**DIABETES PREDICTION USING MACHINE LEARNING ALGORITHM**

**A PROJECT REPORT**

*for*

**DATA MINING TECHNIQUES(ITE2006)**

**B.Tech (Information Technology)**

By

**P. VENKATA RAMANA RAO 19BIT0352**

**C.NIKHIL 19BIT0331**

**EZHILRASI 19BIT0395**

**WINTER SEMESTER 2022**



**School of Information Technology and Engineering**

**APRIL 2022**

**DECLARATION BY THE CANDIDATE**

We here by declare that the project report entitled **“DIABETES PREDICTION USING MACHINE LEARNING ALGORITHM”** submitted by us to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (ITE2006)** is a record of bonafide project work carried out by us under the guidance of **Prof. B.Valarmathi.** We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other course.

Place: Vellore

DATE:26.04.22 SIGNATURE



**School of Information Technology & Engineering [SITE]**

This is to certify that the project report entitled **“DIABETES PREDICTION USING MACHINE LEARNING ALGORITHM”** submitted by **P. VENKATA RAMANA RAO 19BIT0352 , C.NIKHIL 19BIT0331, EZHILRASI 19BIT0395**

to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (ITE2006)** is a record of bonafide work carried out by them under my guidance.

**Prof. B.Valarmathi**

**GUIDE**

**Asso. Professor, SITE**

**DIABETES PREDICTION USING MACHINE LEARNING ALGORITHM**

**P. VENKATA RAMANA RAO, C. NIKHIL, EZHILRASI**

**Department of Information Technology, VIT University, Vellore, Tamil Nadu, India**

**Abstract**

Data mining and machine learning is helping medical professionals make diagnosis easier by bridging the gap between huge data sets and human knowledge. We can begin to apply machine learning techniques for classification in a dataset that describes a population that is under a high risk of the onset of diabetes. Diabetes Mellitus affects 382 million people in the world, and the number of people with type-2 diabetes is increasing in every country. Untreated, diabetes can cause many complications. Given the medical data we can gather about people, we should be able to make better predictions on how likely a person is to suffer the onset of diabetes, and therefore act appropriately to help. Diabetes mellitus is a multifactorial disease that requires long-term care since it involves major changes in both physical and psychosocial dimension of each patient. We can start analyzing data and experimenting with algorithms that will help us study the onset of diabetes in Pima Indians.

**Keywords** – Machine learning, Diabetes prediction, Better Prediction model, Classification algorithms.

**I INTRODUCTION**

Diabetes is a common chronic disease and poses a great threat to human health. The characteristic of diabetes is that the blood glucose is higher than the normal level, which is caused by defective insulin secretion or its impaired biological effects, or both (Lonappan et al., 2007). Diabetes can lead to chronic damage and dysfunction of various tissues, especially eyes, kidneys, heart, blood vessels and nerves (Krasteva et al., 2011). Diabetes can be divided into two categories, type 1 diabetes (T1D) and type 2 diabetes (T2D). Patients with type 1 diabetes are normally younger, mostly less than 30 years old. The typical clinical symptoms are increased thirst and frequent urination, high blood glucose levels (Iancu et al., 2008). This type of diabetes cannot be cured effectively with oral medications alone and the patients are required insulin therapy. Type 2 diabetes occurs more commonly in middle-aged and elderly people, which is often associated with the occurrence of obesity, hypertension, dyslipidemia, arteriosclerosis, and other diseases (Robertson et al., 2011). 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 studying.

**II BACKGROUND**

In this system, the authors used Linear Discriminant Analysis (LDA) to reduce the dimensions and extract the features. In order to deal with the high dimensional datasets, Razavian et al. (2015) built prediction models based on logistic regression for different onsets of type 2 diabetes prediction. Georga et al. (2013) focused on the glucose, and used support vector regression (SVR) to predict diabetes, which is as a multivariate regression problem. Moreover, more and more studies used ensemble methods to improve the accuracy (Kavakiotis et al., 2017). Ozcift and Gulten (2011) proposed a newly ensemble approach, namely rotation forest, which combines 30 machine learning methods. Han et al. (2015) proposed a machine learning method, which changed the SVM prediction rules. Machine learning methods are widely used in predicting diabetes, and they get preferable results. Decision tree is one of popular machine learning methods in medical field, which has grateful classification power. Random forest generates many decision trees. Neural network is a recently popular machine learning method, which has a better performance in many aspects. So in this study, we used decision tree, random forest (RF) and neural network to predict the diabetes.

**III Literature Survey**

**[1]. Machine Learning Methods to Predict Diabetes Complications, 2019**

Authors: Arianna Dagliati, PhD, Simone Marini, PhD, Lucia Sacchi, PhD, Giulia Cogni, MD, Marsida Teliti, MD, Valentina Tibollo, MS, Pasquale De Cata, MD, Luca Chiovato, PhD, and Riccardo Bellazzi, PhD

In this paper, Machine learning algorithms have been embedded into data mining pipelines. Within the EU-funded MOSAIC project, a data mining pipeline has been used to derive a set of predictive models of type 2 diabetes mellitus (T2DM) complications based on electronic health record data of nearly 1000 patients. Center profiling, Predictive models targeting, Predictive models construction, Predictive models validation are the four methods made.

**[2]. Machine Learning and Data Mining Methods in Diabetes Research,2019**

Authors: Ioannis Kavakiotis, Olga Tsave , Athanasios Salifoglou , Nicos Maglaveras,

Ioannis Vlahavas , Ioanna Chouvarda

The aim of this paper is to conduct a systematic review of the applications of machine learning, data mining techniques and tools in the field of diabetes research with respect to a) Prediction and Diagnosis, b) Diabetic Complications, c) Genetic Background and Environment, and e) Health Care and Management. Machine learning tasks and feature selection. PubMed and DBLP are the datasets used.

**[3]. Predicting Diabetes Mellitus With Machine Learning Techniques,2019**

Authors: Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju and Hua Tang

In this study, they have used decision tree, random forest and neural network to predict diabetes mellitus. The dataset is the hospital physical examination data in Luzhou, China. It has 14 attributes. Five-fold cross validation was used to examine the models. Principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) were used to reduce the dimensionality. The results could reach the highest accuracy (ACC = 0.8084)

**[4]. Predictive models for diabetes mellitus using machine learning techniques,2019**

Authors: Hang Lai, Huaxiong H.uang, Karim Keshavjee, Aziz Guergachi and Xin Gao

This paper was to build an effective predictive model with high sensitivity and

selectivity to better identify Canadian patients at risk of having Diabetes Mellitus based on patient demographic data. We built predictive models using Logistic Regression and Gradient BoostingMachine (GBM) techniques. The area under the receiver operating characteristic curve (AROC) was used. Adjusted threshold method and the class weight method were used. The AROC for the proposed GBM model is 84.7% with a sensitivity of 71.6%

**[5]. Predictive models for diabetes mellitus using machine learning techniques,2019**

Authors: Hang Lai, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi and Xin Gao

This paper was to build an effective predictive model with high sensitivity and

selectivity to better identify Canadian patients at risk of having Diabetes Mellitus based on patient demographic data. We built predictive models using Logistic Regression and Gradient BoostingMachine (GBM) techniques. The area under the receiver operating characteristic curve (AROC) was used. Adjusted threshold method and the class weight method were used. The AROC for the proposed GBM model is 84.7% with a sensitivity of 71.6% .

# [6]. Diabetes Prediction using Machine Learning Algorithm, 2019

# Author: Aishwarya Mujumdara, Dr. Vaidehi Vb

# A procedure called, Predictive Analysis, joins an assortment of machine learning algorithms, information mining strategies and factual techniques that utilizes current and past information to track down information and anticipate future occasions. predictive analysis should be possible utilizing machine learning algorithm like Logistic Regression, Random Forest Classifier, AdaBoost Classifier. After applying all machine learning algorithms Logistic Regression gives most elevated exactness of 96%. Use of pipeline gave AdaBoost classifier as best model with precision of 98.8%.

**[7]. A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Technique, 2019**

Author: [Amani Yahyaoui](https://www.researchgate.net/profile/Amani-Yahyaoui-2) , [Jawad Rasheed](https://www.researchgate.net/profile/Jawad-Rasheed-2)

We propose a DSS for diabetes prediction based on Machine Learning (ML) techniques. We compared conventional machine learning with deep learning approaches. For conventional machine learning method, we considered the most commonly used classifiers: Support Vector Machine (SVM) and the Random Forest(RF). On the other hand, for Deep Learning (DL) we employed a fully Convolutional Neural Network (CNN) to predict and detect the diabetes patients. The proposed system is evaluated on publicly available Pima Indians Diabetes database which consisted of total 768 samples each with 8 features. 500 samples were labeled as non-diabetic while 268 were diabetic patients. The overall accuracy obtained using DL, SVM and RF was 76.81%, 65.38% and 83.67% respectively.

# [8]. Implementation of machine learning algorithms to create diabetic patient re-admission profiles, 2019

Author: [Mohamed Alloghani](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0990-x#auth-Mohamed-Alloghani),  [Ahmed Aljaaf](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0990-x#auth-Ahmed-Aljaaf)

# We applied machine learning to the diabetes dataset with the aim of recognizing patterns and combinations of factors that characterizes or explain re-admission among diabetes patients. The classifiers used include Linear Discriminant Analysis, Random Forest, k–Nearest Neighbor, Naïve Bayes, J48 and Support vector machine. Results are Of the 100,000 cases, 78,363 were diabetic and over 47% were readmitted. KNN models lead to the conclusion that fewer number of lab procedures, diagnoses, and medications lead to increased higher re-admission rates.

**[9]. Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm, 2019**

Author: [Mohamed Alloghani](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0990-x#auth-Mohamed-Alloghani), [Ahmed Aljaaf](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0990-x#auth-Ahmed-Aljaaf), [Abir Hussain](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0990-x#auth-Abir-Hussain),

The aim of the paper is tobuild a web application based on the higher prediction accuracy of some powerful machine learning algorithm. We have used a benchmark dataset namely Pima Indian which is capable of predicting the onset of diabetes based on diagnostics manner. With an accuracy of 82.35% prediction rate Artificial Neural Network (ANN) shows a significant improvement of accuracy which drives us to develop an Interactive Web Application for Diabetes Prediction. From different machine learning algorithms Artificial Neural Network (ANN) provide us highest accuracy with Min Max Scaling Method on Indian Pima Dataset.

# [10]. Performance Evaluation of Different Machine Learning Methods and Deep-Learning Based Convolutional Neural Network for Health Decision Making, 2019

Author:Abhaya Kumar Sahoo, Chittaranjan Pradhan ,Himansu Das

To predict the diabetes disease, machine learning and deep learning play major role which uses probability, statistics and neural network concepts, etc. pima indian diabetics dataset is used. The experiment results prove that convolution neural network based deep learning method provides the highest accuracy than other machine learning algorithms.

**[11]. Comparing different supervised machine learning algorithms for diabetes prediction,2019**

Authors: [Shahadat Uddin](https://link.springer.com/article/10.1186/s12911-019-1004-8#auth-Shahadat-Uddin), [Arif Khan](https://link.springer.com/article/10.1186/s12911-019-1004-8#auth-Arif-Khan),  [Md Ekramul Hossain](https://link.springer.com/article/10.1186/s12911-019-1004-8#auth-Md_Ekramul-Hossain) &  [Mohammad Ali Moni](https://link.springer.com/article/10.1186/s12911-019-1004-8#auth-Mohammad_Ali-Moni)

Supervised machine learning algorithms have been a dominant method in the data mining field. Diabetics prediction using health data has recently shown a potential application area for these methods. This study aims to identify the key trends among different types of supervised machine learning algorithms, and their performance and usage for disease risk prediction.

**[12]. Diabetes Prediction Using Different Machine Learning Approaches,2019**

Authors: Priyanka Sonar; K. Jaya Malini

The aim of this analysis is to develop a system which might predict the diabetic risk level of a patient with a better accuracy. Model development is based on categorization methods as Decision Tree, ANN, Naive Bayes and SVM algorithms. For Decision Tree, the models give precisions of 85%, for Naive Bayes 77% and 77.3% for Support Vector Machine. Outcomes show a significant accuracy of the methods.

**[13]. Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster,2019**

Authors: N. Yuvaraj &  K. R. SriPreethaa

 Various Machine Learning algorithms are proposed which automates the working model of health care systems and enhances the accuracy of disease prediction. Hadoop cluster based distributed computing framework supports in efficient processing and storing of extremely large datasets in cloud environment. This work proposes the novel implementation of machine learning algorithms in hadoop based clusters for diabetes prediction. The results show that the machine learning algorithms can able to produce highly accurate diabetes predictive healthcare systems.

**[14]. Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus,2019**

Authors: Md. Faisal Faruque; Asaduzzaman; Iqbal H. Sarker

 Machine learning techniques provide efficient result to extract knowledge by constructing predicting models from diagnostic medical datasets collected from the diabetic patients. Extracting knowledge from such data can be useful to predict diabetic patients. In this work, we employ four popular machine learning algorithms, namely Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN) and C4.5 Decision Tree (DT), on adult population data to predict diabetic mellitus. Our experimental results show that C4.5 decision tree achieved higher accuracy compared to other machine learning techniques.

# [15]. Predictive machine learning model for early detection and analysis of diabetes, 2020

# Author: [D.Jashwanth Reddy](https://www.sciencedirect.com/science/article/pii/S2214785320372382#!), [B.Mounika](https://www.sciencedirect.com/science/article/pii/S2214785320372382#!), [S.Sindhu](https://www.sciencedirect.com/science/article/pii/S2214785320372382#!), [T.Pranayteja Reddy](https://www.sciencedirect.com/science/article/pii/S2214785320372382#!)

The major aim of the paper is designing a model which predicts the diabetes in human with maximum accuracy using machine learning classifiers like Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), Navies Bayes (NB), Gradient Boosting (GB) and Random Forest (RF) Classifier. Analysis is done on Pima Indian Diabetes Database (PIDD), a dataset taken from Kaggle data repository. The performance of all the six classifiers is compared using Accuracy score, Receiver Operating Curve (ROC), Precision, Recall, F-measure evaluated from each model.

# [16]. Artificial Intelligence in Health Care: Predictive Analysis on Diabetes Using Machine Learning Algorithms, 2020

Author:Shruti Wadhwa, Karuna BabberEmail author

It uses the large volume of multimodal patient data to perform correlations between Body Mass Index, Blood Pressure, Glucose levels, Diabetes Pedigree Function and Skin Thickness of people in different age groups with diabetes. Python and data analytic packages are used to predict diabetes among people the results indicate a strong relationship between Blood Pressure, BMI and Glucose levels of people with diabetes whereas a moderate correlation has been found between Age, Skin Thickness and Diabetes Pedigree Function count of people with diabetes. Although present analysis attested many of the previous research findings but getting these inferences matched through analytical tools is a sole purpose of this paper.

**[17]. Predictive modelling and analytics for diabetes using a machine learning approach,2020**

Authors: Harleen Kaur, [Vinita Kumari](https://www.emerald.com/insight/search?q=Vinita%20Kumari)

 In the current research we have utilized machine learning technique in Pima Indian diabetes dataset to develop trends and detect patterns with risk factors using R data manipulation tool. To classify the patients into diabetic and non-diabetic we have developed and analyzed five different predictive models using *R* data manipulation tool. For this purpose we used supervised machine learning algorithms namely linear kernel support vector machine (SVM-linear), radial basis function (RBF) kernel support vector machine, *k*-nearest neighbour (*k*-NN), artificial neural network (ANN) and multifactor dimensionality reduction (MDR).

**[18]. Diabetes Prediction Using Machine Learning Techniques,2020**

Authors: Tejas N. Joshi, Prof. Pramila M. Chawan

The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. This project aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. Diabetes risk prediction with the

help of advanced computational methods and availability of large amount of epidemiological and genetic diabetes risk dataset.

**[19]. Diabetes Prediction using Machine Learning Techniques, 2020**

Authors: Mitushi Soni, Dr. Sunita Varma

In this paper work, early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques. In this work, they used Machine Learning Classification and ensemble techniques on a dataset to predict diabetes. Which are K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT),

Support Vector Machine (SVM), Gradient Boosting (GB) and Random Forest (RF). And

77% classification accuracy has been achieved.

**[20]. Intelligent diabetes mellitus prediction framework using machine learning, 2020**

Authors: Leila Ismail and Huned Materwala

This paper provides novel intelligent diabetes mellitus prediction framework (IDMPF)

using machine learning. The authors implement and evaluate the decision tree (DT)-based random forest (RF) and support vector machine (SVM) learning models for diabetes prediction. The proposed IDMPF is based on the data analytics lifecycle. The proposed framework was using an imbalanced UCI dataset with parameters.

**[21]. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers,2020**

Authors: Md. Kamrul Hasan; Md. Ashraful Alam; Dola Das; Eklas Hossain; Mahmudul Hasan

In this paper, diabetes prediction has been accomplished using the proposed ensemble model from the PID dataset, where the pre processing plays a crucial role in robust and precise prediction where the outlier rejection, filling the missing values, data standardization, feature selection, K-fold cross-validation, and different Machine Learning (ML) classifiers (k-nearest Neighbour, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XG Boost) and Multilayer Perceptron (MLP) were employed.

**[22]. Predicting the Development of Type 2 Diabetes in a Large Australian Cohort Using Machine-Learning Techniques: Longitudinal Survey Study,2020**

Authors: Lei Zhang   ;  Xianwen Shang  , SubhashaanSreedharan   ,  Xixi Yan   , Jianbin Liu  ,  Stuart Keel ,  Jinrong Wu   ,  Wei Peng    ,  Mingguang He

 All machine-learning models predicted BMI as the most significant factor contributing to diabetes onset, which explained 12%-50% of the variance in the prediction of diabetes. The model predicted that if BMI in obese and overweight participants could be hypothetically reduced to a healthy range, the 10-year probability of diabetes onset would be significantly reduced from 8.3% to 2.8% (*P*<.001).

**[23]. Classification and prediction of diabetes disease using machine learning paradigm,2020**

Authors: Md. Maniruzzaman,  Md. Jahanur Rahman,  BenojirAhammed &  Md. Menhazul Abedin

We have used diabetes dataset and adopted four classifiers like naïve Bayes (NB), decision tree (DT), Ada boost (AB), and random forest (RF) to predict the diabetic patients. Three types of partition protocols (K2, K5, and K10) have also adopted and repeated these protocols into 20 trails. The overall ACC of ML-based system is **90.62%**. The combination of LR-based feature selection and RF-based classifier gives **94.25%** ACC and **0.95** AUC for K10 protocol.

**[24]. A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus,2021**

Authors: Haohui Lu,  Shahadat Uddin,  FarshidHajati,  Mohammad Ali Moni & Matloob Khushi

We apply eight machine learning models (Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest, XGBoost and Artificial Neural Network) to the extracted features to predict the chronic disease risk. The extensive experiments show that the proposed framework with machine learning classifiers performance with the Area Under Curve (AUC) ranged from 0.79 to 0.91. The Random Forest model outperformed the other models

**[25]. A comparison of machine learning algorithms for diabetes prediction,2021**

Authors: [Jobeda Jamal Khanam](https://www.sciencedirect.com/science/article/pii/S2405959521000205#!) ,[Simon Y.Foo](https://www.sciencedirect.com/science/article/pii/S2405959521000205#!)

Data mining, machine learning (ML) algorithms, and Neural Network (NN) methods are used in diabetes prediction in our research. We used the Pima Indian Diabetes (PID) dataset and used seven ML algorithms on the dataset to predict diabetes. We found that the model with Logistic Regression (LR) and Support Vector Machine (SVM) works well on diabetes prediction. We built the NN model with a different hidden layer with various epochs and observed the NN with two hidden layers provided 88.6% accuracy.

# [26]. Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications, 2021

# Author: Umair Muneer Butt, Sukumar Letchmunan

# The primary objective of the proposed system is to help the users monitor their vital signs using BLE-based sensor devices with the help of their smartphones. For diabetes classification, three different classifiers have been employed, i.e., random forest (RF), multilayer perceptron (MLP), and logistic regression (LR). For predictive analysis, we have employed long short-term memory (LSTM), moving averages (MA), and linear regression (LR). For experimental evaluation, a benchmark PIMA Indian Diabetes dataset is used. During the analysis, it is observed that MLP outperforms other classifiers with 86.08% of accuracy and LSTM improves the significant prediction with 87.26% accuracy of diabetes.

# [27]. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm, 2021

Author: [Mohamed Alloghani](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0990-x#auth-Mohamed-Alloghani), [Ahmed Aljaaf](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0990-x#auth-Ahmed-Aljaaf), [Abir Hussain](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0990-x#auth-Abir-Hussain)

Prediction of type 2 diabetes (T2D) occurrence allows a person at risk to take actions that can prevent onset or delay the progression of the disease. In this study, we developed a machine learning (ML) model to predict T2D occurrence in the following year (Y + 1) using variables in the current year (Y). The dataset for this study was collected at a private medical institute as electronic health records from 2013 to 2018. To construct the prediction model, key features were first selected using ANOVA tests, chi-squared tests, and recursive feature elimination methods. The resultant features were fasting plasma glucose (FPG), HbA1c, triglycerides, BMI, gamma-GTP, age, uric acid, sex, smoking, drinking, physical activity, and family history. We then employed logistic regression, random forest, support vector machine, XGBoost, and ensemble machine learning algorithms based on these variables to predict the outcome as normal (non-diabetic), prediabetes, or diabetes. The model can provide both clinicians and patients with valuable information on the incidence of T2D ahead of time, which would help patients take measures to mitigate T2D risk, progression, and related complications.

# [28]. A comprehensive review of machine learning techniques on diabetes detection, 2021

Author: [Toshita Sharma](https://vciba.springeropen.com/articles/10.1186/s42492-021-00097-7#auth-Toshita-Sharma) & [Manan Shah](https://vciba.springeropen.com/articles/10.1186/s42492-021-00097-7#auth-Manan-Shah)

# Diabetes can be devastating after a certain period if not detected or diagnosed correctly. Many machine learning methods have been discussed, starting from different basic algorithms such as the LR, SVM, DTs, to further classification including the ID3, C4.5, C5.0, J48 and CART and NB. Ensemble methods, such as bagging, boosting, and RF regressors , are further used to enhance the accuracy and performance of models. These techniques have been implemented on all types of platforms such as Python or MATLAB, and the models have been analyzed using different parameters such as area under curve or confusion matrices or mathematical terms such as the RMSE or MAE.

**[29]. A Machine Learning Approach to Predicting Diabetes Complications, 2021**

Authors: Yazan Jian , Michel Pasquier, Assim Sagahyroon and Fadi Aloul

For this Paper, a dataset collected by the Rashid Center for Diabetes and Research (RCDR) was utilized. The dataset consists of 884 records with 79 features. Some essential preprocessing steps such as data cleaning, data imputation, categorical encoding, data balancing and normalization were applied to handle the missing values and unbalanced data problems. The k-fold cross-validation (KCV) technique is to select a classifier. Feature selection was performed to select the top five and ten features for each complication.

**[30]. Predict Diabetes Mellitus Using Machine Learning Algorithms, 2021**

Authors: Farhana Bano, Munidhanalakshmi K and Dr R.MadanaMohana

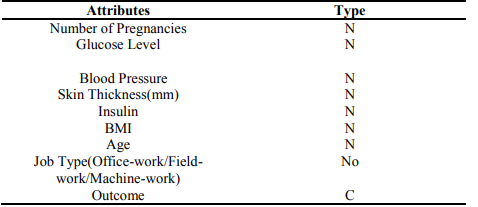
In this paper, they were with different machine learning algorithms, namely SVM, ANN, Decision tree, Logistic regression and Farthest first to predict the accuracy. The data set consist of varied attribute or hazard factor of diabetes of 768 patients. Pre-processing was done at first to clean up the data set. Cloud Application act like a cloud server and storage and we train the dataset model with five algorithms. Farthest First Consist of 99.9 percentage of Accuracy.

| S.NO | Title of the paper and Year | Algorithms used | Datasets being used | Performance measures | Scope for future work |
| --- | --- | --- | --- | --- | --- |
| 1. | Machine Learning Methods to Predict Diabetes Complications, 2019 | random forest and  Logistic Regression | diabetes mellitus (T2DM) complications based on electronic health record data | accuracy = 0.838 |  |
| 2. | A Machine Learning Approach to Predicting Diabetes  Complications  A Machine Learning Approach to Predicting Diabetes  Complications  A Machine Learning Approach to Predicting Diabetes Complications, 2021 | The k- fold cross-validation (KCV), data preprocessing and Feature selection | The dataset consists of 884 records with  79 features. | Accuracy= 97.8% and 97.7% |  |
| 3. | Machine Learning and Data Mining Methods in Diabetes Research,2019 | machine learning and data mining techniques | PubMed and DBLP dataset | acc = 0.88 |  |
| 4. | Predicting Diabetes Mellitus With Machine Learning Techniques,2019 | decision tree, random forest and neural network | The dataset is the hospital physical examination data in Luzhou, China. It contains 14 attributes. | ACC = 0.8084 |  |
| 5. | Diabetes Prediction using Machine Learning Techniques,2020 | K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB) and Random Forest (RF). T | Diabetes risk prediction with the help of advanced computational methods and availability of large amount of epidemiological and genetic diabetes risk dataset. | Accuracy=77% |  |
| 6. | Predictive models for diabetes mellitus using machine learning techniques,2019 | Logistic Regression and Gradient Boosting Machine (GBM) techniques compared these models to other learning machine techniques such as Decision Tree and Random Forest. | They collected demographic data and the laboratory results from Canadian patients | GBM model is 84.7% with a sensitivity of 71.6% and the AROC for the proposed Logistic Regression model is 84.0% with a sensitivity of 73.4%. | These models can be built into an online computer program to help physicians in predicting patients with future occurrence of diabetes and providing necessary preventive interventions. |
| 7. | Diabetes Prediction Using Machine Learning Techniques | SVM, Logistic regression, ANN | epidemiological and genetic diabetes risk dataset | - | The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decision about the disease status. |
| 8. | IDMPF: intelligent diabetes mellitus prediction framework using machine learning | decision tree (DT)-based random forest (RF) and support vector machine (SVM) learning models | Collect data from an online public data repository such as UCI machine learning repository | acc = 0.87 | A larger spectrum of models will be considered in their future work. |
| 9. | A Machine Learning Approach to Predict Diabetes Using Short Recorded Photoplethysmography & Physiological Characteristics,2019 | Linear Discriminant Analysis  ,Data Preprocessing & Feature Extraction and Feature Selection |  | LDA-79%. | - |
| 10. | Predict Diabetes Mellitus Using Machine Learning  Algorithms | SVM, ANN, Decision tree, Logistic regression and Farthest first | They have a inclination to gather diabetes information from the kaggle site. The data set consist of varied attribute or hazard factor of diabetes of 768 patients. | Farthest first-99.9 percentage of Accuracy. | This experimental Results will help to health care to early prediction of the Diabetes. |
| 11. | Diabetes Prediction using Machine Learning  Algorithms with Feature Selection and  Dimensionality Reduction | Support Vector Machine (SVM) & Random Forest (RF), . The Principle Component Analysis | PIMA dataset | Random Forest-83% and Support Vector Machine (SVM)-81.5% | Dimensionality is not of high significance as the dataset is not very large.so in future they are planned to use for very large datasets. |
| 12. | Diabetes Mellitus Prediction using Different  Ensemble Machine Learning Approaches | AdaBoost, Bagging and Random Forest | They collected real time information of both diabetic and non-diabetic people. The dataset contains 464 instances with 22 unique risk factors. | AdaBoost gave 97.84% accuracy, Bagging gave 98.28% accuracy and Random Forest gave 99.35% accuracy | They are goin to develop an expert system with their study therefore, they can predict diabetes more productively and adequately. Also, they need to enhance our sample size (464 instances) and handle the missing data more efficient way. |
| 13. | Presaging The Signs Of Diabetes Using Machine  Learning Algorithms | Logistic Regression, Decision Tree, KNeighbours, and Naive Bayes Classifiers . | The data comes from a study of the Pima Indians in the Phoenix area of Arizona | Logistic Regression model with an Accuracy of 79.2% . |  |
| 14. | Diabetes Prediction Using Different Machine  Learning Algorithms | Random forest and K-NN algorithms | pima Indian diabetes dataset, which is downloaded from Kaggle. | Random forest tree generated 78.4% accuracy whereas K-NN generated 80.8% accuracy. |  |
| 15. | DIABETES PREDICTION USING DIFFERENT  MACHINE LEARNING APPROACHES | Decision Tree, ANN, Naive Bayes and SVM algorithms. | Global dataset  The dataset contains seven sixty eight instances and nine features | Dt-74% | ANN: Gives good prediction and easy to implement.but Difficult with dealing with big data with complex model. Require huge processing time. |
| 16. | Diabetes Mellitus Type-2 Web-based RiskPredictionSystem Using Predictive Analytics &Optimal Features | Naive Bayes, Random Forest, and Support Vector Machine. | dataset was collected from Liaquat University of Medical &Health Sciences (LUMHS) Jamshoro and online surveys inPakistan. The dataset consists of 2053 samples with 10 featuresencompassing significant diabetes risk factors. | l SVM with RBF kernel gave the highest accuracy of 87% |  |
| 17. | Diabetes Prediction using BPSO and Decision Tree  Classifier | BPSO, Decision Tree | PIMA Database | Decision Tree- 96.1 % |  |
| 18. | Diabetes Prediction Using Ensemble Perceptron Algorithm,2017 | boosting algorithm with perceptron algorithm and Ensemble Perceptron Algorithm | NHANES dataset | The proposed algorithm is validated on three different NHANES datasets confirming that AUC value improves from 0.72 to 0.75 by the proposed algorithm. |  |
| 19. | Diabetes prediction in healthcare systems using machine learning  algorithms on Hadoop cluster | Neural network, Decision tree and Naive Bayes | Pima Indians Diabetes Database | Random forest algorithm produces 3% high in accuracy than naïve bayes algorithm and 6% high in accuracy than decision tree algorithm. |  |
| 20. | Benchmarking Machine Learning Algorithms on  Blood Glucose Prediction for Type I Diabetes in  Comparison With Classical Time-Series Models | machine-learning models and Autoregression with Exogenous inputs (ARX) model | The OhioT1DM Dataset | - |  |

* **DATASET DESCRIPTION & SAMPLE DATA**

DATASET NAME - PIMA-INDIAN-DIABETES.

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of this diabetes dataset is to predict whether patient has diabetes or not. The dataset consists of several medical predictor (Independent) variables and one target variable (outcome). Predictor variables include Pregnancies, Glucose, Blood pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age. Our target variable is Outcome.



| Number of instances | 768 |
| --- | --- |
| Number of attributes | 9 |
| Year published | 2016 |

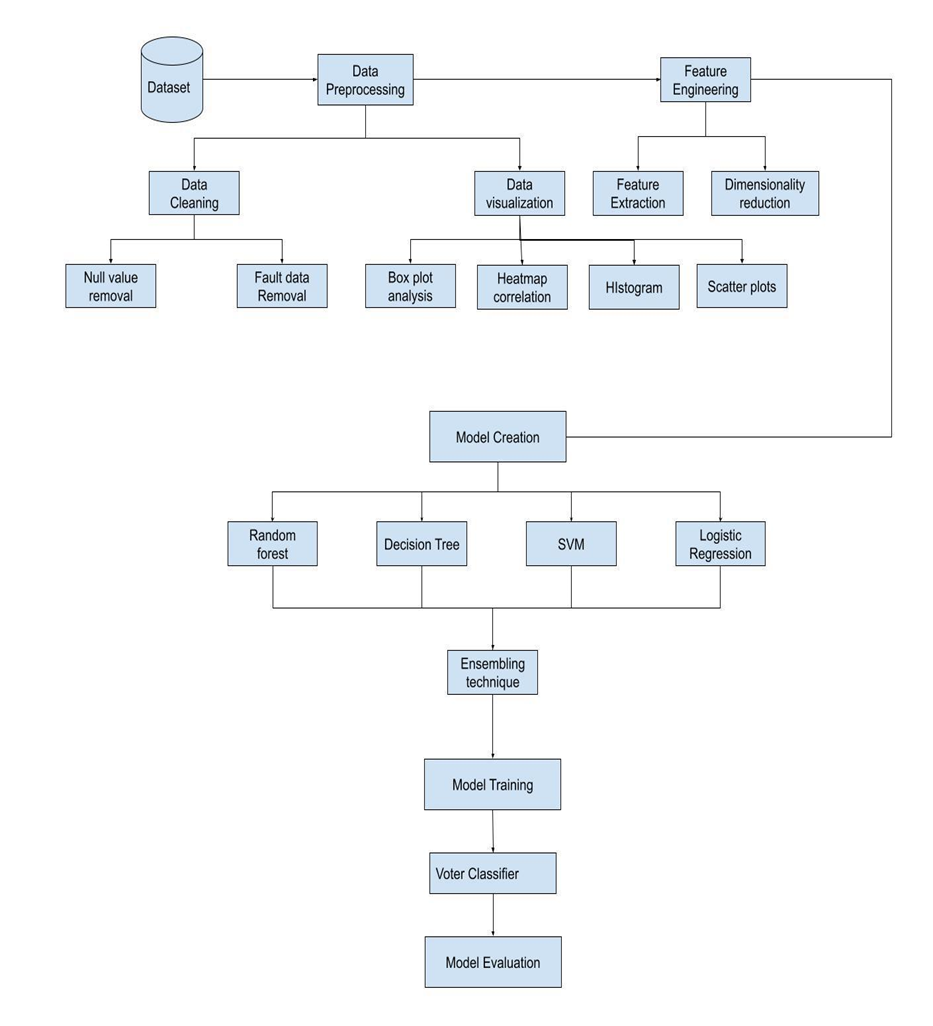
## SAMPLE DATA FOR OUR PROJECT:

| Pregnancies | Glucose | BP | Skin Thickness | Insulin | BMI | Diabetes pedigree function | Age | Outcome |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 0 | 137 | 40 | 35 | 168 | 43.1 | 0.288 | 33 | 1 |

**IV PROPOSED ALGORITHM WITH FLOWCHART**

* Our proposed system shows the higher prediction accuracy than the existing system.
* Our system or model assures less time consuming.
* This model is not so complicated so that accessing will be very easier to the people who wanted to use.

The main objective of this study is to develop a machine learning (ML)-based system for predicting diabetic patients. In this detection of diabetes using machine learning techniques such as Random Forest, Logistic Regression, Naïve Bayes, Decision tree, Support vector machine, KNN are used to predict the risk of diabetes. We will also be using boosting methods and deep learning to increase the accuracy of the system. In all these above-mentioned techniques the patient records are classified and predicted continuously. We will be using different software metrics like Accuracy, Precision, Recall, F1 score, ROC curve, Specificity and compare the above mentioned models and m predict which one performs better than others. The proposed system act as a decision support system and will prove to be an aid for the physicians with the diagnosis. The patient activity is monitored continuously, if there are any changes occur, then the risk level of disease is informed to the patient and doctor. The doctors are able to predict diabetes at an earlier stage because of machine learning algorithms and with the help of computer technology.

****

**ALGORITHMS USED:**

**Logistic Regression:**

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

**Steps in Logistic Regression:**

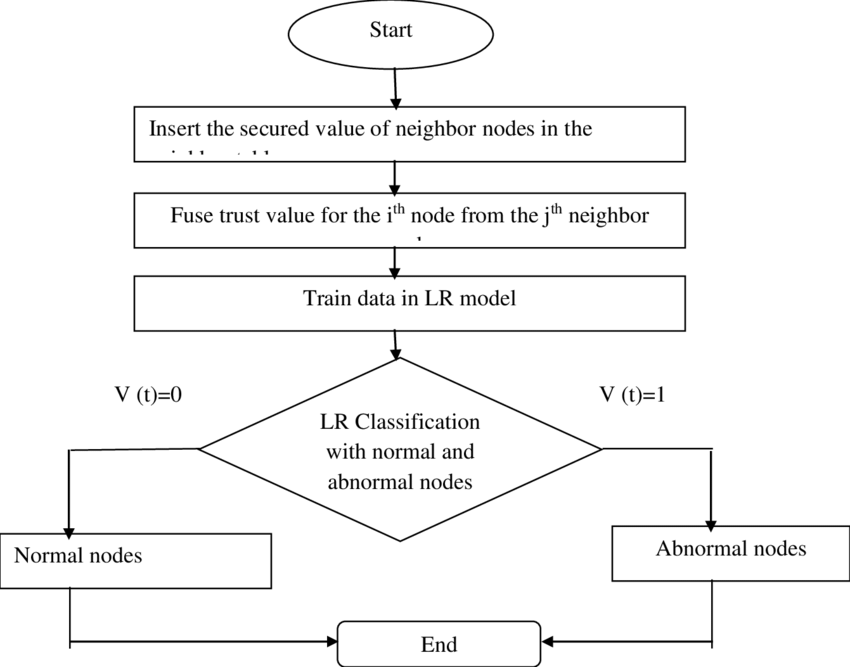
1. Data Pre-processing step

2. Fitting Logistic Regression to the Training set

3. Predicting the Test Result

4. Test Accuracy of the result

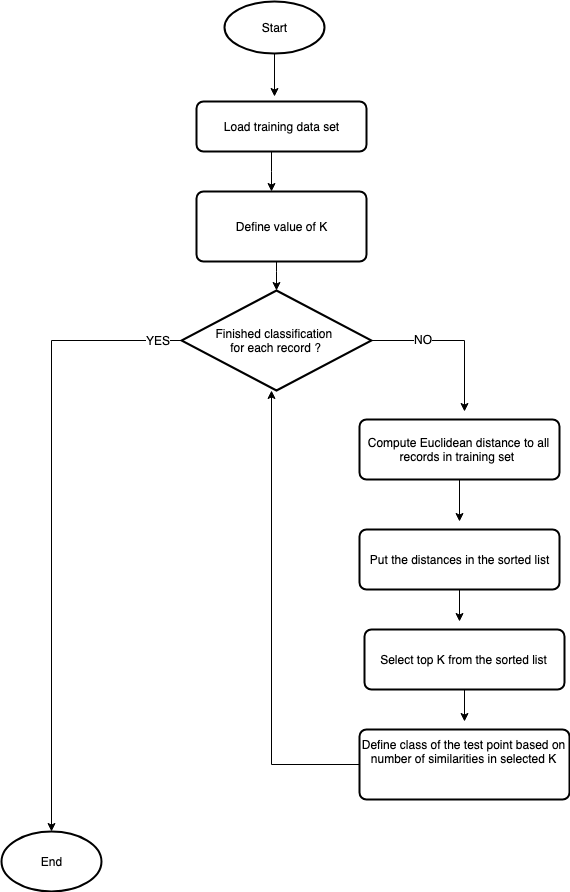
5. Visualizing the training set result



**KNN:**

Nearest neighbor method is a instance based classification technique that remembers all the instances. When the new instance is encountered, it uses previous instances as a model and compares it with the new instance.

* **Step-1:** Select the number K of the neighbors
* **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
* **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
* **Step-4:** Among these k neighbors, count the number of the data points in each category.
* **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum



**Naive bayes classifier:**

Classification based on Bayes Theory is known as Bayesian Classification. Naïve Bayes classifier is a statistical based classifier which is based on Bayes Theory. It assumes that attributes are statistically independent.

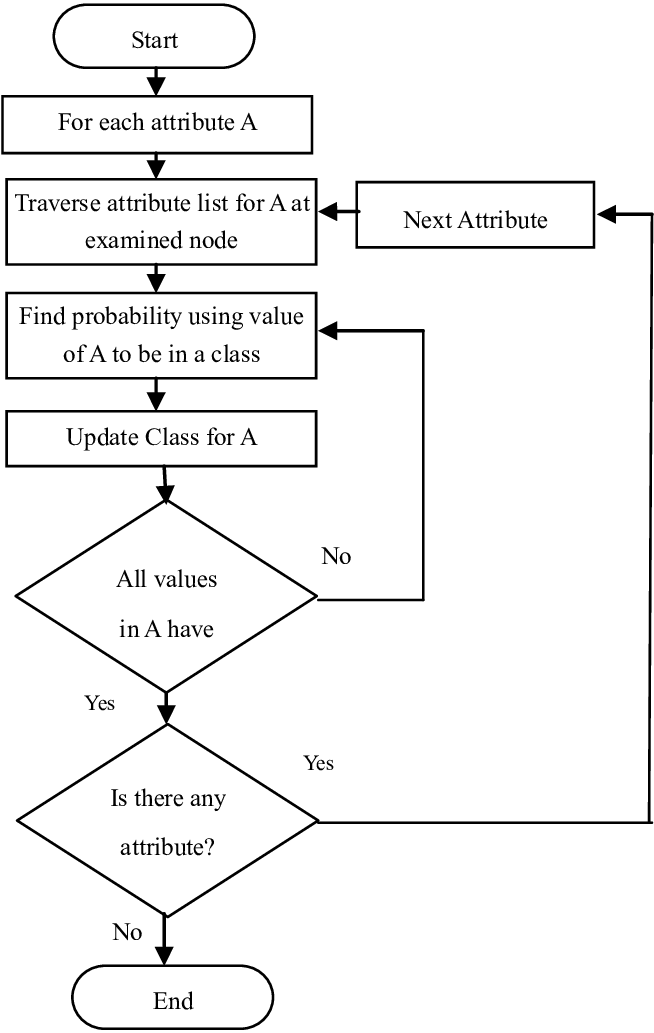
### Step 1: Separate By Class

### Step 2: Summarize Dataset

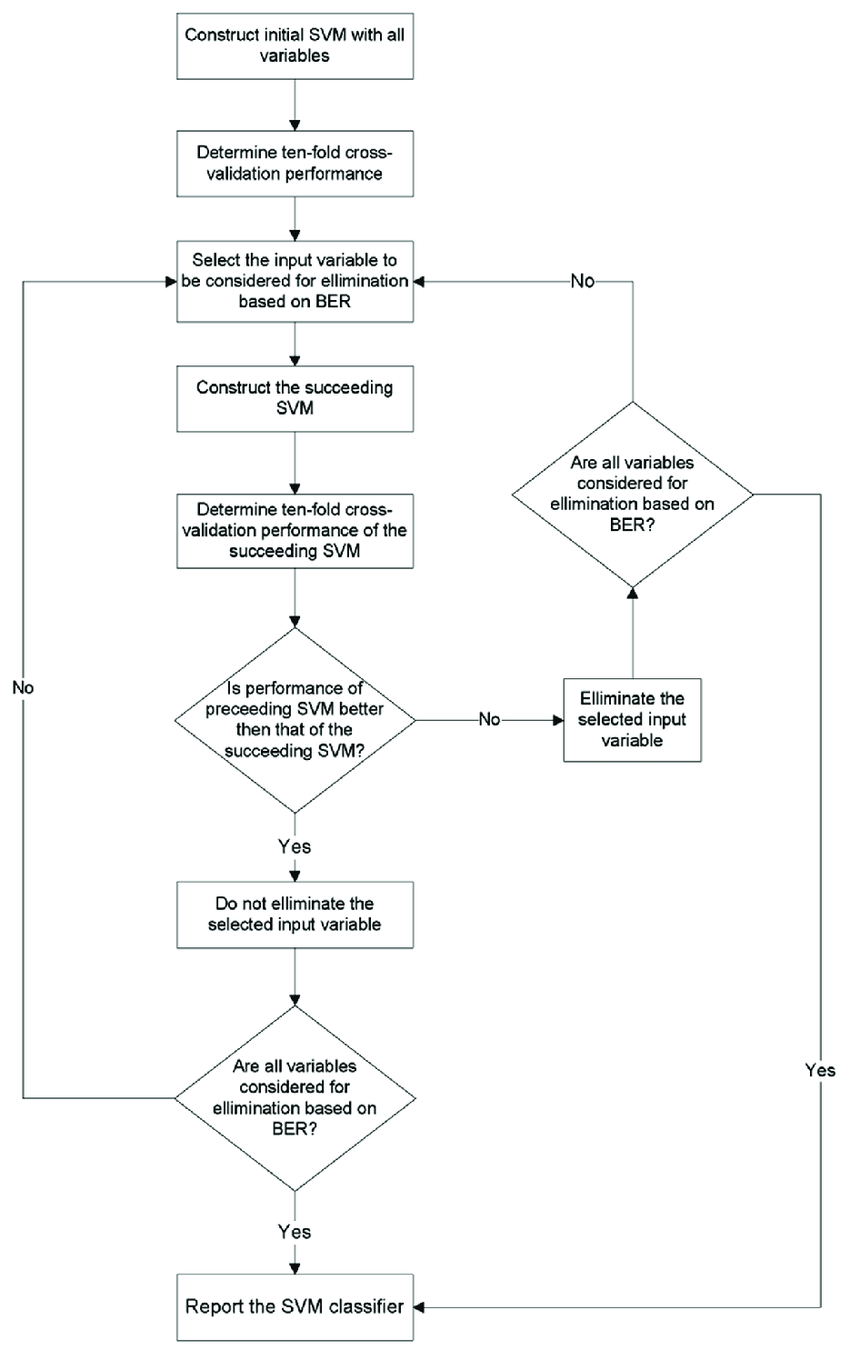
### Step 3: Summarize Data By Class

### Step 4: Gaussian Probability Density Function

### Step 5: Class Probabilities



**Support vector machine(SVM):**

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. 

**Decision tree:**

The decision tree is a structure that includes root node, branch and leaf node. Each internal node denotes a test on an attribute, each branch denotes the outcome of test and each leaf node holds the class label. The first node in the tree is the root node.

Step 1: Determine the Root of the Tree.

Step 2: Calculate Entropy for The Classes.

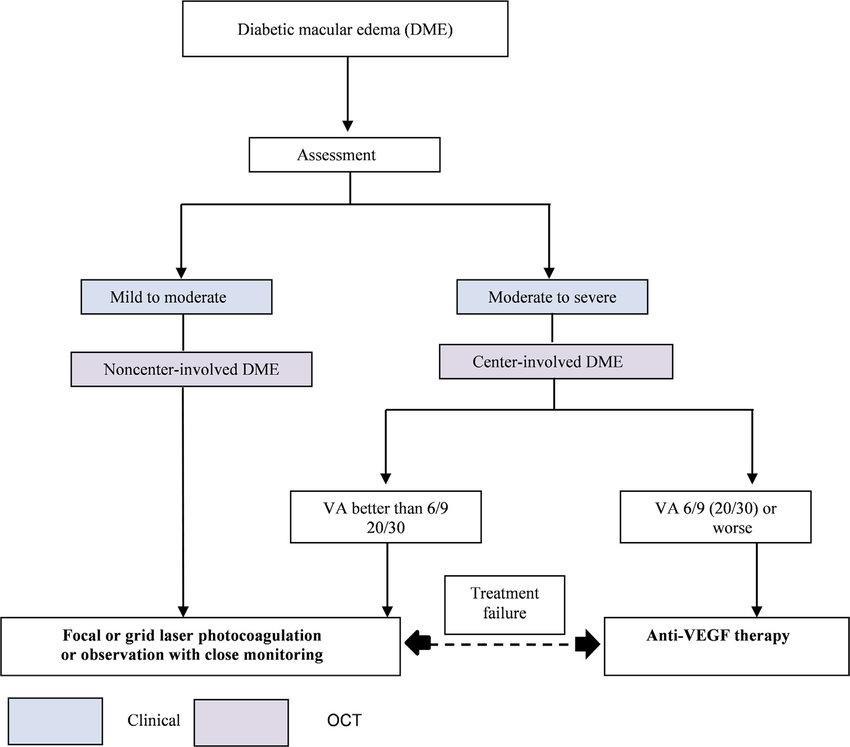
Step 3: Calculate Entropy After Split for Each Attribute.

Step 4: Calculate Information Gain for each split.

Step 5: Perform the Split.

Step 6: Perform Further Splits.

Step 7: Complete the Decision Tree.



**Random Forest:**

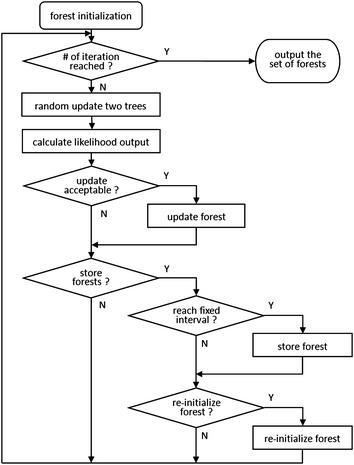
Random Forest is a supervised learning algorithm. Random forest can be used for both classification and regression problems, by using random forest regressor we can use random forest on regression problems. But we have used random forest on classification in this project so we will only consider the classification part.

Step 1 − First, start with the selection of random samples from a given dataset.

Step 2 − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

Step 3 − In this step, voting will be performed for every predicted result.

Step 4 − At last, select the most voted prediction result as the final prediction result.



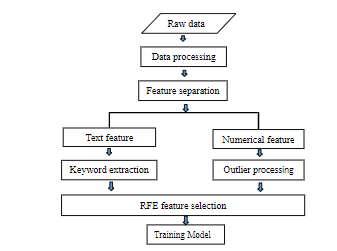
**XG BOOSTER:**

### Step 1: Load all the libraries

### Step 2 : Load the dataset

### Step 3: Data Cleaning & Feature Engineering

### Step 4: Tune and Run the model

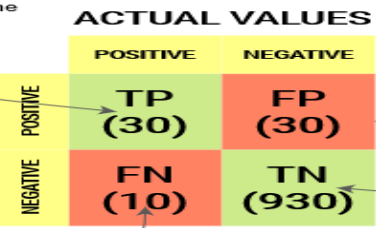


Formulas:

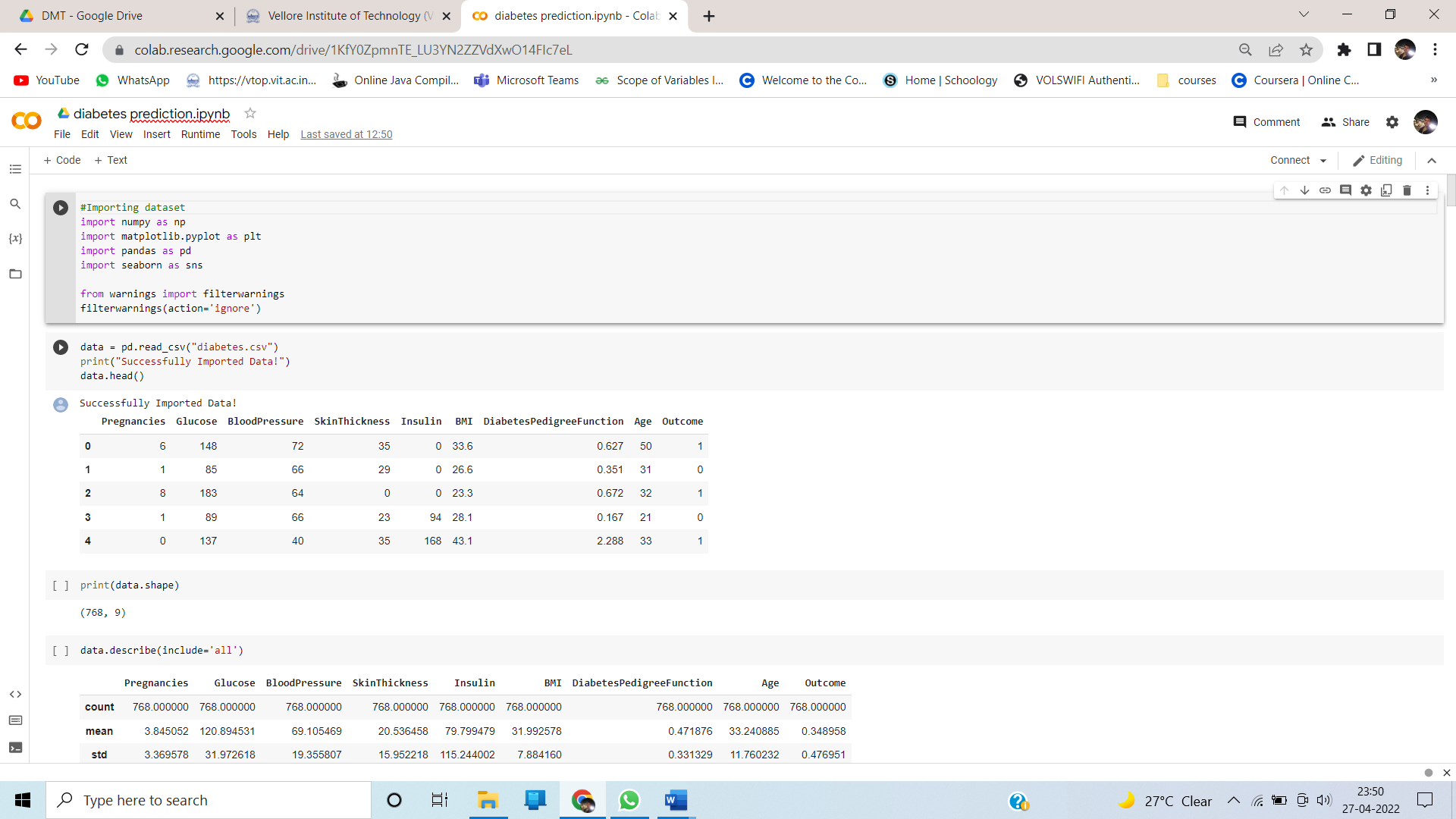
Accuracy = (TP + TN)/All

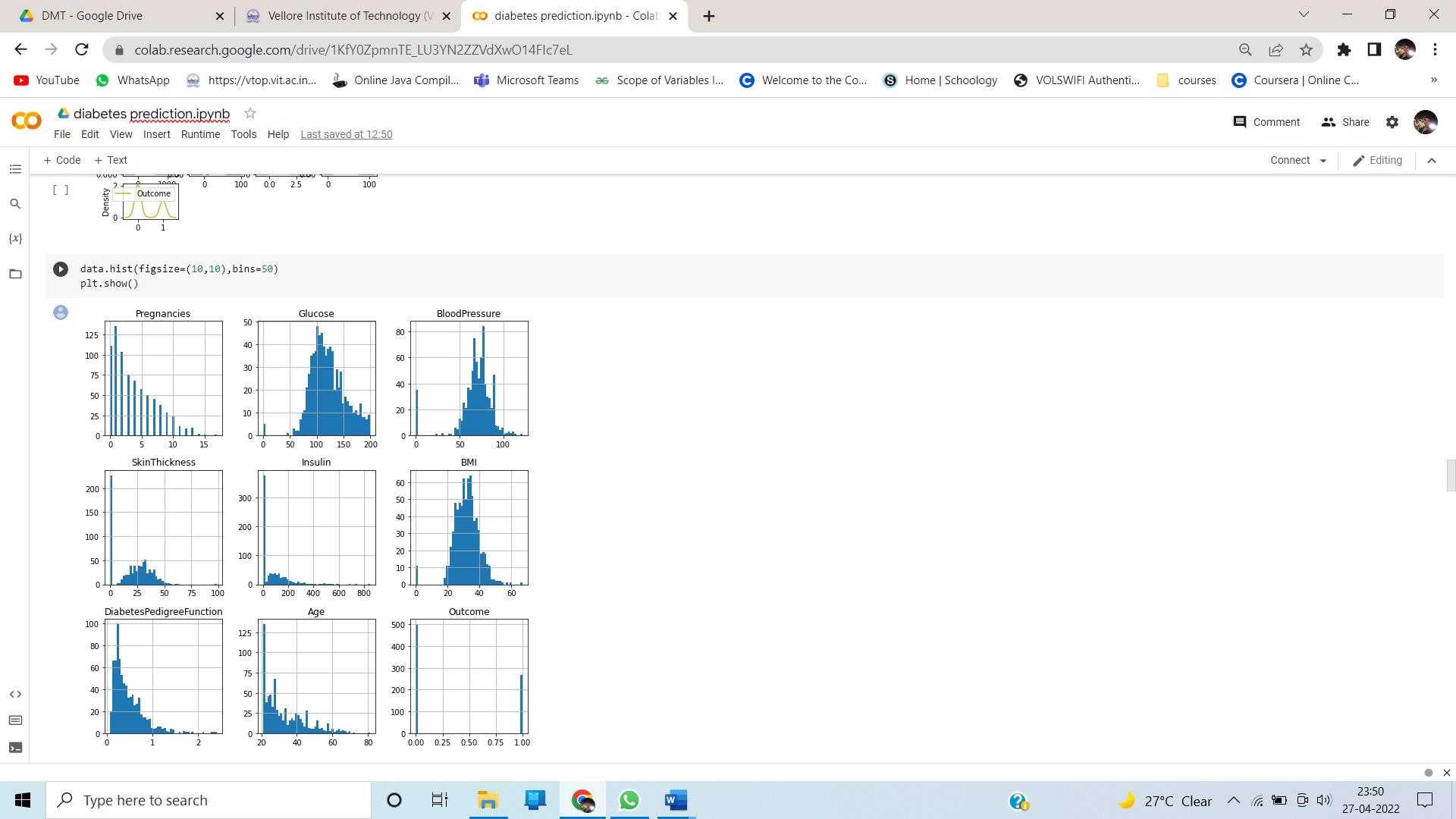


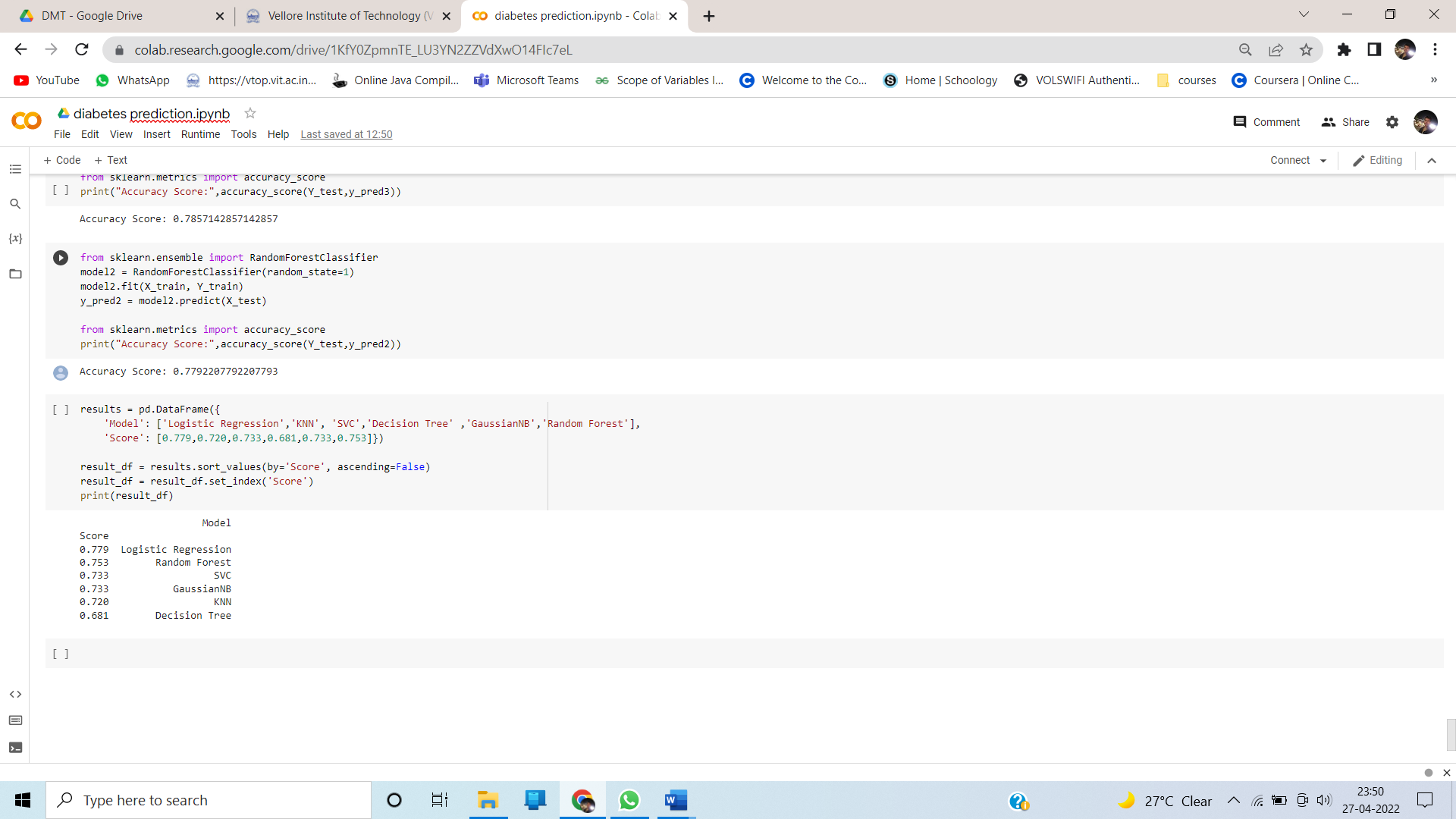
**Confusion matrix:**



**VI EXPERIMENTS RESULTS**

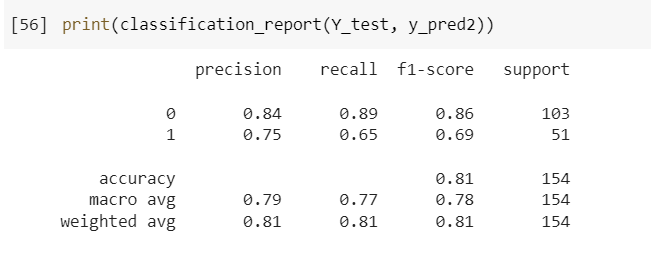




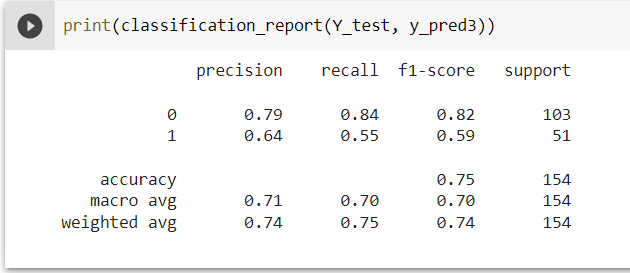


For each model:

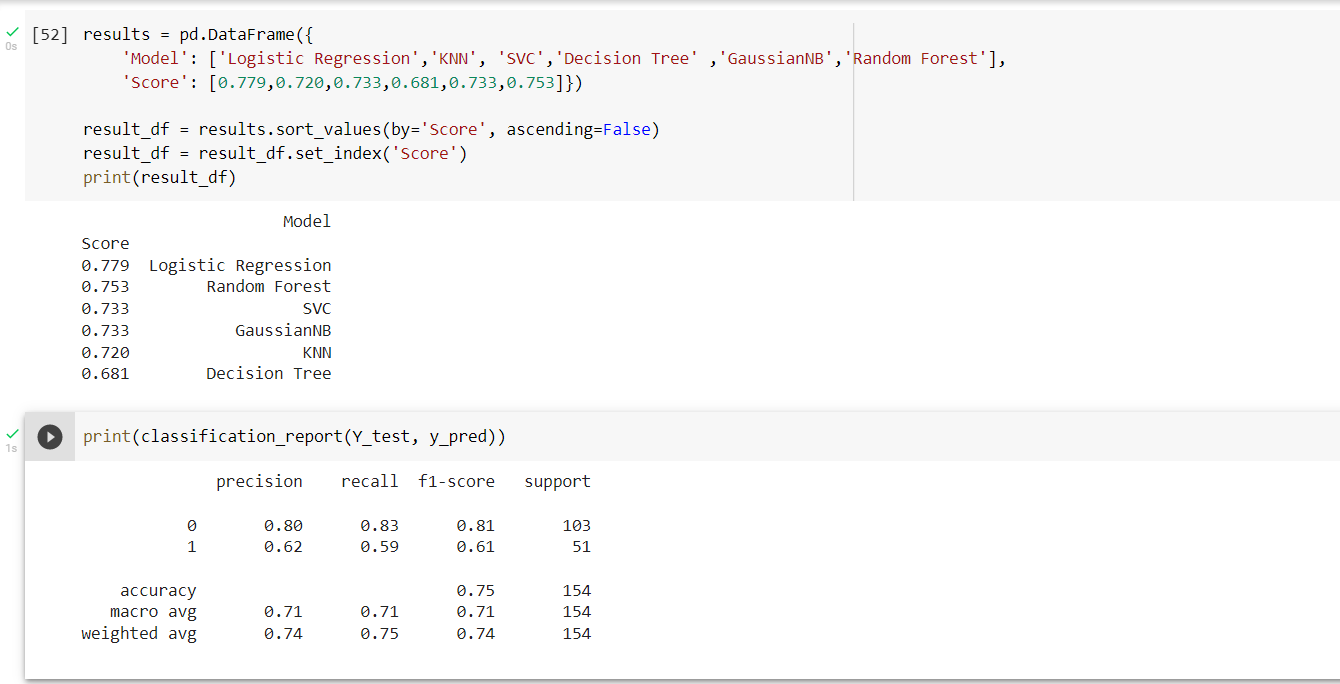
For RandomForestClassifier



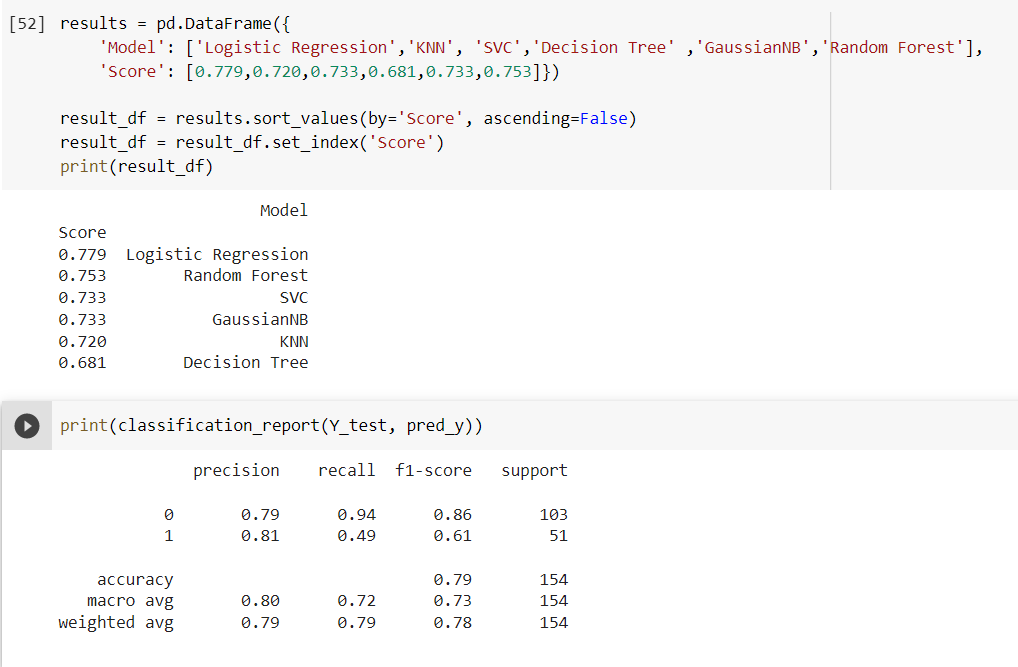
For GaussianNB:



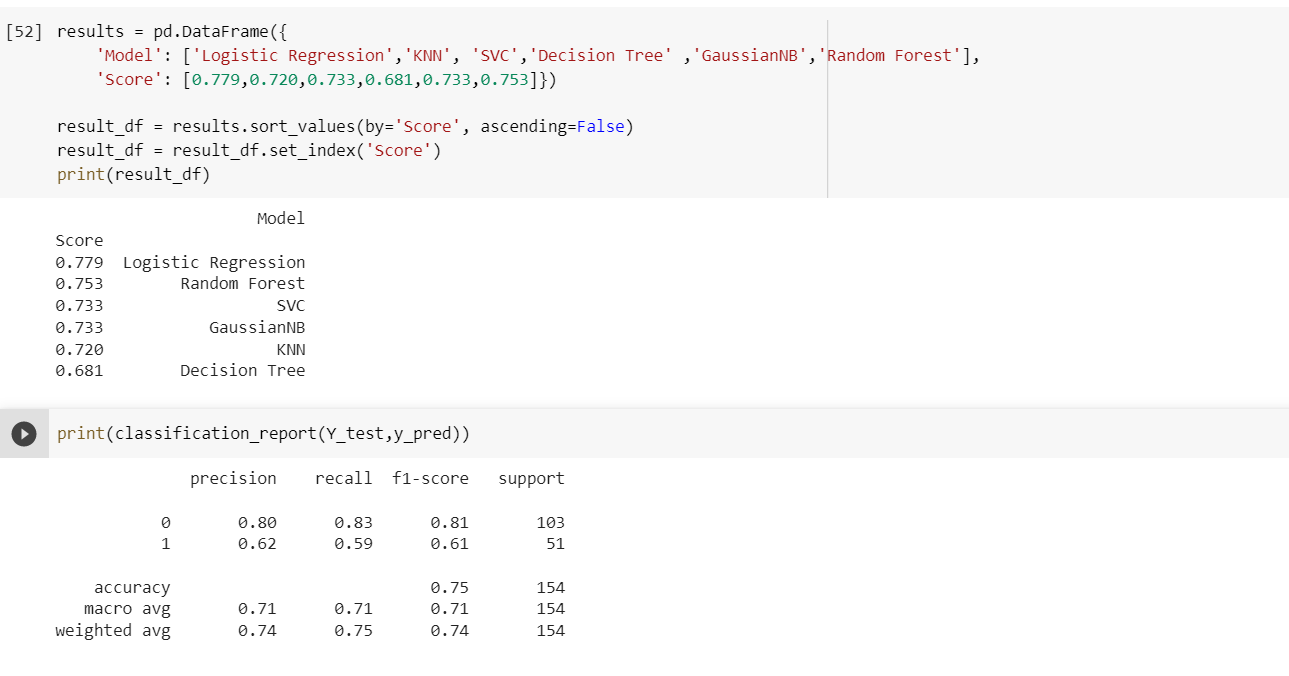
For DecisionTreeClassifier:



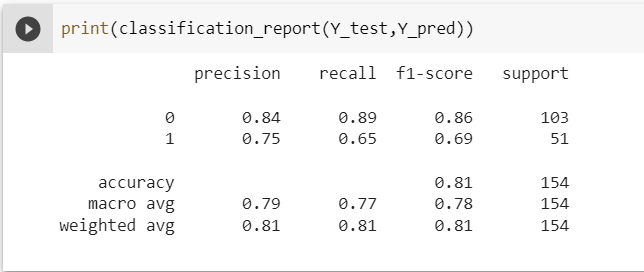
**For SVC**

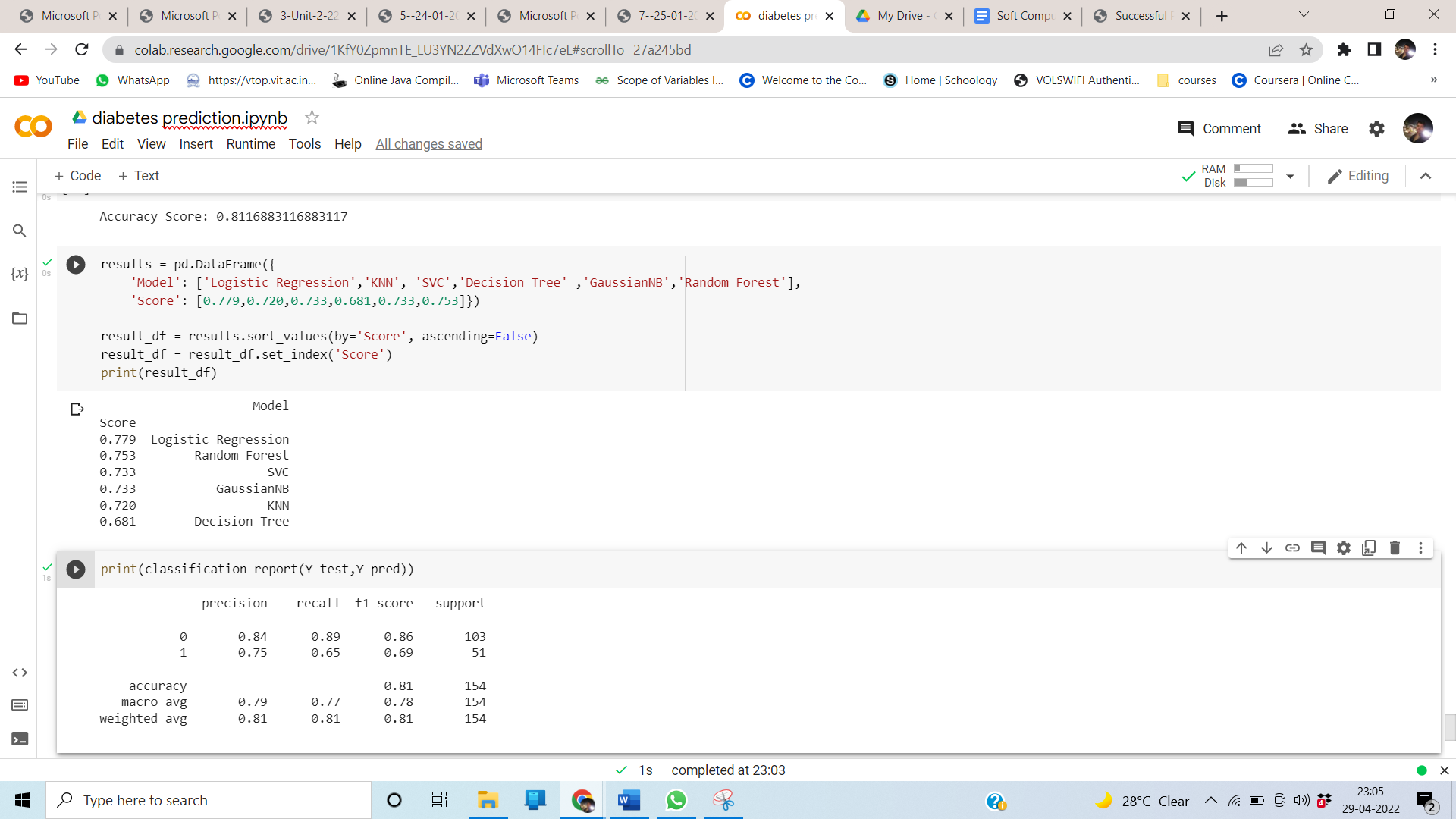


For KNeighborsClassifier



For logistic regression:





Output:

| **MODEL** | **ACCURACY** |
| --- | --- |
| Logistic Regression | 0.779 |
| Random Forest | 0.753 |
| Support vector machine | 0.733 |
| GaussianNB | 0.733 |
| KNN | 0.722 |
| Decision Tree | 0.681 |

**Confusion matrix:**

|  | Diabetic | NonDiabetic |
| --- | --- | --- |
| Diabetic | 93 | 12 |
| NonDiabetic | 22 | 27 |

**Googledrivefile:** [**https://drive.google.com/drive/folders/1rj1MarQB-6dkJX5XHdcRQbv3jV4oN1oT?usp=sharing**](https://drive.google.com/drive/folders/1rj1MarQB-6dkJX5XHdcRQbv3jV4oN1oT?usp=sharing)

**VIII CONCLUSION AND FUTURE WORK**

Diabetes is a most common disease in India. Early detection of Diabetes will increase the survival rate hence this prediction of Diabetes is intended to predict, whether the patient has Diabetes or not with the help of clinical dataset named Pima Indian dataset which will assist the diagnosis process. In this project we have discussed some of the effective techniques such as Logistic Regression, K nearest neighbors Classifier (KNN), Naive-Bayes Classifier, Support vector machine (SVM), Decision tree (DT), Random Forest (RF) and so on. And also we have used some Classifier evaluation metrics like Accuracy, Precision, F1 score, Recall, specificity and ROC curve. An important challenge in building the Diabetes Prediction is to build precise and computationally efficient classifiers for medical application and the performance of Random Forest shows high level compared with other classifier Algorithms. So we conclude that Logistic Regression is the best classifier Algorithm with Accuracy Score of 77.9%.

**IX REFERENCES**

1. Arianna Dagliati, PhD, Simone Marini, PhD, Lucia Sacchi, PhD, Giulia Cogni, MD, Marsida Teliti, MD, Valentina Tibollo, MS, Pasquale De Cata, MD, Luca Chiovato, PhD, and Riccardo Bellazzi, PhD (2019). Machine Learning Methods to Predict Diabetes Complications: Journal of diabetes science and technology.
2. Yazan Jian , Michel Pasquier, Assim Sagahyroon and Fadi Aloul(2021). A Machine Learning Approach to Predicting Diabetes Complications: Research gate.
3. Ioannis Kavakiotis, Olga Tsave , Athanasios Salifoglou , Nicos Maglaveras, Ioannis Vlahavas , Ioanna Chouvarda(2019). Machine Learning and Data Mining Methods in Diabetes Research,2019: National centre for biotechnology information.
4. Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju and Hua Tang(2019). Predicting Diabetes Mellitus With Machine Learning Techniques: Frontiers in genetics.
5. Tejas N. Joshi, Prof. Pramila M. Chawan(2020). Diabetes Prediction Using Machine Learning Techniques: International journal of engineering research and technology.
6. Hang Lai, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi and Xin Gao(2019). Predictive models for diabetes mellitus using machine learning techniques: Biomediacal center for endocrine disorders.
7. Mitushi Soni, Dr. Sunita Varma (2020).Diabetes Prediction using Machine Learning Techniques: International journal of engineering research and technology.

9.Chirath Hettiarachchi, Charith Chitraranjan (2019).A Machine Learning Approach to Predict Diabetes Using Short Recorded Photoplethysmography & Physiological Characteristics: ResearchGate

10. Farhana Bano, Munidhanalakshmi K and Dr R.MadanaMohana (2021). Predict Diabetes Mellitus Using Machine Learning Algorithms: Journal of physics.

11. S. Sivaranjani, S. Ananya, J. Aravinth and R. Karthika, "Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, pp. 141-146, doi: 10.1109/ICACCS51430.2021.9441935

12. M. T. Islam, M. Raihan, N. Aktar, M. S. Alam, R. R. Ema and T. Islam, "Diabetes Mellitus Prediction using Different Ensemble Machine Learning Approaches," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-7, doi: 10.1109/ICCCNT49239.2020.9225551.

13. A. Kaushik, A. Sehgal, S. Vora, V. Palan and S. Patil, "Presaging The Signs Of Diabetes Using Machine Learning Algorithms," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-5, doi: 10.1109/ICCCNT51525.2021.9579669.

14. S. K. Reddy, T. Krishnaveni, G. Nikitha and E. Vijaykanth, "Diabetes Prediction Using Different Machine Learning Algorithms," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 1261-1265, doi: 10.1109/ICIRCA51532.2021.9544593

15. P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 367-371, doi: 10.1109/ICCMC.2019.8819841.

16.H. A. Qureshi, A. Dewani, M. U. Arain, M. Sami, N. Dewani and M. Rasheed, "Diabetes Mellitus Type-2 Web-based Risk-PredictionSystem Using Predictive Analytics & Optimal Features–A Case Study of LUMHS, Jamshoro," 2021 International Conference on Artificial Intelligence (ICAI), 2021, pp. 9-15, doi: 10.1109/ICAI52203.2021.9445244.

17.D. R. Nemade and R. K. Gupta, "Diabetes Prediction using BPSO and Decision Tree Classifier," 2nd International Conference on Data, Engineering and Applications (IDEA), 2020, pp. 1-5, doi: 10.1109/IDEA49133.2020.9170744.

18.R. Mirshahvalad and N. A. Zanjani, "Diabetes prediction using ensemble perceptron algorithm," 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN), 2017, pp. 190-194, doi: 10.1109/CICN.2017.8319383.

19.Yuvaraj, N.; SriPreethaa, K.R. “Machine Learning Methods to Predict Diabetes Complications

Cluster Computing, 16 January 2019, 22 Language: English. Springer New York LLC DOI: 10.1007/s10586-017-1532-x

20. Xie, Jinyu; Wang, Qian.” IEEE Transactions on Biomedical Engineering. Nov2020, Vol. 67 Issue 11, p3101-3124. 24p. DOI: 10.1109/TBME.2020.2975959