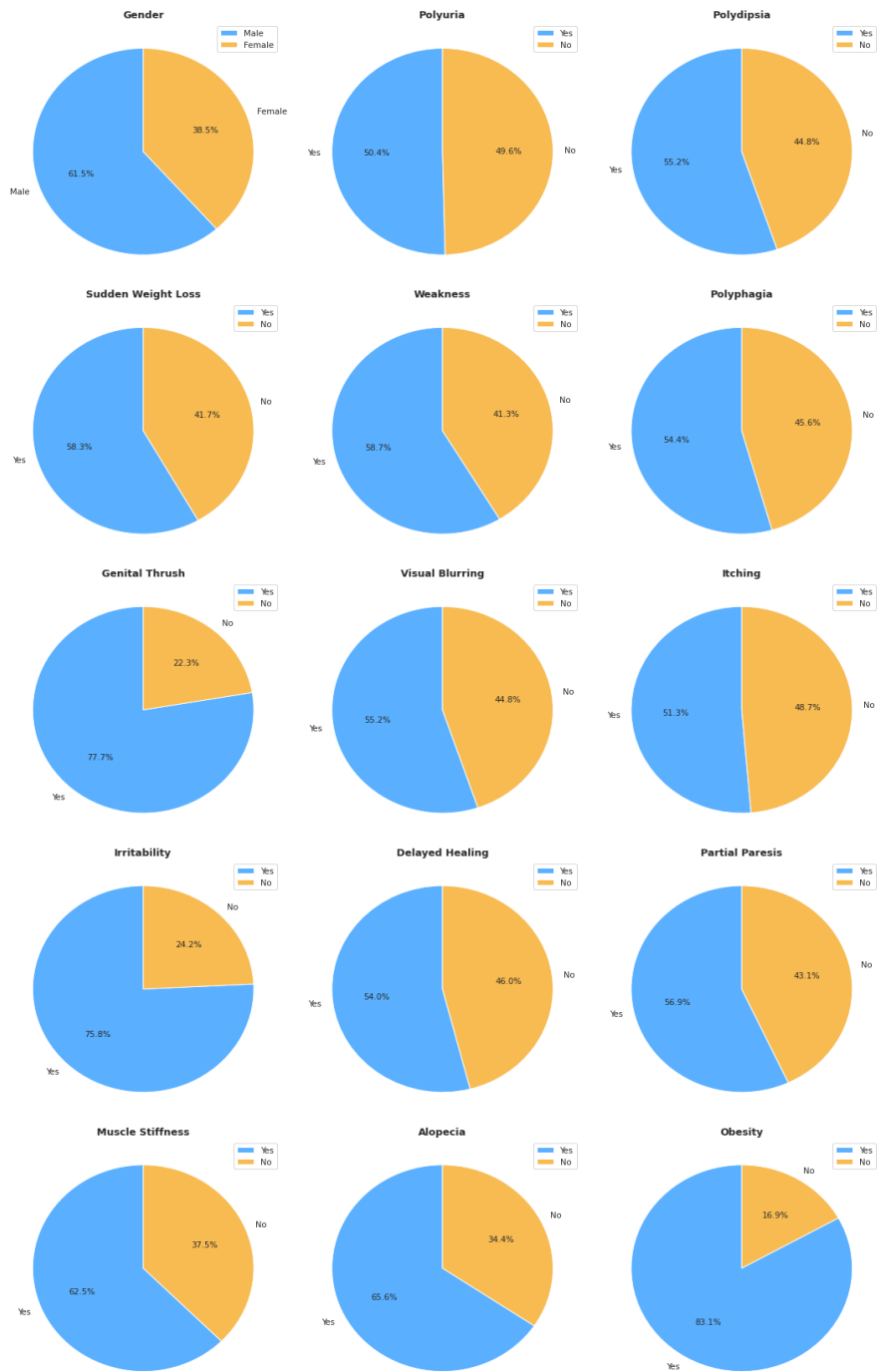


Figure 4: Categorical Features Distribution

5.3 Dataset Preprocessing

Preprocessing aids in the transformation of the dataset so that a more accurate expert system may be generated. The quality of data has an impact on the model performance used for classification and prediction because raw data is frequently incomplete and has some noise. Data preparation task turns raw data into a logical or understandable format such that domain values are consistent and features are comparable. Data cleaning, data standardization, data transformation, and data reduction are all part of data preparation. Preprocessing the dataset before training it in the model is critical for better learning the dataset's hidden patterns.

5.3.1 Label Encoding

In this study, the data collected by questionnaires is a combination of both categorical and continuous variables. Categorical variables will be difficult for most machine learning algorithms to comprehend or deal with. When data is provided as a number rather than category to a model for training and testing, intelligent retrieval techniques perform better. The categorical value has been substituted with a numeric value between 0 and the number of classes minus 1 in Python label encoding. This dataset contains “Male/Female”, “Yes/No” nominal values. These values have been transformed to 0/1 values. Figure 5 shows the encoded labels of nominal values where 0 is the label for No and Female, 1 is the label for Yes and Male respectively.

Gender	Polyuria	Polydipsia
Male	No	Yes
Male	No	No
Male	Yes	No
Male	No	No
Male	Yes	Yes

(a) Nominal Scale

Gender	Polyuria	Polydipsia
1	0	1
1	0	0
1	1	0
1	0	0
1	1	1

(b) Label Encoding

Figure 5: Nominal scale values to encoded values

5.3.2 Feature Selection

Pearson's correlation approach is more familiar for determining the most important features or dimensions. It's a way to sum up the strength of a linear relationship between two sets of data. It is the normalization of the covariance between the two variables to give an interpretable score. This approach calculates the correlation coefficient which is related to the output and input qualities. The value of the coefficient remains between -1 and 1. A significant correlation exists if a value greater than or less than 0.5, whereas no correlation is indicated by a value of zero. The equation for calculating the coefficient as follows:

$$coefficient = \frac{covariance(X, Y)}{stdv(X) * stdv(Y)}$$

To determine the Pearson's correlation coefficient between two data samples of same length, pearsonr() SciPy method can be used. In this work, this coefficient has been used to evaluate the relationship between variables. This calculation produces a correlation matrix.

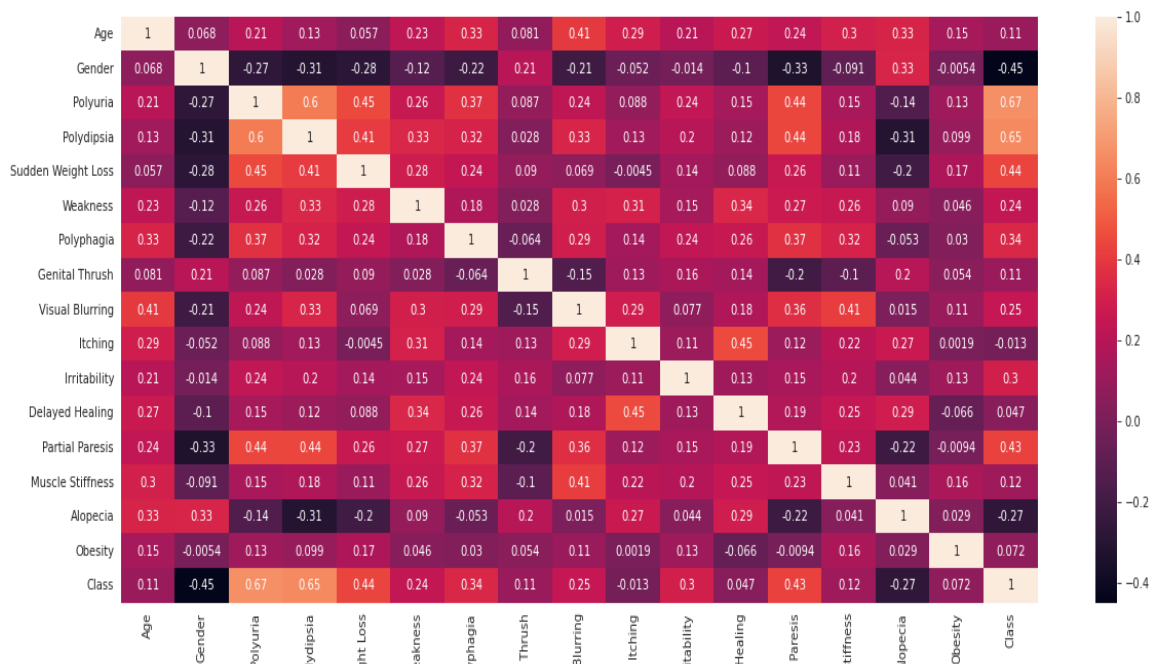


Figure 6: Correlation Matrix

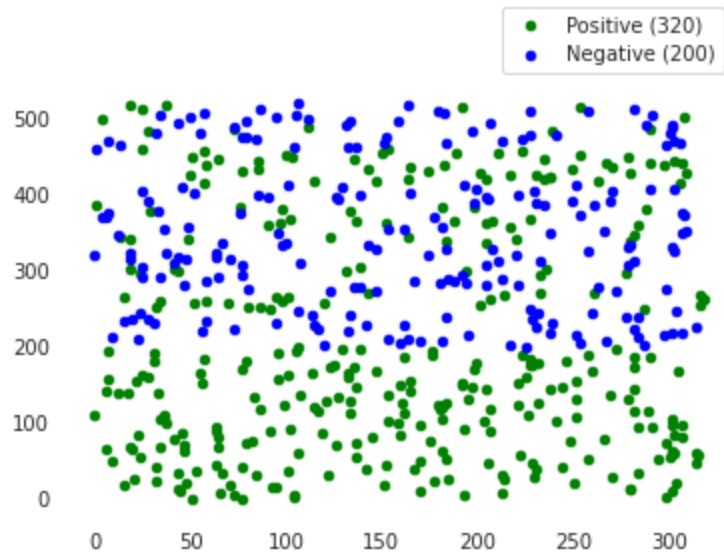
From the above figure, it is visible that all the variables are not strongly correlated with the target class. Following features have been considered to boost up the efficiency of the process.

Table 3: Selected Features

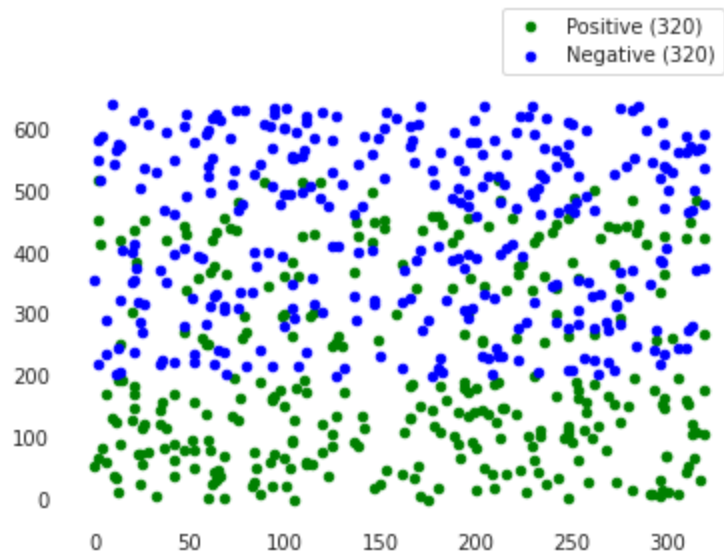
Columns				
Polydipsia	Sudden Weight Loss	Polyuria	Irritability	Weakness
Partial Paresis	Visual Blurring	Polyphagia	Age	Gender

5.3.3 SMOTE Resampling Method

In real life scenarios, classes aren't always evenly distributed and data imbalances might emerge when one class is a minority and the other is a majority. The fundamental difficulty with imbalanced data is that it often leads to misclassification, with the minority class being misclassified the most. Undersampling and oversampling are the two most common ways for balancing. Oversampling techniques are preferred over undersampling techniques in most circumstances. The reason for this is that in undersampling, many occurrences that can be excluded may contain crucial information. Oversampling the examples in the minority class has been followed in this work. This task can be carried through successfully by duplicating minor occurrences in the training dataset before utilising the model in the system. This can help to balance the class distribution, but it doesn't give the model any extra information. The Synthetic Minority Oversampling Technique (SMOTE) is the most extensively used method for creating new samples. For the minority class, it creates virtual training records using linear interpolation. SMOTE starts by finding close-together instances in the feature set, drawing a line in the subspaces between the examples, and going to put a new sample along that line. Following figure will help to visualize how the SMOTE oversampling method balances the minority class.



(a) Before Applying SMOTE



(b) After Applying SMOTE

Figure 7: Normal Vs Resampling

5.3.4 Normalization

The quality of the data put into the machine learning algorithm is critical to its performance. Outliers, missing values, incorrect data types, irrelevant features, and non-standardized data are common in real-world data. These will hinder the intelligent retrieval system from learning accurately. As a result, converting the raw dataset into a usable and valid format is a crucial part of this process. When it comes to pre-processing data, one approach you'll come across repeatedly is normalization. Data normalization is a technique that involves converting various range numeric columns to a base scale so that all values will be on the similar scale. It is very usual that some input variables' value range differ from others while considering real life problems. This type of learning process will be dominated by traits with higher values. However, this does not imply that those factors are more essential in predicting the model's conclusion. Data normalization is the process of converting multi scaled data to a single scale. When normalization has been applied on any dataset, all columns have an interchangeable impact on the system, enhancing the learning algorithm's firmness and efficacy.

The Z-score method has been used here to normalize the dataset. The information is analyzed into a normal distribution with a mean of 0 and a standard deviation of 1 using the method. By subtracting the mean of the associated characteristic and then dividing by the standard deviation, each standardized value is calculated.

$$x_{std} = \frac{x - \mu}{\sigma}$$

In this research work, StandardScaler has been applied to calculate z-score which is already included in the aforementioned python library. Z-score calculation is performed by the .transform() method, which employs the parameters supplied by the .fit() method. The StandardScaler equation as follows:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (y_i - \mu)^2}{N}}$$

Here N is the count of instances in the population.

5.4 Dataset Train and Test Splitting Technique

The dataset is prepared to train and test after it has been cleaned and preprocessed. But there can be some problems to achieve the best result from the trained models. Overfitting is a common problem while training a model. When a model performs exceptionally well on the data that has been used to train it, but fails to generalize well to new, previously unseen data points, this phenomenon happens. This can occur when the data includes noise or the model learns to anticipate certain inputs rather than the predictive factors that could help it make more accurate predictions. In general, the more complicated a model is, the more likely it is to be overfitted. Underfitting scenario happens when the system performs poorly even though the same data was used to train it. Underfitting happens when the model isn't appropriate for the problem you're trying to solve. This usually signifies that the model isn't as complicated as it has to be in order to learn the parameters that have been shown to be predictive.

The most typical strategy for identifying these difficulties is to create separate data that can be used as training and testing models. The most straightforward technique to divide the modeling dataset into training and testing sets is to assign two-thirds of the data points to the former and one-third to the latter.

Another methodology used in pattern recognition applications to assess a machine learning model's skill on unknown data is cross-validation. It's a strategy for assessing

models on a tiny data set that uses resampling. The algorithm has only one parameter (k), which determines how many groups a given data sample should be divided into. Consequently, k -fold cross-validation is commonly used for the procedure.

In this study, to perform the comparative result analysis of multiple machine learning models, 10-fold cross-validation and the 80/20% train/test splitting approach are utilized individually. The dataset was randomly splitted into the training and testing sets using the train/split method. The data is separated into 10 folds in the K cross-validation approach. Testing has been conducted with one fold and training is done with the remaining $K-1$ folds. The technique will be repeated until each and every K fold is a test set. The average of all recorded K th test scores is used to assess performance.

5.5 Applied Classifiers

5.5.1 Decision Tree (DT)

Decision Tree, a supervised machine learning approach is used to solve categorization problems. J. R. Quinlan's ID3 method, which performs a top-down, greedy search across the universe of possible branches with no backtracking. The tree structure has been used, and the tree starts with a single node representing the training samples. The node becomes the leaf if all of the specimens belong to a certain class, and the class is being utilized to identify it. Otherwise, the discriminatory property is chosen as the root node of the tree by the algorithm. The training dataset is separated into many subsets, each of which forms a branch, based on the value of the current decision node attribute and there are several values that form multiple branches. In every phase, this decision tree considers each and every node by calculating the highest information gain among all the attributes [42].

Entropy: The entropy of any information system is a measure of its randomness. The greater the entropy, the more complicated it is to make any conclusions from the system.

Entropy for single attribute:

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

Here, S = Current state, P_i = Probability of event i of State S .

Entropy for multiple attributes is as follows:

$$E(T, M) = \sum_{c \in M} P(c) E(c)$$

Here, M = Selected Field, T = Current State,

Information Gain: Information gain, a quantitative attribute used in statistics and quantifies how skillfully a given attribute separates training examples based on their classification target. Looking for a property which has highest information gain and the lowest entropy is the precondition to constructing an effective decision tree. Basically Information gain is a decrease in entropy.

$$IG(T, M) = E(T) - E(T, M)$$

5.5.2 Random Forest (RF)

Random Forest is a multifunctional technique that populates numerous decision trees in order to classify the target class. Each tree in this classifier produces individual classification results and votes when the forest is predicting a new object according to several qualities and the total outcome of the forest will be the maximum number of taxonomies. Random forests are a way of decreasing variation that calculates the average of multiple deep decision trees trained on different parts of the same training set [43]. That's why it doesn't suffer from the overfitting problem.

How nodes branch on a decision tree can be calculated using following equation:

$$G = 1 - \sum_{x=1}^N P_x^2$$

Here, P_x represents the frequency of the target class of the dataset and N is total class count. This method estimates Gini of each branch on a node based on the class and probability, identifying which branch is more likely to occur.

5.5.3 K-Nearest Neighbours (KNN)

The K-NN approach stores all available dataset and categorizes new feature points depending on homogeneousness of the current data. This means that utilizing the K-NN approach, fresh data can be smoothly categorized into a well-defined category. This classifies objects by using the point's closest distance from the training data to classify them as much as data [44]. It keeps track of all available points and classifies new ones using similar distance functions.

K-NN has been regarded as a lazy learning method. As no assumptions of distribution have been made, that's why it is called a "non-parametric" method. It can be reiterated as the model structure is according to the dataset. When any model generation algorithm does not require any training data points, that can be referred to as a "lazy" algorithm. So, the testing phase has gone through all available training data. This boosts up the training process while slowing down and expanding the expense of testing.

In K-NN, K is the number of nearest neighbors. Generally an odd number is chosen as the value of K, when the class count is 2. The algorithm is referred to as the nearest neighbor algorithm when the value of K is 1. K-NN engages the Euclidean distance between data points to find out the nearest neighbors. That's why data must be scaled.

$$\text{Euclidean distance between point } x \text{ and } y = \sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

Here, k = number of related neighbors.

Usually this algorithm should be run multiple times with different values of k to choose an optimal value which will minimize the error.

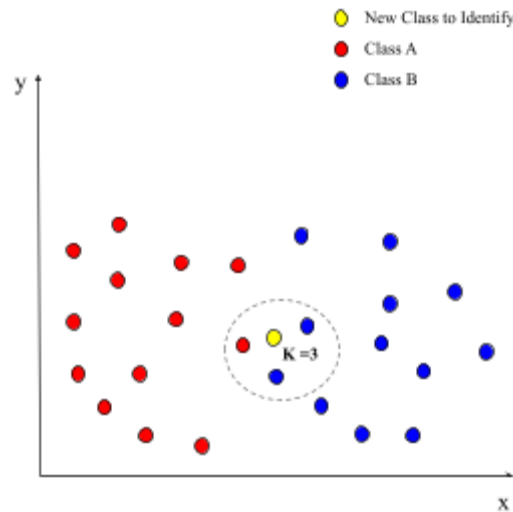


Figure 8: 3-nearest neighbours

5.5.4 Support Vector Machine (SVM)

This classifier's purpose is to locate a line or (n-1) dimension hyper-plane that divides the two classes in n-dimensional space [13]. Support Vector Machine distributes the training

examples in such a way, so that there exists a significant distance between the two categories and the objective of this classifier is to maximize the distance as much as possible. Unknown instances are then categorized into the same space and classified according to the side of the breach on which they stumble on. SVMs utilize a kernel trick that maps the feature variables into high-dimensional feature spaces.

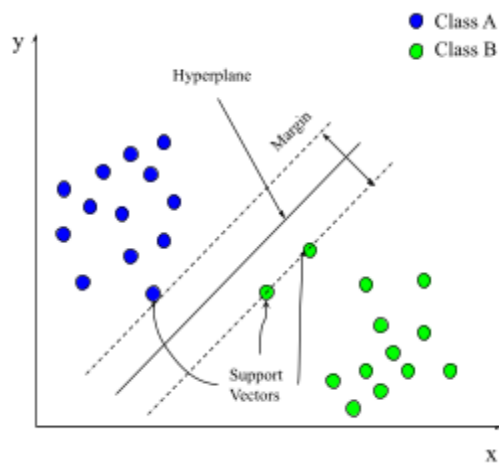


Figure 9: Support Vectors, Hyperplane, Margin

Support Vectors: The support vectors are the nearest data points to the hyperplane. This system first computes the margins and those points will better define the separation line. These points are more crucial to this algorithm's construction process .

Hyperplane: It is a decision plane which makes the separation between various class variables.

Margin: The distance between the two groups on the class points that are closest to each other. A good margin covers a large area, on the other hand a smaller distance is considered a bad margin.

Usually various kernel functions transform a less dimensional input space into a greater dimensional space in this case. Following are some kernel functions that can be used with the algorithm as hyperparameter..

- Linear
- Polynomial
- Radial Basis Function

5.5.5 Naive Bayes (NB)

The Naive Bayes is one of the most basic supervised learning algorithms which is treated as quick, accurate and trustworthy. These classifiers outperform other algorithms in accuracy and speed on a huge dataset. This classifier estimates a feature's importance on a class taking into consideration that each class is independent of other features. Because this assumption makes calculation easier, it is regarded as naive. Class conditional independence is the term for this assumption. The posterior probability $P(a|C)$ can be calculated from $P(a)$, $P(C)$ and $P(C|a)$ using the following equation.

$$P(a|C) = \frac{P(C|a)P(a)}{P(C)}$$

Here,

$P(a)$ = Likelihood of hypothesis a being true.

$P(C)$ = Probability of data.

$P(a|C)$ = Probabilistic ratios of hypothesis a given C .

$P(C|a)$ = Probability of C given that hypothesis a was true.

5.5.6 Logistic Regression (LR)

It is the most elementary and vastly empowered knowledge engineering method for two-class classification. The relation between one dependent variable and independent features is described and estimated using this. It's a method for predicting binary classes based on statistics. The term “dichotomous” states that there can be only two potential classes. It determines the probabilistic data of an event happening. When the target variable is categorical, then Linear Regression has been used to predict the occurrence to happen. The logit function aids the logistic regression to predict the likelihood of a binary event occurring [46].

Linear Regression Equation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

Here, y is a variable that depends on others.

$X_1, X_2 \dots$ and X_n are explanatory variables.

Sigmoid Function as follows:

$$p = \frac{1}{(1 + e^{-y})}$$

Apply this function on linear regression:

$$p = \frac{1}{(1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)})}$$

5.6 Comparative Analysis

Several performance indicators have been introduced here to assess the performance of Machine Learning classifiers. Performance measures related to classifications are presented here as the work solely deals with classification difficulties. Accuracy, F-Score, Recall, Precision and ROC are the metrics that are utilized for comparing the efficiency of each classifier in this work. Table-4 defines the accuracy matrix [47].

Table 4: Performance Metrics

Metrics	Definitions	Formulas
Accuracy (A)	It represents the count of correctly classified instances over the total instances count.	$\frac{(TN + TP)}{(TN + TP + FN + FP)}$
Precision (P)	The fraction of accurately predicted positive records to the total predicted positive instances.	$\frac{TP}{(TP + FP)}$
Recall (R)	The Proportion of faultlessly predicted positive outcomes over all observations in actual class.	$\frac{TP}{(TP + FN)}$
F1 score	Weighted average of P & R	$\frac{2 * (Precision * Recall)}{(Precision + Recall)}$
Receiver Operating Curve (ROC)	The Curves are to differentiate the fruitfulness of tests.	

Here,

TP: Count of identified positive instances in the positive set.

TP: Count of classified negative instances in the negative group.

FP: Count of identified positive records in the negative set.

FN: Count of identified negative records in the positive group.

Six different machine learning algorithms - Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbours, Naive Bayes & Logistic Regression are applied in this research work. 80% of the dataset has been used as train dataset and the remaining has been used for testing the mentioned algorithms. Then 10-fold cross validation is applied to check the performance of each classifier. This dataset has gone through multiple cleaning tasks in the entire process. After applying SMOTE resampling technique the dataset contains 640 instances instead of 520. Dimensions also have been reduced to 10 from 16 by using feature selection techniques. Table 5 states the processed dataset.

Table 5: Processed Dataset

	No of Attributes	No of Instances
Processed Dataset	10	640

Table 6: Comparison of correct & incorrect instances using cross validation and splitting

Classifiers	K fold Cross Validation		Train/Test Splitting	
	Correct	Incorrect	Correct	Incorrect
Decision Tree	614	26	123	5
Random Tree	615	15	125	3
K Nearest Neighbors	550	90	115	13
Logistic Regression	588	52	121	7
Naive Bayes	572	68	121	7
Support Vector Machine	557	83	125	3

10 fold cross validation and 80/20 % train/test split method have been applied on the processes dataset. Each classifier has been tested with both test methods individually. The classified instances at hand, Table-8 evaluates the performance of classifiers. Out of a total number of cases, the performance of each method is measured in consideration of correctly classified instances & incorrectly classified instances. All algorithms perform well in considered methodologies. However, the Random Forest predicts more instances as correct than other algorithms. It classifies 615 records as diabetic positive on fold cross validation and 125 records using % split technique. The Decision Tree classifier gives the second best accuracy level in cross validation technique. It predicts 614 instances correctly and 26 instances incorrectly.

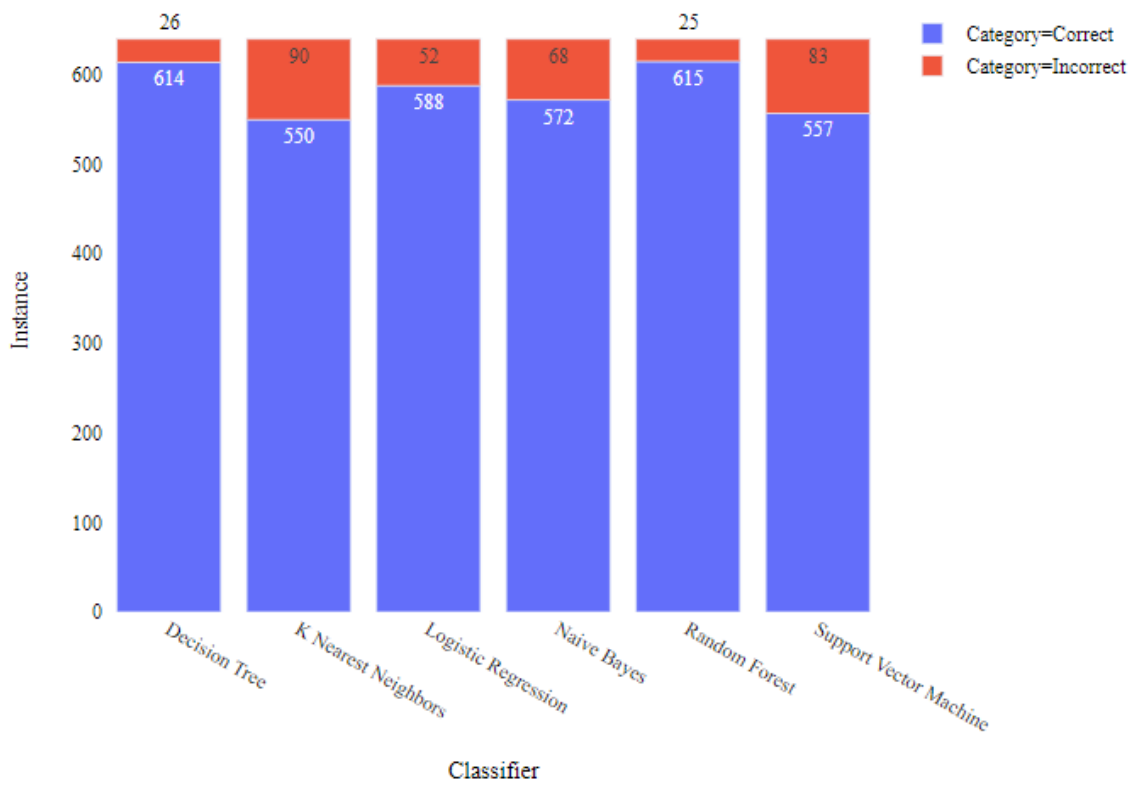


Figure 10: Classifiers' Performance using cross validation

Figure 9 and Figure 10 are the visual representation of performance of different classifiers using both evaluation techniques.

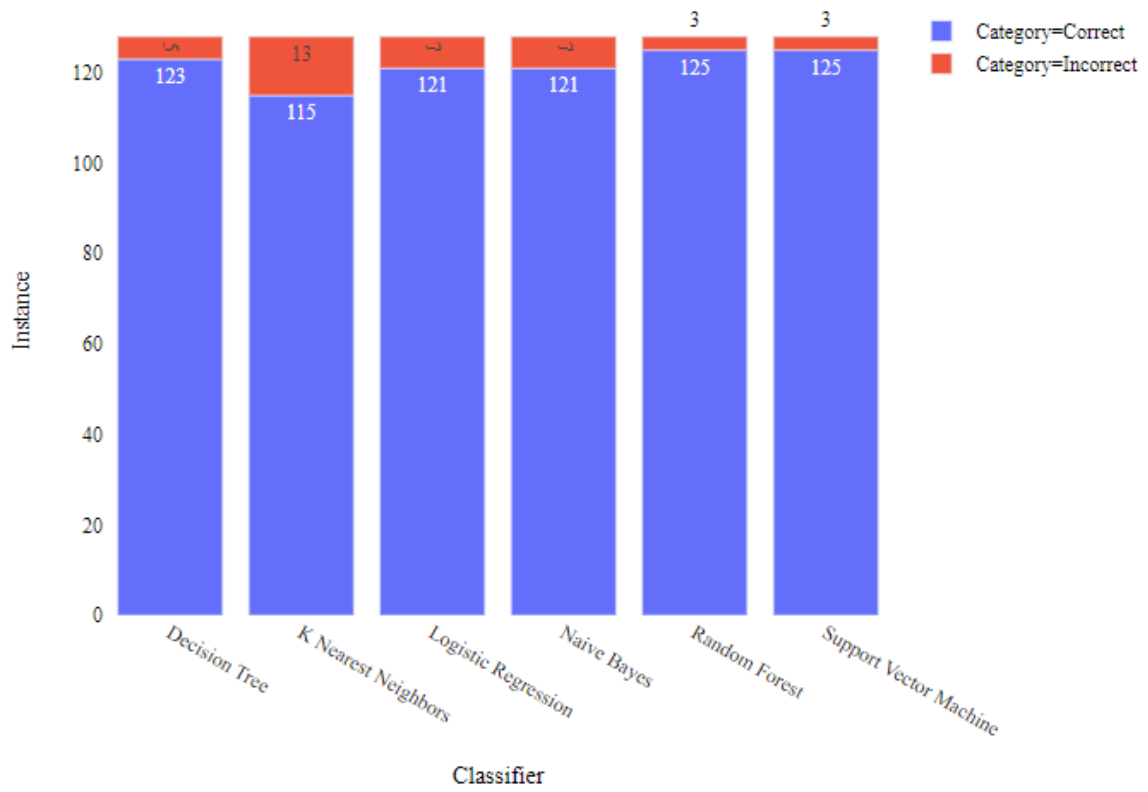


Figure 11: Classifiers' Performance using split

Table 7: Comparison of performance metrics

Classifiers	Split Accuracy	Cross Validation Accuracy	Precision	Accuracy	F1 Score	ROC
Decision Tree	0.960938	0.949170	0.969231	0.954545	0.961832	0.961144
Random Forest	0.984375	0.962821	0.984848	0.983849	0.982412	0.984360
K-Nearest Neighbors	0.882812	0.894796	0.947368	0.818182	0.878049	0.884897
Logistic Regression	0.937500	0.906335	0.953125	0.924242	0.938462	0.937928
Naive Bayes	0.945312	0.884842	0.953846	0.939394	0.946565	0.925503
Support Vector Machine	0.976562	0.949284	0.984615	0.969697	0.977099	0.976784

Table 7 shows the detailed comparison of various classifier's efficiency using both split techniques. RF gives the highest accuracy level of 98.43% using split techniques whereas it decreases to 96.28% in cross validation. The Support Vector Machine classifier with

specific kernel function (rbf) achieves 97.66% accuracy in percentage split and 94.93% in cross validation techniques respectively.

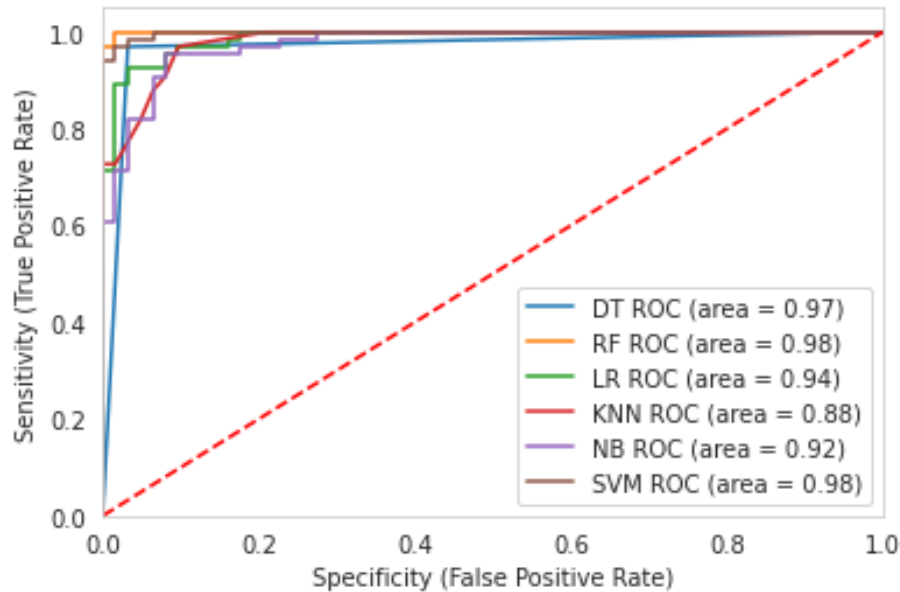


Figure 12: ROC curves of six algorithms

It can be advantageous to sum up the accomplishment of each classifier into a single measure when comparing different classifiers. The Receiver Operating Characteristics, or ROC, is a graphical representation of sensitivity against (1- Specificity) or a comparison of true positive rate and false positive rate. It is used to visualize a classifier's performance at various thresholds in order to discover the classifier's best threshold point. Figure 11 is the graphical representation of the FP rate vs the TP rate for a number of threshold values between 0 and 1. Basically, it plots the false alarm rate vs the hit rate. As Random Forest produces curves adjacent to the y-axis, it stipulates better accuracy.

Confusion matrix is a snapshot of predicted outcomes of the classification problems. The number of exact and wrong predictions is summed and divided by class using count values. It's a statistic used to quantify the performance of an expert classifier. Recall, Specificity, Accuracy, and Precision are considered as most significant predictive analysis and those are visualized by confusion matrices. In the literature [48], both variants are available, where each row of the matrix represents examples in an actual class, on the other hand each column represents instances in a predicted class. These are crucial in this manner as they open a path for values such as True Positives, False Positives, True Negatives, and False Negatives to place side by side in a straightforward manner. This enables for a more in-depth examination than merely looking at the percentage of correct classifications. The confusion matrix of both evaluation techniques applied in this work, have been visualized on figure 12 and 13.

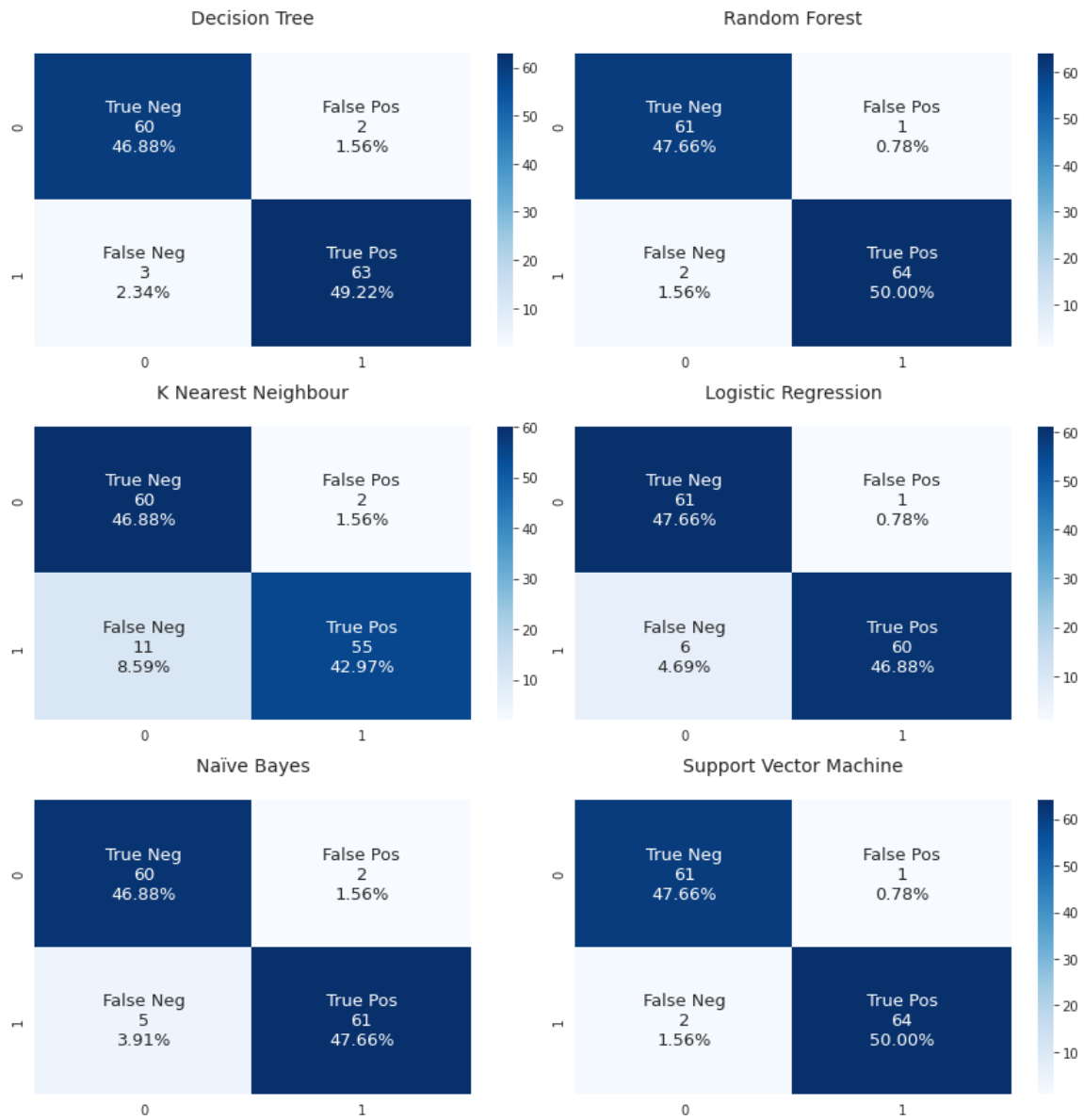


Figure 13: Confusion matrix using split technique

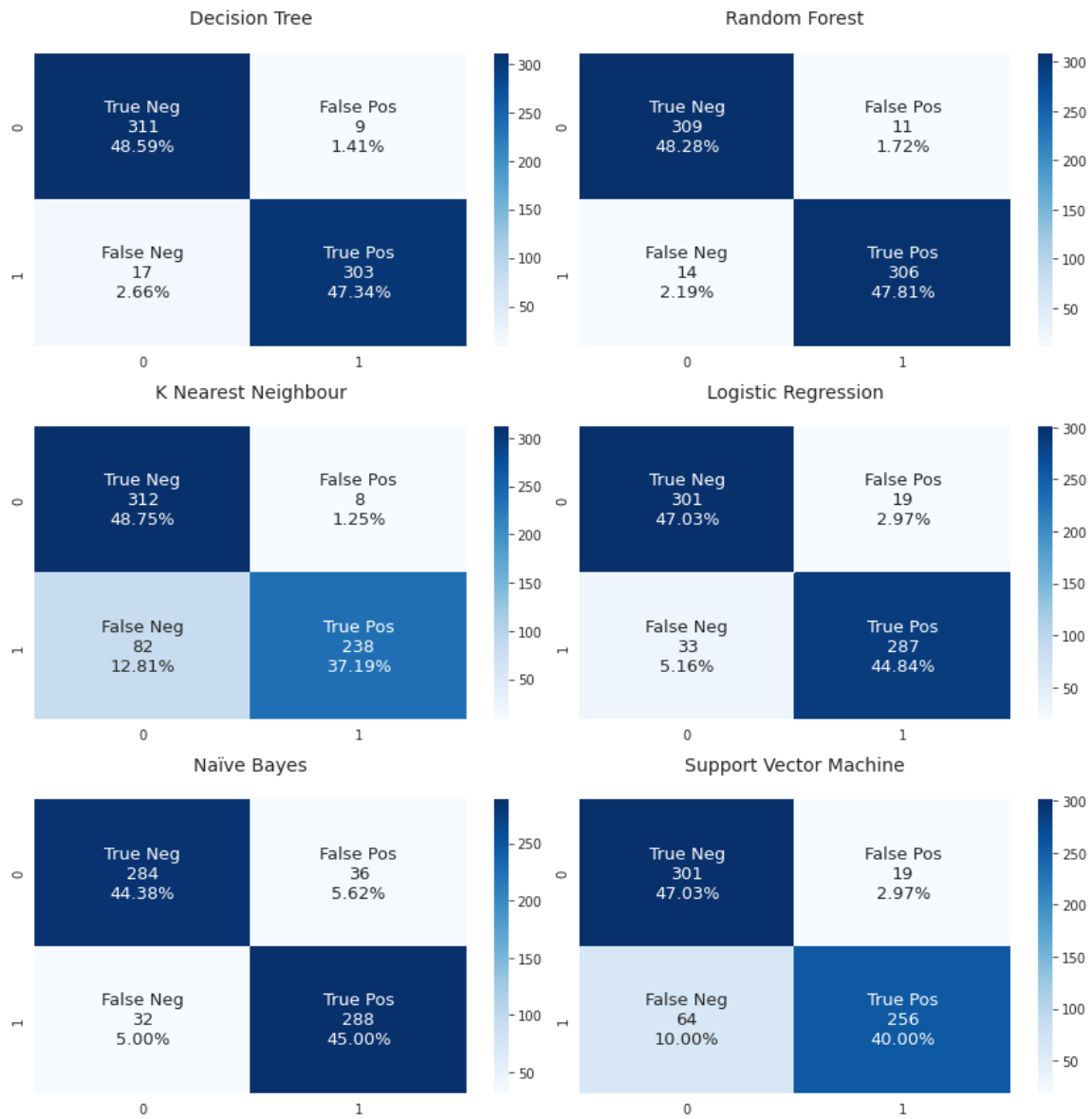


Figure 14: Confusion matrix of six classifiers using cross validation method

5.7 Conclusion

Different classifiers can perform well in various situations under many constraints. In this research, multiple factors have been considered to find out the efficacy of some specific machine learning techniques. F-score is one of the useful performance measurement metrics. It is visible on Table 7 that Random Forest has the highest F1 score among all applied classifiers. Support vector machine algorithm is close behind with a score of 0.98. From the above depicted two confusion matrices, a conclusion can be drawn that Random Forest outperforms other classifiers that are mentioned in this work. It correctly predicts 125 patients among 128 instances in split technique. In contrast, K nearest neighbours could detect 115 out of 128 instances. Random Forest classifier has detected the diabetic patients using only 10 features among 16 features with accuracy level 98.38%. Another supervised machine learning algorithm, Support Vector Machine also predicts diabetes with lower accuracy level 97.97% than the former one, but it is not negligible in real world scenarios. In this work, the support vector machine has been used with radial kernel function. K nearest neighbors algorithm performs poorly among all classifiers with 81.81% accuracy level and 0.88 ROC score. So, comparing all the experimental results, the best combination to predict diabetes is the Random Forest algorithm.

CHAPTER 6

CRITICAL APPRAISAL

6.1 Introduction

Critical appraisal is one step in the process of evidence-based clinical practice. We require critical appraisal abilities to evaluate what is the "best" evidence. These skills will help us grasp study methodologies and outcomes, as well as assess the research's quality. Most research isn't faultless, and critical evaluation isn't an exact science; it won't always provide us with the "correct" answer. However, it can assist us in determining whether a piece of research that has been reported is suitable for decision-making. It is a method of evaluating outcomes as well as the utility of published research papers in a systematic fashion.

6.2 SWOT Analysis

In this work, critical appraisal has been conducted through SWOT analysis. It stands for Strengths, Weaknesses, Opportunities, and Threats, it describes the project's strengths and weaknesses, as well as the opportunities and threats it confronts. SWOT analysis can be used to develop and expand on the findings of an external environment inquiry. It aims to identify external dangers and opportunities, as well as the strengths and weaknesses of current resources and activities that could be exploited to seize opportunities or avoid threats. The analysis allows for informed conclusions about the existing records service's ability to contribute to the development of a new program. SWOT analysis is a strong tool for understanding how an ongoing plan will perform.

6.2.1 Strengths

The study does not rely on statistics data, but rather on the replies of participants and their subjective experiences with the issue. The participants were selected through purposive convenience sampling, with no incentives. They needed to answer only those questions that can be expressed as “Yes/No” values. It was easy for them to share the answers without any clinical test. This dataset is also validated by the doctor. This is a strong motivation for choosing this dataset for this research.

6.2.2 Weakness

The dataset on hand contains only 520 instances. More valid dataset ensures a more accurate result. This can be considered as a weakness of this research. Moreover, this dataset is not balanced. Imbalanced dataset may lead to a wrong prediction result. If the data can be collected through a systemic and central approach, more valid data can be processed to gain the best efficiency of the work.

6.2.3 Opportunities

This can be observed, practically all of the system's flaws arise from a lack of relevant data. However, if enough hospitals adopt the system and correctly save their data, systems like this one can be used more efficiently and effectively. In the future, a predictive model like this one could be utilized in clinical information systems for patient prediction and treatment planning. This work can help doctors to make automatic preliminary diagnoses of issues, allowing them to focus more time on patients who are more likely to develop complications. Furthermore, tools can be developed to assist patients in gaining an understanding of their current situation without having to contact the doctor on a regular basis.

6.2.4 Threats

This research can be extended with the help of other concerned organizations. In the meantime, if any renowned organization conducts the same type of research, then that can be considered as a threat for this work.

6.3 Conclusion

Critical appraisal emphasizes objective evaluation of the utility of information - skills in critical appraisal are used to evaluate published research, but all evidence should be evaluated to determine its value. When critical assessment reveals a dearth of strong evidence, it can be discouraging. This research is well organised and produces useful outcomes. The objectives and background information are perfect for this study, which aids other researchers with lots of reasons to carry it out. The facts are quite well, and the commentary supports them.

CHAPTER 7

CONCLUSION

7.1 Conclusion

Early detection is a better approach rather than curing any diseases. Diabetes is one of the few diseases for which there is currently no cure. Furthermore, several causative and confounding factors, such as obesity, contributed to the formation of this deadly disease. Obesity has been linked to diabetes, as evidenced by numerous studies [49,50]. Hypertension is frequent in diabetic people, and DM aids in the facilitation of hypertension [51,52]. Diabetes and high blood pressure complement one another since they share physiological characteristics [37].

People in our country are so preoccupied with their everyday lives that they have little time to think about their health, despite the fact that a healthy lifestyle is one of our basic demands. In Bangladesh, 3 in 5 patients with diabetes were uninformed that they had it and only a 1/3 were receiving adequate care [22]. To improve diabetes outcomes, additional efforts are sought to generate diabetes awareness, treatment, and control. This work is intended to detect diabetes at an early stage because the number of diabetic patients in our country is rapidly increasing.

In this research, UCI repository dataset has been considered as a dataset which has been collected by direct questionnaires of diabetes patients. Some processing tasks have been applied to extract information which can be useful to classify the dataset. The dataset on hand has unequal instances of target class; To overcome this imbalance issue, multiple resampling methods were examined. Here, SMOTE for oversampling the minority class has been utilized to balance the dataset. Then multiple machine learning classifiers like Decision Tree, Support Vector Machine, Random Forest, Logistic Regression, Naive Bayes have been evaluated here. Moreover, k-fold cross-validation where k is 10 and eighty-twenty percentage split techniques were utilized. In both splitting techniques, the

Random Forest algorithm can predict diabetes more accurately and k nearest neighbors shows poor performance. The second accurate classifier is Support Vector Machine which also scores near about the former.

7.2 Further Suggested Works

There are numerous significant elements to consider when predicting diabetes. There could be other risk factors that are overlooked here. As this dataset has been constructed by questionnaires, most of the question's answer contains only "Yes/No" value. Other significant factors include gestational diabetes, insulin, glucose, and skin thickness, BMI etc. which may contain some numerical values, should be incorporated to get more accurate results. Moreover, this dataset contains only 520 records of diabetic patients. More instances may contribute more for upgrading efficiency of this system.

This research's goal was to assist diabetes people in living a healthier life. Furthermore, while there has been a lot of research done on this topic, there have only been a few real-life systems. As a result, rather than limiting the notion to research, it is conceivable to construct an entire application in the near future using correct techniques. This research investigates the subject of machine learning as a means of predicting diabetes, with the goal of eventually developing a complete clinical system that will benefit both patients and doctors. Automatic diabetes analysis as well as applying some other classifiers on this process will be the expansion and enhancement of this work.

REFERENCES

- [1] T.M. Alam, et al., Informatics in medicine unlocked a model for early prediction of diabetes, *Inform. Med. Unlocked* 16 (2019) 100204.
- [2] D. Sisodia, D.S. Sisodia, Prediction of diabetes using classification algorithms, *Procedia Comput. Sci.* 132 (2018) 1578–1585.
- [3] N.P. Tigga, S. Garg, Predicting type 2 Diabetes using Logistic Regression accepted to publish in: *Lecture Notes of Electrical Engineering*, Springer.
- [4] Salim Amour Diwani, Anael Sam, Diabetes forecasting using supervised learning techniques, *Adv. Comput. Sci.: Int. J.* [S.I.] (ISSN:2322-5157) (2014) 10–18, Available at: <http://www.acsij.org/acsij/article/view/156>.
- [5] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang, Predicting Diabetes Mellitus with Machine Learning Techniques, Vol. 9, *Frontiers in genetics*, 2018, p. 515, <http://dx.doi.org/10.3389/fgene.2018.00515>.
- [6] World Health Organization. Diabetes country profiles, 2016: Bangladesh. France: WHO (2016).
- [7] Rakibul M. Islam, Md. Nuruzzaman Khan, Prevalence of diabetes and prediabetes among Bangladeshi adults and associated factors: Evidence from the Demographic and Health Survey, 2017-18.
- [8] Shariful Islam SM, Lechner A, Ferrari U, et al Healthcare use and expenditure for diabetes in Bangladesh *BMJ Global Health* 2017;2:e000033
- [9] Nai-Arun, N., Sittidech, P., 2014. Ensemble Learning Model for Diabetes Classification. *Advanced Materials Research* 931-932, 1427–1431. doi:10.4028/www.scientific.net/AMR.931-932.1427.
- [10] Orabi, K.M., Kamal, Y.M., Rabah, T.M., 2016. Early Predictive System for Diabetes Mellitus Disease, in: *Industrial Conference on Data Mining*, Springer. Springer. pp. 420–427.
- [11] Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K., 2016. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. *Procedia Computer Science* 82, 115–121. doi:10.1016/j.procs.2016.04.016.
- [12] Hina S., Anita S. and Sohail Abul Sattar “Analyzing diabetes datasets using data mining” *Journal of Basic & Applied Sciences*, 13, pp 466-471, 2017.

- [13] Prema N. S., Varshith V. and Yogeswar J. “Prediction of diabetes using ensemble techniques” International Journal of Recent Technology and Engineering (IJRTE), 2277-3878, Volume-7, Issue-6S4, April 2019.
- [14] Kadhm M. S., Ikhlas W. G. and Duaa E. M. “An accurate diabetes prediction system based on K-means clustering and proposed classification approach” International Journal of Applied Engineering Research, Volume 13, pp. 4038-4041, Number 6, 2018.
- [15] Nilasi, M., Ibrahim O., Dalvi M., Ahmedi H. and Shahmoradi L., “Accuracy improvement for diabetes disease classification: a case on a public medical dataset” Fuzzy Information Engineering, vol. 9, pp. 345-357, 2017.
- [16] Lee, B.J., Kim, J.Y.: Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. *IEEE J. Biomed. Health Inform.* 20(1), 39–46 (2016).
- [17] Lee BJ, Ku B, Nam J, Pham DD, Kim JY. Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes. *IEEE J Biomed Health Inform.* 2014 Mar; 18(2):555-61. doi: 10.1109/JBHI.2013.2264509. PMID: 24608055.
- [18] Diagnosis and classification of diabetes mellitus; American Diabetes Association; *Diabetes Care*, 32 (Suppl. 1) (2009), pp. S62-S67
- [19] Miller, Y. D. & Dunstan, D. W. The effectiveness of physical activity interventions for the treatment of overweight and obesity and type 2. diabetes. *J Sci Med in Sport* 7(1), 52–59 (2004).
- [20] Hayes, C. & Kriska, A. Role of physical activity in diabetes management and prevention. *J Am Diet Assoc.* 108(4), S19–S23 (2008).
- [21] Herbst, A. et al. Effects of regular physical activity on control of glycemia in pediatric patients with type 1 diabetes mellitus. *Arch Pediatr Adolesc Med.* 160(6), 573–577 (2006).
- [22] Prevalence of diabetes and prediabetes among Bangladeshi adults and associated factors: Evidence from the Demographic and Health Survey, 2017-18.
- [23] Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care.* 1997;20:1183–97.
- [24] World Health Organization. Global report on diabetes. France: WHO, <https://apps.who.int/iris/handle/10665/204871> (2016).

- [25] Atlas, Diabetes International diabetes Federation. 10th edi. IDF diabetes Atlas, 2019.
- [26] Guariguata L, Whiting DR, Hambleton I, et al. Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes res clin pract* 2014;103:137–49.doi:10.1016/j.diabres.2013.11.002.
- [27] Ramachandran A, Snehalatha C, Ma RCW; Diabetes in south-east Asia: an update. *Diabetes Res Clin Pract*2014;103:231–7.doi:10.1016/j.diabres.2013.11.011pmid:http://www.ncbi.nlm.nih.gov/pubmed/24300015.
- [28] Chowdhury MZI, Anik AM, Farhana Z, *et al* Prevalence of metabolic syndrome in Bangladesh: a systematic review and meta-analysis of the studies. *BMC Public Health* 2018;18:308. doi:10.1186/s12889-018-5209-zpmid:http://www.ncbi.nlm.nih.gov/pubmed/29499672.
- [29] Jayawardena R, Ranasinghe P, Byrne NM, *et al*. Prevalence and trends of the diabetes epidemic in South Asia: a systematic review and meta-analysis. *BMC Public Health* 2012;12:380. doi:10.1186/1471-2458-12-380pmid:http://www.ncbi.nlm.nih.gov/pubmed/22630043.
- [30] Shaw JE, Sicree RA, Zimmet PZ . Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res Clin Pract* 2010;87:4–14.doi:10.1016/j.diabres.2009.10.007pmid:http://www.ncbi.nlm.nih.gov/pubmed/19896746.
- [31] International Diabetes Federation. IDF Diabetes Atlas, 9th edn. *Brussels. Belgium*:2019. Available at: <https://www.diabetesatlas.org>, accessed 18 November 20202019.
- [32] International centre for diarrhoeal disease research Bangladesh. Available:<https://www.icddr.org/news-and-events/press-corner/media-resources/non-communicable-diseases> [Accessed 06 Sep 2019].
- [33] Cho NH, Whiting D, Forouhi N, Guariguata L, Hambleton I, Li R, et al. IDF DIABETES ATLAS. 7th ed. Hallado en: <http://www.idf.org/diabetesatlas/5e/es/prologo>: International diabetes federation; 2015.
- [34] Cho NH, Kirigia J, Mbanya JC, Ogurstova K, Guariguata L, Rathmann W, et al. IDF DIABETES ATLAS. 8th ed; 2017.
- [35] World Health Organization. Bangladesh health system review. Manila: WHO Regional Office for the Western Pacific; 2015.
- [36] Bangladesh Bureau of Statistics SaID, Ministry of Planning. national accounts statistic, 2018.

- [37] Anwer, Z. et al. Hypertension management in diabetic patients. *Eur Rev Med Pharmacol Sci.* 15(11), 1256–1263 (2011).
- [38] International Diabetes Federation Task Force on Diabetes Health Economics, *Diabetes health economics: facts, figures, and forecasts*, International Diabetes Federation, Brussels, 1997.
- [39] R. Shobhana, P.R. Rao, A. Lavanya, R. Williams, V. Vijay, A. Ramachandran, Expenditure on health care incurred by diabetic subjects in an industrializing country — a study from southern India, *Diabetes Res. Clin. Pract.* (2000) in press.
- [40] Han J., Kamber M. and Pei J.(3rd ed.), *Data mining: concepts and techniques*; The Morgan Kaufmann series in data management systems (2011).
- [41] M. M. Faniqul Islam, R. Ferdousi, S. Rahman, H.Y. Bushra, Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. DOI: [10.1007/978-981-13-8798-2_12](https://doi.org/10.1007/978-981-13-8798-2_12)
- [42] Iyer, A., S, J., Sumbaly, R., 2015. Diagnosis of Diabetes Using Classification Mining Techniques. *International Journal of Data Mining & Knowledge Management Process* 5, 1–14.
- [43] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). *The Elements of Statistical Learning* (2nd ed.). Springer. ISBN 0-387-95284-5.]
- [44] Zou Q., Qu K., Luo Y., Yin D., Ju Y. and Tang H. “Predicting diabetes mellitus with machine learning techniques” *Front. Genet.* 9:515. doi: 10.3389/fgene.2018.00515, 2018.
- [45] *Machine Learning Mastery*, available at <https://machinelearningmastery.com/>, last accessed on 11-12-2021 at 01:13 PM.
- [46] *Data Science Tutorials*, available at <https://www.datacamp.com/community/tutorials>, last accessed on 27-11-2021 at 09:47 PM.
- [47] Swapna G., Vinayakumar R. and Soman K.P. “Diabetes detection using deep learning algorithms” *ICT Express* 4, pp 243–246, 2018.
- [48] Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". *Journal of Machine Learning Technologies.* 2 (1): 37–63. S2CID 55767944.
- [49] Al-Goblan, A. S., Al-Alfi, M. A. & Khan, M. Z. Mechanism linking diabetes mellitus and obesity. *Diabetes Metab Syndr Obes.* 7, 587 (2014).

[50] Eckel, R. H. et al. Obesity and type 2 diabetes: what can be unified and what needs to be individualized? *J Clin Endocrinol Metab* 96(6), 1654–1663 (2011).

[51] Song, J. et al. Management of hypertension and diabetes mellitus by cardiovascular and endocrine physicians: a China registry. *J Hypertens*. 34(8), 1648 (2016).

[52] Petrie, J. R., Guzik, T. J. & Touyz, R. M. Diabetes, hypertension, and cardiovascular disease: clinical insights and vascular mechanisms. *Can J Cardiol*. 34(5), 575–584 (2018).

Turnitin Originality Report

Processed on: 25-Jan-2022 13:02 +06
 ID: 1747705908
 Word Count: 12806
 Submitted: 1

211-25-954 By Mithun Mondal

Similarity Index

18%

Similarity by Source

Internet Sources: 12%
 Publications: 10%
 Student Papers: 9%

2% match (student papers from 28-Mar-2018)

Class: Article 2018
 Assignment: Journal Article
 Paper ID: [937594737](#)

1% match (Internet from 07-Apr-2021)

http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5423/162-15-8250%20%2824_%29.pdf?isAllowed=y&sequence=1

1% match (Internet from 29-Aug-2020)

http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/4181/P15391%20%2828_%29_.pdf?isAllowed=y&sequence=1

1% match (Internet from 02-Apr-2021)

http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5364/162-15-739%20%2830_%29.pdf?isAllowed=y&sequence=1

1% match (Internet from 16-Sep-2019)

<http://ikee.lib.auth.gr/record/292819/files/loannis%20Kavakiotis%20et%20al.pdf>

1% match (publications)

[Mariette Awad, Rahul Khanna. "Efficient Learning Machines", Springer Nature, 2015](#)

1% match (Internet from 25-Jul-2021)

<https://bmjopen.bmj.com/content/10/9/e036086.full>

1% match ()

[Afsana Afroz, Khurshid Alam, Liaquat Ali, Afsana Karim et al. "Type 2 diabetes mellitus in Bangladesh: a prevalence based cost-of-illness study", BMC Health Services Research](#)

< 1% match (Internet from 07-Apr-2021)

http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5394/162-15-8222%20%2823_%29.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 19-Jan-2022)

<http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/6815/161-15-7655%20%2819%25%29%20clearance.pdf?isAllowed=y&sequence=1>

< 1% match (Internet from 03-Jan-2020)

<http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/3342/152-15-565%3d21%25.pdf?isAllowed=y&sequence=1>

< 1% match (Internet from 07-Dec-2020)

<http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5339/172-15-9961%20%2829%25%29.pdf?isAllowed=y&sequence=1>

< 1% match (Internet from 07-Apr-2021)

http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5144/152-15-5619%20%2820_%29.pdf?isAllowed=y&sequence=1

https://www.turnitin.com/newreport_printview.asp?eq=1&eb=1&esm=10&oid=1747705908&sid=0&n=0&m=2&svr=52&r=22.63080576144205&lang=e... 1/19