Diabetes Prediction using Machine Learning

V Siva Supradeep[1], P. Sai Phanidhar[2] , G. Parimala[3]

[1,2] Dept. of CSE, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamilnadu, India

vs8724@srmist.edu.in, pe8135@srmist.edu.in, parimalg@srmist.edu.in

Abstract: Diabetes is considered to be one of the worst illnesses in the world. Diabetes is caused by a combination of variables, including obesity, excessive blood glucose levels, and other causes. It does this by altering the insulin hormone, which in turn causes an irregular metabolism in the crab and raises its blood sugar levels. This program's primary objective is to lessen the risk that people may acquire diabetes by making forecasts for them and urging them to take more care of their diet and lifestyle in the years to come. The key goals of this research were to develop and execute a method for predicting diabetes using machine learning techniques, as well as investigate the strategies that would be used to achieve success in this Endeavour. The suggested technique makes use of a wide variety of classification and ensemble learning algorithms, some examples of which include Knn, Label Encoder, and train test split. The results of the research may provide information that will help medical professionals make more accurate early predictions and judgments in order to better manage diabetes and save lives. The method first extracts information from a dataset, such as certain symptoms that may be utilized to gain further knowledge about diabetes, and then validates that information using other data. This project's objective was to build classification models for the diabetes data set, develop models that can determine whether or not a person is sick, and get the greatest possible validation scores in the models that were developed. Massive datasets may be found in the healthcare business. By investigating enormous datasets in this manner, we may uncover previously unknown information and trends, which will enable us to draw conclusions based on the data and make accurate forecasts. We categorize the dataset using random techniques since our major goal in doing this research is to determine the method that is the most accurate for predicting diabetes. This will be accomplished by integrating machine learning, data visualization, and data interpretation. The use of machine learning, which is becoming more important in the modern healthcare sector, will be the focus of this research. Massive datasets may be found in the healthcare business.

I. INTRODUCTION:

The field of medicine has been quick to embrace the idea of "machine learning," which stands for "machine learning." The predictions made by the research community and the analysis of medical data sets contribute to the prevention of disease by indicating the most effective treatments and preventative measures. In the field of machine learning, algorithms that are able to assist in decision making and prediction. In addition, we investigate a wide variety of applications of machine learning in the area of medicine, with a special emphasis on the use of machine learning for diabetes prediction. Diabetes is one of the illnesses that are spreading at the quickest rate throughout the globe, and it is imperative that it be constantly managed at all times. In order to investigate this, we investigate a wide variety of machine learning algorithms that have the potential to assist in early sickness prediction. This article delves into a wide range of subjects pertaining to machine learning, including the several classifications of algorithms that provide assistance with decision making and prediction. People are able to prevent becoming ill by taking the appropriate care and precautions thanks to the forecasts and analyses provided by the scientific community for medical datasets. Explore the many different uses of machine learning in the medical field, with a particular emphasis on utilizing machine learning to predict diabetes. This article delves into a wide range of subjects pertaining to machine learning, including the several classifications of algorithms that provide assistance with decision making and prediction. People are able to prevent becoming ill by taking the appropriate care and precautions thanks to the forecasts and analyses provided by the scientific community for medical datasets. Explore the many uses of machine learning in medicine, with a particular focus on making predictions about diabetes using machine learning. This article examines a variety of concerns pertaining to machine learning, such as the several classifications of algorithms that might be of assistance in decision making and prediction. Patients are given assistance in adopting the appropriate preventative measures and treatment thanks to the scientific community's forecasts and analyses of medical datasets, which help patients, avoid becoming unwell. Explore the many uses of machine learning in medicine, with a particular focus on making predictions about diabetes using machine learning. Because of this, there is a higher concentration of glucose in the urine. Diabetes may lead to organ failure, cardiovascular disease, and disturbances in a variety of physiological systems if it is not well managed. The World Health Organization (WHO) identifies diabetes as one of the four most significant non-communicable diseases (NCDs) that are now having an effect on the global population (World Health Day, 2016). The most recent results from the WHO are quite concerning. As was mentioned before, diabetes may result in a wide range of major problems with the cardiovascular system. According to the World Health Organization (WHO), diabetes and cardiovascular illnesses would be responsible for the deaths of 3.7 million people before they reach the age of 70. A blood glucose level that is not under control is the primary contributor to diabetes. Diabetes is a serious illness that affects a significant portion of the world's population. Diabetes may be brought on by a variety of factors, including advancing age, being overweight, experiencing a rapid loss of weight, overeating (polyphagia), irritability, and muscular stiffness, amongst others. We proposed a model for the prediction of diabetes as a means of improving the classification of diabetes. This model combines a number of diabetes-specific exterior characteristics with regular aspects such as delayed healing, partial paresis, irritation, itching, and visual blurring, among other things. People would be able to take better care of themselves if diabetes could be foreseen. Diabetes develops when the body is unable to produce an adequate amount of insulin. Diabetes affects around 422 million people all over the globe, most significantly in low- and middle-income nations, and it is anticipated that this number will rise to 490 billion by the year 2030. The completion of this task has the potential to save a significant number of lives.

II. LITERATURE SURVEY

1. An Innovative Approach to Diabetes Prediction

Algorithm Used: Naive Bayes Classificationn

Drawbacks: Its estimates can be off in some circumstances, so don't take its probability values too seriously.

It is assumed that each individual feature can stand on its own. If we come across terms in the test data for a particular class that aren't in the training data, we run the risk of arriving at the conclusion that there is no probability associated with that class. The fundamental objective of this research is to examine a database of diabetic patients in order to develop a method that accurately forecasts the onset of diabetes at an earlier stage. Within the framework of this proposed experiment system, the Naive Bayes Classification points to diabetes. The process of collecting data from a dataset and changing it into a usable structure so that it may be used in the future is known as "information mining." According to the findings, the one-of-a-kind method that was developed has a higher accuracy rate (0.96), when compared to conventional methods or methods that already exist. The suggested system would include a web interface that would show the repercussions of having diabetes as well as the repercussions of not having diabetes depending on a variety of input criteria such as insulin level, age, and so on. This contributes to an increase in the system's precision.

The Naive Bayes approach is a supervised learning strategy that utilizes the Bayes theorem to provide solutions to classification problems that may arise. Text classification is one of the most common applications for this technique since it calls for a significant amount of data to be used during the training phase.

2. Diabetes Prediction Using Machine Learning Algorithms:

Diabetes raises the risk of developing a number of different ailments, including cardiovascular disease, renal disease, stroke, vision problems, and nerve damage, amongst others. The standard procedure followed in hospitals these days is to first conduct a battery of tests in order to gather the information necessary for a diabetic diagnosis, and only after that would they prescribe the right treatment. It is impossible to place an adequate amount of emphasis on the significance of big data analytics in the field of healthcare. Massive datasets may be found in the healthcare business. With the use of big data analytics, it is possible to scour through enormous datasets in search of concealed information and patterns in order to gain insight and anticipate results.

The accuracy of categorization and prediction provided by the existing technique is subpar. Standard components such as glucose, body mass index (BMI), age, insulin, and so on are included in the diabetes prediction model that we present in this research for the purpose of improving diabetes classification. In addition to these components, we add a few external characteristics associated with diabetes. In addition, a pipeline model for diabetes prediction was implemented with the intention of achieving higher levels of accuracy in categorization.

3. Diabetes Prediction Using Machine Learning Techniques:

4. The purpose of this research is to develop a model that has a greater capacity for accurately predicting diabetes. In order to forecast diabetes, we tested a variety of classification and ensemble methods. The information that follows provides an overview of the time period in question. Diabetes prognosis via the random Forest method. using the Random Forest algorithm as part of a machine learning approach, your goal should be to design a system that is capable of earlier and more accurate detection of diabetes in a patient. The findings that the suggested model produces are the most accurate diabetic prediction results, and the data show that the prediction system is able to accurately, effectively, and—most importantly—immediately anticipate the diabetes illness.

5. Academics are becoming more interested in the topic of diabetes prediction because they want to train a computer to determine whether or not a patient has diabetes by applying a good classifier to a dataset and teaching it to do so. Proved. The prediction of diabetes is a significant challenge in the field of computing. hence, a system is necessary to solve the problems that were presented by prior research.

6. Diabetes Prediction using Learning Algorithm Techniques:

The primary goals of this effort are to increase the accuracy of diabetic and non-diabetic data categorization and to categorize data as diabetic or non-diabetic. When doing a variety of classification tasks, the accuracy of the classification will suffer according to the number of samples that are picked. In many situations, the performance of the algorithm in terms of speed is outstanding, but the accuracy of the data categorization is rather bad. Acquiring a high level of accuracy should be considered our primary goal in developing this model. The accuracy of classification may be improved by first training on a large portion of the data set and then validating the model on a more limited subset of the data. For the purpose of this research, both diabetic and non-diabetic data was analyzed using a variety of categorization schemes. As a consequence of this, several methods, such as Logistic Regression, Support Vector Machine, and Artificial Neural Network, have been shown to be the most suitable for the development of a Diabetes prediction system.

III.METHODOLOGY

A. System model

1) Data Acquisition:
"The process of sampling signals that measure physical situations in the actual world and changing the samples that are obtained as a consequence into digital numeric values that can be controlled by a computer is referred to as data acquisition," and "sampling" refers to the act of measuring. Data acquisition systems, also known as DAS or DAQ, are responsible for converting the physical circumstances that are represented by analogue waveform into digital values that may then be stored, analysed, and processed. The phrase "data acquisition" consists of two separate words: "data" and "acquisition." While "acquisition" refers to the process of collecting data for a specific goal, "data" refers to the unstructured or organised facts and statistics that are the subject of study. Data acquisition refers to the act of gathering information from pertinent sources for the purposes of storing, cleansing, and preparing it before it is used in subsequent procedures. It is the process of collecting critical business information, translating it into the proper business format, and then

putting it into the appropriate system. The majority of a data scientist's workday is spent searching for, cleaning, and analysing data sets. As the use of machine learning becomes more widespread, a variety of applications are struggling due to an inadequate supply of tagged data. Even the most sophisticated machine learning algorithms are doomed to failure in the absence of sufficient data and an adequate data cleansing process. In addition, the approaches of deep learning need enormous amounts of data since, in contrast to machine learning, the algorithms involved in deep learning produce features on their own. If it weren't the case, we'd have trash coming in and garbage leaving. As a consequence, data acquisition or collecting is a vital component.

2) Pre-Processing:

This section provides an explanation of how we use recordings in our artwork. In order to proceed with the classification process, it is necessary for us to now extract pertinent characteristics from our training set. The data is made consistent and correct before being uploaded into the network. Following normalisation, the price falls somewhere in the range of 0 to 1. The act of transforming raw data into a form that can be used by a machine learning model is referred to as data preparation. In the process of constructing a model for machine learning, this step is the first and most crucial one.

When we are working on a project that involves machine learning, we do not always come across data that is clean and well-prepared. In addition, the data must be thoroughly cleansed and organised before any action can be taken on it. As a consequence of this, we turn to the work of data preparation to do this out. Collecting the dataset, importing libraries and datasets, determining whether data are missing, encoding categorical data, dividing the dataset into training and test sets, and scaling features are the phases that make up this process.

3) Classification:

The train-test split is a method for gauging the efficacy of a machine learning system. It can be used with any supervised learning technique, from regression to classification. One component of the method involves splitting a dataset into two subsets. The training dataset was the primary data used during the model fitting process. The second subset is not used to train the model, but instead serves as the input for the model, which then makes predictions and checks those predictions against the original dataset. The second group of numbers is the "test dataset." The machine learning model is being trained using this dataset as input. A test dataset for future use. Method for judging the machine learning model's accuracy in representing the data. In order to gauge how well the machine learning model performs, this project will use data that was not included in the model's training process. This is how we intend to put the plan into action. Another way of putting it is that we want to generate predictions based on future conditions that have inputs and outputs that are not yet known but do have target values, and then we want to fit them to previously collected data with known inputs and outputs. When working with a large enough dataset, the train-test methodology becomes an option.

B.HARDWARE AND SOFTWARE USED:

Hardware:

1. Processor: Intel i3 or above.
2. RAM: 6GB or more

Software:

1. Operating System : Windows 7/8/10
2. python
3. Anaconda
4. Jupyter notebook

IV. FLOW OF ACTIVITY

A. IMPLEMENATATION

The term "implementation" refers to all of the work that is done to transition from the old system to the new one. The current system is based on manual processes and operates in a manner that is substantially unlike to the system that is being considered as a successor. The establishment of a dependable system that caters to the requirements of organizations calls for the use of implementation expertise. The success of the computerized system may be jeopardized by an improper installation.

When it comes to dealing with the installation and subsequent conversion from the previous electronic system to the new one, there are various different approaches that may be used.

Running both the old and new systems at the same time is the method that ensures the smoothly transition from the old to the new system. A worker may do their duties using either the traditional manual processing method or the modern computerized approach. This method offers an exceptionally high level of protection as a result of the fact that we may fall back on the manual system in the event that the automated system malfunctions. On the other hand, the cost of operating two systems in parallel is too high. This has a greater impact than the benefits. Going from the manual to the computerized system in one seamless transition is yet another popular choice. The transition might take as little as one day or as long as a week. There aren't any other events happening at the same time as this one. However, there is no solution available in the event that there is a problem. Careful and thorough preparation is required for this procedure. It is also possible to construct a functional version of the system in one sector of the organization, with workers acting as test subjects and contributing any necessary enhancements. However, considering that the whole system has been destroyed, this method is one that should be avoided whenever possible.

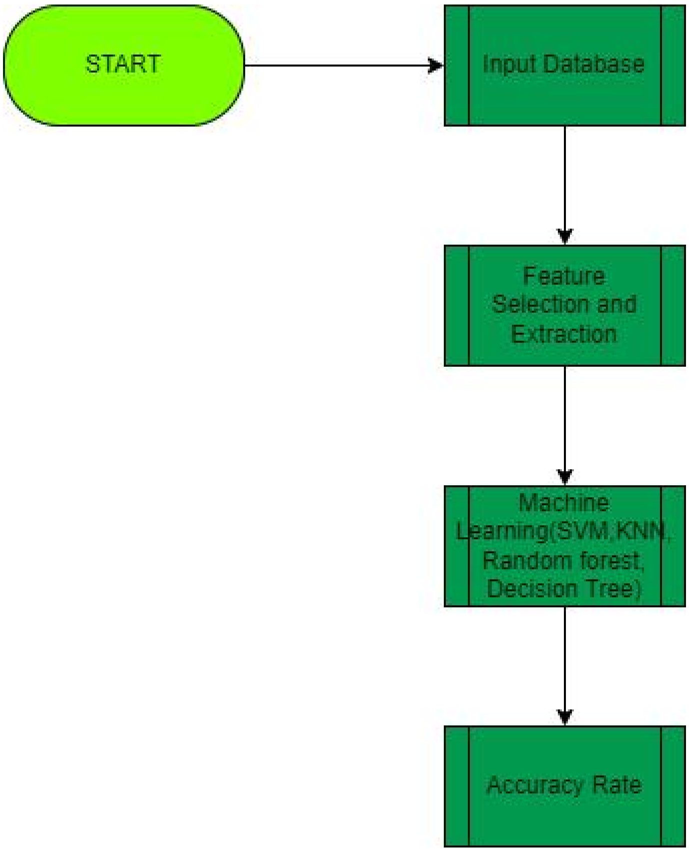

START
Input Database
Feature Selection and Extraction
Machine Learning(SVM,KNN, Random forest, Decision Tree)
Accuracy Rate

Fig 1: Data flow Diagram (level 1)

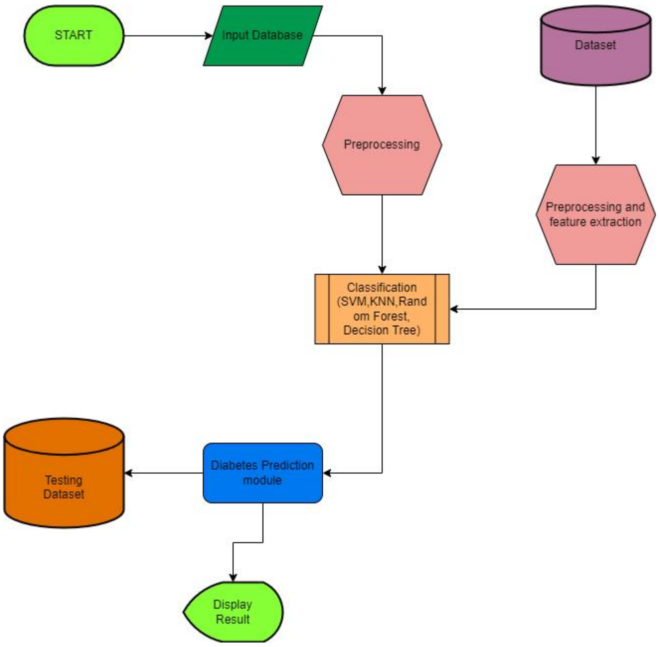

START
Input Database
Dataset
Preprocessing
Preprocessing and feature extraction
Classification (SVM,KNN,Rand om Forest, Decision Tree)
Testing Dataset
Diabetes Prediction module
Display Result

Fig 2: Data flow Diagram (level 2)

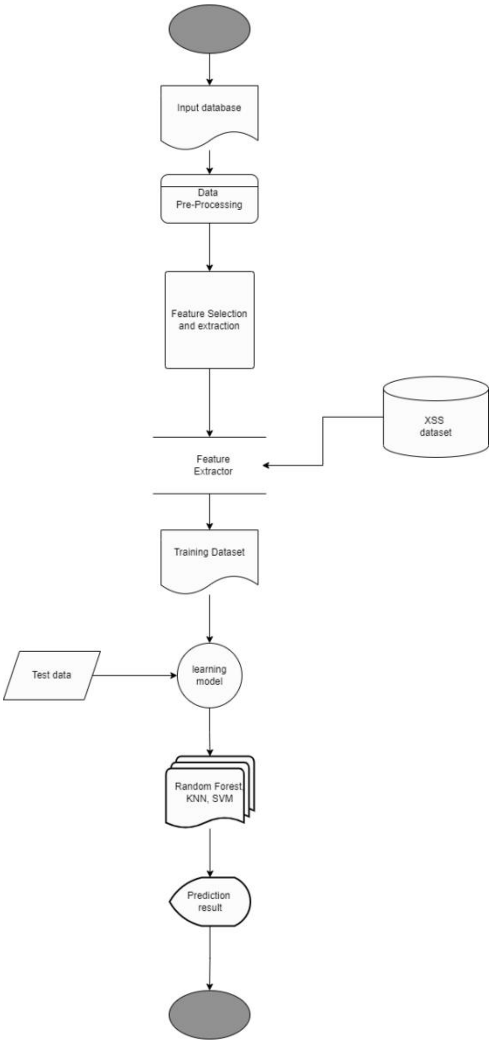


Fig 3: Activity Diagram

V. SOFTWARE MODULES

1) Python:

Python is a high-level programming language that can be interpreted and is object-oriented. Python also has dynamic semantics. Because of its high-level built-in data structures, dynamic type, and dynamic binding, this language is particularly well suited for usage as a scripting or glue language to connect together already existing components. Readability is put first by Python's condensed and straightforward syntax, which results in decreased expenses associated with programmed maintenance. Support for modules and packages are built into Python, which promotes the modularity of programs and the reusability of code. The interpreter of Python as well as its huge standard library is both downloadable in source and binary form for all of the main operating systems.

Because of the significant productivity boost it offers, programmers often choose to work in Python. The edit-test-debug cycle moves at a breakneck speed since there is no compilation step. The process of debugging Python programs is quite basic since a segmentation fault will never be produced by a bug or incorrect input. An exception will be thrown if the interpreter finds a mistake in the code. The interpreter will provide a stack trace in the event that the program does not successfully catch the exception. A source level debugger provides a number of different capabilities, including the ability to examine both local and global variables, evaluate arbitrary expressions, create breakpoints, and go through the code one line at a time. The fact that the debugger is developed in Python demonstrates the capability of the language to do introspection. On the other hand, adding a few print statements to the source code is often the easiest approach to debug a programming statement. As a result of the quick edit-test-debug cycle, this basic technique is quite effective.

2) Anaconda:

It is a free and open source distribution of the computer languages Python and R, with the primary goal of simplifying package management and deployment in applications related to data science and machine learning (large-scale data processing, predictive analytics, scientific computing). The package management system known as Conda is in charge of handling the different versions of each package. Over 6 million people use the Anaconda package, which comprises over 250 of the most popular data science programs for Windows, Linux, and MacOS. These programs may be used on any of these operating systems.

3) KNN Classifier:

The K-NN method begins with the presumption that the new case or data is comparable to the cases that have come before it and assigns the new case to the category that is the most comparable to the categories that are already in place. This indicates that the K-NN approach may be used to efficiently categorize new data into the right category in a short amount of time. The K-NN method may be used for classification as well as regression; however, the application of this method for classification is much more popular.

4) Random Forest:

The approach known as ensemble learning may be used for both classification and regression. It achieves its results by first training a large number of decision trees, and then outputting either the class, also known as the mode of the classes, or the regression of the individual trees. According to what the name suggests, "Random Forest" is a classification method that "consists of a number of decision trees on different subsets of the supplied dataset and takes the average to raise the projected accuracy of that dataset."

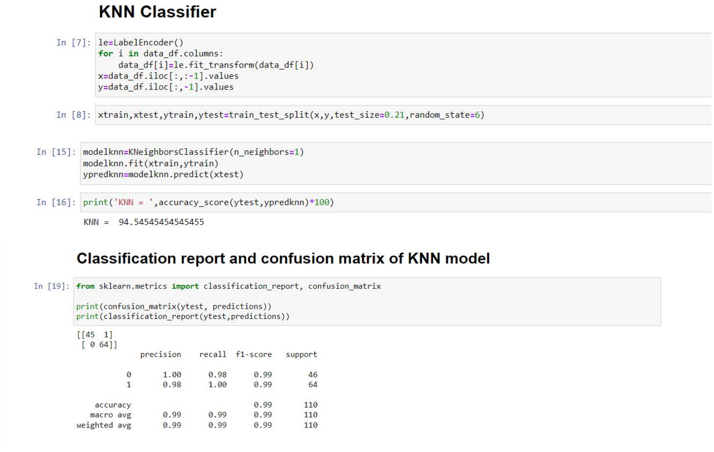
5) SVM:

The concept of supervised learning serves as the foundation for the SVM (Support Vector Machine). A training set and labels are necessary components of an SVM. If test data is provided to the model after it has been trained, the model will classify it into one of two categories. It is effective when used to linear classification. It is also capable of doing well in nonlinear classification by applying a kernel technique to transform the inputs into a high-dimensional feature space. This allows it to execute nonlinear classification. A categorization hyper plane is produced as a result. The hyper plane is chosen such that the distance between the data points that are next to it on each side is as little as possible.

6) Decision Tree:

The Decision Tree is a technique to supervised learning that may be used for both classification and regression problems; however, the classification problem is the one that is utilized with it the most often. It is a tree-structured classifier, with leaf nodes that reflect the result, core nodes that store dataset properties, and branches that represent decision rules. It is a graphical representation of all the potential answers to a question or choices that may be made depending on a set of criteria. Because it, like a tree, starts with the root node and then branches out to create a tree-like structure, a decision tree derives its name from its similarity to the structure of a tree.

VI. RESULTS



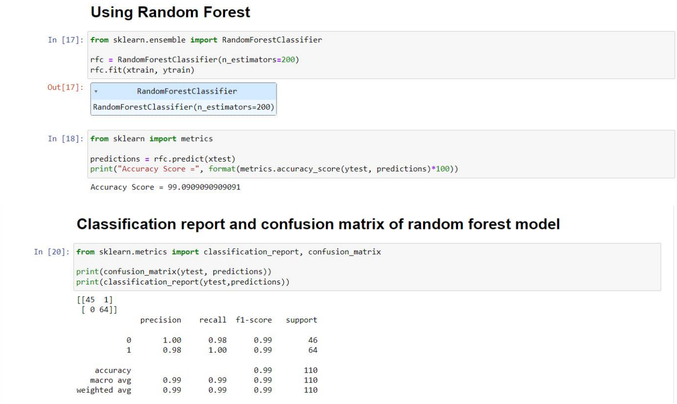
KNN Classifier

```
In [7]: le=LabelEncoder()
        for i in data_df.columns:
            data_df[i]=le.fit_transform(data_df[i])
        x=data_df.iloc[:,:-1].values
        y=data_df.iloc[:,-1].values

In [8]: xtrain,xtest,ytrain,ytest=train_test_split(x,y,test_size=0.21,random_state=6)

In [15]: modelknn=KNeighborsClassifier(n_neighbors=1)
         modelknn.fit(xtrain,ytrain)
         ypredknn=modelknn.predict(xtest)

In [16]: print('KNN = ',accuracy_score(ytest,ypredknn)*100)

         KNN =  94.54545454545455
```

Classification report and confusion matrix of KNN model

```
In [19]: from sklearn.metrics import classification_report, confusion_matrix

         print(confusion_matrix(ytest, predictions))
         print(classification_report(ytest,predictions))

[[45  1]
 [ 0 64]]
              precision    recall  f1-score   support

           0       1.00      0.98      0.99        46
           1       0.98      1.00      0.99        64

    accuracy                           0.99       110
   macro avg       0.99      0.99      0.99       110
weighted avg       0.99      0.99      0.99       110
```

Fig 1: KNN Classifier Accuracy



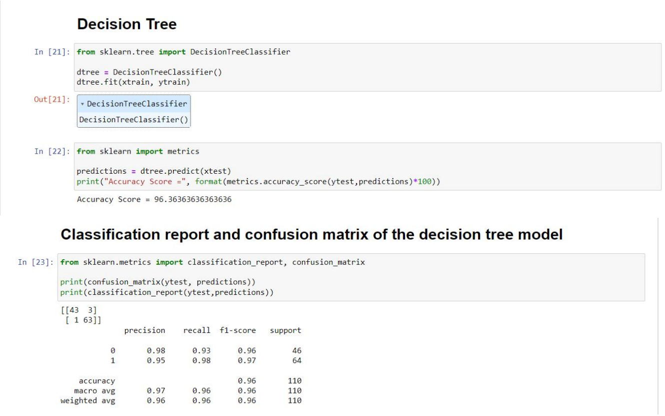
Using Random Forest

```
In [17]: from sklearn.ensemble import RandomForestClassifier

         rfc = RandomForestClassifier(n_estimators=200)
         rfc.fit(xtrain, ytrain)

Out[17]: ▾        RandomForestClassifier
         RandomForestClassifier(n_estimators=200)

In [18]: from sklearn import metrics

         predictions = rfc.predict(xtest)
         print("Accuracy Score =", format(metrics.accuracy_score(ytest, predictions)*100))

         Accuracy Score = 99.0909090909091
```

Classification report and confusion matrix of random forest model

```
In [20]: from sklearn.metrics import classification_report, confusion_matrix

         print(confusion_matrix(ytest, predictions))
         print(classification_report(ytest,predictions))

[[45  1]
 [ 0 64]]
              precision    recall  f1-score   support

           0       1.00      0.98      0.99        46
           1       0.98      1.00      0.99        64

    accuracy                           0.99       110
   macro avg       0.99      0.99      0.99       110
weighted avg       0.99      0.99      0.99       110
```

Fig 2: Random Forest Accuracy



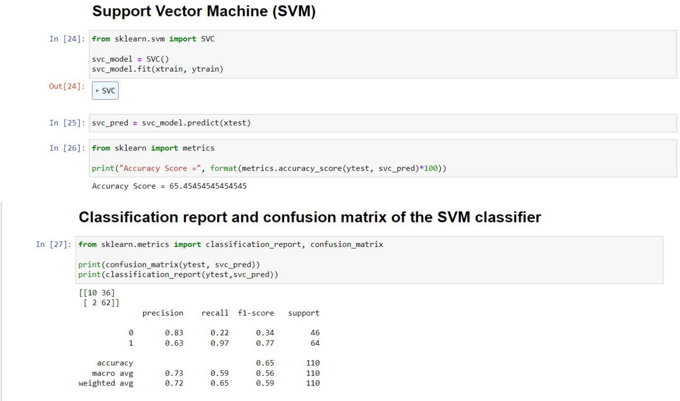
Decision Tree

```
In [21]: from sklearn.tree import DecisionTreeClassifier

         dtree = DecisionTreeClassifier()
         dtree.fit(xtrain, ytrain)

Out[21]: ▾ DecisionTreeClassifier
         DecisionTreeClassifier()

In [22]: from sklearn import metrics

         predictions = dtree.predict(xtest)
         print("Accuracy Score =", format(metrics.accuracy_score(ytest,predictions)*100))

         Accuracy Score = 96.36363636363636
```

Classification report and confusion matrix of the decision tree model

```
In [23]: from sklearn.metrics import classification_report, confusion_matrix

         print(confusion_matrix(ytest, predictions))
         print(classification_report(ytest,predictions))

[[43  3]
 [ 1 63]]
              precision    recall  f1-score   support

           0       0.98      0.93      0.96        46
           1       0.95      0.98      0.97        64

    accuracy                           0.96       110
   macro avg       0.97      0.96      0.96       110
weighted avg       0.96      0.96      0.96       110
```

Fig 3: Decision Tree Accuracy



Support Vector Machine (SVM)

```
In [24]: from sklearn.svm import SVC

         svc_model = SVC()
         svc_model.fit(xtrain, ytrain)

Out[24]: ▸ SVC

In [25]: svc_pred = svc_model.predict(xtest)

In [26]: from sklearn import metrics

         print("Accuracy Score =", format(metrics.accuracy_score(ytest, svc_pred)*100))

         Accuracy Score = 65.45454545454545
```

Classification report and confusion matrix of the SVM classifier

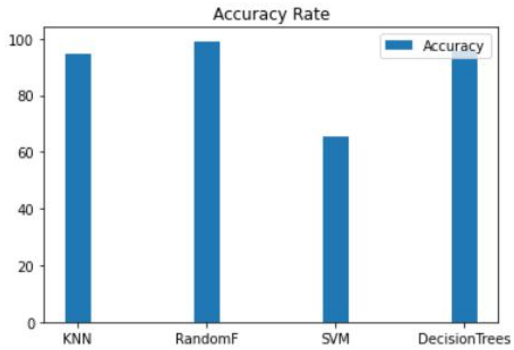
```
In [27]: from sklearn.metrics import classification_report, confusion_matrix

         print(confusion_matrix(ytest, svc_pred))
         print(classification_report(ytest,svc_pred))

[[10 36]
 [ 2 62]]
              precision    recall  f1-score   support

           0       0.83      0.22      0.34        46
           1       0.63      0.97      0.77        64

    accuracy                           0.65       110
   macro avg       0.73      0.59      0.56       110
weighted avg       0.72      0.65      0.59       110
```

Fig 4: SVM Accuracy

VII. CONCLUSUION

As a direct result of this, with an accuracy score of 98% Random Forest is the most appropriate model for making this prediction. One of the most well-known techniques for supervised machine learning is called Random Forest. In the field of machine learning, it may be used to tackle issues involving classification and regression.

It is predicated on the idea of ensemble learning, which is a method of combining a large number of different classifiers in order to solve a difficult issue and enhance the performance of the model.



REFERENCES

[1] “Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus” Md Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, IEEE 2019.

[2] “A Comprehensive Exploration to the Machine Learning Techniques for Diabetes Identification” Sidong Wei1, Xuejiao Zhao, Chunyan Miao Shanghai Jiao Tong University, China.

[3] “Association Rule Extraction from Medical Transcripts of Diabetic Patients” Lakshmi K S, G Santhosh Kumar, 2014.

[4] “Diabetes Care Decision Support System” 2nd International Conference on Industrial and Information Systems IEEE 2010.

[5] “An Intelligent Mobile Diabetes Management and Educational System for Saudi Arabia: System Architecture” M.M. Alotaibi, R.S.H. Istepanian, A.Sungoor and N. Philip, IEEE 2014.

[6] “Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases” by BerinaAlic, Lejla Gurbeta, IEEE 2017.

[7] “Performance Analysis of Classification Approaches for the Prediction of Type II Diabetes” by M. Durgadevi, M. Durgadevi, IEEE 2017.

[8] “Cloud-Based Diabetes Coaching Platform for Diabetes Management” Elliot B. Sloane Senior Member IEEE, Nilmini Wickramasinghe, Steve Goldberg 2016.

[9] Minyechil Alehegn and Rahul Joshi, “Analysis andprediction of diabetes diseases using machine learning algorithm”:International Research Journal of Engineering and Technology Volume: 04 Issue: 10 | Oct -2017.

[10] P. Suresh Kumar and V. Umatejaswi, “Diagnosing Diabetes using Data Mining Techniques”,International Journal of Scientific and ResearchPublications, Volume 7, Issue 6, June 2017 705 ISSN 2250-3153.

[11] “Clustering Medical Data to Predict the Likelihood of Diseases” by Razan Paul, Abu Sayed Md. Latiful Hoque, IEEE 2010.

[12] “Robust Parameter Estimation in a Model for Glucose Kinetics in Type 1 Diabetes Subjects” Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, Aug 30-Sept 3, 2006.

[13] Anjali C And Veena Vijayan V, Prediction and Diagnosis of Diabetes Mellitus, “A Machine Learning Approach” ,2015 IEEE in Intelligent Computational Systems (RAICS) | Trivandrum.

[14] Ridam Pal ,Dr. Jayanta Poray, and Mainak Sen, ,“Application of Machine Learning Algorithms on Diabetic Retinopathy”, 2017 2nd IEEE International Conference On Recent Trends In Electronics Information & Communication Technology, May 19-20, 2017, India.

[15] Dr. M. Renuka Devi and J. Maria Shyla, “Analysis of Various Data Mining Techniques toPredict Diabetes Mellitus”, International Journal ISSN 0973-4562 Volume 11, Number 1 (2016) pp 727-730.

[16] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar,” Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop”, International Conference On I-SMAC,978-1-5090-3243-3,2017.

[17] Ayush Anand and Divya Shakti,” Prediction of Diabetes Based on Personal Lifestyle Indicators”, 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.

[18] B. Nithya and Dr. V. Ilango,” Predictive Analytics in Health Care Using Machine Learning Tools and Techniques”, International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.

[19] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S,” Predictive Methodology for Diabetic Data Analysis in Big Data”, 2nd International Symposium on Big

Data and Cloud Computing,2015.

[20] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

[21] P. Suresh Kumar and S. Pranavi "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.

[22] Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, Volume 120 - Number 8,2015.

[23] K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

[24] Humar Kahramanli and Novruz Allahverdi,"Design of a Hybrid System for the Diabetes and Heart Disease", Expert Systems with Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.

[25] B.M. Patil, R.C. Joshi and Durga Toshniwal,"Association Rule for Classification of Type-2 Diabetic Patients", ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010.

[26] Dost Muhammad Khan1, Nawaz Mohamudally2, "An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm ", Journal Of Computing, Volume 3, Issue 12, December 2011
.