Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.776923

# A NOVEL APPROACH FOR DIABETESDISEASES PREDICTION USING DIFFERENT MACHINE LEARNIN G CLASSIFIERS

**Mr. Rahul K. Sharma[1],Dr. Paresh Tanna[2]**
[1]PhD Scholar , Department Of Computer Engineering, School of Engineering,
R.K.University, Rajkot, Gujarat, India.
[1]rahul_sharma1818@yahoo.com, rsharma261@rku.ac.in
[2]Professor, Department Of Computer Engineering, , School of Engineering, R.K.University,
Rajkot, Gujarat, India, [2]paresh.tanna@rku.ac.in

## ABSTRACT

Diabetes is an ineradicable disease that can be found in most of the people nowadays. Due to hectic schedulespeople are unable to focus on their health. The food we are consuming is fragmented into glucose; thesefragments will be delivered into blood. The pancreas releases a hormone named as insulin when the glucoselevels are high. This insulin plays a vital role in transporting the glucose to cells that can be used as energy. Tomaintain a sustainable life detection of diabetes in early stage will be beneficial. Machine learning algorithmswill be a productive approach as it will be trained & test with vast data and it enhances itself with upcomingfuture predictions. In this article, various algorithms like KNN , Naive Bayes are used, Decision Treeand trained with our collected dataset. Among these three algorithms it was observed that Decision Tree producedaccurateresults.

**Keywords:** Machine Learning, K-NN, Naïve Bayes, Decision Tree.

## I. INTRODUCTION

Today diabetes is a very common disease. Earlier diabetes was observed in adults and old age. But these daysdiabetes was reported in teen age also. These are some aspects for developing diabetes such as family history,age,foodhabits,andhighbloodpressure, obese.

In general, there are two types of diabetes.In type 1 diabetes production of insulin is less due to this insulinproducing cells in pancreas will be affected by immune system. Types 1 diabetes develops due to family history.In the second type of diabetes the body is found to act as a resistant to insulin therefore this results in need ofless insulin than required this type of irregularity is found due to procrastination of exercising, unwholesomefood and obesity.Type 2 diabetes develops due to obese, family history and inactive lifestyle. It was found thatthere are multiple risks involved if the diabetes is not controlled / detected in an early stage. During a researchresult from reputed article, it was found that younger aged people are suffering from type-1 diabetes, womenare giving birth to a child that is weighted over 9 pounds, due to unhealthy diet, people are experiencingoverweight and obesity, people are also suffering from polycystic overy syndrome etc., all these side effects anddiseasesare founddueto unhealthydietandunwholesomefood.

During development stage of our model various literature review papers were researched and found thattraditional approacheswere already implemented. So, to implement a model that is unique and futuristicmodernised dataset is considered i.e., is a digital medical dataset that consists readings of people eating junkfood, average people taking for exercising and the day-to-day lifestyle. This is totally different from traditionalapproach where a fixed set of values are taken using prediction this results in inaccurate results. Therefore,efficient results were not found in their approach. To overcome this problem including digital dataset, machinelearning algorithms like KNN , Decision Trees & Naïve Bayes are applied. And out of these algorithmsDecision Tree performedwellwithourdigitaldatasetandtheresultswerefoundoutbebetterandefficient.

## II. LITERATURESURVEY

KM Jyothi Rani Proposed a system for predicting diabetes based on Machine learning algorithms. In this paperthey have used the dataset which contains 9 features and 2000 entries out of which outcome describes 0 meansno diabetes, 1 means diabetes. They have used 5 machine learning algorithms in this paper out of these 5algorithmsDecisionTreealgorithm providesgood accuracy.

Raja Krishnamoorthi proposed a diabetes healthcare disease prediction framework using machine learningtechniques. The dataset contains 768 rows and 9 columns and 90% of the data is used for training and 10%used for the testing purpose and they performed hyper-parameter tuning to evaluate the Machine Learningmodelsand used to increase theaccuracy.Out of 5algorithmsbestone isidentified and hyper parametertuninghasbeen appliedtoprovide betteraccuracyasaresultof 86%

Desmond Bala Bisandu proposed a system for diabetes prediction using data mining techniques. In this paperthere are 5 parameters based on which diabetes is predicted and data is pre-processed to remove noise and toremove null values and classification and prediction was done using Naive Bayes Classifier and efficiency wasaround95%

B. Suvarnamukhi proposed a big data processing system which uses machine learning techniques for predictingdiabetes. Due to rapid increase in technology the data is stored in the form of electronic records (EHR) and thisdata is processed using big data and for prediction of diabetes ELM is used and compared with other algorithmsanddiabeteswhichispredictedof3types

Mitush Soni proposed machine learning algorithms for providing better accuracy in diabetes prediction. In thispaperthedatasetcontains500negativeoutcomesmeansnodiabetesand268positiveoutcomesmeansdiabetes and For Predicting accurately they have used 6 machine learning algorithms and among these 6algorithmsKNN algorithmpredictswith77%accuracy

N. Sneha1 and Tarun Gangil has designed a model for Analysis of diabetes mellitus for early prediction usingoptimal features selection The dataset consists of 2500 entries and 15 attributes and 768 items used for testingandtheyhaveused 5algorithmsout ofwhichsupport vectormachineprovides77% accuracy.

Abdullah A. Aljumah and M.G Ahmad proposed a data mining application to predict diabetes in young and oldpatients using regression-based mining technique. The dataset is used is a NCD risk factor report from Ministryofhealthreport,SaudiArabiaandusingdataminingtanalysisondatasettheyhavepredictedtheeffectivenessinyoungandoldgroupfor differenttreatments.

Salliah Shafia and Prof. Gufran Ahmad Ansari designed a model for Early Prediction of Diabetes Disease &Classification of Algorithms Using Machine Learning Approach.this research uses the WEKA tools to predictdiabetes in patients from Pima India Diabetes Data Set consists of 7 attributes and 767 entries and in this paper,theyhaveused3classificationalgorithmsoutof whichNaïvebayesprovides74% accuracy.
R M Anjana prepared a report on Prevalence of diabetes and prediabetes (impaired fasting glucose and/orimpaired glucose tolerance) in urban and rural India. In this report they conducted a survey on urban and ruralparts of india to estimate prevalence of diabetes and prediabetes and in the report, Chandigarh was found to behavehighestdiabetespercentage.

## III. PROPOSEDSYSTEM

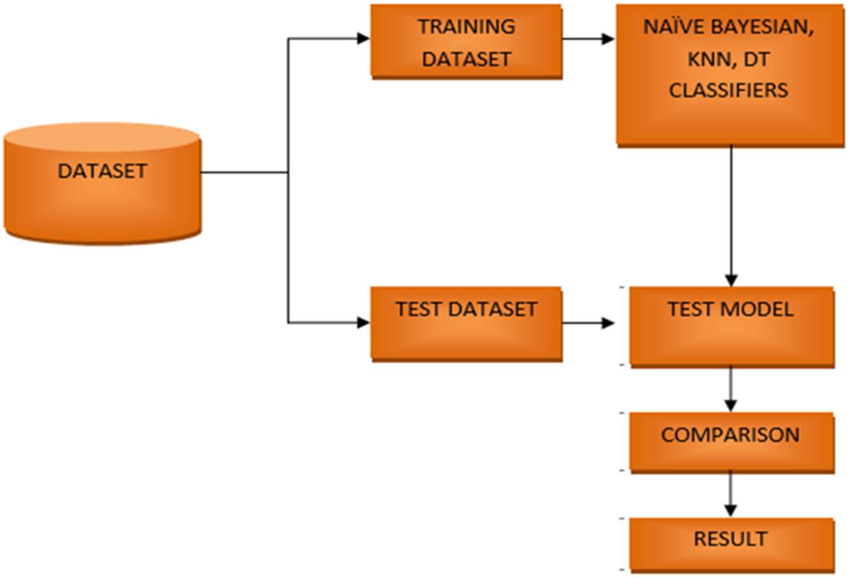


Fig1:Overviewoftheprocess

Thismodelhelpstopredictdiabeteswithbetteraccuracy.Weexperimentedwithdifferentclassificationalgorithms.

### 1. DatasetDescription-

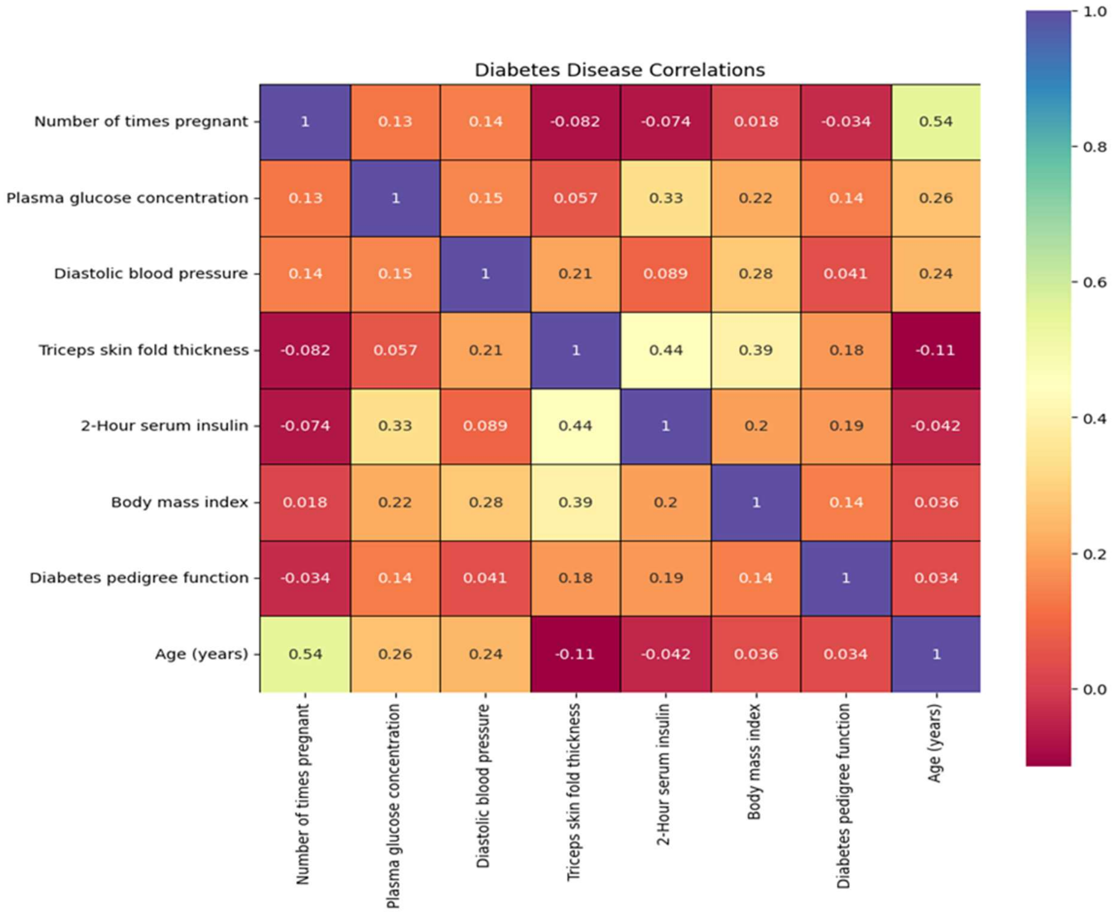
ThedataisgatheredfromKagglewebsitewhichisnamedasDiabetesHealthIndicators Dataset.It Containsof253679 entries ofdataandeach recordconsistsof22columns.

**Table1:DatasetDescription**

| SNo. | Attributes |
|---|---|
| 1 | Diabetes_012 |
| 2 | HighBp |
| 3 | HighChol |
| 4 | Cholcheck |
| 5 | BMI(Bodymassindex) |
| 6 | Smoker |
| 7 | Stroke |
| 8 | HeartDiseaseorAttack |
| 9 | PhysActivity |

| | |
|---|---|
| 10 | fruits |
| 11 | Veggies |
| 12 | HvyAlcoholConsume |
| 13 | AnyHealthCare |
| 14 | NoDocbcCost |
| 15 | GenHealth |
| 16 | MenHealth |
| 17 | PhysHealth |
| 18 | DiffWalk |
| 19 | Sex |
| 20 | Age |
| 21 | Education |
| 22 | Income |

Correlation Matrix: It is used to demonstrate the relation between the attributes and displays matrix as aoutput



2. Data Pre-processing -This phase model handles inconsistent data, missing values and other impurities thatcouldcauseeffectivenessofdata. DataPre-processingisdone toimprovethequalityandtoobtainaccurateresults.

A. Missing values removal - Instances with zero as worth are removed. Through eliminating irrelevantinstances,we make featuresubset and this process is called features subset selection,which help to workfaster.

B. Splitting of data - After removal of irrelevant instances, data is normalized in training and testing themodel. When data is splitted then we train the efficient algorithm on the training data set and keep test data setaside.

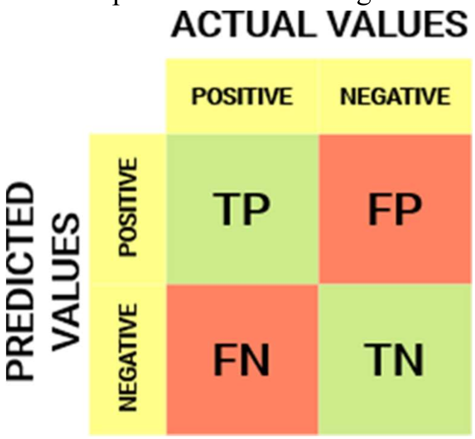
C. Apply Machine Learning – After pre-processing of the data we will split the data into training and testingparts, 80% of the data for training part and 20% of the data for the testing part and now we will train the datausing machine learning classification algorithms.These algorithms include KNN , Decision Trees,Naïve Bayes. We will train the data using these algorithms and after training the data we will measure theaccuracyusingtestdata.

3. Evaluation-

InThisstepweevaluatethepredictionresultsusingdifferentperformancemetricssuchasconfusionmatrix,accuracy,precision,recall andf1score.

ConfusionMatrix–

Confusionmatrixisusedtodescribetheperformanceofthe Algorithmsandgivesmatrixas anoutput.



| | | ACTUAL VALUES | |
|---|---|---|---|
| | | POSITIVE | NEGATIVE |
| PREDICTED VALUES | POSITIVE | TP | FP |
| | NEGATIVE | FN | TN |

Where, TP – True PositiveFP– FalsePositive

FN – False NegativeTN–TrueNegative

Accuracy-

Itistheratioofnumberofcorrectedpredictionstobytotalnumberofobservations.Themodelisbest if ithashighaccuracy

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$$

**Precision**–Itistheratioof correctlypredictedpositiveresultsbythetotalpredictivepositiveresults

$$\text{Precision}=\frac{TP}{TP+FP}$$

**Recall**-–Itistheratioofcorrectlypredictedpositiveresultsbythetotalresultsinactualclass

$$\text{Recall}=\frac{TP}{TP+FN}$$

**F1Score**–It isweightedaverageofPrecisionand Recall.

$$F1=2*\frac{\text{Precision*Recall}}{\text{Precision+Recall}}$$

### 4. Savethemodel-

Inthisstagewewillcomparetheaccuraciesofeachmodelandwewillsavethemodelwithhighestaccuracy andwe will usethatmodelfor predicting thedisease.

## IV. EXPERIMENTANALYSIS

Confusionmatrixused todescribethe performanceofthealgorithmsandhere wewillseethe confusionmatrixfor3algorithms

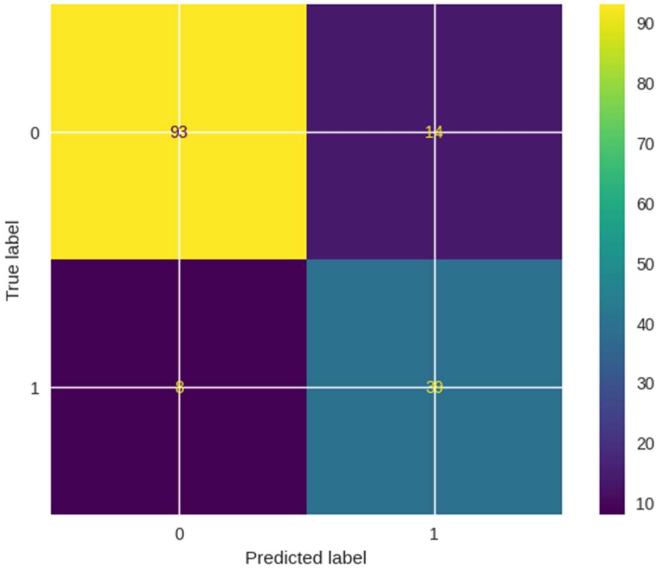


Fig2:Confusionmatrixfor NaiveBayesclassification

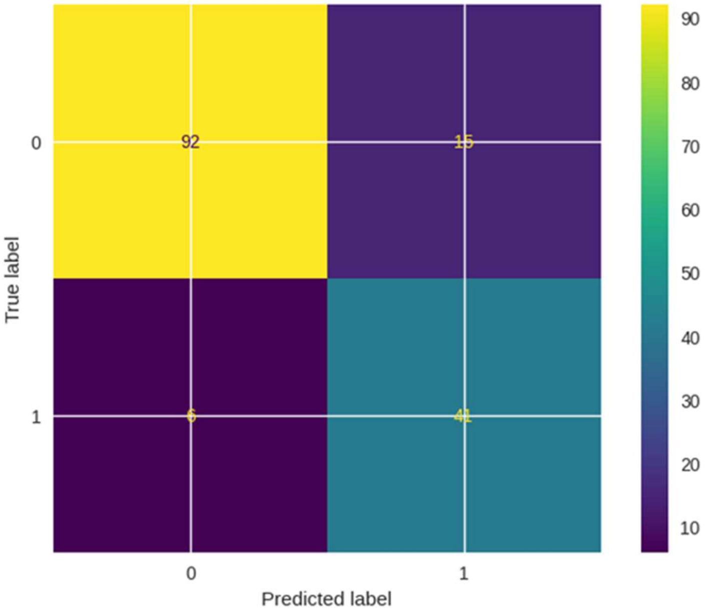


Fig3:ConfusionmatrixforDecisionTreeclassification

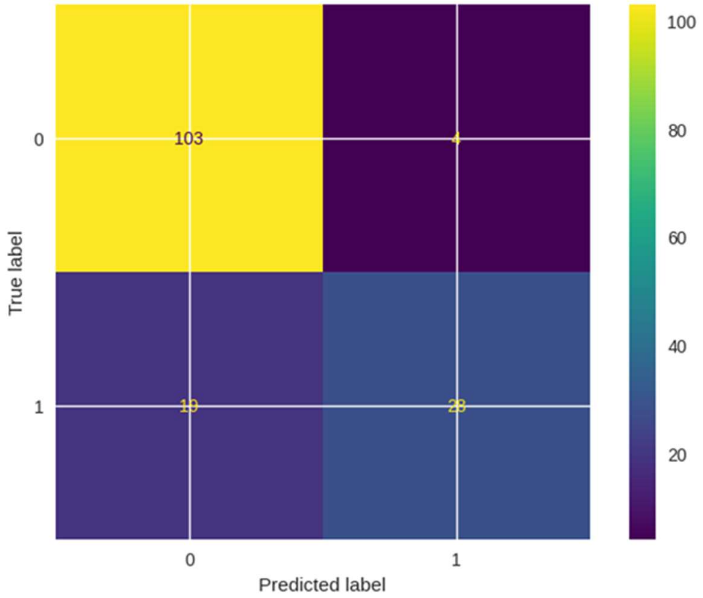


Fig4:ConfusionmatrixforKNN Classification

**ComparisonofTable–**

In this research we have used 3 algorithms and the above table describes the performance metrics of thesealgorithms and out of those 3 algorithms Decision Tree gives better results in terms of accuracy, F1 Score,RecallScoreandprecision Score.

## V. CONCLUSION

We have successfully built a model where it will predict whether a patient has diabetes or not using 3 machinelearning algorithms which are Decision Tree classifier, Naïve Bayes and KNN. Out of these 3algorithmsDecision Tree gives86%accuracy.

## VI. FUTUREWORK

The above model is used to predict whether a person has diabetes or not using their health records and infuture we can build a perfect model using deep learning techniques and providing best accuracy and further wecan also build a Web application using flask so that users can give the parameters and based on those attributesthemodel willpredict.

## VII. REFERENCES

[1] KMJyotiRani,"DiabetesPredictionUsingMachineLearning,"InternationalJournalofScientificResearchinComputerScienceEngineeringandInformationTechnology,volume.6,pp.294-305,2022.

[2] Raja Krishnamoorthi, Shubham Joshi, and Hatim Z. Almarzouki, “A Novel Diabetes Healthcare DiseasePrediction Framework using Machine Learning Techniques,” Journal of Healthcare Engineering, pp. 1-102022.

[3] DesmondBalaBisandu,GodwinThomas"DiabetesPredictionusingDataminingTechniqu

es,"InternationaljournalofresearchandInnovationinAppliedSciences,volume4,pp.103-111,2022.

[4] B.Suvarnamukhi , M. Seshashayee, "Big Data Processing System for Diabetes Prediction using MachineLearning Techniques," in International Journal of Innovative Technology and Exploring Engineering,volume8,pp.4478–4483,2022.

[5] MitushiSoni,Dr.SunitaVarma,"DiabetesPredictionusingMachineLearningTechniques" ,InternationalJournalofEngineeringResearch& Technology,Volume9,pp.921-925,2022.

[6] N.Sneha,Tarun Gangil, "Analysisof diabetesmeelitusforearlypredictionusing optimalfeaturesselection",Journal of Bigdata,pp.1-19, 2022.

[7] Abdullah A. Aljumah, M.G Ahmad, M.K Siddiqui, Application of data mining: Diabetes health care inyoungandold patients,JournalofKing Sauduniversity,volume25,pp.127-136,2022.

[8] SalliahShafi, Prof. GufranAhmad Ansari, "EarlyPrediction ofDiabetesDisease& Classification ofAlgorithms Using Machine Learning Approach", International Conference on Smart Data Intelligence,2021.

[9] R M Anjana, R pradeep , "Prevelance of Diabetes and prediabetes (impaired fasting and/or impairedglucose tolerance) in urban and rural india, Indian Council of Medical Research, volume 54, pp 3022-3027,2022.