A picture containing text

Description automatically generated

UNIVERSITY OF HERTFORDSHIRE

School of Physics, Engineering and Computer Science

Advanced Computer Science Masters Project

Title

**Early Detection of Diabetes Using Machine Learning**

**Final Project Report (FPR)**

Name:- Venkatesh Tamminaina

Student Id- 19062859

Course:- MSc Data Science and Analytics Course( Sandwich)

Module:- 7COM1039-0109-2021-Advanced Computer Science Masters Project

Supervised by:- Professor Iain Werry

**Abstract**

Machine Learning has a versatile application and especially in healthcare and medical science. Machine learning is used in medical science for automation, especially for detecting diseases. Different diseases can be predicted automatically with the input of the symptoms. In this study, machine learning has been applied to detect diabetes. A Dataset with 768 observations of patients with and without diabetes has been found from Kaggle, which is considered for this research. The crucial symptoms have been chosen using pairwise correlation, and the data has been prepared with those symptoms. Five classifiers have been selected and applied to the actual and resampled data, and finally, the detection outcome has been achieved. The detection accuracies have been achieved and compared from the application of classifiers. The comparison shows that the decision tree has produced the highest accuracy in detecting diabetes disease with an accuracy rate of 98.76%, which is also higher than the existing models.

MSc Final Project Declaration

This report is submitted in partial fulfilment of the requirement for the degree of Master of Science in 7COM1039-0109-2021-Advanced Computer Science Masters Project at the University of Hertfordshire (UH).

It is my own work except where indicated in the report. I did not use human participants in my MSc Project.

I hereby give permission for the report to be made available on the university website, provided the source is acknowledged.

**List of Abbreviations**

LR: LogisticRegression

SVM: Support Vector Machine

PSO: Particle Swarm Optimization

KNN: K-Nearest Neighbours

ANN: Artificial Neural Network

**List of Figures**

[Figure 1 Diabetes Detection Method 28](#_Toc106289978)

[Figure 2 Data Reading 37](#_Toc106289979)

[Figure 3 Encoding Target Features 37](#_Toc106289980)

[Figure 4 Data Information and Missing values 38](#_Toc106289981)

[Figure 5 Analysis of Age on Diabetic Possibility 38](#_Toc106289982)

[Figure 6 Analysis of Glucose and BMI on Diabetic Possibility 39](#_Toc106289983)

[Figure 7 Average Age of patients 39](#_Toc106289984)

[Figure 8 Average Glucose Level 39](#_Toc106289985)

[Figure 9 Average BMI Level 40](#_Toc106289986)

[Figure 10 Average Insulin Level 40](#_Toc106289987)

[Figure 11 Feature Correlation 41](#_Toc106289988)

[Figure 12 Classification Report for Decision Tree with Resampled Data 44](#_Toc106289989)

[Figure 13 Comparison of Accuracies 45](#_Toc106289990)

[Figure 14 Comparison of Precisions 45](#_Toc106289991)

[Figure 15 Comparison of Recalls 46](#_Toc106289992)

[Figure 16 Comparison of F1-Scores 46](#_Toc106289993)

[Figure 17 Comparison with Previous Approaches 47](#_Toc106289994)

**List of Tables**

[Table 1 Research Comparison 26](#_Toc106289967)

[Table 2 Result of Diabetes Detection with Original Data 42](#_Toc106289968)

[Table 3 Confusion Matrixes for all classifiers for original data 43](#_Toc106289969)

[Table 4 Result of Diabetes Detection with Resampled Data 43](#_Toc106289970)

[Table 5 Confusion Matrixes for all classifiers for resampled data 44](#_Toc106289971)

Table of Contents

[Chapter-1 Introduction 10](#_Toc109417323)

[1.1 Overview of Research 10](#_Toc109417324)

[1.2 Diabetes Disease 10](#_Toc109417325)

[1.2.1 Causes of Diabetes 11](#_Toc109417326)

[1.2.2 Symptoms of Diabetes 11](#_Toc109417327)

[1.2.3 Risk Issues 12](#_Toc109417328)

[1.3 Research Question 12](#_Toc109417329)

[1.4 Hypotheses 13](#_Toc109417330)

[1.5 Research Aim 13](#_Toc109417331)

[1.6 Research Objectives 13](#_Toc109417332)

[1.7 Research Issues 14](#_Toc109417333)

[1.7.1 Legal Issues 14](#_Toc109417334)

[1.7.2 Ethical Issues 14](#_Toc109417335)

[1.7.3 Social Issues 14](#_Toc109417336)

[1.7.4 Security Issues 14](#_Toc109417337)

[1.8 Research Planning 15](#_Toc109417338)

[Chapter-2 Review of Previous Papers 16](#_Toc109417339)

[2.1 Effects of Diabetes 16](#_Toc109417340)

[2.2 Detection of Diabetes Disease 16](#_Toc109417341)

[2.2.1 Diagnosis Process 16](#_Toc109417342)

[2.2.2 Machine Learning 17](#_Toc109417343)

[2.3 Previous Research on Disease Detection 17](#_Toc109417344)

[2.4 Previous Research to Detect Diseased Concerning Symptoms 19](#_Toc109417345)

[2.5 Early Detection of Diabetes Disease 22](#_Toc109417346)

[2.6 Comparative Discussions of Previous Models 24](#_Toc109417347)

[Chapter-3 Research Methodology 26](#_Toc109417348)

[3.1 Proposed Methodology 26](#_Toc109417349)

[3.1.1 Proposed Method for Diabetes Detection 26](#_Toc109417350)

[3.2 Diabetes Data 27](#_Toc109417351)

[3.2.1 Data Description 27](#_Toc109417352)

[Chapter-4 Research Components and Artefact 29](#_Toc109417353)

[4.1 Programming Tool 29](#_Toc109417354)

[4.1.1 Tool Selection 29](#_Toc109417355)

[4.1.2 Reasons for Preferences 29](#_Toc109417356)

[4.2 Selection of Tool and Libraries 30](#_Toc109417357)

[4.3 Planned Technology to be Used 30](#_Toc109417358)

[4.3.1 Data Analysis 30](#_Toc109417359)

[4.3.2 Predictive Learning 30](#_Toc109417360)

[4.4 Planned Algorithms to be Applied 31](#_Toc109417361)

[4.4.1 Logistic Regression 31](#_Toc109417362)

[4.4.2 Adaptive Boosting Tree 31](#_Toc109417363)

[4.4.3 MLP Classifier 31](#_Toc109417364)

[4.4.4 Decision Tree Classifier 31](#_Toc109417365)

[4.4.5 Support Vector Machine 31](#_Toc109417366)

[4.5 Performance Evaluation Process 32](#_Toc109417367)

[4.5.1 Accuracy 32](#_Toc109417368)

[4.5.2 Precision and Recalls 32](#_Toc109417369)

[4.5.3 F1-Score 32](#_Toc109417370)

[4.6 Research Artefact 32](#_Toc109417371)

[4.6.1 Reading Data and Formulating Target 32](#_Toc109417372)

[4.6.2 Treating Missing Values 32](#_Toc109417373)

[4.6.3 Functions for Visualization 33](#_Toc109417374)

[4.6.4 Selecting Important Features 33](#_Toc109417375)

[4.6.5 Preparing Data and Split it 33](#_Toc109417376)

[4.6.6 Assigning and Tuning Algorithms 34](#_Toc109417377)

[4.6.7 Prediction 34](#_Toc109417378)

[4.6.8 Result Presentation 34](#_Toc109417379)

[Chapter-5 Analysis and Result 35](#_Toc109417380)

[5.1 Data Reading 35](#_Toc109417381)

[5.2 Explanatory Analysis 36](#_Toc109417382)

[5.2.1 Analysis of Age on Diabetic Possibility 36](#_Toc109417383)

[5.2.2 Analysis of Glucose and BMI on Diabetic Possibility 36](#_Toc109417384)

[5.2.3 The average age of patients 37](#_Toc109417385)

[5.2.4 Average Glucose Level 37](#_Toc109417386)

[5.2.5 Average BMI Level 37](#_Toc109417387)

[5.2.6 Average Insulin Level 38](#_Toc109417388)

[5.3 Feature Selection Process 38](#_Toc109417389)

[5.3.1 Feature Selection Process 38](#_Toc109417390)

[5.3.2 Feature Correlation and Heatmap 39](#_Toc109417391)

[5.3.3 Finally, Selected Features (Symptoms) 39](#_Toc109417392)

[5.4 Creating Training and Testing Set 40](#_Toc109417393)

[5.5 Diabetes Detection 40](#_Toc109417394)

[5.5.1 With Original Data 40](#_Toc109417395)

[5.5.2 With Resampled data 41](#_Toc109417396)

[5.6 Comparative discussion and comparison 42](#_Toc109417397)

[5.6.1 Comparison of Accuracies 42](#_Toc109417398)

[5.6.2 Comparison of Precisions 43](#_Toc109417399)

[5.6.3 Comparison of Recalls 43](#_Toc109417400)

[5.6.4 Comparison of F1-Scores 43](#_Toc109417401)

[5.7 Comparison with Previous Models 44](#_Toc109417402)

[Chapter-6 Discussion and Conclusion 45](#_Toc109417403)

[6.1 Discussion on Research Questions 45](#_Toc109417404)

[6.2 Discussion on Hypothesis 46](#_Toc109417405)

[6.3 Research Strength 46](#_Toc109417406)

[6.4 Research Limitations 46](#_Toc109417407)

[6.5 Future Scopes 46](#_Toc109417408)

[6.6 Conclusion 47](#_Toc109417409)

[References 48](#_Toc109417410)

[Appendix 53](#_Toc109417411)

# Introduction

## Overview of Research

Diabetes is a metabolic disease which escalates blood sugar levels in the body. The insulin hormone uses to move the sugar from the blood into the cells of humans. There the sugar is stored and used to generate energy. So, humans will achieve power for their work through this process. If anyone is affected by diabetes, insulin cannot be produced sufficiently, making them store less sugar in the blood (Asif et al., 2020). So, those people get less energy for their work and eventually get tired after some time. Different types of diabetes can be seen in humans for various reasons and ages. The pancreas is the hormonal gland where insulin is produced. In Type 1 diabetes, due to immune system attacks, the cells in the pancreas are destroyed. This will stop or reduce producing insulin from the pancreas. So, if the people are affected with this type of diabetes, they use to get tired after some work has been done (Costea, et al., 2021). This type of diabetes is seen in the aged, and almost 10% of total diabetes-affected people have this type of disease. In the context of Type-2 diabetes, the human body becomes resistant to insulin production. This causes the complete stop of the transaction of sugar from the blood to cells. Thus, sugar grows in the blood, which causes a rise in blood sugar. Prediabetes is generated at the time when the level of blood sugar rises in the blood below the level of Type-2 (Mohan & Jain, 2020). Gestational diabetes occurs in female patients during pregnancy due to high blood sugar.

Other reasons for diabetes are a rise in bad cholesterol in the blood converted to glucose and increased sugar levels in the blood. So, different factors influence the cause and level of diabetes (Cıhan & Coşkun, 2021). These problems are referred to as the symptoms of diabetes and can be seen at the early stage. Therefore, if observed minutely, diabetes can be detected precisely. Furthermore, as diabetes has no proven drug, detection at the early stage is essential so that the disease can be prohibited from growing in the body (Katarya & Jain, 2020).

## Diabetes Disease

Diabetes is the most common and general clinical term used for Diabetes Mellitus. It lies in that group of diseases that determines the level of a body's usage of sugar or glucose. Glucose is a vital source of energy supply to the body that helps build muscles and tissues (NIDDK, 2022). It also acts as an essential source of brain activators. Diabetes mellitus is a clinical condition of the body when the proper functioning of the pancreas gets hampered in the form of a lower insulin production rate, thereby causing hindrance in the glucose level controlling activity. The symptoms of diabetes are not very absurd. Still, some of the common ones like an Increase in thirst, frequent need for urination, excessive loss in weight, extreme feeling of fatigue, slow healing rates for wounds or cuts, prone to infections, etc. diabetes also causes a prolonged impact on the heart, kidneys, eyes, etc. The most significant characteristic of Diabetes is Hyperglycemia, or an increase in blood glucose level. The different types of diabetes include Type 1 diabetes, an autoimmune disease where the body's insulin-producing cells get destroyed, and the body invades itself. Type 2 diabetes is where the body neither produces nor responds to adequate glucose levels. Prediabetes (not officially confirmed as type 2 diabetes) and gestational diabetes (this mainly happens during pregnancy). The other uncommon forms of diabetes include Cystic fibrosis-associated diabetes, drug or chemical-caused diabetes, or monogenic diabetic syndrome. Nowadays, it occurs among nearly 34.2 million of the total population, and approximately 26% of adults above 65 years of age are suffering with diabetes.

### Causes of Diabetes

The process of digestion includes breaking down food into simpler substances. When we eat carbohydrates, our body breaks down into glucose. With the help of insulin, glucose enters our body's cells. Glucose provides energy to our body and helps properly function the organs. If the insulin production in the pancreas is inadequate or the body cannot take glucose into the cells, it builds extra sugar in our bloodstream. This causes diabetes (Cleveland Clinic, 2022). Diabetes can be controlled through medication, insulin injection, a balanced diet, and exercise. If poorly controlled, diabetes can damage many organs and tissues such as the heart, kidneys, nerves, and eyes. It may lead to coronary heart diseases, stroke, and high cholesterol. Kidney damage can lead to kidney failure, kidney transplant, or dialysis. Eye damage can cause blindness. It can also lead to hearing loss, depression, dementia, and dental problems.

The causes of diabetes differ for diabetes types.

* Type 1 diabetes- It destroys the pancreas, which helps digestion. Thus, most of the body's glucose is not processed correctly, leading to a lack of glucose in the blood and human cells.
* Type 2 diabetes and prediabetes- in these types of diabetes, the body's cells do not allow the glucose to enter the cells. The pancreas also loses the ability to produce sufficient insulin in the body, raising glucose levels in the blood.

### Symptoms of Diabetes

Increased thirst, tiredness, numbness, sudden weight loss, frequent urination at night, dry mouth, and blurred vision are general diabetes symptoms. However, symptoms may vary for diabetes types.

* Type 1 diabetes- symptoms include nausea, vomiting, stomach pain, yeast infection, and urinary tract infection. It can develop over a few weeks or months (Cleveland Clinic, 2022).
* Type 2 diabetes- people of all age groups suffer from type 2 and prediabetes. It may not have any particular symptoms as it develops over several years.
* Gestational diabetes- happens during pregnancy. Therefore, the doctor may ask for a gestational diabetes test between 24 and 28 weeks of pregnancy.

### Risk Issues

The risk factors involved vary as per the type of diabetes:

For type 1 diabetes, the risk factors are mainly:

1. A family history usually prevails or becomesnto a genetic disorder.
2. Prevalence of autoantibodies that attacks the self-body of the individual.
3. Pancreatic injuries like tumours, surgery, etc. (Cleveland Clinic, 2022).

Type 2 diabetes and prediabetes risk factors are:

1. Overweight or obesity
2. High blood pressure
3. Physical inactiveness
4. Occurrence of gestational diabetes and a cardiac disorder or stroke history.

## Research Question

The aim of the research has been discussed in the last section. It defines the process of creating and applying the machine learning model that will predict diabetes disease in patients (Kumari, et al., 2021). To fulfil the aim, the research will be conducted, and the research question has been framed so that the critical outcomes of the study can be interpreted and addressed. So, the research questions are as follows:

1. How successfully can diabetes disease be detected using the application of machine learning?
2. How can the most important symptoms of diabetes be identified?
3. Which algorithm will be the most effective in detecting diabetes disease?
4. What will the research achieve, and how it improve the detection process of diabetes at the early stage?

## Hypotheses

Diabetes is one of the most commonly seen diseases in the world. Diabetes can be seen for many reasons, which may vary by the person (Dunbray, et al., 2021). So, the determination of diabetes needs to be done by emphasising the patients' symptoms and information. Regarding the facts, the hypothesis has been framed, and those are as follows:

**Null Hypothesis (H0)**

There is no influence of the amount of glucose on the possibility of diabetes.

**Alternative Hypothesis**

**H1**: The amount of glucose in the blood increases the possibility of diabetes.

## Research Aim

The project aims to detect diabetes disease by applying machine learning classifiers. In this project, the target is to see diabetes by taking the user input about the symptoms and identifying whether that person is affected by diabetes or not (Chauhan, et al., 2021). So, to do this, the machine learning model should be applied through which the detection will be done, and the machine learning model will conduct the testing. Finally, the most effective model will be used for the real-time testing of diabetes disease (Katarya & Jain, 2020).

## Research Objectives

1. To emphasise the research approaches from the existing papers and gather ideas regarding the detection of diseases, especially diabetes
2. To select the dataset that will contain the patient records regarding symptoms of diabetes and study the features.
3. To choose the algorithms of machine learning (classifiers) by taking the insight into the reviews of the existing approaches.
4. To tune the classifiers, prepare the model, and apply the data to classify and detect diabetes; observe the performances; If the performance is not satisfactory, resample the data and execute the detection process again.
5. Observe the models' accuracies and compare to get the best one out of the selected classifiers.
6. To compare the present strength of detection with the previous models and find the achievement of the research.

## Research Issues

The growing and enhancing world population and an increase in the economical rate have caused a severe enhancement not only in the number of people being affected but also in individuals at a much younger age.

### Legal Issues

With a substantial expansion in the number of employment availabilities, the discovery of insulin has helped in the removal of many legal as well as behavioural limitations as well. So, again, it has been made mandatory by the government for schools, offices, and other institutions to ensure a safe working or learning environment for diabetic individuals. These mandates and care plans are prepared by the Section 504 Plan or (IEP) Individualized Education Program to ensure the best diabetes need of a child.

### Ethical Issues

Ethical issues determine the philosophical faiths and beliefs associated with the right or wrong moral values in diabetes detection. The rise in research conducted in terms of diagnosing and treatment of the disease will eventually cause a rise in ethical issues. The most fundamental moral issue associated with diabetes is the deprivation of the patients in achieving a cost-friendly treatment due to the presence of attractive market and business opportunities in the field of treatment.

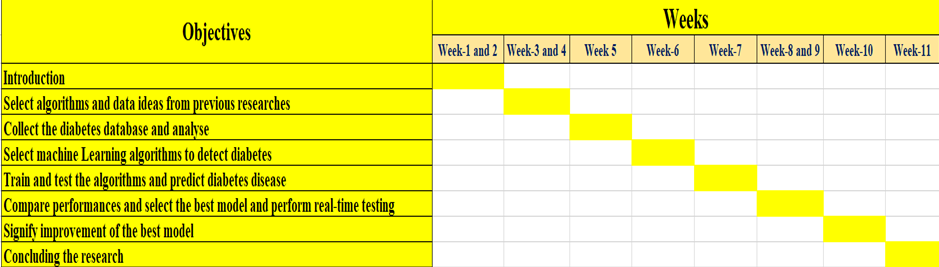
### Social Issues

The most vital social issues include an enhanced treatment cost as well as employment-related complications, and also a hampered productivity rate, and also a hindrance in achieving educational encouragement, thereby causing a negative impact on the advancement of society.

### Security Issues

The proper maintenance of security management and privacy of patients suffering from diabetes is the first and foremost duty of the healthcare sector. But the increased risk of less stringent security protections, there must be provisions for ensuring patients' safety and security of the information collected and stored.

## Research Planning



# Review of Previous Papers

## Effects of Diabetes

Diabetes is the kind of disease that makes it harder for the body to control blood pressure as well as the level of cholesterol. Even this disease seems to lead to heart attack, stroke and several other problems. This mainly happens as it becomes harder for the level of blood to flow towards the legs and the feet (The Victorian Government, 2021). Nerves inside our body can easily get damaged, which might be the reason behind causing pain, burning, tingling, and loss of feeling as well. Keeping the blood pressure, blood glucose level, and cholesterol level within the stipulated range can be helpful in reducing the risk of long-term effects of diabetes. In addition, keeping a healthy weight, eating in healthy proportions, reducing alcohol intake, and keeping smoking at bay will help reduce the risk level. Diabetes is a severe disorder that might lead to heart disease, blindness, and kidney and nerve failure. Apart from restrictions on food habits, there seems to be no cure or treatment for diabetes. On the contrary, with time, it becomes more aggressive as the kidneys start deteriorating, which leads to failure.

## Detection of Diabetes Disease

### Diagnosis Process

Following are the tests that are used for the process of diagnosing the level of diabetes:

* A test that measures fasting plasma glucose in the blood glucose after a patient seems to have gone through at least 8 hours without eating. This test appears to be detecting the level of diabetes or prediabetes (Dansinger, 2020)
* A tolerance level related to an oral glucose test helps measure the blood sugar after the person has gone at least eight hours without food and two hours after drunk any beverage containing glucose. This test is commonly used for the diagnosis of diabetes or prediabetes.
* In a random test of plasma glucose, the doctor checks the blood sugar level without any association with the time when the last meal has been taken. This test, along with a symptoms assessment, is used for diagnosing diabetes, but not prediabetes.
* A haemoglobin A1c (HbA1c) test can be done without fasting and can be used for performing diagnosis or confirming either prediabetes or diabetes.

### Machine Learning

Diabetes mellitus is mainly characterised by hyperglycemia which might cause many complications. According to the growing rate of morbidity in recent years, diabetic patients all over the world will reach about 642 million by the year 2040, which means that one in ten adults tends to be suffering from diabetes. There remains no doubt that these Statistics need great attention (Cıhan & Coşkun, 2021). Machine learning seems to be helping out people to make a preliminary judgment regarding diabetes mellitus based on the data related to their daily physical examination, which might serve as a doctor's reference. Machine learning methods are widely used to predict diabetes and give preferable results. Through this study, the Decision tree has been founisopular methods in machine learning in the medical field, which has grateful classification power. The Random forest method generates many decision trees, whereas the Neural network is the recent popular machine learning method that seems to have a better performance in many aspects. So in this study, the decision tree, random forest (RF), and the neural network has been profoundly used for making predictions regarding diabetes.

## Previous Research on Disease Detection

According to Brankovic & Zamani (2019), Steatosis or Fatty liver disease (FLD) has become a common condition in today's society. FLD is not limited to people who have a background of continued alcohol consumption. It denotes fat buildup that could cause inflammation, stiffness, and cirrhosis. Early identification of FLD is critical because it avoids additional liver degeneration, such as liver cancer and acute liver failure, and an increased risk of cardiovascular events. The suggested system's simplicity and the consistency and dependability of the outputs are advantages over current CAD systems. The first findings show that the suggested method can identify FLD with greater than 97% accuracy for the generated torso design.

As stated by Qin & Chen (2020), Chronic kidney disease (CKD) is an international public health issue that affects around 10% of the worldwide population. In the early phases of CKD, there are no visible symptoms. As a result, the illness may not even be recognised until the kidney has lost around 25% of its functioning. Furthermore, CKD has a significant morbidity and death rate and a worldwide influence on the human system. Random forest outperformed the other machine learning models, with 99.7% diagnostic accuracy. Furthermore, this concept may be relevant to clinical data from other disorders. After studying the errors produced by the existing approaches, researchers developed an integrated model which integrates random forest and LR utilising perceptron, with an accuracy rate of 99.8% after ten simulations.

According to Sivasangari & Krishna Reddy (2020), Considering the nuances of the signs in the early stages, liver disease is complex to diagnose. Problems with liver disorders are frequently not discovered until everything is beyond the end of the road, as the liver continues to function also when damaged. Early intervention may save a person's life. Liver illness is brutal to identify depending on the unique character of its signs. Liver disease prediction came after the data preparation stage, in which information was obtained from a public database and preprocessed for -1 value substitution. The complete variety of data is divided into education and research. Finally, quantitative analytical criteria like accuracy, precision, and recall are applied to distinct machine learning techniques.

As stated by Singh & Gourisaria (2021), Chronic liver illness is the most dangerous illness and the leading cause of death in humans. The liver is one of our body's most prominent and influential structures. The liver, as an exocrine gland, produces bile in the gut. During the previous two years, liver illness has been one of the top Twelve risky causes of mortality, and it is one of the leading causes of death in individuals aged 45 to 55. On features extracted from the set of data, researchers applied different machine learning methods such as LR, Random forest, KNN, and Decision tree to anticipate liver disease, and it has been discovered that Random Forest achieved the best performance among all approaches, gaining good accuracy and precision and performing astonishingly in all metric assessments.

According to Kumari & Mehta (2021), In cardiac illness, the heart is frequently incapable of pumping enough blood to the other central systems, limiting the body's functions. As a result, the heart eventually stops functioning or fails. Researchers used seven machine learning systems to determine heart illness and evolutionary algorithms like AdaBoost and the polling ensemble approach to increase the efficiency of poor functioning algorithms. Linear Discriminant Evaluation's achievement is excellent compared to other techniques; its average accuracy rate is close to 84%, and an absolute error is 16%.

As stated by Kuzhippallil & Joseph (2020), Machine Learning, a subset of AI Technology, enables the machine to learn without explicitly knowing anything. Trained algorithms employ human deliverables for training and forecasting precision and are consequently used in various classification tasks. As a result, the use of ML has expanded to include healthcare. Another of the most outstanding healthcare issues is the growing number of individuals suffering from liver disease. To increase performance, the information is cleared using several approaches such as imputation of incomplete data using the mean, labelling coding to transform categorical information into numeric information for simple analysis, duplicated value removal, and outliers are removed using Isolation forest. Finally, in this research, the authors have obtained the highest accuracy using the Stacking classifier (an ensemble model) by 83%.

According to Amirgaliyev & Shamiluulu (2018), Chronic Kidney Disease (CKD) is a kind of kidney illness that causes a progressive decrease in kidney functioning. Because of the various living situations of patients, this phenomenon might be noticed over years or months. CKD, commonly known as chronic kidney failure, affects 10% of worldwide people, as per recent medical data. In 2005, there had been around 58,000 thousands deaths globally. Several researchers are interested in automated illness detection. As a result, timely illness identification is critical, especially in underdeveloped nations where illnesses are typically identified in the final stages. Coming up with a solution to the challenges mentioned above and overcoming limitations became a powerful motivator for doing this investigation. Clinical histories, physical exams, and laboratory testing create the chronic kidney disease database. The experimental findings demonstrated a success rate of over 93% using the SVM.

As stated by Gudeti & Mishra (2020), The kidney inside the mammalian system is responsible for collecting and excreting all hazardous and unnecessary elements, general trash, from the body via the egestion and elimination processes. Around one million instances of Chronic Kidney Disease (CKD) are diagnosed in India. It is harmful to the kidneys and causes a progressive decrease in renal functioning. Within the constraints of this medical situation, the findings show that the SVM method diagnoses Chronic Kidney Disease superior to K-Nearest Neighbors and LR. This technique’s advantage is that the forecast procedure requires much less time, allowing clinicians to begin therapy for individuals with CKD as soon as possible, ancategorisingng more significant populations of people in a shorter amount of time.

## Previous Research to Detect Diseased Concerning Symptoms

According to Shakeel & Ahmad (2020), agriculture is based on the number and grade of harvested crops, usually plants. Several scientists have used image processing technologies to simplify activities associated with identifying plant illnesses, such as aberrant growth or sickness. Agriculture is often regarded as the foundation of any nation. Cotton leaf infections are comparable to Macula and Alternaria on plants in 80 to 95% of cases. The K-means clustering technique is utilised to divide pictures into groupings. The hybrid approach for textural and colour feature extraction is used for extracting features. Eventually, Cercospora cotton leaves are classified using the SVM. Recall and precision function assessment criteria are utilised to assess accuracy, and this is discovered that around 96% accuracy is obtained.

As stated by Almubark et al. (2019), Alzheimer's disease (AD) is the leading source of dementia in elderly adults, contributing 60 to 80% of the total of all dementia occurrences. Alzheimer's disease is the sixth most significant cause of mortality in the United States. Presently, 5,800 thousand Americans have Alzheimer's disease; by 2050, this figure is predicted to climb to about 14 million. Individuals with moderate Alzheimer's disease or mild cognitive impairment were lumped together in the AD group. Four distinct machine learning techniques have been used to differentiate patients from healthy controls employing information from neuropsychological testing, cognitive tasks, etc. The original study findings demonstrated that machine learning algorithms can assist in diagnosing Alzheimer's disease utilutilisingropsychological information and that by integrating cognitive and neuropsychological information, classification accuracy may be further enhanced.

According to Kohli & Arora (2018), the Centers for Medicaid and Medicare Services, half of all Americans suffer from numerous chronic conditions. Timely identification of common illnesses such as cancer, tumour, coronary artery disease, and diabetes might regulate and minimise the person's risk of death. Many classifiers and clustering methods are being utilised with the progress of artificial intelligence and machine learning. The suggested technique has a prediction accuracy of 87.1% in Heart Disease identification employing LR, 85.71% in Diabetes diagnosis operating SVM, and 98.57% in Breast Cancer identification applying AdaBoost classifier. The firm's future extension and enhancement will include automated stages, including data munging, extraction of features, and structure model for the most fantastic predictive performance.

As stated by Basha & Kumar (2019), Individuals are stressed and anxious due to their hectic schedules and usual responsibilities. Furthermore, some individuals are hooked on chronic habitual behaviour, such as the intake of Cigarettes and Gutka. These individuals suffer from chronic cardiovascular disease, cancers, liver, kidney, etc. Thus a healthcare or hospital set of data is acquired from the Kaggle site to study and apply the information on several techniques to assess the average accuracy, sensitivities, and specificity of the main feature of heart disease individuals. Researchers examine the suggested modelling for heart disease individuals using a different algorithm, confirming several essential qualities. The KNN method proves to be the most efficient and effective in terms of average accuracy in heart disease prediction, with an accuracy of 85%.

According to Yahyaoui & Yumuşak (2021), Although many studies on the issue of chest ailments, its identification and treatment remain uncertain, perplexing, and rugged for professionals. Chest illness is a severe health issue that has remained an intriguing area of study. The timely and precise prognosis of lung illnesses has become a critical requirement for saving people's life and facilitating physicians' jobs. These methodologies are tested using a private information collection from Diyarbakir Hospital's pulmonary disorders division in Turkey. The findings demonstrated the usefulness of these approaches in detecting pulmonary disorders, notably the KNN method, which had a diagnostic accuracy of 95% and 94.3% when employing the DNN technique.

As stated by Wang & Lee (2020), Correctly identifying Parkinson's disease (PD) at such a preliminary phase is unquestionably critical for halting its progression and giving individuals access to disease-modifying medication. The premotor stage of Parkinson's disease must be closely studied to achieve that goal. Depending on premotor traits, an unique deep learning approach is proposed to swiftly determine if a person has Parkinson's disease or not. A contrast of the suggested deep learning model with twelve ensemble learning and machine learning approaches based on comparatively limited information sets, including many average persons and early Parkinson's disease patients, reveals that the developed model achieves high accuracy, 96.45% on average. Researchers present the selected features of the PD detection procedure depending on the Boosting approach and identify the PD.

According to Katarya & Srinivas (2020), Just after the brain, the heart is among the essential components of the human system. The heart's principal duty is to circulate blood throughout the body. Cardiovascular disease refers to any condition that can impair the heart's activity. Clinics and other institutions provide costly treatments and surgeries to address cardiovascular disease. In today's rapidly changing world, CAD affects 2% of the global population, with 10% of those affected being over the age of 65. There are many forms of heart illness around the globe; the most frequent heart disorders are heart failure (HF) and coronary artery disease (CAD). This report detailed several of the expertise automation machines. Selection of features and predictions are critical components of any automated process. Researchers can improve our prediction of heart disease by selecting characteristics wisely.

As stated by Rahman & Siddiqua (2019), Chronic Kidney Condition, also known as chronic renal dysfunction, is a disease in which the kidneys begin to deteriorate. It is a persistent disorder that leads the kidney to degenerate and lacks its capacity to operate correctly, eventually leading to the fifth and final, deadly phase, Late Phase Renal Disease, wherein the kidney function decreases to around 10% to 15% of their healthy capabilities. The validation phase yielded promising findings, as the model correctly classified users as usual or renal patients. The investigation discovered an accuracy of 97.6% utilising both variables QT and RR gap, which was higher than the accuracy observed when only one of the characteristics was employed.

## Early Detection of Diabetes Disease

Diabetes Mellitus (DM) advance identification is a mechanism for detecting the likelihood of someone getting DM. The subject being investigated is the clinical signs of diabetes mellitus that is only seen in atypical concerns with data that may be classified as a hot coding condition, with evidence in the form of 'Yes' or 'No.' (Aofa, et al., 2018) compared the Standart Backpropagation Neural Network (SBNN), SBNN+ALR, SBNN with PSO (SBNN+PSO), and SBNN with PSO and Adaptive Learning Rate (SBNN+PSO+ALR) for timely identification of DM in this research. Indications and conditions that promote diabetes (a total of nine variables) were considered in this investigation. Clinical information at the Brebes Health Center (Puskesmas) was used to compile the data for this analysis. The K-fold Cross-Validation technique was applied to analyse the distribution of training and sample datasets. The optimal architecture had been found to be SBNN+PSO+ALR, according to the findings. In just 30 sessions, the SBNN+PSO+ALR architecture achieved an overall accuracy of 88.75% and a sensitivity of 82.5%. The mechanism also attained a specificity of 95% and an MSE of 0,02939.

(Gupta, et al., 2021) emphasised applying a range of preprocessing strategies to generate the best outputs when constructing robust classification methods for the diagnosis of diabetes, heart disease, and liver illness. The models were developed using resources from the UCI Machine Learning Repository. This work employs a variety of preprocessing methodologies, including feature engineering, data trimming, oversampling for unbalanced collections, imputation of void information, categorised variable embedding, and feature scaling. These methods significantly improve the accuracy of classification algorithms such as Random Forest, KNN, and SVM, among others. Hyperparameter adjustment improves the functionality of these algorithms much further, resulting in much higher accuracy ratings. Heart disease, liver problems, and diabetes predictions all had the highest accuracies of 90.16%, 73%, and 93.23%, accordingly. The report also highlighted the positives of detecting these conditions earlier on, which might make a significant impact in a lot of situations. Measurements like Accuracy, Balanced Accuracy, and F-1 score have been analysing the performance of various classifiers. To explore the optimum findings, dovisualisationlization and assessment of the performance of the classification methods.

Diabetes is a condition under which the body's glucose, or blood sugar, is not properly processed, causing the glucose level to rise to dangerously high proportions. Ordinarily, the hormone insulin helps manage the quantity of glucose in one's system; however, persons with diabetes either don't create enough insulin or do not respond to it appropriately. Type 2 diabetes accounts for nearly 90% of overall diabetes occurrences. The major focus of this study by Abdulhadi & Al-Mousa (2021) was to use numerous machine learning approaches to forecast the occurrence of diabetes, particularly in females, at a premature time. Early identification of diabetes can certainly helps to manage the problem from progressing and lessen the risk of catastrophic consequences like heart and renal disease. Implementing the appropriate lifestyle adjustments can also help minimise diabetes, including all the conditions that come with it. As a result, a technique that may require better aid clinicians in detecting this dangerous disease at an initial period and thereby stopping its spread is critical. Finally, using the random forest classifier approach, this prototype obtained an accuracy of 82%.

(Ahmed, et al., 2022)proposed a diabetes forecasting framework based on a fused machine learning technique. There are two categories of modelling in the theoretical model: SVM and ANN models. These algorithms researched the issue to see if a diabetes diagnosis was affirmative or negative. The sample utilised in this study was split into two parts: training data and assessment dataset, with a 70:30 split. The outcome of this modelling is used as the fuzzy model's input inclusion functionality, and the fuzzy logic has been used to evaluate if a diabetes assessment is positive or negative. The fused representations are saved in a cloud storage facility for subsequent application. The fused approach determines whether or not the person is diabetic, dependent on the person's real-time health records. The accuracy rate of the developed fused ML algorithm has been 94.87%, which is superior to earlier documented techniques.

Akula et al. (2019) have detected diabetes disease with the application of machine learning classifiers. As a result, researchers consolidated all of the methods into a weighted mean or soft voting aggregation approach, in which each algorithm contributes to a clear judgment on if an individual has diabetes or not. The Ensemble modelling-based Practice Fusion has an accuracy of 85%; this composite technique is by far novel in this domain. The weighted average ensemble strategy did not simply outperform the competition in terms of overall performance, but it notably improved in the recovery of erroneous assumptions and the precise identification of Type 2 diabetes. The suggested accurate, innovative model may be deployed as an advanced warning system for people to seek clinical attention.

To identify diabetes sufferers, a lot of scholars have devised a myriad of prediction frameworks built on Machine Learning (ML) approaches. Hasan et al. (2020) used different Feature Transformation (FT) methodologies, including Robust (RS) Scaler, Standard Scaler (SS), and Min-Max Scaler, to modify an early-stage diabetic probability assessment dataset. Additionally, different classification algorithms have been tested on these modified records, and it was discovered that Extra Trees (ET) performed highest on the z-score converted dataset. Furthermore, using the z-score processed sample, multiple Feature Selection Techniques (FSTs) have been used to accomplish a classification job using ET, and it has been discovered that the PCC was effective in choosing the most essential features. This demonstrates that using this approach to categorise diabetes individuals at an initial stage could be achievable.

The research of Dunbray et al. (2021) revealed that In today's technological era, machine learning and data mining constitute two developing areas. Researchers may use such strategies to monitor previous data behaviours and anticipate potential occurrences to a degree using these approaches. This gave prominence to the phrase 'prediction models,' which is now well-known in the IT industry. Predictive methods for diabetes could be developed by researchers. This aids in the prompt diagnosis of diabetes, allowing it to be diagnosed as soon as practical in order to avoid problems. They can precisely forecast if the individual is diabetic and avert additional health complications by choosing the algorithm with the best accuracy. The approaches utilised to construct a distinctive predictive model for diabetes detection were also highlighted in this research work.

Rubaiat et al. (2018) used a neural network to construct a method for autonomously predicting type 2 diabetic Mellitus (T2DM). This research focuses on determining which form gives better results for detecting diabetes. In this collection, two methodologies were used to conduct the analysis. Data Recovery has been the initial step, accompanied by feature selection. They entered these characteristics into the MLP neural network classification model, which had an accuracy of 85.15%. They used a noise reduction mechanism built on k-means and proceeded by selecting features in the subsequent strategy. Random Forest, LR, and MLP classifiers have been used to combine the characteristics acquired. The highest level of accuracy recorded by these classifiers equals 77.08%.

## Comparative Discussions of Previous Models

The comparative discussion of the existing approaches will be discussed in the below table regarding the algorithms that they have used and the accuracy that they have achieved for different types of disease detection:

Table 1 Research Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year** | **Authors** | **Purpose of Research** | **Algorithm** | **Accuracy** |
| 2019 | (Brankovic & Zamani, 2019) | Fatty liver disease detection | Supervised Machine Learning | 97% |
| 2020 | (Kuzhippallil & Joseph, 2020) | liver disease detection | Stacking classifier | 83% |
| 2018 | (Amirgaliyev & Shamiluulu, 2018) | Chronic Kidney Disease detection | SVM | 93% |
| 2020 | (Shakeel & Ahmad, 2020) | Plant Disease detection detection | SVM | 96% |
| 2018 | (Kohli & Arora, 2018) | Heart Disease detection | AdaBoost classifier | 98.57% |
| 2019 | (Basha & Kumar, 2019) | heart disease detection | KNN | 85% |
| 2021 | (Yahyaoui & Yumuşak, 2021) | pulmonary disorders detection | KNN | 95% |
| 2020 | (Wang & Lee, 2020) | Parkinson's disease detection | Deep Ensemble Network | 96.45% |
| 2019 | (Rahman & Siddiqua, 2019) | Chronic Kidney Disease detection | Supervised Machine Learning | 97.6% |
| 2018 | (Aofa, et al., 2018) | Diabetes Disease Detection | Standard Backpropagation Neural Network | 88.75% |
| 2021 | (Gupta, et al., 2021) | Diabetes Disease Detection | SVM | 93.23% |
| 2021 | (Abdulhadi & Al-Mousa, 2021) | Diabetes Disease Detection | random forest classifier | 82% |
| 2022 | (Ahmed, et al., 2022) | Diabetes Disease Detection | ANN | 94.87 |
| 2019 | (Akula, et al., 2019) | Diabetes Disease Detection | Ensemble classifier | 85% |
| 2018 | (Rubaiat, et al., 2018) | Diabetes Disease Detection | MLP (ANN) | 85.15% |

# Research Methodology

The approach that will be taken to detect diabetes disease using machine learning will be discussed in this chapter. The chapter will contain a discussion of the proposed method, the dataset that will be chosen for this research and the tool that will be selected for analysis. Additionally, the algorithms and the evaluation methods will also be emphasised in this research.

## Proposed Methodology

### Proposed Method for Diabetes Detection

The proposed methodology for the detection of diabetes disease is shown below:

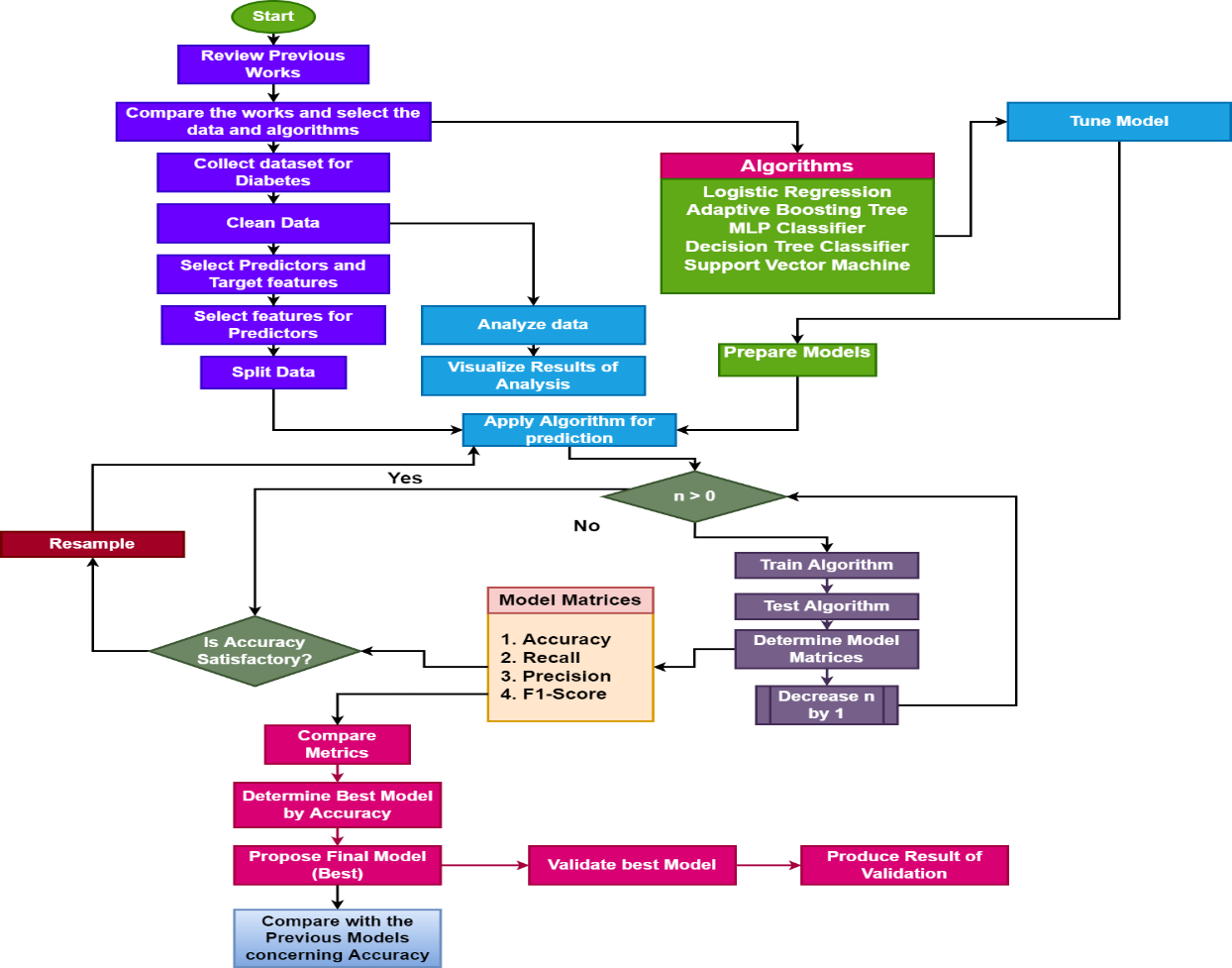


Figure 1 Diabetes Detection Method

## Diabetes Data

The data is the main and essential component of the project. By taking the idea of the data from the previous approaches, the dataset is extracted from Kaggle named Pima Indian Diabetes Dataset (Kaggle, 2016).



**Fig-2: Pima Indian Diabetes Dataset**

### Data Description

The data consists of 268 observations of diabetes patients out of 768 observations, and the rest 500 observations are of non-diabetes patients. For determining diabetic probability, nine features illustrate the symptoms and historical records of patients. The features labelled as "Outcome" will be used as the target feature, and based on the rest features, the outcome will be classified to identify diabetes.

The feature of the dataset are discussed below:

* Pregnancies: This feature of the dataset is specially collected for the female patient to denote the number of pregnancies they have faced. A more number of pregnancies implies the storage of extra fat in the body that may raise the possibility of blood sugar and will have an impact on diabetes.
* Glucose: This feature signifies the level of glucose in the blood of a human. As glucose is the main source of energy, it is one of the main factors of the data. This is because the increase in the level of glucose rises the possibility of diabetes.
* Blood Pressure: This implies the blood pressure values of the human from whom the observations have been taken.
* Skin Thickness: This implies the thickness of the skin of the human from whom the observations have been taken
* Insulin: This feature implies the level of insulin in the human blood. As insulin is the main hormone to transport blood glucose to the human cells, it is another main factor of the research. If the level of glucose goes down, less amount of blood glucose will be transported to cells, and this will raise the amount of blood glucose. So, this has a direct impact on the possibility of diabetes.
* BMI: This is the measure of fat in the body compared to the height and weight of a human. This is termed Body Mass Index, and if the body mass index is increased, the obsession may be seen in humans that may raise the possibility of diabetes due to the storage of a higher amount of fat in the body.
* Diabetes Pedigree Function: This feature provides information about the probability of diabetes in humans due to heredity. Diabetes Pedigree Function is the measure of the likelihood probability of the human for diabetes and signifies its heredity of it.
* Age: This feature implies the age of the patients.
* Outcome: This is the target feature of the dataset that contains two labels, 1 and 0. Label-1 implies those records belong to the people who have diabetes. On the other hand, label-0 implies those records belong to people who do not have diabetes.

# Research Components and Artefact

## Programming Tool

### Tool Selection

Python is mainly a programming language, and it is used by both non-developers as well as developers to build different websites, programs, applications, and software (Akula et al., 2019). Apart from that, Python is also used in developing data analysis, automation visualisation, automated programs, and activities. It has been widely useful because of its easy-to-write, read and learn features (Upgrad, 2021). Python is an interpreted, general purpose-oriented, and high-level programming language. As it is very easy to learn and read, Python is used by various non-programmers like scientists and accountants for operating financial activities etc.

### Reasons for Preferences

Python has also brought various advantages to programmers as well as non-programmers to use Python as a robust programming language for operating their various automation, application-based, and software-based tasks. The main benefits that Python provides to the users include:

**Easy to learn to read and write**

This is a high level of integrated programming language, which includes English-based syntax. It has made Python easy to learn and read. The users can also understand the codes easily (Upgrad, 2021). The users can also easily pick up this language, and that is why many people try to learn Python. The programmers also need to write a few lines of code in this language, which also has a massive benefit for the programmers.

**An enhanced productivity**

This is very protective language. Python is very easy to use and simple, which is why the developers and programmers can solve various problems without spending much time understanding the syntax and behaviours of the programming language. By writing less code, the programmers can get their activities and things done quickly.

**Interpreted language**

Python can execute the codes directly line by line with the help of its feature of being an interpreted language. When the programmers make any error, Python can stop its execution and report the error that the programmer has created.

## Selection of Tool and Libraries

For programming, I consider python would be perfect as it employs with several libraries. Python is advantageous for many researchers in their investigations and for programmers to develop. Python is an eligible programming language for machine learning applications. In this project. I use Jupyter Notebook for IDE to Devlop. The following libraries will be used:

1. Numpy, Pandas
2. Matplotlib, Seaborn, Plotly Express
3. Scikit Learn

## Planned Technology to be Used

### Data Analysis

Data analysis is very crucial for detecting diabetes. Data analysis refers to the method of cleaning, inspecting, modelling, and transforming various data and information and the main aim is to discover the necessary information for helping in the session making and drawing a conclusion. To detect diabetes using machine learning, performing data analysis is very crucial. Various information and data about the previous cases of diabetic patients and their symptoms can be analyzed with the data analysis process (Wakefield, 2022). It can help in developing strategies to detect diabetic patients and give them proper treatment.

### Predictive Learning

Predicted learning is based on predictive analytics, which is used in the form of advanced mathematical calculations. Predictive analytics is used to identify current and past patterns by analysing the data (Wakefield, 2022). It helps in making projections about the future. With the help of predictive learning, doctors and healthcare practitioners can identify the patterns and previously taken measures to treat diabetic patients. It can also help them make future predictions through which they can detect diabetes effectively. This is essential to improve the health care services and treatment processes.

## Planned Algorithms to be Applied

### Logistic Regression

Logistic Regression refers to one of the most crucial and popular algorithms used in machine learning. This is one of the supervised learning methods (Abdulhadi & Al-Mousa, 2021). LR is generally used to predict the variables which are categorically dependent with the help of a set of independent variables (Brankovic & Zamani, 2019). It also can predict the outputs of the categorical dependent variables by giving a probabilistic value. To solve various classification problems, LR also has a huge contribution as this is very similar to 'linear regression'.

### Adaptive Boosting Tree

The adaptive boosting tree is also known as an Adaboost. This is an ensemble method used in machine learning techniques (Kohli & Arora, 2018). This type of algorithm generally uses a decision tree that has one level. It means that the decision tree has only one split. This type of decision tree is known as a decision stump.

### MLP Classifier

MLP classified or multilayer perceptron classifier refers to an algorithm used as a supplement to the feed-forward neural network (Rahman & Siddiqua, 2019). MLP has three layers, including output input and a hidden layer. It is used for solving classification prediction problems in which the inputs are connected to a level or class. From that, it is also used in solving regression prediction issues in which the real value quantities are projected with the help of the inputs (Wang & Lee, 2020).

### Decision Tree Classifier

The decision tree classifier is also a very significant algorithm used in machine learning (Basha & Kumar, 2019). It can help in creating classification models with the help of creating a decision tree. In this decision tree, the nodes can specify a test on the attributes. Each of the branches in the decision tree descends from the nodes, which correspond to the probable values of that attribute (Rahman & Siddiqua, 2019).

### Support Vector Machine

It is another supervised missing learning method. SVM is also used in solving issues and problems based on classification and Regression. Simple SVMs are used for solving linear Regression and linear classification problems (Tavasoli, 2022). kernel SVMs transformer data sets for helping the nonlinear decisions to transform the linear equations into a high number of dimensions. SVMs mainly help in finding a set of data in different dimensional areas and can help in making an observation-based summary (Amirgaliyev & Shamiluulu, 2018).

## Performance Evaluation Process

### Accuracy

Machine learning model accuracy refers to the measurement process which is used for determining which machine learning models are best suitable for understanding the data set patterns and relationships or analysing the information (Shung, 2018). The ratio of correct expected information is termed accuracy. Accuracy is very important while detecting diabetes with the assistance of machine learning.

### Precision and Recalls

Precision refers to positive predictive values that can be expected from data analysis. The expected ratio or a fraction of correct values is called precision (Shung, 2018). The recall is how many true positives have been found by any search. These can help in maintaining a proper way to detect diabetes.

### F1-Score

F1 score refers to the harmonic mean between recall and precision processes. It also refers to a statistical measure for calculating the rate of the performance of the detection process (Shung, 2018).

## Research Artefact

The artefact of the project and the execution process will be discussed in this section.

### Reading Data and Formulating Target

The diabetes data will be read in Python using the read\_csv() method, where the parameter will be the file name that is diabetes.csv.

### Treating Missing Values

After reading the data, the missing values will be found and treated. In this context, the standard syntax that is isna().sum() will be used, and if missing values are found, those will be replaced.

### Functions for Visualization

Two functions will be designed to plot the visualisation.

The first function will plot any kind of chart where the data will be prepared through cross-tabulation. The details of the function are:

Name: create\_chart(df,v1,v2,ctp)

Parameters:

Df: name of the data frame (prepared through cross-tabulation)

V1: first features

V2: target or reference feature

Ctp: type of chart that will be produced

The second function will plot the bar chart where the data will be supplied with features, and the average values will be calculated. The details of the function are:

Name: create\_bar(df1,df2,v1)

Parameters:

Df1: First Data frame

Df2: Second Data frame

V1: Feature to calculate mean

### Selecting Important Features

The correlation will be performed to find the correlation values of the features with the target attribute. After getting the correlation result, the top 6 features will be taken with the highest correlation values.

### Preparing Data and Split it

The predictor data will be prepared by taking the selected features, and the target data will be prepared by taking “Outcome” into consideration. The data will be split into train and test with a ratio of 80%-20%. The operation will be performed two times, first with the original data and next using resampled data, where the resampling will be done if the accuracy of the original data will not be found higher. In that case, ten times the rows will be generated.

### Assigning and Tuning Algorithms

The algorithms of machine learning will be assigned in a list with tuned parameters. The list technique has been used to apply all the selected algorithms at a time in a loop.

### Prediction

The prediction will be done both on the original adat and resampled data. In this context, the following steps will be observed:

1. Reading the model
2. Training the model
3. Testing and Predicting the Model
4. Storing the classification metrics

### Result Presentation

While predicting through the original data and resampled data, the comparison will be made concerning the stored metrics to find the most effective model. The comparison results will be visualised using a bar chart with value indicators.

# Analysis and Result

The application of machine learning classifiers have been made to detect diabetes disease, and the outcomes will be discussed in term of the performance metrics which has been discussed in the methodology chapter. Outcome of the research about the strength of the detection of diabetes will be illustrated in contrast to the previous models in this chapter.

## Data Reading

The diabetes dataset has been read in Python primarily for the purpose of analysis and detecting diabetes disease. After loading it, the data looks as follows:

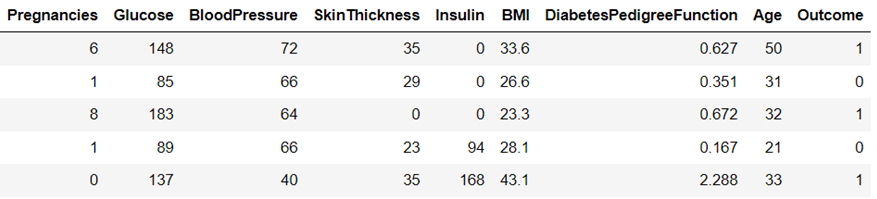


Figure 2 Data Reading

The target or outcome feature has been found with the labels 1 and 0. Those labels have been encoded to Diabetic and Non-Diabetic, respectively, and the data looks as follows:

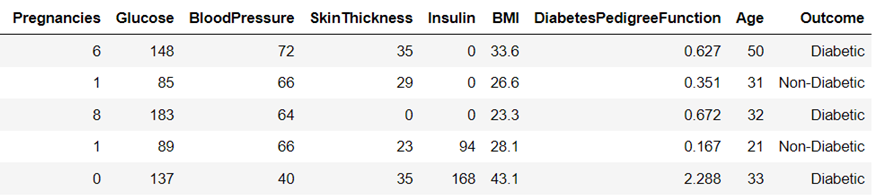


Figure 3 Encoding Target Features

The information of the data has been taken using coding, and this is shown below:

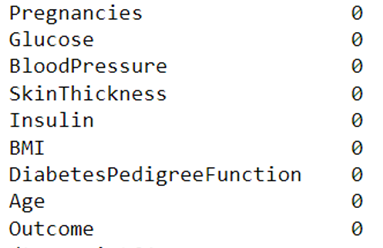
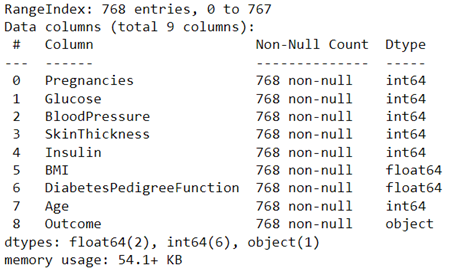


Figure 4 Data Information and Missing values

So, it can be seen that the data is clean, and thus, it is ready for analysis. In the next section, the explanatory analysis will be conducted on the data, and the results will be interpreted.

## Explanatory Analysis

The explanatory analysis is fruitful in understanding the data statistics, which will provide insights into the data. Thus, the features of the data have been analysed to get clear views of the diabetes symptoms and other information.

### Analysis of Age on Diabetic Possibility

The analysis of the age factor shows that the tendency of diabetic attacks will be higher in the age range between 40 and 50 years. The outcome is depicted below:

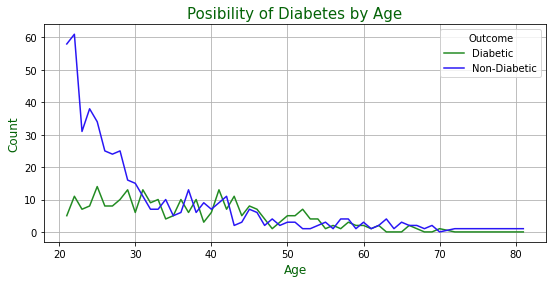


Figure 5 Analysis of Age on Diabetic Possibility

### Analysis of Glucose and BMI on Diabetic Possibility

As the level of glucose and body, mass index BMI are two important factors in determining diabetes possibility, so analyses have been done on those features. It has been found that glucose level between 120 to 150 has been seen with the maximum attack of diabetes. On the other hand, a BMI level between 30 to 50 is highly responsible for diabetes. The results are shown below:

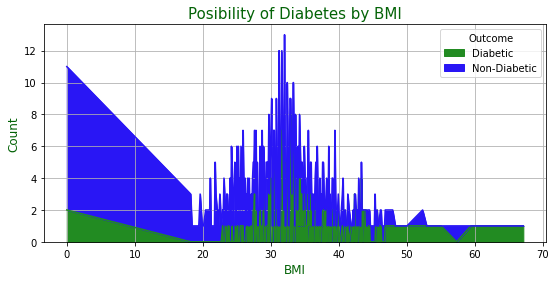
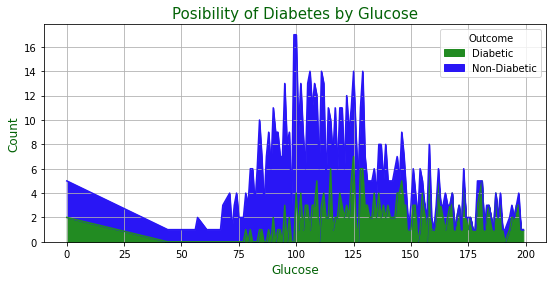


Figure 6 Analysis of Glucose and BMI on Diabetic Possibility

### The average age of patients

In the historical data, it has been found that the average age of diabetic people is 37, which indicates that at that age, most of the patients have been found with diabetes, as shown below:

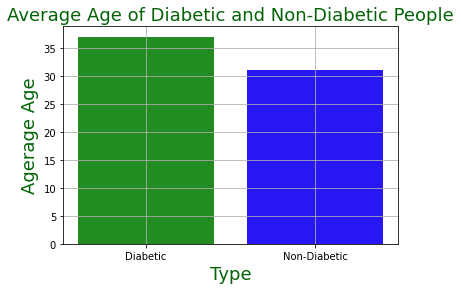


Figure 7 Average Age of patients

### Average Glucose Level

The analysis shows the fact that the average glucose level of diabetic patients is around 140, whereas, for healthy people, it is 110. The result of the analysis is shown below:

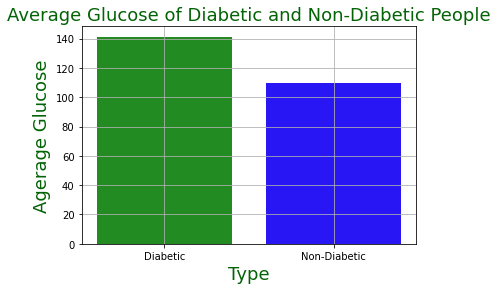


Figure 8 Average Glucose Level

### Average BMI Level

The analysis shows the fact that the average BMI level of diabetic patients is around 35 whereas, for healthy people, it is 30. The result of the analysis is shown below:

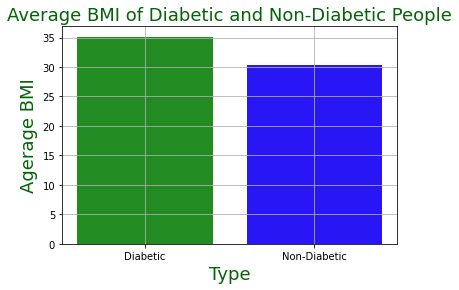


Figure 9 Average BMI Level

### Average Insulin Level

The analysis shows the fact that the average insulin level of diabetic patients is 100, whereas, for healthy people, it is around 70. The result of the analysis is shown below:

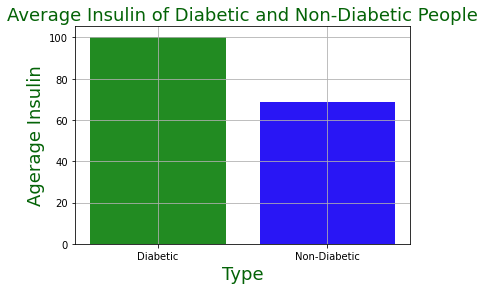


Figure 10 Average Insulin Level

## Feature Selection Process

### Feature Selection Process

The selected data contains different features based on which the target feature can be detected for diabetes. The rule has been designed to select important features which will be treated as the essential symptoms of diabetes. The rule is discussed as follows:

Rule-1: Apply Correlation (Pairwise Correlation) to the data to find the correlation value of the features with the outcome feature.

Rule-2: Setting the critical value of filtering the features with correlation. The stoping filter with the critical values is shown below:

These stoping filters will be applied to reject the features, and the rest features will be accepted as the final predictors. So finally, the features with a higher positive correlation and higher negative correlation will be taken as the final predictors. The features with zero correlation or very less correlation will be rejected.

### Feature Correlation and Heatmap

The heatmap of the feature correlation has been produced to get a clear view of the correlation with values. The heatmap is shown below:

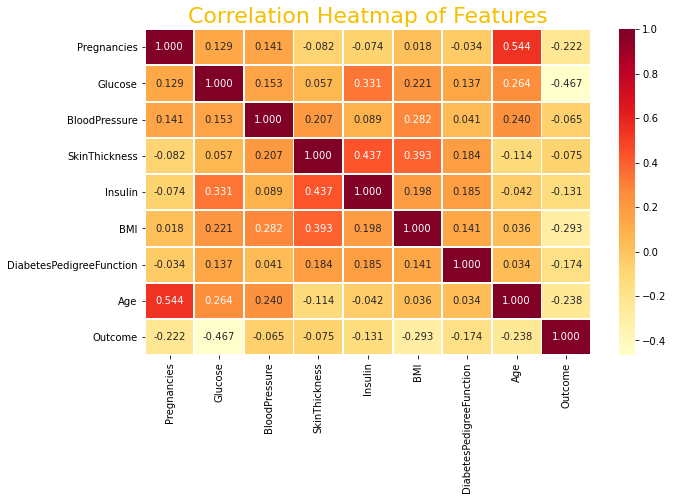


Figure 11 Feature Correlation

### Finally, Selected Features (Symptoms)

Now, as per the feature correlation and by applying the rule sets, the following features have been finally selected as the predictor to detect diabetes. Those features will be treated as important symptoms as well. Those are shown below:

* Glucose
* BMI
* Age
* Pregnancies
* Diabetes Pedigree Function
* Insulin

## Creating Training and Testing Set

In this experiment, data has been prepared with the finally selected features. The algorithms will be applied to the prepared data to detect diabetes. So, the algorithms need to be trained to gather the required statistical information on the data. Later those will be tested to get the prediction result. So, for the first case, the training set is required, and for the last case, the testing set will be applied. Thus, the data has been split to train and test sets with a ratio of 80%-20%. After splitting the prepared data, the algorithms have been applied.

## Diabetes Detection

The detection of diabetes disease has been done, and the results will be presented and discussed in this section.

### With Original Data

After splitting the data, the train set contained 614 instances and the test set obtained 154 instances. Now, the algorithms have been applied to the data to detect diabetes disease. Hence the detection result has been presented below:

Table 2 Result of Diabetes Detection with Original Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifiers** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| **Decision Tree Classifier** | 75.32 | 75 | 75 | 75 |
| **SVM** | 75.97 | 77 | 76 | 74 |
| **Adaptive Boosting Classifier** | 74.68 | 74 | 75 | 74 |
| **MLP Classifier** | 67.53 | 67 | 68 | 65 |
| **LR** | 62.34 | 61 | 62 | 52 |

It can be seen from the result that the decision tree produces the highest accuracy in detecting diabetes, with an accuracy rate of 75.32%. The lower value of accuracy has been obtained because of the low number of data instances for which the algorithms may not generate the statistics properly. Hence, the confusion matrixes of the classifiers are shown below:

Table 3 Confusion Matrixes for all classifiers for original data

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A  C  T  U  A  L |  | **Decision Tree Classifier** | | **SVM** | | **Adaptive Boosting Classifier** | | **MLP Classifier** | | **LR** | |
| **Predicted** | | | | | | | | | |
| **D** | **ND** | **D** | **ND** | **D** | **ND** | **D** | **ND** | **D** | **ND** |
| **D** | 36 | 23 | 26 | 30 | 32 | 27 | 21 | 38 | 4 | 55 |
| **ND** | 15 | 80 | 9 | 88 | 12 | 83 | 12 | 83 | 3 | 92 |

***\* D à Diabetic ND à Non-Diabetic***

So, from the experiment, it can be seen that the decision tree has detected 116 instances in the test set out of 154 test data, followed by an SVM which has a detection rate of 114 out of 154. However, as the accuracy is not satisfactory, the data will be resampled, and the same detection processes will be executed again.

### With Resampled data

As per the methodology proposed, the data has been resampled with ten times the present row, and the data has been split with the same 80%-20% split ratio. So, now the train set contains 6144 instances, and the test set contains 1536 instances.

Now, the algorithms have been applied to the data to detect diabetes disease. Hence the detection result has been presented below:

Table 4 Result of Diabetes Detection with Resampled Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifiers** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| **Decision Tree Classifier** | 98.76 | 99 | 99 | 99 |
| **Adaptive Boosting Classifier** | 83.92 | 84 | 84 | 84 |
| **SVM** | 77.86 | 77 | 78 | 77 |
| **MLP Classifier** | 74.74 | 74 | 75 | 73 |
| **LR** | 68.75 | 67 | 69 | 66 |

It can be seen that the accuracy of the decision tree has increased from 75.32% to 98.76% with the data resampling. The accuracy has also been increased for Adaptive boosting from 74.68% to 83.92%. However, the accuracies and performances of the classifiers such as LR, MLP classifier and SVM have not been improved. So, the tree-based algorithm that is the decision tree has performed the best or most effective in detecting diabetes with the resampled data. Similar to earlier, the confusion matrixes have been arranged in the table, and these are shown below:

Table 5 Confusion Matrixes for all classifiers for resampled data

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A  C  T  U  A  L |  | **Decision Tree Classifier** | | **SVM** | | **Adaptive Boosting Classifier** | | **MLP Classifier** | | **LR** | |
| **Predicted** | | | | | | | | | |
| **D** | **ND** | **D** | **ND** | **D** | **ND** | **D** | **ND** | **D** | **ND** |
| **D** | 519 | 13 | 297 | 235 | 378 | 154 | 258 | 274 | 167 | 365 |
| **ND** | 6 | 998 | 105 | 889 | 93 | 911 | 114 | 890 | 115 | 889 |

So, it can be seen that with the data resampling, the decision tree has gained 98.76% accuracy, and the detection rate has increased significantly. It has detected 519 instances out of 532 available instances doe diabetic patients and 998 instances out of 1004 instances of non-diabetic patients. The precision, recall and f1-scores have been observed for the decision tree by 99%. The classification report for the decision tree is shown below:

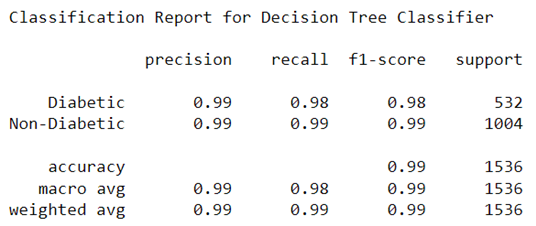


Figure 12 Classification Report for Decision Tree with Resampled Data

So, the Decision tree can be considered to be the most effective classifier to detect diabetes disease based on the symptoms.

## Comparative discussion and comparison

The comparison of metrics of classification will be compared and presented in this section.

### Comparison of Accuracies

The comparison of accuracies for the actual data and the resampled data is shown below:

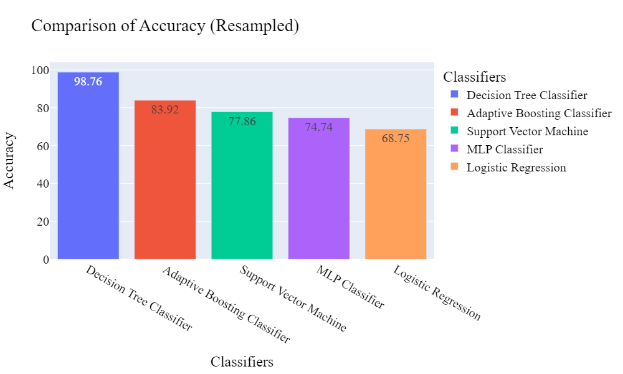
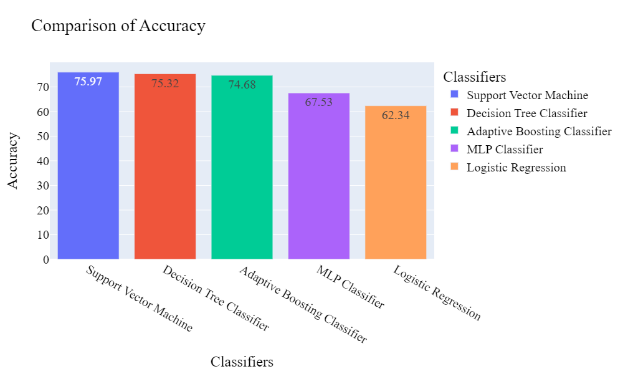


Figure 13 Accuracies of Detection

### Comparison of Precisions

The comparison of precisions for the actual data and the resampled data is shown below:

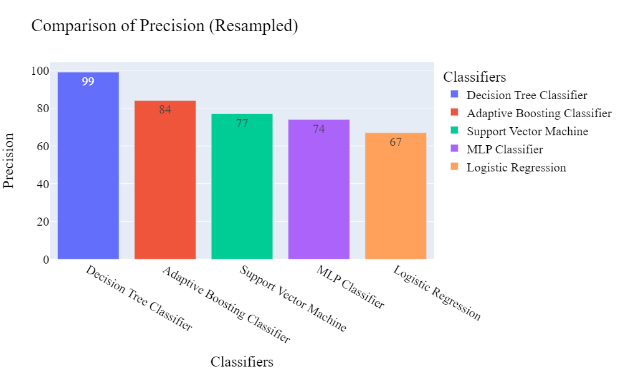
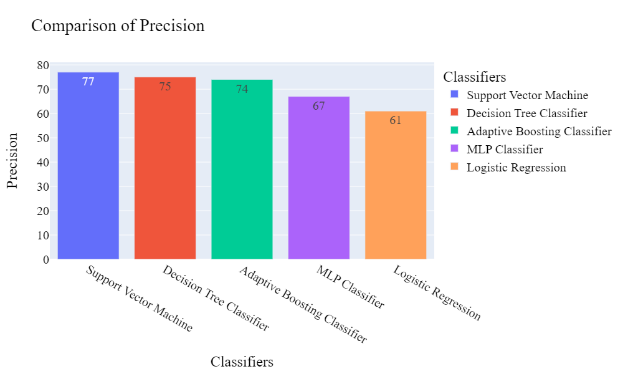


Figure 14 Precisions of Detection

### Comparison of Recalls

The comparison of recalls for the actual data and the resampled data is shown below:

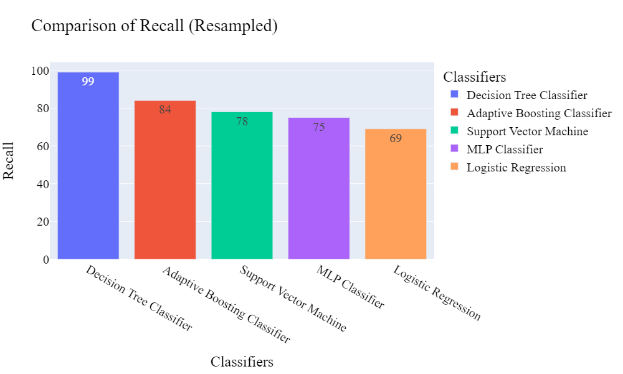
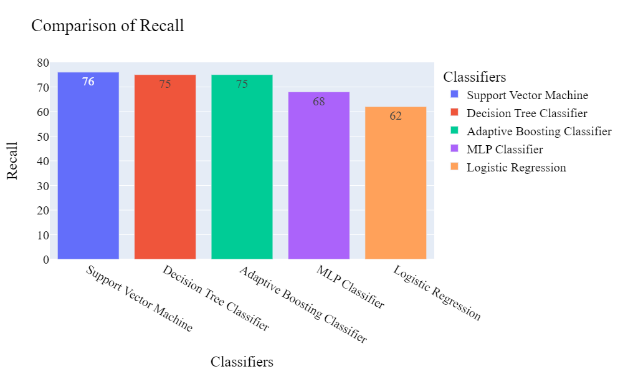


Figure 15 Recalls of Detection

### Comparison of F1-Scores

The comparison of f1-scores for the actual data and the resampled data is shown below:

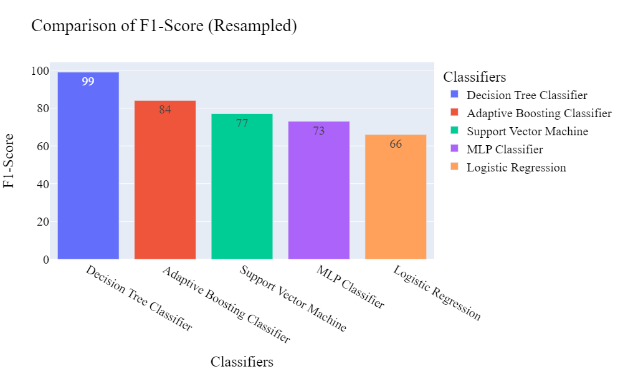
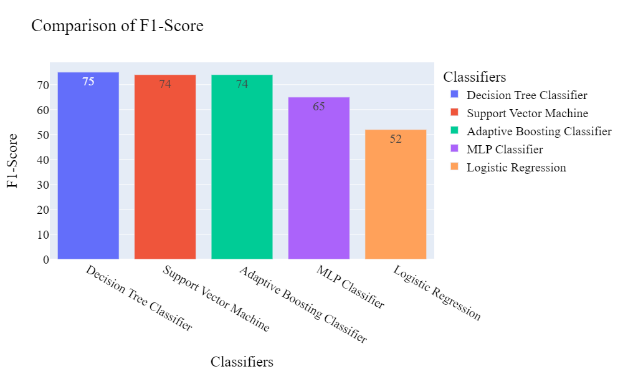


Figure 16 F1-Scores of Detection

## Comparison with Previous Models

The comparison of the previous approach has been discussed in section 2.6. From that table, the models related to diabetes detection have been graphically compared and presented below:

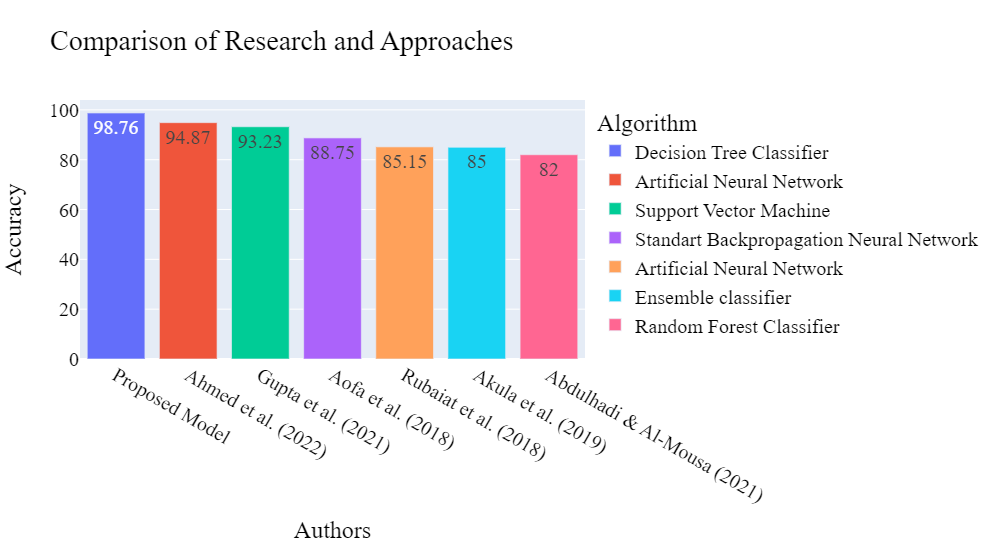


Figure 17 Comparison with Previous Approaches

From this comparison, it can be seen that the present approach has secured 4% more accuracy in detecting diabetes disease compared to the highest seen accuracy in the previous model (Ahmed, et al., 2022). So, the present research has the improvement to detect diabetes disease concerning the symptoms.

# Discussion and Conclusion

With the application of machine learning classifiers aims to detect diabetes disease have been fulfilled. The entire research has been done by following the research approach and concerning the objectives.

## Discussion on Research Questions

***How successfully the diabetes disease can be detected using the application of machine learning?***

For achieving moderate accuracies, the machine learning classifiers are applied to the original data. It has been seen that the Decision Tree Classifier has produced 75.32% accuracy, SVM has produced 75.97% accuracy, Adaptive Boosting Classifier has produced 74.68% accuracy, MLP Classifier has produced 67.53% accuracy, and LR has produced 62.34% accuracy. As the accuracies are not higher, so the data has been resampled, and the same process of detection has been conducted again. With this experiment, the decision tree has produced 98.76% accuracy in detecting diabetes. Thus, the first research question has answered that diabetes can be detected successfully using machine learning.

***How the most important symptoms of diabetes can be identified?***

The selected data contains different features based on which the target feature can be detected for diabetes. We can filter features as few of them has less significance in supporting our research. So, the rule has been designed to select important features which will be treated as the important symptoms of diabetes. By applying the rule to select features using correlation, the features or symptoms such as Glucose, BMI, Age, Pregnancies, Diabetes Pedigree Function and Insulin have been seen to be the most important symptoms.

***Which classifier or algorithm in machine learning is most effective in detecting diabetes disease?***

From the application of classifiers, it has been seen that the decision tree has produced 98.76% accuracy in detecting diabetes. The decision tree also detected diabetes disease with very few false positive and negative rates. So, this classifier has been taken as the most effective for detecting diabetes disease based on the important symptoms.

***What will the research achieve, and how it improve the detection process of diabetes at the early stage?***

In this project, the previous approaches have been studied, and several approaches and algorithms have been found to be applied by researchers. Out of those, the researchers such as (Aofa, et al., 2018), (Gupta, et al., 2021), (Abdulhadi & Al-Mousa, 2021), (Ahmed, et al., 2022), (Akula, et al., 2019) and (Rubaiat, et al., 2018) chosen machine learning and deep learning in detecting diabetes. The present research has been compared with the previous approaches. From this comparison, it can be seen that the present approach has secured 4% more accuracy in detecting diabetes disease compared to the highest seen accuracy in the previous model by (Ahmed, et al., 2022). So, the present research has the improvement to detect diabetes disease concerning the symptoms.

## Discussion on Hypothesis

From the analysis of the data and applying the correlation rule, it has been found that the level of glucose has significant effects on diabetes disease. It means that these two factors influence diabetes disease significantly. So, in that analytical outcome, it can be said that the Null Hypothesis is not true and rather, the alternative hypothesis can be accepted.

## Research Strength

Research has been done to detect diabetes disease, and the strength of the present model has been compared with the previous approaches. In the literature review chapter, the comparison of the approaches has been made with two types of segregation, namely disease detection of different types and diabetes disease detection. In the detection of disease, Kohli & Arora (2018) have been seen to achieve 98.57% accuracy. On the other hand, (Ahmed, et al., 2022) achieved 94.87% accuracy in detecting diabetes disease. So, compared to all previous models, it can be said that the present model based on the decision tree has gained the highest accuracy, with an accuracy rate of 98.76%.

## Research Limitations

1. The selected dataset contains limited instances or records of the patients.
2. In this overall research, only five classifiers from machine learning have been applied, and no other classifiers have been applied.
3. Deep learning classifiers have not been applied in the research.
4. To select the features, only pairwise correlation has been applied.

## Future Scopes

1. The project can be extended by applying deep learning models such as the convolutional neural network.
2. The application can be designed for the detection of diabetes detection by taking the symptoms from users.

## Conclusion

The project has been conducted based on the aim that has been taken in addition to the objectives that have been prepared. The process of execution has been followed by the proposed approach, which has been designed based on the objectives. A dataset has been extracted from Kaggle for the purpose of research. The data consists 268 observations of diabetes patients out of 768 observations and the rest 500 observations are of non-diabetes patients. For determining diabetic probability, nine features which illustrates the symptoms and historical records of patients. Next, the algorithms have been chosen based on the reviews of previous research, and those have been applied to the data to detect diabetes. In the first phase of the experiment with the original data, the accuracy was seen to be highest for the decision tree by 75.97%, which is lower to declare it a good model. So, as per the proposed approach, the data has been resampled, and the experiment has been run again. From the experiment on the resample data, it has been seen that the accuracy of the decision tree has increased from 75.32% to 98.76% with the data resampling. The accuracy has also been increased for Adaptive boosting from 74.68% to 83.92%. However, the accuracies and performances of the classifiers such as LR, MLP classifier and SVM have not been improved. So, the tree-based algorithm that is the decision tree has performed the best or most effective in detecting diabetes with the resampled data. So, the decision tree has gained 98.76% accuracy, and the detection rate has increased significantly. It has detected 519 instances out of 532 available instances doe diabetic patients and 998 instances out of 1004 instances of non-diabetic patients. Thus, finally, the decision tree has been taken as the final model to detect diabetes, and the performance has been compared with the previous models. From this comparison, it can be seen that the present approach has secured 4% more accuracy in detecting diabetes disease compared to the highest seen accuracy in the previous model by (Ahmed, et al., 2022). So, the present research has the improvement to detect diabetes disease concerning the symptoms.

# References

Abdulhadi, N. & Al-Mousa, A., 2021. Diabetes Detection Using Machine Learning Classification Methods. *In the 2021 International Conference on Information Technology (ICIT). IEEE.,* pp. pp. 350-354.

Ahmed, U. et al., 2022. Prediction of Diabetes Empowered With Fused Machine Learning. *IEEE Access, 10.,* pp. 8529-8538..

Akula, R., Nguyen, N. & Garibay, I., 2019. Supervised machine learning based ensemble model for accurate prediction of type 2 diabetes. *In 2019 SoutheastCon. IEEE.,* pp. pp. 1-8.

Almubark, I., Chang, L.-C. & Nguyen, T., 2019. Early Detection of Alzheimer’s Disease Using Patient Neuropsychological and Cognitive Data and Machine Learning Techniques. *IEEE International Conference on Big Data (Big Data),* pp. 5971-5973.

Amirgaliyev, Y. & Shamiluulu, S., 2018. Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods. *12th International Conference on Application of Information and Communication Technologies (AICT),,* pp. 1-4.

Aofa, F., Sasongko, P. S. & Adzani, W. A., 2018. Aofa, F., Sasongko, P. S., & Adzani, W. A. (2018, October). Early detection system of diabetes mellitus disease using ANN backpropagation with adaptive learning rate and PSO. *In 2018 2nd International Conference on Informatics and Computational Sciences (ICICoS). IEEE.,* pp. pp. 1-5.

Asif, M. A.-A.-R. et al., 2020. Computer Aided Diagnosis of Thyroid Disease Using Machine Learning Algorithms. *11th International Conference on Electrical and Computer Engineering (ICECE),* pp. 1-4.

Basha, N. & Kumar, P. A., 2019. Early Detection of Heart Syndrome Using Machine Learning Technique. *4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT),* pp. 387-391.

Brankovic, A. & Zamani, A., 2019. Electromagnetic Based Fatty Liver Detection Using Machine Learning. *13th European Conference on Antennas and Propagation (EuCAP),* pp. 1-3.

Chauhan, T., Rawat, S., Malik, S. & Singh, P., 2021. Supervised and Unsupervised Machine Learning based Review on Diabetes Care. *7th International Conference on Advanced Computing and Communication Systems (ICACCS),* pp. 1-5.

Cıhan, P. & Coşkun, H., 2021. Performance Comparison of Machine Learning Models for Diabetes Prediction. *29th Signal Processing and Communications Applications Conference (SIU),* pp. 1-5.

Cleveland Clinic, 2022. *Diabetes: An Overview.* [Online]

Available at: https://my.clevelandclinic.org/health/diseases/7104-diabetes-mellitus-an-overview

Costea, N. E., Moisi, E. V. & Popescu, D. E., 2021. Comparison of Machine Learning Algorithms for Prediction of Diabetes. *16th International Conference on Engineering of Modern Electric Systems (EMES),* pp. 1-5.

Dansinger, M., 2020. *Diagnosis of Diabetes.* [Online]

Available at: https://www.webmd.com/diabetes/guide/diagnosis-diabetes

Dunbray, N. et al., 2021. A Novel Prediction Model for Diabetes Detection Using Gridsearch and A Voting Classifier between Lightgbm and KNN. *In 2021 2nd Global Conference for Advancement in Technology (GCAT). IEEE.,* pp. pp. 1-7.

Dunbray, N. et al., 2021. A Novel Prediction Model for Diabetes Detection Using Gridsearch and A Voting Classifier between Lightgbm and KNN. *2nd Global Conference for Advancement in Technology (GCAT),* pp. 1-5.

Guide, B. & Mishra, S., 2020. A Novel Approach to Predict Chronic Kidney Disease Using Machine Learning Algorithms. *4th International Conference on Electronics, Communication and Aerospace Technology (ICECA),* pp. 1630-1635.

Gupta, S. et al., 2021. Data-driven Preprocessing Techniques for Early Diagnosis of Diabetes, Heart and Liver Diseases. *In 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT). IEEE.,* pp. pp. 1-8.

Hasan, S. M., Rabbi, M. F., Champa, A. I. & Zaman, M. A., 2020. A Machine Learning-Based Model for Early Stage Detection of Diabetes. *In 2020 23rd International Conference on Computer and Information Technology (ICCIT). IEEE.,* pp. pp. 1-6.

Kaggle, 2016. *Pima Indians Diabetes Database.* [Online]

Available at: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[Accessed 2021].

Katarya, R. & Jain, S., 2020. Comparison of Different Machine Learning Models for diabetes detection. *IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE),* pp. 1-4.

Katarya, R. & Srinivas, P., 2020. Predicting Heart Disease at Early Stages using Machine Learning: A Survey. *International Conference on Electronics and Sustainable Communication Systems (ICESC),* pp. 302-305.

Kohli, P. S. & Arora, S., 2018. Application of Machine Learning in Disease Prediction. *4th International Conference on Computing Communication and Automation (ICCCA),* pp. 1-4.

Kumari, A. & Mehta, A. K., 2021. A Novel Approach for Prediction of Heart Disease using Machine Learning Algorithms. *Asian Conference on Innovation in Technology (ASIANCON),* pp. 1-5.

Kumari, L. R. et al., 2021. Machine Learning based Diabetes Detection. *International Conference on Communication and Electronics Systems (ICCES),* pp. 1-6.

Kuzhippallil, M. A. & Joseph, C., 2020. Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients. *6th International Conference on Advanced Computing and Communication Systems (ICACCS),* pp. 778-782.

Mohan, N. & Jain, V., 2020. Performance Analysis of SVM in Diabetes Prediction. *4th International Conference on Electronics, Communication and Aerospace Technology (ICECA),* pp. 1-4.

NIDDK, 2022. *What is Diabetes?.* [Online]

Available at: https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes#:~:text=Diabetes%20is%20a%20disease%20that,to%20be%20used%20for%20energy.

Qin, J. & Chen, L., 2020. A Machine Learning Methodology for Diagnosing Chronic Kidney Disease. *IEEE Access, vol. 8,* pp. 20991-21002.

Rahman, T. M. & Siddiqua, S., 2019. Early Detection of Kidney Disease Using ECG Signals Through Machine Learning Based Modelling. *International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST),* pp. 319-323.

Rubaiat, S. Y., Rahman, M. M. & Hasan, M. K., 2018. Important feature selection & accuracy comparisons of different machine learning models for early diabetes detection. *In 2018 International Conference on Innovation in Engineering and Technology (ICIET). IEEE.,* pp. pp. 1-6.

Shakeel, W. & Ahmad, M., 2020. Early Detection of Cercospora Cotton Plant Disease by Using Machine Learning Technique. *30th International Conference on Computer Theory and Applications (ICCTA),* pp. 44-48.

Shung, K. P., 2018. *Accuracy, Precision, Recall or F1?.* [Online]

Available at: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

Singh, V. & Gourisaria, M. K., 2021. Performance Analysis of Machine Learning Algorithms for Prediction of Liver Disease. *IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON),* pp. 1-7.

Sivasangari, A. & Krishna Reddy, B. J., 2020. Diagnosis of Liver Disease using Machine Learning Models. *Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC),* pp. 627-630.

Tavasoli, S., 2022. *Top 10 Machine Learning Algorithms For Beginners: Supervised, Unsupervised Learning and More.* [Online]

Available at: https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article

The Victorian Government, 2021. *Diabetes - long-term effects.* [Online]

Available at: https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/diabetes-long-term-effects

Upgrade, 2021. *Top 10 Reasons Why Python is So Popular With Developers in 2022.* [Online]

Available at: https://www.upgrad.com/blog/reasons-why-python-popular-with-developers/#:~:text=The%20python%20language%20is%20one,faster%20than%20other%20programming%20languages.

Wakefield, K., 2022. *Predictive Modeling Analytics and Machine Learning.* [Online]

Available at: https://www.sas.com/en\_gb/insights/articles/analytics/a-guide-to-predictive-analytics-and-machine-learning.html

Wang, W. & Lee, J., 2020. Early Detection of Parkinson’s Disease Using Deep Learning and Machine Learning. *IEEE Access, vol. 8,* pp. 147635-147646.

Yahyaoui, A. & Yumuşak, N., 2021. Deep And Machine Learning Towards Pneumonia And Asthma Detection. *International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT),* pp. 494-497.

# Appendix

**# <span style='color:#2916F5'> Importaing Libraries </span>**

**import warnings**

**warnings.filterwarnings("ignore")**

**import numpy as np**

**import pandas as pd**

**import os**

**import matplotlib.pyplot as plt**

**import seaborn as sns**

**import plotly.express as px**

**from sklearn.linear\_model import LogisticRegression**

**from sklearn.ensemble import AdaBoostClassifier**

**from sklearn.neural\_network import MLPClassifier**

**from sklearn.tree import DecisionTreeClassifier**

**from sklearn.svm import LinearSVC**

**from sklearn.pipeline import make\_pipeline**

**from sklearn.preprocessing import StandardScaler**

**from sklearn.utils import resample**

**from sklearn.utils import shuffle**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.metrics import precision\_score, recall\_score, f1\_score, classification\_report,accuracy\_score**

**# <span style='color:#2916F5'> Data Reading </span>**

**## <span style='color:#2916F5'>Loading Data </span>**

**##pip install plotly**

**diab=pd.read\_csv("diabetes.csv")**

**diab.head()**

**## <span style='color:#2916F5'>Formatting Target label </span>**

**diab.Outcome=diab.Outcome.replace(diab.Outcome.unique(),["Diabetic","Non-Diabetic"])**

**diab.head()**

**## <span style='color:#2916F5'>Data Information </span>**

**diab.info()**

**diab.Outcome.value\_counts()**

**## <span style='color:#2916F5'>Determining Missing Values</span>**

**diab.isna().sum()**

**# <span style='color:#2916F5'> Data Analysis and Visualization </span>**

**txtcol="#046307"**

**tcol=["#228B22","#2916F5"]**

**## <span style='color:#2916F5'>Function to Create Chart</span>**

**def create\_chart(df,v1,v2,ctp):**

**pd.crosstab(df[v1],df[v2]).plot(kind=ctp,figsize=(9,4),color=tcol)**

**plt.title('Posibility of Diabetes by {}'.format(v1), fontsize=15,color=txtcol)**

**plt.xlabel('{}'.format(v1), fontsize=12,color=txtcol)**

**plt.ylabel('Count', fontsize=12,color=txtcol)**

**plt.grid()**

**plt.show()**

**## <span style='color:#2916F5'>Function to Create Bar Chart</span>**

**def create\_bar(df1,df2,v1):**

**avg1=df1[v1].mean()**

**avg2=df2[v1].mean()**

**cat=["Diabetic","Non-Diabetic"]**

**val=[avg1,avg2]**

**plt.figure(figsize=(6,4))**

**plt.title("Average {} of Diabetic and Non-Diabetic People".format(v1),fontsize=18,color=txtcol)**

**plt.xlabel("Type",fontsize=18,color=txtcol)**

**plt.ylabel("Agerage {}".format(v1),fontsize=18,color=txtcol)**

**plt.bar(cat,val,color=tcol)**

**plt.grid()**

**plt.show()**

**create\_chart(diab,"Age","Outcome","line")**

**create\_chart(diab,"Pregnancies","Outcome","bar")**

**create\_chart(diab,"Glucose","Outcome","area")**

**create\_chart(diab,"BMI","Outcome","area")**

**dbp=diab[diab['Outcome']=="Diabetic"]**

**ndbp=diab[diab['Outcome']=="Non-Diabetic"]**

**create\_bar(dbp,ndbp,"Age")**

**create\_bar(dbp,ndbp,"Glucose")**

**create\_bar(dbp,ndbp,"BMI")**

**create\_bar(dbp,ndbp,"Insulin")**

**## <span style='color:#2916F5'>Diabetes Detection</span>**

**### <span style='color:#2916F5'>Feature Selection</span>**

**diab1=diab.copy()**

**diab1['Outcome']=diab1['Outcome'].replace(diab1['Outcome'].unique(),[i+1 for i in range(len(diab1['Outcome'].unique()))])**

**crdiab=diab1.corr()**

**plt.figure(figsize=(10,6))**

**plt.title("Correlation Heatmap of Features", fontsize=22,color="#F6BE00")**

**sns.heatmap(crdiab,annot=True,cmap='YlOrRd',fmt='.3f',linewidths=1)**

**plt.show()**

**selected=crdiab['Outcome'].sort\_values()[:-3].index.tolist()**

**print("Selected Features:\n",\*selected, sep="\n")**

**### <span style='color:#2916F5'>Initiating Predictors and Target Feature and Split the data</span>**

**X=diab[selected]**

**y=diab['Outcome']**

**x\_train,x\_test,y\_train,y\_test=train\_test\_split(X,y, train\_size=0.8, random\_state=10)**

**print("===========================================================")**

**print("==================== Data Splitting =======================")**

**print("\tSplit Ratio (Train : Test): {}% : {}%".format(round((len(x\_train)/len(X))\*100),round((len(x\_test)/len(X))\*100)))**

**print("===========================================================")**

**print("\t Train Set: {}".format(round(len(X)\*(0.8))))**

**print("\t Test Set: {}".format(round(len(X)\*(1-0.8))))**

**print("===========================================================")**

**### <span style='color:#2916F5'>Diabetes Detection using ML</span>**

**clf\_diab=[**

**MLPClassifier(hidden\_layer\_sizes=(80,),activation='relu',solver="sgd",learning\_rate="adaptive",learning\_rate\_init=0.0001),**

**LogisticRegression(penalty="elasticnet",solver="saga",tol=0.0004, C=0.6,l1\_ratio=0.8),**

**AdaBoostClassifier(n\_estimators=100,learning\_rate=0.9),**

**DecisionTreeClassifier(criterion='gini',max\_features='sqrt',splitter='best',min\_samples\_split=4,max\_depth=12),**

**make\_pipeline(StandardScaler(), LinearSVC(random\_state=0, tol=0.001)),**

**]**

**clf\_diab\_names=[**

**"MLP Classifier",**

**"Logistic Regression",**

**"Adaptive Boosting Classifier",**

**"Decision Tree Classifier",**

**"Support Vector Machine"**

**]**

**diabperf=[[],[],[],[]]**

**backupscvores1=[]**

**print("\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_")**

**for i in range(len(clf\_diab)):**

**print(" {} ".format(clf\_diab\_names[i]))**

**print("\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_")**

**cvscore=[[],[],[],[],[],[]]**

**for ts in range(10):**

**clf\_diab[i].fit(x\_train,y\_train)**

**y\_pred=clf\_diab[i].predict(x\_test)**

**cvscore[0].append(round(accuracy\_score(y\_test,y\_pred)\*100,2))**

**cvscore[1].append(round(precision\_score(y\_test, y\_pred, average='weighted'),2)\*100)**

**cvscore[2].append(round(recall\_score(y\_test, y\_pred, average='weighted'),2)\*100)**

**cvscore[3].append(round(f1\_score(y\_test, y\_pred, average='weighted'),2)\*100)**

**cm=pd.crosstab(y\_test, y\_pred, rownames=['True'], colnames=['Predicted'], margins=True)**

**cvscore[4].append(cm.iloc[:2,:2])**

**cvscore[5].append(classification\_report(y\_test, y\_pred))**

**backupscvores1.append(cvscore[0])**

**mxaccidx=cvscore[0].index(max(cvscore[0]))**

**diabperf[0].append(cvscore[0][mxaccidx])**

**diabperf[1].append(cvscore[1][mxaccidx])**

**diabperf[2].append(cvscore[2][mxaccidx])**

**diabperf[3].append(cvscore[3][mxaccidx])**

**print("\nAccuracy: {}%\n".format(cvscore[0][mxaccidx]))**

**print("\nClassification Report for {} \n\n{}".format(clf\_diab\_names[i],cvscore[5][mxaccidx]))**

**print("\nConfusion Matrix for {} \n\n{}\n".format(clf\_diab\_names[i],cvscore[4][mxaccidx]))**

**print("\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_")**

**### <span style='color:#2916F5'>Performance Comparison for Normal Data</span>**

**diab\_df=pd.DataFrame({**

**"Classifiers":clf\_diab\_names,**

**"Accuracy":diabperf[0],**

**"Precision":diabperf[1],**

**"Recall":diabperf[2],**

**"F1-Score":diabperf[3]**

**})**

**for i in diab\_df.columns.tolist()[1:]:**

**diab\_df=diab\_df.sort\_values(by=i,ascending=False)**

**fig = px.bar(diab\_df, y=i, x="Classifiers",text=i,color="Classifiers",**

**title="Comparison of {}".format(i),height=600,width=1000)**

**fig.update\_layout(**

**font=dict(**

**family="Times New Roman, Bold",**

**size=20,**

**color="black"**

**)**

**)**

**fig.show()**

**diab\_df.to\_csv("diab\_df.csv")**

**## <span style='color:#2916F5'>Experimenting Using Resampled Data</span>**

**diabres=resample(diab, replace = True, n\_samples = len(diab)\*10, random\_state = 10)**

**print(diabres.shape)**

**X1=diabres[selected]**

**y1=diabres['Outcome']**

**x\_train1,x\_test1,y\_train1,y\_test1=train\_test\_split(X1,y1, train\_size=0.8, random\_state=10)**

**print("===========================================================")**

**print("==================== Data Splitting =======================")**

**print("\tSplit Ratio (Train : Test): {}% : {}%".format(round((len(x\_train1)/len(X1))\*100),round((len(x\_test1)/len(X1))\*100)))**

**print("===========================================================")**

**print("\t Train Set: {}".format(round(len(X1)\*(0.8))))**

**print("\t Test Set: {}".format(round(len(X1)\*(1-0.8))))**

**print("===========================================================")**

**diabperf1=[[],[],[],[]]**

**backupscvores2=[]**

**print("\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_")**

**for i in range(len(clf\_diab)):**

**print(" {} ".format(clf\_diab\_names[i]))**

**print("\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_")**

**cvscore1=[[],[],[],[],[],[]]**

**for ts in range(10):**

**clf\_diab[i].fit(x\_train1,y\_train1)**

**y\_pred1=clf\_diab[i].predict(x\_test1)**

**cvscore1[0].append(round(accuracy\_score(y\_test1,y\_pred1)\*100,2))**

**cvscore1[1].append(round(precision\_score(y\_test1, y\_pred1, average='weighted'),2)\*100)**

**cvscore1[2].append(round(recall\_score(y\_test1, y\_pred1, average='weighted'),2)\*100)**

**cvscore1[3].append(round(f1\_score(y\_test1, y\_pred1, average='weighted'),2)\*100)**

**cm=pd.crosstab(y\_test1, y\_pred1, rownames=['True'], colnames=['Predicted'], margins=True)**

**cvscore1[4].append(cm.iloc[:2,:2])**

**cvscore1[5].append(classification\_report(y\_test1, y\_pred1))**

**backupscvores2.append(cvscore1[0])**

**mxaccidx=cvscore1[0].index(max(cvscore1[0]))**

**diabperf1[0].append(cvscore1[0][mxaccidx])**

**diabperf1[1].append(cvscore1[1][mxaccidx])**

**diabperf1[2].append(cvscore1[2][mxaccidx])**

**diabperf1[3].append(cvscore1[3][mxaccidx])**

**print("\nAccuracy: {}%\n".format(cvscore1[0][mxaccidx]))**

**print("\nClassification Report for {} \n\n{}".format(clf\_diab\_names[i],cvscore1[5][mxaccidx]))**

**print("\nConfusion Matrix for {} \n\n{}\n".format(clf\_diab\_names[i],cvscore1[4][mxaccidx]))**

**print("\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_")**

**### <span style='color:#2916F5'>Performance Comparison for Resampled Data</span>**

**diab\_df1=pd.DataFrame({**

**"Classifiers":clf\_diab\_names,**

**"Accuracy":diabperf1[0],**

**"Precision":diabperf1[1],**

**"Recall":diabperf1[2],**

**"F1-Score":diabperf1[3]**

**})**

**for i in diab\_df1.columns.tolist()[1:]:**

**diab\_df1=diab\_df1.sort\_values(by=i,ascending=False)**

**fig = px.bar(diab\_df1, y=i, x="Classifiers",text=i,color="Classifiers",**

**title="Comparison of {} (Resampled)".format(i),height=600,width=1000)**

**fig.update\_layout(**

**font=dict(**

**family="Times New Roman, Bold",**

**size=20,**

**color="black"**

**)**

**)**

**fig.show()**

**diab\_df1.to\_csv("diab\_df1.csv")**

**diab\_df1**

**auths=["Aofa et al. (2018)","Gupta et al. (2021)",**

**"Abdulhadi & Al-Mousa (2021)","Ahmed et al. (2022)",**

**"Akula et al. (2019)","Rubaiat et al. (2018)","Proposed Model"]**

**acc=[88.75,93.23,82,94.87,85,85.15,diab\_df1['Accuracy'].tolist()[0]]**

**algn=["Standart Backpropagation Neural Network ","Support Vector Machine","Random Forest Classifier ",**

**"Artificial Neural Network ","Ensemble classifier","Artificial Neural Network",diab\_df1['Classifiers'].tolist()[0]]**

**athdf=pd.DataFrame({"Authors":auths,"Accuracy":acc,"Algorithm":algn})**

**athdf=athdf.sort\_values(by="Accuracy",ascending=False)**

**fig = px.bar(athdf, x="Authors", y="Accuracy",text="Accuracy",color="Algorithm",**

**title="Comparison of Research and Approaches".format(i),height=550,width=1000)**

**fig.update\_layout(**

**font=dict(**

**family="Times New Roman, Bold",**

**size=20,**

**color="black"**

**)**

**)**

**fig.show()**