1. **INTRODUCTION**

Machine learning techniques are a great way to train a system for the prediction and classification of anything. Machine learning entails deducing structures from a given input. Machine learning has recently proven to be quite effective in the design, development, and training of models for disease prediction. Machine learning has gotten a lot of interest in the medical profession since it takes less time to process data and requires less user interaction, which saves money on health care.

Diabetes is a chronic illness (life-long). Diabetes is caused by the failure of the human body to create enough insulin to manage glucose levels after eating. Diabetes is caused by a lack of ability to create enough insulin and a lack of insulin tolerance. Diabetes causes a slew of health issues.

One of the most common methods for predicting or discovering an underlying process is through machine learning. Machine training is thus focused on the method or algorithm that produces intelligent results by recognizing detailed patterns. The underlying data mechanism is assumed to have properties such as models or predictions. A learner can use examples (data) to capture the interesting features of their unknown underlying probability distribution. Data might be viewed as illustrations of possible correlations among the variables observed.

Machine learning has designed different patterns to make intelligent decisions about input data. The biggest challenge in machine learning is the behavior of the inputs trained in the observed examples. Therefore, they are trained to output efficiently and sensitively at all possible inputs. The next chapter outlines various machine learning methods for detecting diabetes, outlines their strengths and weaknesses, and searches a wide range of literature with the precision achieved by several researchers.

Diabetes is a varied set of conditions that can lead to an increase in glucose in the blood and a shortage of glucose in the urine. It is not a genetic disorder. Diabetes is caused by a combination of genetics, lifestyle, and environment. Being overweight and following a risky weight loss strategy both contribute to the development of diabetes. High blood sugar levels can cause renal disease and coronary heart disease. Blood sugar levels that are too high can destroy your body's small blood vessels. Diabetes symptoms include blurry vision, intense hunger, unexpected weight loss, frequent urination, and thirst. Glucose, Blood Pressure, Pore and Skin Thickness, Insulin, and Age are the characteristics employed in this article to locate diabetes. The healthcare industry generates a large number of statistical units. Those data sets are a compilation of diabetic patient information from hospitals. Big data analytics is a type of processing that examines data units and displays hidden data. This data comes from the National Institute of Diabetes and Digestive Diseases' Pima Indians Diabetes Database (PIDD). The dataset's goal is to determine whether or not a patient has diabetes based on diagnostic metrics in the collection. The enormous database provided several constraints.

Diabetes is a harmful disease all over the world. Diabetes is caused by obesity and high blood sugar. It affects the hormone insulin, which causes abnormal metabolism of cancer and improves blood sugar levels. Diabetes occurs when the body does not produce enough insulin. According to the World Health Organization (WHO), about 422 million people have diabetes, especially in low-income or low-income countries. And this could increase to 490 billion by 2030. However, the prevalence of diabetes is found in different countries such as Canada, China, and India. The actual number of diabetics in India is 40 million, as India's population is currently over 100 million.

Diabetes is one of the leading causes of death worldwide. Early prediction of illnesses such as diabetes can save lives. To achieve this, this task examines diabetes predictions using various diabetes-related attributes. To do this, we use the Pima Indian Diabetes Dataset and apply various classifications and ensemble machine learning techniques to predict diabetes. Machine learning is a method used to explicitly train a computer or machine. Various machine learning techniques can be used to create different classification and ensemble models from the collected datasets, resulting in efficient results for gathering knowledge. The data collected in this way helps predict diabetes. You can make predictions with a variety of machine learning techniques, but it is difficult to choose the best one. Therefore, for this purpose, we apply common classification and ensemble techniques to predictive datasets.

**1.1EXISTING SYSTEM:**

Diabetes is a metabolic disorder that prevents the human body from processing blood sugar levels, known as blood sugar levels. The disease is characterized by hyperglycemia due to defects in insulin secretion, insulin action, or both. An absolute lack of insulin secretion causes type 1 diabetes (T1D). Diabetes spreads dramatically because patients cannot use the insulin produced. Talk about type 2 diabetes (T2D). Both types increase rapidly, but the rate of increase in T2D is higher than that in T1D. 90-95% of diabetic cases are T2D.

According to the World Health Organization (WHO), about 1.6 million people die from diabetes each year. Diabetes is a type of illness that occurs when the body's blood sugar/glucose levels are very high. According to health experts, diabetes occurs when the human pancreas is unable to produce enough insulin (type 1 diabetes) and the cells of the body cannot use the insulin produced (type 2 diabetes). When we eat food, glucose is released after the digestive process. Insulin is a blood hormone that tells cells to enter cells from the blood and use blood sugar levels to convert them into energy. If the pancreas cannot produce enough insulin, the cells cannot take up glucose and glucose stays in the blood. Therefore, blood sugar/blood sugar levels in the blood rise to very unacceptable levels. Due to high blood sugar, some symptoms appear in the human body, such as extreme hunger, severe thirst, and frequent urination. The normal blood sugar level in the human body is 70-99 mg per deciliter. If your blood glucose is above 126 mg/dl, this indicates diabetes. A person is considered to have prediabetes if the body's glucose concentration is 100-125 mg/dl.

**DISADVANTAGES OF EXISTING SYSTEMS**

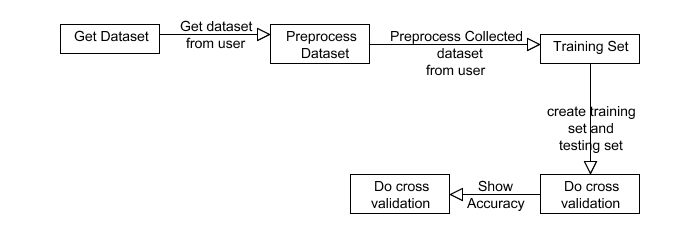
* Using machine learning the accuracy of detection is less.
* High false positives.
* There is no interactive tool for users to predict

**1.2 PROPOSED SYSTEMS:**

The proposed diabetes prediction system has two main phases that work together to achieve the desired results. The first step in the proposed system is data preparation and the second step is classification. However, the input to the system is the PID dataset and the output is a class that represents health or diabetes.

The purpose of the studies conducted is to classify the available data into diabetic or non-diabetic with the help of monitored learning algorithms. The dataset is divided into training sets and test sets. More data needs to be trained for accuracy. Next, perform a comparative analysis of the results of the diabetes early-detection algorithm. Models such as support vector machines and logistic regression have proven to be most useful in detecting diabetes in patients. The central goal of our model is to achieve higher accuracy and overall improvement in the early detection of diabetes.

**Proposed system framework**



**1.2** Proposed framework

The purpose of this task is to investigate whether the model can more accurately predict diabetes. We experimented with various classification and ensemble algorithms to predict diabetes. The following is a brief description of the phases.

**Dataset Description**- the data is gathered from the UCI repository which is named Pima Indian Diabetes Dataset. The dataset has many attributes of 768 patients.

Table 1.2.1: Dataset Description

|  |  |
| --- | --- |
| S No. | Attributes |
| 1 | Pregnancy |
| 2 | Glucose |
| 3 | Blood Pressure |
| 4 | Skin thickness |
| 5 | Insulin |
| 6 | BMI(Body Mass Index) |
| 7 | Diabetes Pedigree Function |
| 8 | Age |

The 9th attribute is the class variable of each data point. This class variable shows the outcomes 0 and 1 for diabetics which indicates positive or negative for diabetics.

**Distribution of Diabetic patients**- We have created a model that predicts diabetes, but the dataset is slightly imbalanced, with about 500 classes showing no diabetes with a negative mark of 0 and 268 classes showing diabetes with a mark of 1. It shows that.

**Data Preprocessing**- Data preprocessing is the most important process. In most cases, health-related data contains missing values ​​and other impurities that can reduce the effectiveness of the data. Data preprocessing is performed to improve the quality and effectiveness achieved after the mining process. For the effective application of machine learning techniques to datasets, this process is essential for accurate results and successful predictions. For Pima Indian Diabetes datasets, pretreatment is required in two steps.

**Missing Values removal**- Delete all instances that have a value of zero (0). It cannot have zero as a value. Therefore, this instance will be deleted. Create a feature subset by eliminating irrelevant features/instances. This process, called feature subset selection, reduces the dimensions of the data and speeds up the operation of the data.

**Splitting of data**- After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train the algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Normalization aims to bring all the attributes under the same scale.

**Apply Machine Learning-** When your data is ready, use machine learning technology. We use different classification and ensemble techniques to predict diabetes. The method was applied to Pima Indian diabetes records. The main goal is to apply machine learning techniques to analyze the performance of these methods, determine their accuracy, and find responsible/important characteristics that play a key role in prediction.

**Support Vector Machine**- Support vector machines, also known as SVMs, are supervised machine learning algorithms. Sam is the most popular classification method. SVM creates a hyperplane that separates the two classes. You can create hyperplanes or sets of hyperplanes in high-dimensional space. This hyperplane can also be used for classification or regression. Some can also distinguish instances of a particular class and classify entities that are not supported by the data. Separation is done through a hyperplane that separates to the nearest training point in any class.

**K-Nearest Neighbor-** KNN is also a supervised machine learning algorithm. KNN helps solve both classification and regression problems. ANN is a lazy forecasting method. KNN assumes that similar things are close to each other. Similar data points are often very close together. KNN helps group new works based on a measure of similarity. The ANN algorithm records all data records and classifies them according to similarity. Use a tree-like structure to find the distance between points. To predict a new data point, the algorithm finds the next data point in the nearest training dataset. Where K = the number of neighbors, always a positive integer. Neighbor values ​​are selected from several classes. Proximity is primarily defined by the Euclidean distance. The Euclidean dis- tance between two points P and Q i.e. P (p1,p2, Pn) and Q (q1, q2,..qn) is defined by the following equation:-

Algorithm-

* Take a sample dataset of columns and rows named Pima Indian Diabetes data set.
* Take a test dataset of attributes and rows.
* Find the Euclidean distance with the help of formula-
* Then, Decide a random value of K. is the no. of nearest neighbors
* Then with the help of these minimum distance and Euclidean distance find out the nth column of each.
* Find out the same output values.

If the values are the same, then the patient is diabetic, otherwise not.

**Decision Tree**- Decision trees are the basic method of classification. It is a supervised learning method. Decision trees are used when the response is of categorical type. Decision trees have a structure-based model, such as a tree that describes a classification process based on input characteristics. Input variables can be of any type, such as graph, text, discrete, continuous, and so on.

**Logistic Regression-**Logistic regression is also a supervised learning classification algorithm. It is used to estimate the probability of a binary answer based on one or more predictors. They can be continuous or unobtrusive. Logistic regression is used to classify or distinguish some data items into categories.

It classifies the data in binary form means only in 0 and 1 which refer to the cases to classify positive patients or negative for diabetes.

The main aim of logistic regression is to best fit which is responsible for describing the relationship between target and predictor variable. Logistic regression is based on the Linear regression model. The logistic regression model uses a sigmoid function to predict the probability of positive and negative classes.

Sigmoid function P = 1/1+e – (a+bx) Here P = probability, a and b = parameter of Model.

**Ensembling**- Ensembling is a machine learning technique Ensemble means using multiple learning algorithms to- gether for some task. It provides better prediction than any other individual model that’s why it is used. The main cause of the error is noise bias and variance, ensemble methods help to reduce or minimize these errors. There are two popular ensemble methods as Bagging, Boosting, ada-boosting, Gradient boosting, voting, averaging, etc. Here In this work, we have used Bagging (Random forest) and Gradient boosting ensemble methods for predicting diabetes.

**Random Forest** It is a type of ensemble learning method and is also used for classification and regression tasks. The accuracy it gives is greater than compared to other models. This method can easily handle large datasets. Random Forest is developed by Leo Bremen. It is a popular - ensemble Learning Method. Random Forest Improve Performance of Decision Tree by reducing variance. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.

**Gradient Boosting** – Gradient Boosting is the most powerful ensemble technique used for prediction and it is a classification technique. It combines week learners to make strong learner models for prediction. It uses the Decision Tree model. it classifies complex data sets and it is a very ef- effective and popular method. In gradient boosting model performance improve over iterations.

MODEL BUILDING

This is the most important phase which includes model building for the prediction of diabetes. In this, we have implemented various machine learning algorithms which are discussed above for diabetes prediction.

**2. REQUIREMENT ANALYSIS**

Requirement Analysis, also known as Requirement Engineering, is the process of defining user expectations for a new software being built or modified. In software engineering, it is sometimes referred to loosely by names such as requirements gathering or requirements capturing. Requirements analysis encompasses those tasks that go into determining the needs or conditions to meet for a new or altered product or project, taking account of the possibly conflicting requirements of the various stakeholders, analyzing, documenting, validating, and managing software or system requirements. Here are the objectives for performing requirement analysis in the early stage of a software project.

**2.1 Hardware requirements:**

• OS: Windows 7 with SP1; Recommended: Windows 10

• CPU: Intel or AMD processor with 64-bit support; Recommended: 2.8 GHz or

faster processor

• Disk Storage: 8 GB minimum

• Monitor Resolution: 1280x800; Recommended: 1920x1080

• Internet: Internet connection required for software activation

**2.2 Software requirements:**

• Python version 3.6 (any Python version 3. x will be fine).

• VSCode can be used to edit the code and to run the code as well.

• Anaconda Python 3 for installing Python and the required modules.

• You can use any OS—macOS, Windows, and Linux-based OS—with this book. We

recommend you have at least 4 GB RAM in your system

3. **GLOSSARY/MODULES**

**GLOSSARY**

A glossary of terms commonly associated with diabetes.

##### Beta cells

Beta cells are found in the pancreas that produces, store, and release insulin.

##### Cardiovascular disease (CVD)

Diseases and injuries of the circulatory system: the heart, the blood vessels of the heart, and the system of blood vessels throughout the body and to (and in) the brain. [CVD](https://www.idf.org/our-activities/care-prevention/cardiovascular-disease.html)generally refers to conditions that involve narrowed or blocked blood vessels.

##### Diabetes complications

Acute and chronic conditions are caused by diabetes. Acute [complications](https://www.idf.org/aboutdiabetes/complications.html) include diabetic ketoacidosis (DKA), hyperglycaemic hyperosmolar syndrome (HHS), hyperglycaemic diabetic coma, seizures or loss of consciousness, and infections. Chronic microvascular complications include [retinopathy](https://www.idf.org/our-activities/care-prevention/eye-health.html) (eye disease), [nephropathy](https://www.idf.org/our-activities/care-prevention/diabetes-and-the-kidney.html) (kidney disease), neuropathy (nerve disease), and periodontitis (inflammation of the tissue surrounding the tooth), whereas chronic macrovascular complications are [cardiovascular disease](https://www.idf.org/our-activities/care-prevention/cardiovascular-disease.html) (disease of the circulatory system), diabetic encephalopathy (brain dysfunction) and [diabetic foot](https://www.idf.org/our-activities/care-prevention/diabetic-foot.html) (foot ulceration and amputation).

##### Diabetes (Mellitus)

A condition that arises when the pancreas does not produce enough insulin or when the body cannot effectively use insulin. The three most common [types of diabetes](https://www.idf.org/component/content/?id=450&Itemid=922) are type 1, type 2, and gestational.

##### Diabetic foot

A foot that exhibits any disease that results directly from diabetes or a complication of diabetes.

##### Gestational diabetes mellitus (GDM)

Hyperglycaemia (high blood glucose level) that is first detected during pregnancy is classified as either [gestational diabetes mellitus (GDM)](https://www.idf.org/our-activities/care-prevention/gdm.html) or diabetes mellitus in pregnancy. Women with slightly elevated blood glucose levels are classified as having GDM and women with substantially elevated blood glucose levels are classified as women with diabetes in pregnancy.

##### Glucagon

A hormone produced in the pancreas. If blood glucose levels decrease, it triggers the body to release stored glucose into the bloodstream.

##### Glucose

Also called dextrose or blood sugar. The main sugar the body produces to store energy from proteins, fats, and carbohydrates. Glucose is the major source of energy for living cells and is carried to each cell through the bloodstream. However, the cells cannot use glucose without the help of insulin.

##### Glycogen

A form of glucose that is used for storing energy in the liver and muscles. If blood glucose levels decrease, the hormone glucagon triggers the body to convert glycogen to glucose and release it into the bloodstream.

##### Glycosylated hemoglobin A1c (HbA1c)

Hemoglobin to which glucose is bound. Glycosylated hemoglobin is tested to determine the average level of blood glucose over the past two to three months.

##### Hyperglycaemia

A raised level of glucose in the blood. It occurs when the body does not have enough insulin or cannot use the insulin it does have to turn glucose into energy. Signs of hyperglycemia include excessive thirst, dry mouth, and the need to urinate often.

##### Hypoglycemia

A lowered level of glucose in the blood. This occurs when a person with diabetes has injected too much insulin, eaten too little food, or has exercised without extra food. A person with [hypoglycemia](https://www.idf.org/our-activities/care-prevention/hypoglycaemia.html) may feel nervous, shaky, weak, or sweaty, and have a headache, blurred vision, and hunger.

##### Impaired fasting glucose (IFG)

Blood glucose is higher than normal blood glucose, but below the diagnostic threshold for diabetes after fasting (typically after an overnight fast).

##### Impaired glucose tolerance (IGT)

Blood glucose that is higher than normal blood glucose, but below the diagnostic threshold for diabetes after ingesting a standard amount of glucose during an oral glucose tolerance test.

##### Insulin

A hormone produced in the pancreas. If blood glucose levels increase, insulin triggers cells to take up glucose from the bloodstream and convert it to energy, and the liver takes up glucose from the bloodstream and store it as glycogen.

##### Monogenic diabetes

A less common type of diabetes, which arises as a result of a genetic mutation.  Examples include Maturity-Onset Diabetes of the Young (MODY) and Neonatal Diabetes Mellitus.

##### Nephropathy

Damage, disease, or dysfunction of the kidney, which can cause the kidneys to be less efficient or to fail.

##### Neuropathy

Damage, disease, or dysfunction of the peripheral nerves, which can cause numbness or weakness.

##### Pancreas

An organ located behind the stomach produces several important hormones, including insulin and glucagon.

##### Periodontitis

Also known as gum disease. An inflammatory disease affects the tissues that surround and support the teeth.

##### Retinopathy

A disease of the retina of the [eye](https://www.idf.org/our-activities/care-prevention/eye-health.html), which may cause visual impairment and blindness.

##### Secondary diabetes

A less common type of diabetes, which arises as a complication of other diseases (e.g. hormone disturbances or diseases of the pancreas).

##### Type 1 diabetes

People with type 1 diabetes cannot produce insulin. The disease can affect people of any age, but onset usually occurs in children or young adults.

##### Type 2 diabetes

People with type 2 diabetes cannot use insulin to turn glucose into energy. Type 2 diabetes is much more common than type 1 and occurs mainly in adults although it is now also increasingly diagnosed in children and young adults.

****

3.1 Representation of various implementation

**MODULES**

**System module:**

1. System will monitor the data.

2. System will update the data sets.

3. It will preprocess the data.

4. It will extract the data.

5. Segmentation as data.

6. Classification as SVM.

7. Identifying the news is true or fake.

**User module:**

1. User will enter the values from the data which are of diabetes data set.

2. User will see the result that the data is either a person having a diabetic or not.

**Numpy**

Numpy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed. This tutorial explains the basics of NumPy such as its architecture and environment. It also discusses the various array functions, types of indexing, etc. An introduction to Matplotlib is also provided. All this is explained with the help of examples for better understanding.

Numpy is a Python package. It stands for 'Numerical Python. It is a library consisting of multidimensional array objects and a collection of routines for processing of array.

**Numeric**, the ancestor of NumPy, was developed by Jim Hugunin. Another package Numarray was also developed, having some additional functionality. In 2005, Travis Oliphant created the NumPy package by incorporating the features of Numarray into the Numeric package. There are many contributors to this open-source project.

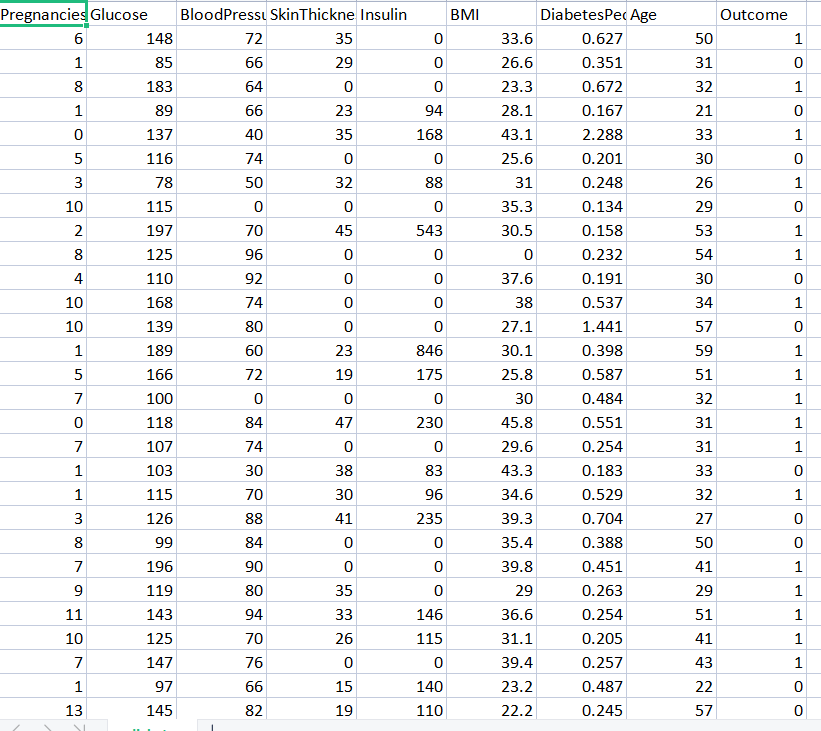
Operations using NumPy

Using NumPy, a developer can perform the following operations −  Mathematical and logical operations on arrays.

 Fourier transforms and routines for shape manipulation.

 Operations related to linear algebra. NumPy has in-built functions for linear algebra and random number generation.

**Different modules of the project**

****

**3.2. Data set**

**Sklearn :**

Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering, and dimensionality reduction.

**Components of sci-kit-learn:**

Scikit-learn comes loaded with a lot of features. Here are a few of them to help you

understand the spread:

• **Supervised learning algorithms:** Think of any supervised machine learning algorithm you might have heard about and there is a very high chance that it is part of sci-kit-learn. Starting from Generalized linear models (e.g Linear Regression), Support Vector Machines (SVM), Decision Trees to Bayesian methods – all of them are part of the sci-kit-learn toolbox. The spread of machine learning algorithms is one of the big reasons for the high usage of sci-kit-learn. I started using scikit to solve supervised learning problems and would recommend that to people new to scikit / machine

learning as well.

• **Cross-validation:** There are various methods to check the accuracy of supervised models on unseen data using sklearn.

• **Unsupervised learning algorithms:** Again there is a large spread of machine learning algorithms in the offering – starting from clustering, factor analysis, principal component analysis to unsupervised neural networks.

• **Various toy datasets:** This came in handy while learning scikit-learn. I had learned SAS using various academic datasets (e.g. IRIS dataset, Boston House prices dataset). Having them handy while learning a new library helped a lot.

• **Feature extraction:** Scikit-learn for extracting features from images and text (e.g. Bag of words)

**Pickle:**

The Python pickle module is another way to serialize and deserialize objects in Python. It differs from the JSON module in that it serializes objects in a binary format, which means the result is not human readable. However, it’s also faster and it works with many more Python types right out of the box, including your custom-defined objects.

**Pandas:**

Pandas is defined as an open-source library that provides high-performance data manipulation in Python. It is built on top of the NumPy package, which means **Numpy** is required for operating the Pandas. The name of Pandas is derived from the word **Panel Data**, which means **Econometrics from Multidimensional data**. It is used for data analysis in Python and was developed by **Wes McKinney in 2008**. Before Pandas, Python was capable of data preparation, but it only provided limited support for data analysis. So, Pandas came into the picture and enhanced the capabilities of data analysis. It can perform five significant steps required for processing and analysis of data irrespective of the origin of the data, i.e., **load, manipulate, prepare, model, and analyze**.

**Flask framework:**

Flask is a web application framework written in Python. Armin Ronacher, who leads an international group of Python enthusiasts named Pocco, develops it. Flask is based on Werkzeug WSGI toolkit and Jinja2 template engine. Both are Pocco projects.

4. **SCOPE** **OF PROJECT**

Diabetes prediction is an area that has received considerable attention. The objective of this study is to predict the risk of diabetes for everyone without the need for a blood test or going to a hospital. The study also aims to encourage and promote the good health of people. In addition, the diabetes prediction will be created as a simple diagnosis application and will be published by a website. However, this application is only an initial diagnosis. People who found that they are in the diabetes risk group should go to see a doctor for a formal diagnosis to prevent themselves from serious diabetes.

With the development of living standards, diabetes is increasingly common in people’s daily life. Therefore, how to quickly and accurately diagnose and analyze diabetes is a topic worthy of study. In medicine, the diagnosis of diabetes is according to fasting blood glucose, glucose tolerance, and random blood glucose levels. The earlier diagnosis is obtained, the much easier we can control it. Machine learning can help people make a preliminary judgment about diabetes according to their daily physical examination data, and it can serve as a reference for doctors.

Health condition diagnosis is an essential and critical aspect for healthcare professionals. Classification of a diabetes type is one of the most complex phenomena for healthcare professionals and comprises several tests. However, analyzing multiple factors at the time of diagnosis can sometimes lead to inaccurate results. Therefore, the interpretation and classification of diabetes are very challenging tasks. Recent technological advances, especially machine learning techniques, are incredibly beneficial for the healthcare industry. Numerous techniques have been presented in the literature for diabetes prediction.

Since diabetes is a long-lasting disease and import permanent damage to the limbs and vital organs in the body, using artificial intelligence tools can enhance the detection methods and disease control which will be of great help to the physicians. According to the Diabetes Research Centre, it has been shown that early diagnosis of patients at risk can prevent 80 percent of lasting complications of type II diabetes or defer them.

The early intervention of diabetes can reduce the prevalence of diabetes and hence the economic burden due to it. Machine Learning techniques play an important role in treatment plan workout, rehabilitation, chronic diseases management plan, etc. A long-term follow-up plan may be easily guided and keen supervision is possible. The systems may help in reducing of cost of patient management by avoiding unnecessary investigations and patients follow up. These prediction systems will add accuracy and time management. Computer-based patient support systems benefit patients by providing informational support that increases their participation in health care.

**5. SYSTEM DESIGN**

**5.1 INTRODUCTION**

This is a visual portrayal of an assortment of thoughts that are essential for the design, including its qualities, components, and parts. Here we present the proposed model diagram for the calculation of results.

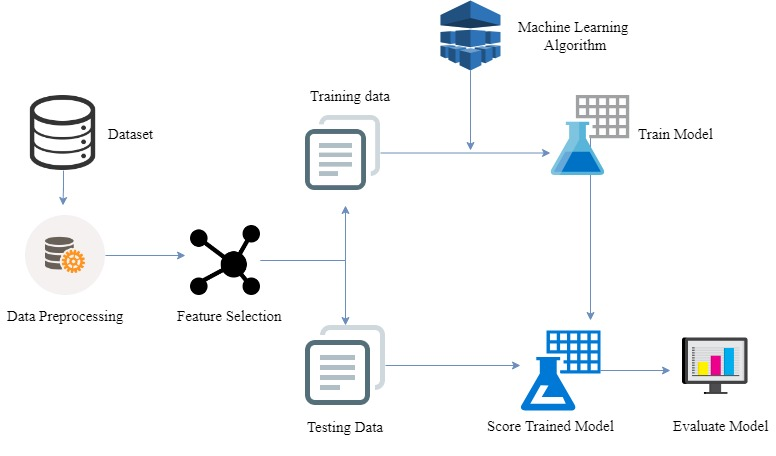


Fig: 5.1 Schematic Representation

In the design and development of the architecture for the diabetes prediction system, the requirements and design analysis of the system were based on discussions with collaborators from the diet type of patients was determined to be an essential approach suitable for the diabetes prediction system. The following functionalities were mentioned: Scheduling and reminding diabetic patients to take their medication and blood glucose readings, recommending healthy meals for diabetics to keep their blood glucose levels in check, encouraging and tracking the activity of diabetic patients, providing a visual interface to help them make meaning of their readings and establishing a sufficient connection between the doctor and the diabetic patient for making a person more healthy from predicting diabetes.

Providing the diabetic patient with a data visualization tool to display the data in tables, charts, and an educational program for newly diagnosed and ongoing diabetes treatment is valuable for the treatment and management of diabetes.

As we have mentioned earlier our goal is to develop A web-based application that can predict diabetes based on Patients’ health data. To find the most suitable ML algorithm that is capable of predicting diabetes more precisely, we have tested some sort of powerful machine learning models like SVM, KNN, Naive Bayes, and ANN. For the successful evaluation of these models, we have used some machine learning libraries such as Scikit-learn, NumPy,

matplotlib, pandas, and tensorflow.js. To get rid of the overfitting problem we split up the dataset into two subparts: one is for testing and the other is for training. Based on different training dataset sizes we have achieved the accuracy rate for every defined Model. However, preprocessing of the Indian Pima Dataset can produce higher prediction accuracy. Therefore, we also calculate the accuracy for our defined model to improve the prediction accuracy. Data normalization is an effective way to increase the accuracy of certain machine learning models and some machine learning model does not perform well without data Normalization.

**5.1.1ALGORITHM**

KNN ALGORITHM

Step1: Import required libraries, Import diabetes dataset.

Step2: Pre-process data to remove missing data.

Step3: Perform a percentage split of 80% to divide dataset as Training set and 20% to Test set.

Step4: Select the machine learning algorithm i.e. K- Nearest Neighbor, Support Vector Machine, Decision Tree, Logistic regression, Random Forest, and Gradient boosting algorithm.

Step5: Build the classifier model for the mentioned machine learning algorithm based on the training set.

Step6: Test the Classifier model for the mentioned machine learning algorithm based on the test set.

Step7: Perform a Comparison Evaluation of the experimental performance results obtained for each classifier.

Step8: After analyzing based on various measures conclude the best performing algorithm.

**KNN algorithm** is a supervised machine learning algorithm that deals with similarity. KNN stands for K-Nearest Neighbors. It’s a classification algorithm that will make a prediction of a class of a target variable based on a defined number of nearest neighbors. It will calculate distance from the instance you want to classify to every instance of the training dataset, and then classify your instance based on the majority classes of k nearest instances.

**KNN algorithm for classification**:

To classify a given new observation (new\_obs), the k-nearest neighbor’s method starts by identifying the k most similar training observations (i.e. neighbors) to our new\_obs and then assigns new\_obs to the class containing the majority of its neighbors.

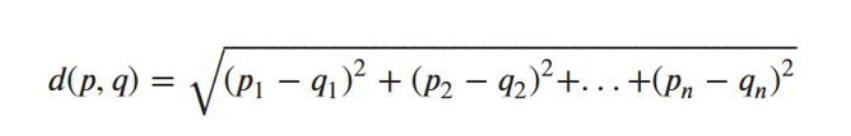
**KNN algorithm for regression**:

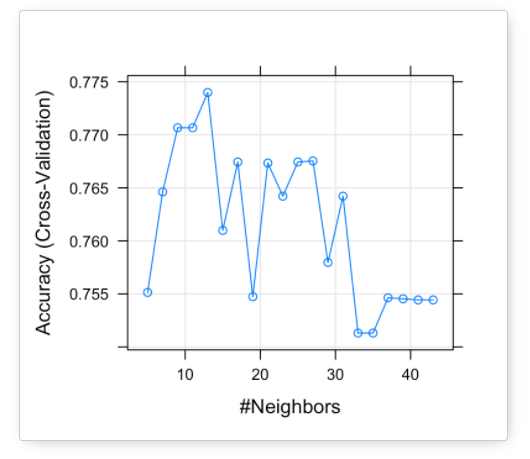
Similarly, to predict a continuous outcome value for a given new observation (new\_obs), the KNN algorithm computes the average outcome value of the k training observations that are the most similar to new\_obs and returns this value as new\_obs predicted outcome value.

**Similarity measures**:

Note that, the (dis)similarity between observations is generally determined using [Euclidean distance measure](http://www.sthda.com/english/articles/26-clustering-basics/86-clustering-distance-measures-essentials/), which is very sensitive to the scale on which predictor variable measurements are made. So, it’s generally recommended to standardize (i.e., normalize) the predictor variables for making their scales comparable.

The formula which is used for this algorithm is:

****

****

Graphical representation

**5.2 UML DIAGRAMS**

**5.2.1 USE CASE DIAGRAM**

Use case diagrams are a set of use cases, actors, and their relationships. They represent the use case view of a system. Use case diagrams are used to gather the requirements of a system including internal and external influences. These requirements are mostly design requirements. Hence, when a system is analyzed to gather its functionalities, use cases are prepared and actors are identified. When the initial task is complete, use case diagrams are modeled to present the outside view. In brief, the purposes of use case diagrams can be said to be as follows:

• Used to gather the requirements of a system.

• Used to get an outside view of a system.

• Identify the external and internal factors influencing the system.

• Shows the interaction among the requirements.

The Use Case diagram of the project diabetes prediction using machine learning

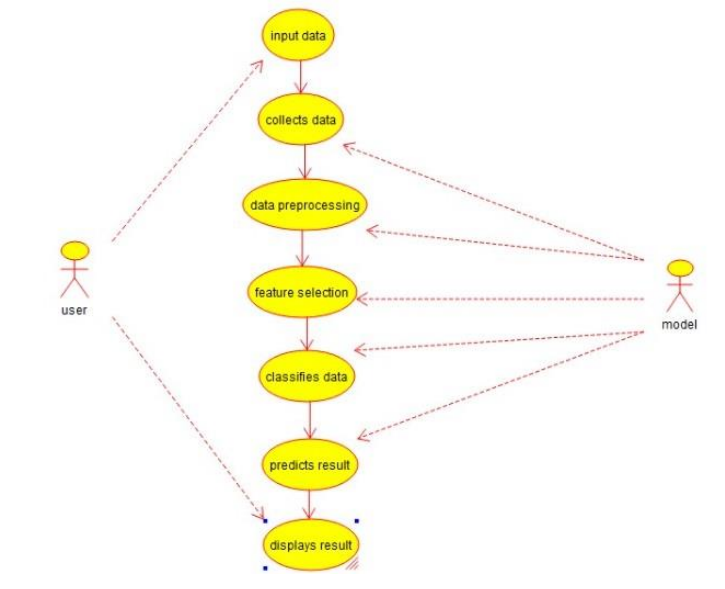
consists of all the various aspects a normal use case diagram requires. This use case

diagram shows how from starting the model flows from one to another step, like he/she

enters all the details that go into the system, compares with the prediction model, and

if true is predicted the appropriate results. Here the use case diagram of all the entities

is linked to each other where the user gets started with the system.

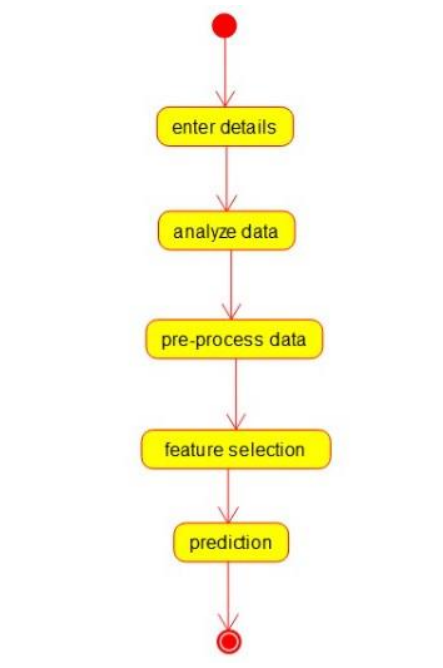
****

**5.2.1 Use case Diagram**

**5.2.2 ACTIVITY DIAGRAM**

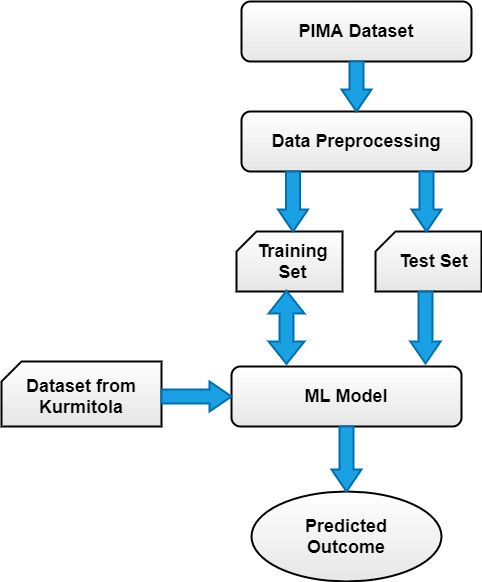
The activity diagram describes the flow of control in a system. It consists of activities and links. The flow can be sequential, concurrent, or branched. Activities are nothing but the functions of a system. Numbers of activity diagrams are prepared to capture the entire flow in a system.

Activity diagrams are used to visualize the flow of controls in a system. This is prepared to have an idea of how the system will work when executed. Activity is a particular operation of the system. Activity diagrams are not only used for visualizing the dynamic nature of a system but are also used to construct the executable system by using forward and reverse engineering techniques. The only missing thing in the activity diagram is the message part. First, we enter details into the system, then the data is analyzed. Next, the data is preprocessed and we select the feature that predicts the data.



5.2.2. Activity diagram

5.3 FLOWCHART



6**. SYSTEM IMPLEMENTATION**

**6.1 Selected Software**

This application is written in Sublime text editor or VScode editor and uses anaconda prompt for the execution of the program. It uses the sklearn module and flask framework.

Scikit-learn is probably the **most useful library for machine learning**in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering, and dimensionality reduction.

**Ide’s used:**

**Sublime text3:**

A sublime Text editor is a sophisticated text editor which is widely used among developers. It includes wide features such as Syntax Highlight, Auto Indentation, File Type Recognition, Sidebar, Macros, Plug-in, and Packages that make it easy for working with the codebase. This tutorial gives you comprehensive coverage of concepts of Sublime Text and makes you comfortable using it in your software development projects.

**VSCode:**

**Visual Studio Code** is a code **editor** redefined and optimized for building and debugging modern web and cloud applications.

**Anaconda ide:**

Anaconda is **a distribution of Python for scientific** computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment.

**6.2 SAMPLE CODE**

1. Make necessary imports

from flask import Flask, render\_template, request

import pickle

import NumPy as np

import pandas as pd

import pickle

data=pd.read\_csv(r"diabetes.csv")

x=data.iloc[:,:-1].values

y=data.iloc[:,-1].values

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

xtrain,xtest,ytrain,ytest=train\_test\_split(x,y,test\_size=0.21,random\_state=6)

from sklearn.neighbors import KNeighborsClassifier

modelknn=KNeighborsClassifier(n\_neighbors=3)

modelknn.fit(xtrain,ytrain)

ypredknn=modelknn.predict(xtest)

filename = 'diabetes-prediction-rfc-model.pkl'

pickle.dump(modelknn, open(filename, 'wb'))

# Importing essential libraries

from flask import Flask, render\_template, request

import pickle

import NumPy as np

# Load the KNeighbour CLassifier model

filename = 'diabetes-prediction-rfc-model.pkl'

modelknn = pickle.load(open(filename, 'rb'))

app = Flask(\_\_name\_\_)

@app.route('/')

def home():

return render\_template('index.html')

@app.route('/predict', methods=['POST'])

def predict():

if request.method == 'POST':

preg = int(request.form['pregnancies'])

glucose = int(request.form['glucose'])

bp = int(request.form['bloodpressure'])

st = int(request.form['skinthickness'])

insulin = int(request.form['insulin'])

bmi = float(request.form['bmi'])

dpf = float(request.form['dpf'])

age = int(request.form['age'])

data = np.array([[preg, glucose, bp, st, insulin, bmi, dpf, age]])

my\_prediction = modelknn.predict(data)

return render\_template('result.html', prediction=my\_prediction)

if \_\_name\_\_ == '\_\_main\_\_':

app.run(debug=True)

The data collection is a framework collection which is already a predefined data set that is taken the PIMA diabetes data set which are the collections of several forms of the parameters that are in the considerations. In the data set formation, there are dependent variables as X and independent variables as Y.

When the dependent variables depend on the data includes in the various implementations of the considerations of several pregnancies, glucose level, skin thickness, insulin, body mass index, diabetic pedigree function, blood pressure these are all dependent variables that indicate the symbol of X. Independent variables is nothing but the result obtained by the dependent variables. The methods which are involved in this prediction are using the implementation of the KNN algorithm.

**7. TESTING**

**7.1 INTRODUCTION**

Software testing is defined as an activity to check whether the actual results match the expected results and to ensure that the software system is Defect free. It involves the execution of a software component or system component to evaluate one or more properties of interest.

Validation is a complex process with many possible variations and options, so specifics vary from database to database, but the general outline is:

**** Requirement Gathering

* + - The Sponsor decides what the database is required to do based on regulations, company needs, and any other important factors.
    - The requirements are documented and approved.

**** System Testing

* + - Procedures to test the requirements are created and documented.
    - The version of the database that will be used for validation is set up.
    - The Sponsor approves the test procedures.
    - The tests are performed and documented.
    - Any needed changes are made. This may require another, shorter round of testing and documentation.

**** System Release

* + - The validation documentation is finalized.
    - The database is put into production.

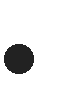
**7.2 TESTING LEVELS**

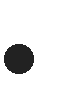
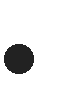
**White Box Testing**

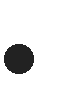
White Box Testing is defined as the testing of a software solution's internal structure, design, and coding. In this type of testing, the code is visible to the tester. It focuses primarily on verifying the flow of inputs and outputs through the application, improving design and usability, strengthening security. White box testing is also known as Clear Box testing, Open Box testing, Structural testing, Transparent Box testing, Code-Based testing, and Glass Box testing. It is usually performed by developers.

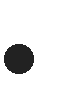
It is one of two parts of the **"Box Testing" approach** to software testing. Its counterpart, **Blackbox testing**, involves testing from an external or end-user type perspective. On the other hand, Whitebox testing is based on the inner workings of an application and revolves around internal testing.

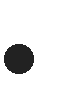
The term "WhiteBox" was used because of the see-through box concept. The clear box or WhiteBox name symbolizes the ability to see through the software's outer shell (or "box") into its inner workings. Likewise, the "black box" in "[Black Box](https://www.guru99.com/black-box-testing.html) [Testing](https://www.guru99.com/black-box-testing.html)" symbolizes not being able to see the inner workings of the software so that only the end-user experience can be tested.

 Internal security holes

 Broken or poorly structured paths in the coding processes  The flow of specific inputs through the code

 Expected output

 The functionality of conditional loops

 Testing of each statement, object, and function on an individual basis

The testing can be done at system, integration, and unit levels of software development. One of the basic goals of white-box testing is to verify a working flow for an application. It involves testing a series of predefined inputs against expected or desired outputs so that when a specific input does not result in the expected output, you have encountered a bug.

**SYSTEM TESTING**

**System Testing** is a level of software testing where complete and integrated software is tested. The purpose of this test is to evaluate the system’s compliance with the specified requirements. The process of testing an integrated system to verify that it meets specified requirements.

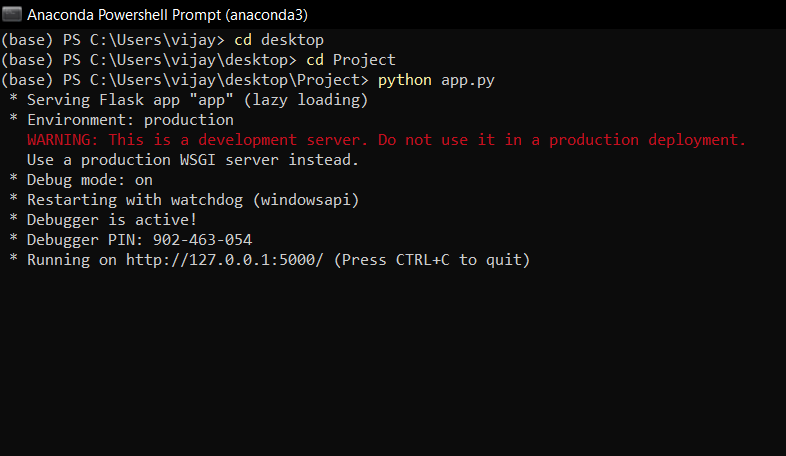
**ACCEPTANCE TESTING**

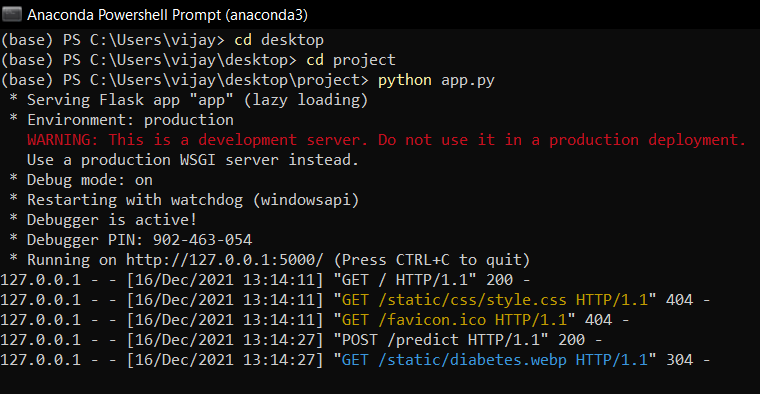
**Acceptance Testing** is a level of software testing where a system is tested for acceptability. The purpose of this test is to evaluate the system’s compliance with the business requirements and assess whether it is acceptable for delivery. Formal testing concerning user needs, requirements, and business processes is conducted to determine whether or not a system satisfies the acceptance criteria and to enable the user, customers, or other authorized entity to determine whether or not to accept the system.

**8. SCREENS AND REPORTS**

**8.1 SCREENSHOTS**

Run the prompt with command as python app.py

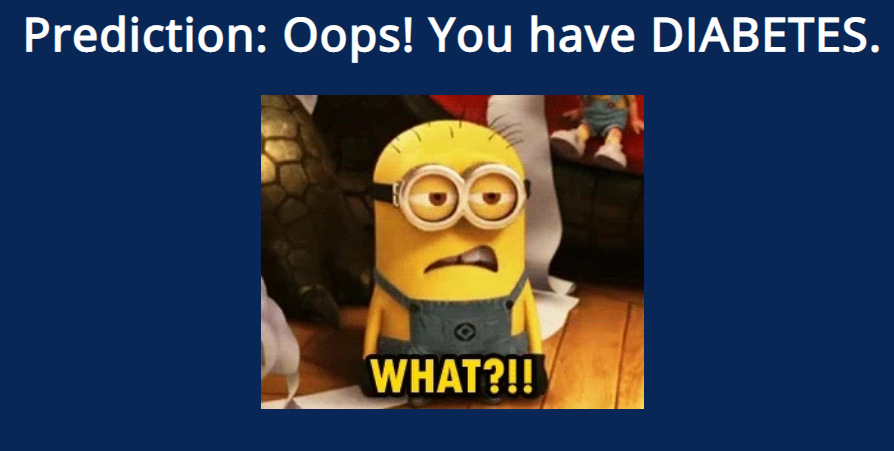
****



1)

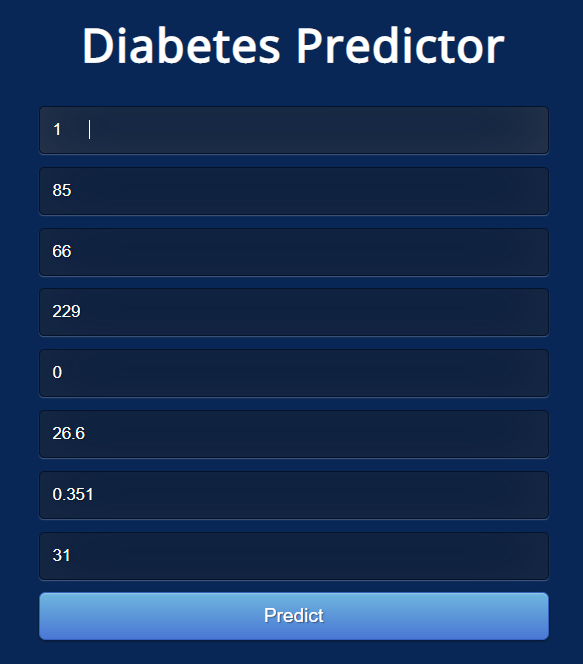
****

****

****

**2)**

****

****

****

**8.2 REPORT**

* The proposed approach uses different classification and ensemble methods and is implemented using python. These methods are standard Machine Learning methods used to obtain the best ac- curacy from data. In this work, we see that the random forest classifier achieves better compared to others.
* Overall we have used the best Machine Learning techniques for prediction and to achieve high-performance accuracy.

**9. CONCLUSION AND FUTURE SCOPE**

**9.1 CONCLUSION**

* KNN is even above the level of SVMs. The proposed algorithms share the advantages of KNN (no training required, ideal for fast adaptation, natural handling of the multi-class case) and its drawbacks (requires large memory, slow testing). However, our latest result also indicates the possibility of substantially reducing the reference set in memory without losing accuracy. This suggests that the algorithm indeed captures essential information in the data and that our initial intuition on the nature of the ﬂaw of KNN may well be at least partially correct.
* In this research study, we have used different types of machine learning algorithms for the Detection of diabetes. We implemented machine learning algorithms on the dataset and performed classification to signify the best machine learning algorithm for diabetes prediction based on old data available. The higher accuracy the better the prediction rate we will achieve. The random forest algorithm obtained the best accuracy and Roc. While on the other hand logistic regression had the lowest score. ROC of 76% and accuracy of 84%. The overall experimentation displayed that random forest is better than other algorithms in diabetes prediction. In our future work, we will work on the diagnosis of diabetes in young adults.
* The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning methods in which SVM, KNN, Random Forest, Decision Tree, Logistic Regression, and Gradient Boosting classifiers are used. And 77% classification accuracy has been achieved.
* The Experimental results can be asst health care to take early prediction and make early decision to cure diabetes and save humans life.

**9.2 FUTURE SCOPE**

Healthcare professions found it hard to find healthcare data and perform analysis on them due to a lack of tools, resources. But using ML, we can overcome this and can perform analysis on real-time data leading to better modeling, predictions. This enhances and improves overall healthcare services. Now, it's being integrated with ML to make smart healthcare devices that sense if there is any change in the person’s body, health data when he uses the device and this will notify the person regarding this through an app. This helps in easy monitoring, advanced prediction, and analysis thereby reducing errors, saving time and life of people.

10. **BIBLIOGRAPHY**

* Mahgoub, A. A. (2021, March 30). Diabetes prediction system with KNN algorithm. Medium. <https://medium.com/codex/diabetes-predication-system-with-knn-algorithm-e040999229f7>
* Diabetes prediction using machine learning techniques*. (2020, April 10). IJERT –* International Journal of Engineering Research & Technology. <https://www.ijert.org/diabetes-prediction-using-machine-learning-techniques>
* Predicting diabetes mellitus with machine learning techniques*. (n.d.).* Frontiers*.* <https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full>
* Machine learning-based diabetes classification and prediction for healthcare applications*. (*2021, October )*.* Publishing Open Access research journals & papers | Hindawi. <https://www.hindawi.com/journals/jhe/2021/9930985/>
* Analysis of diabetes mellitus for early prediction using optimal features selection*.* (2019, February 6). SpringerOpen. <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0175-6>