Prediction Of Diabetes Mellitus Using Measure Of Insulin Resistance: A Combined Classifier Approach

J. Omana[1], Dr.M. Moorthi[2]

[1]Assistant Professor, Prathyusha Engineering College, Part-Time Scholar, Anna University
[2]Professor, Saveetha Engineering College
[1]omanajayakodi@gmail.com, [2]moorthidmp@gmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: Diabetes is a cluster of diseases that are categorized by hyper glycaemia which is the result of defects in insulin secretion by the pancreas, the action of the insulin over the carbohydrates that we consume, or both of the conditions. Diabetes holding hyper glycaemia is resulting with failure of organs at a longer term rate such as eyes, kidney and heart. Since the disease is creating havoc in the human race it's important to identify the cause with great accuracy and precision. Using data mining and machine learning algorithms we try to find the accuracy of classifying the same. The diabetes dataset is a binary classification problem and its main objective is to analyze if a patient is affected by the disease or not. We concentrate here on the various classifiers and their accuracy results in identifying the presence and absence of diabetes. The study is conducted on classifiers like Decision Tree, SVM, Logistic Regression, Linear Regression, K- Nearest Neighbor, Random Forest and Naïve Bayes algorithms. An in depth analysis is made on the contribution of the attributes in the classification problem.

Keywords: Machine learning, feature extraction, support vector machine, Decision tree, Linear regression heat map, logistic regression.

1. Introduction

Diabetes is generally orchestrated into the going with four classes: Type 1 diabetes which is the delayed consequence of the β-cell destruction, and the result is that preeminent insulin inadequacy is caused. Type 2 diabetes is caused as a result of insulin release blemish by the pancreas. The third one is the Gestational diabetes mellitus such a diabetes is regularly investigated in the second or third trimester of pregnancy. Various establishments for diabetes are, neonatal diabetes and improvement starting diabetes of the young, sicknesses of the exocrine pancreas, and drug or manufactured incited diabetes, for instance, in the therapy of COVID19. The out of date models of type 2 diabetes happening simply in adults and type 1 diabetes simply in children are not, now accurate, as the two infections occur in the two accomplices. A captivating late end is that there is an association between Covid-19 and diabetes and moreover it is bidirectional. The chief case is that diabetes is related to an improved threat for genuine Covid-19. Next is that the fresh start diabetes and extraordinary metabolic disarrays of past diabetes, including diabetic ketoacidosis and hyperosmolarity for which remarkably high segments of insulin are supported, have been found in patients with Covid-19. 19–related diabetes. In this paper we will analyze on the various classifiers that can be applied on the prima Indian diabetic dataset and how absolutely it portrays the same**.**

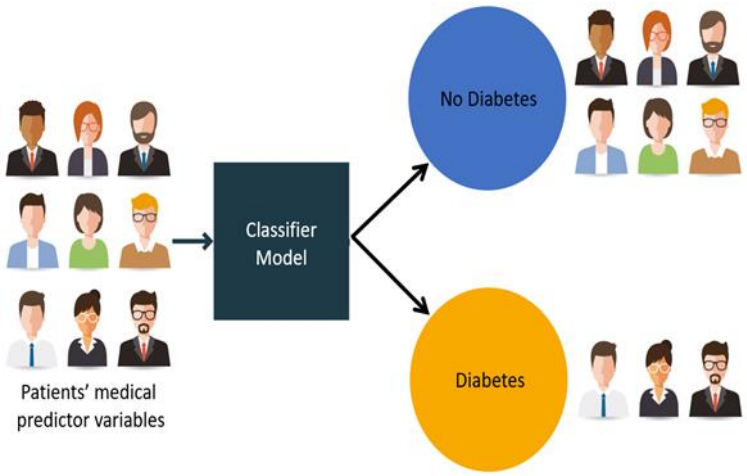


Fig 1 Classifier model of Prediction

2 literature survey

For Type 2 Diabetes Mellitus (T2DM) it is important to guess the long period difficulties risk for medical assessment process. Guidelines for T2DM patient to stop from Cardiovascular Disease (CVD) possibility by inducting proper treatment. The investigation of use of MLT'S in an improvement direction of revised replicas to envisage T2DM patient form CVD occurrence is the main goal of this study. By applying unusual cooperative schemes, the essential task of managing unstable environment of open dataset is addressed. By following a sub-sampling technique HWNNs and SOMs create the primary schemes for constructing troupes. By applying various models, the results of primary schemes are pooled and evaluated. For progress and estimating purpose the five years recorded data of more than 550 patients with T2DM is used. The best results are achieved in terms of AUC by considering primary schemes outputs which based on HWNNs and SOMs. BLR model is used to validate the requirements to apply sophisticated methods in satisfactory way to provide unfailing CVD possibility marks. The future models are higher than the BLR model [1]. In medical research field prevalent and significant approach is machine learning. Using medical registers of cardio respiratory fitness for expecting diabetes the performance of various machine learning methods like Logistic Model Tree, Decision tree and Random Forests are examined in this study. To expose possible clairvoyants of diabetes author also applied some other technique. This study used 5 years follow –up of more than 32,000 patients who are not having heart problems and who undertook the treadmill stress test at Henry Ford Health Systems. At the end of the 5th year 5,000 patients had diabetes. The collection of data had 62 characteristics which divided into 4 types. Those are history of medication used, history of disease, and characteristics of demographic, signs of stress test. Using 13 of those characteristics author established an ensembling-based analytical method. The established model had negative outcome which was controlled by SMOTE (Synthetic Minority Oversampling Technique). The analytical model complete performance was enhanced by the decision trees of Ensemble machine learning approach and accomplished great prediction accurateness as 0.92 AUC. For guessing diabetes by using cardio respiratory fitness data, the capacity of ensembling and SMOTE approaches was shown in this study. Even though a large number of researches has collected to design models to guess the diabetes, the machine learning approach has gain continues attention of healthcare society. In the prediction of incident diabetes, the SMOTE approach showed the major development [2][17].

Diabetes is a metabolic disorder. There were Type1 diabetes and Type2 diabetes. In Type1 diabetes body's immune system kills level of insulin generating beta cells. Where Type2 diabetic is a condition where your blood sugar levels or glucose levels are too high. This occurs when your body can not use glucose you get from food as a consequence glucose forms in the body. Insulin cannot respond to body. Finally, the body does not generate insulin completely. Because of this the body undergone to different complications. This can damage the arteries to make 2-3 times likely a heart attack, stroke or develop vascular dementia, blindness, kidney failure, nerve damage. Providing some valuable information to the patients for predicting some significant complications in Type2 diabetes is the aim of this study. A set of known algorithms have been accomplished and tested over 1,000 patients' dataset to know the best algorithm to predict the complications in Type2 diabetes. The Random Forest algorithm and the Naïve Bays classifier algorithms are resulted as the best and worst algorithms respectively. Maximum previous research done to fine the finest algorithm to predict the complications of Type2 diabetic and few attempts are made by increasing the dataset to reduce the error rate of the prediction whereas the future techniques attempts to follow mutual aspects. For supporting physicians to take decisions based on the health information system this study presents a method with the observation of complications of diabetes patients, family history of patients and BMI index and so on [3].

One of the most health problem around the world is Diabetes Mellitus (DM) which causes domestic financial problem and short life period. Because of this there is a necessity to prevent and detect diabetes. Should improve the diabetes treatment and control. The performance of assessment of recent glucose prediction methods are the aim of this study. To implement data analytics in a wireless body area network system, based on the evaluation a best fit method is suggested. Recommended glucose prediction algorithm is established on ARX model which considered BP, LDL, CGM data, TC and HDL as inputs. To estimate the performance of the recommended algorithm over MAE, R2, RMSE above 440 diabetic patients' dataset was used. The tentative results determine that the estimated accurateness of glucose can be improved by recommended prediction algorithm. For next level improvement of prediction algorithm methods probable research work and dares are listed out. The patient's data which include medication, way of life and common information handled by self-observation and leads a large dataset which demands well prediction, prevention, detection and treatment of diabetes system. For further work this research can be protracted to examine and enhance the glucose prediction algorithm performance [4][15].

Diabetes associated characteristics of complications have analysed with the main objective of successfully preventing dangerous complications. More than 49,000 non diabetic patients and nearly 8000 related diseases of diabetic patients were examined. Hypertension, hyperlipemia, heart disease and cerebral infarction were four major complications. Based on the associated relationship between these complications the characteristics of these four

complications in male, female and different age groups were examined. Statistical analysis was performed over 599 medical exposure indices in diabetic patients who were having and not having these four complications. It showed that there was frequent occurrence of these complications in diabetic women having age between 65 to 70 years. Women before reaching this age should change their diet and life style by doing exercise and food control to prevent from these complications. This study showed that cerebral infarction was 2.5% in diabetic patients. In other case cerebral infarction extended to 10% if diabetic patient had huge heart diseases. Compared to women heart diseases were occurred frequently in men having diabetes [5]. T2DM is a metabolic disease which develops several complications. A significant clinical value is the primary identification of a personal at menace for complications after being identified with T2DM. In this study, the author presents an extrapolative method to envisage the above significant clinical value. The author executes general experimentations on patient data mined from a large electronic health record claims database, to measure the presentation of these approaches. The outcomes showed efficiency of the multi-task framework above demonstrating every complication individually. A quantity of upcoming research guidelines is available. International Classification of Diseases ciphers and primary demographic data were used in evaluation. An expectation performance of complication can improve by integrating new types. It is essential to detect the significant related risk features long with the obstacle measures. In conclusion the author also fascinated to get used these methods to other kinds of electronic health record data and other metabolic diseases [6].

There are many patients who are suffering from comorbidity like T2D, heart diseases, cancer, infectious diseases, BP and dementia, eating disorders, anxiety disorders, and substance abuse. Most of the medical expenditure in US is for management of patients with these several parallel disorders. The promotion of Pre-emptive caring and cost reduction of a single patient can be achieved by Guessing prospective comorbid conditions of that patient. The author guided that for future work the tree-based trajectory model can be improved, however the present trajectories are having high prevalent conditions and also possible that sieving trajectories with low confidence ranks. For assessing the excellence of projected trajectories more tests will be considered and executed. Also, the trajectory result used in evaluation of whether here any significant sex or age particular changes in trajectories. The advance trajectory one hand expects the possible prevalence on the other hand it expects the intermediary situations, so it can be considered sa a standard form of a cooperative expectation model. The author planned to apply helpful expectation time series features in the upcoming and also planned to combine the two models into one single model [7][16]. This study shows usefulness of mobiab system for patients and also show topographies of the system generally used. Few users who used mobiab systems are selected. Which the data collected for this study was chosen from One year period. The users who were involved in this study was the users of mobiab system who was suffering with diabetes mellitus and also the users of non-diabetic. The outcomes showed that mobiab system is convenient for 70% of the users and they used maximum components of the system. 30% of the users used limited components of the system only. A user use mobiab system in daily rehearsal. An example a specific diabetic user in 6 months period recorded entries of food, activities, glycaemia, insulin values. Quantity of users were augmented by publicity and non-diabetic users were motivated by maximum rewards to escalate continuing observance. In forthcoming this system compulsory remodel, the mobile applications in appearance of system design to attract the users. The grouping of many functionalities and survival of inspiration and gasification component is the main benefit of mobiab system [8].

Now a Diabetes is treated as worse than cancer and HIV. It causes blindness, kidney failure, and heart disease. In the health care community the avoidance of the disease is a warm topic. To find the cause of the disease and antidote it several investigations have been take place. In this study the author discussed the diabetes likely to be happen based on the persons' life style events. It includes eating and sleeping ways, bodily movement along with other indicators like BMI, waist circumference etc. For a extremely unqualified dataset CART estimate model has been applied with one third precision in this effort. As a major factor of producing diabetes, blood pressure is known. Other causes street food eating, late-night sleeping, heredity, rice intake and bodily events made by a person in a day.so it can be said that any one enjoy each part of his life but a petite attention in ones' daily routine does no hurt [9]. It is needed to estimate the tool to determine a diabetic patient. The Back propagation neural network is one among the different prediction methods which produce accurate result at present. In this paper the author discussed about this tool. To make tool user friendly the GUI is developed to get perfect results in the absence of doctor. This project helps doctors to get the results of patient in second so that he can save the time for next treatment of the patient. This study shows tool implementation and progress in MATLAB. In case of identifying a patient is diabetic or not he BPNN performance is 80% compared to previous work there is an improvement and it is better for patients instead of finger stick which painful in failure of the test a greater number of times. This tool shows the results in binary format i.e.: 0 or 1. The number 0 means non-diabetic and number 1 means diabetic [10].

3 experimental set up

The objective of this paper is to use binary classifier and compare the results of the same. The 8 medical predictor features that are present in the data set listed below:

Pregnancies: Indicating pregnancy times
Glucose: Level of tolerance of glucose plasma concentration
Blood Pressure: Diastolic blood pressure (mm Hg)
Skin Thickness: Triceps skin fold thickness (mm)
Insulin: 2-Hour serum insulin (mu U/ml)
BMI: Body Mass Index (weight in kg/ (height in m)^2)
Diabetes Pedigree Function: Diabetes pedigree function on genetic influence and hereditary risk
Age: Age (years)

3.1 Random forest classifier:

To check the reputation of feature Random Forest Classifier was taken as implementation. Among the total parameters the three main predictor features are BMI, Glucose and Age.

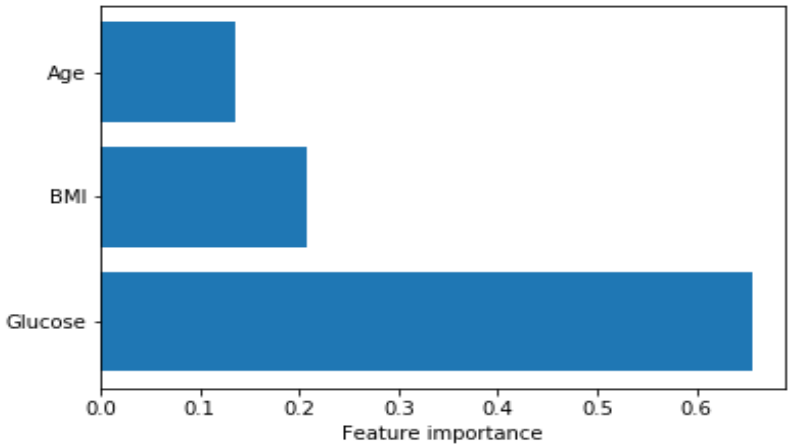


Fig 2 Prediction of feature importance

Based on the feature importance graph it is evident that three factors contribute more for the successful extraction of the results. They are age, BMI and the glucose level. Even though other factors also contribute to the classification the above three plays a major role in classification.

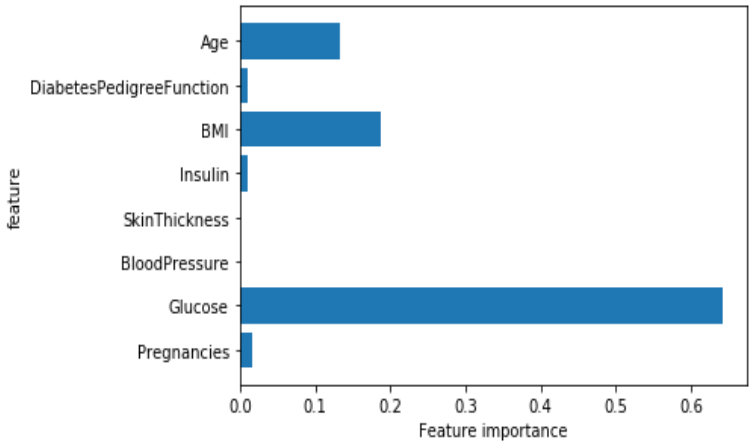


Fig 3 Overall parameters for prediction

The graph above depicts the feature importance of all the attributes in the dataset. From the above it is evident that pregnancies, insulin and diabetes pedigree along with the age, BMI and glucose plays an important role in the classification.

3.2 logistic regression

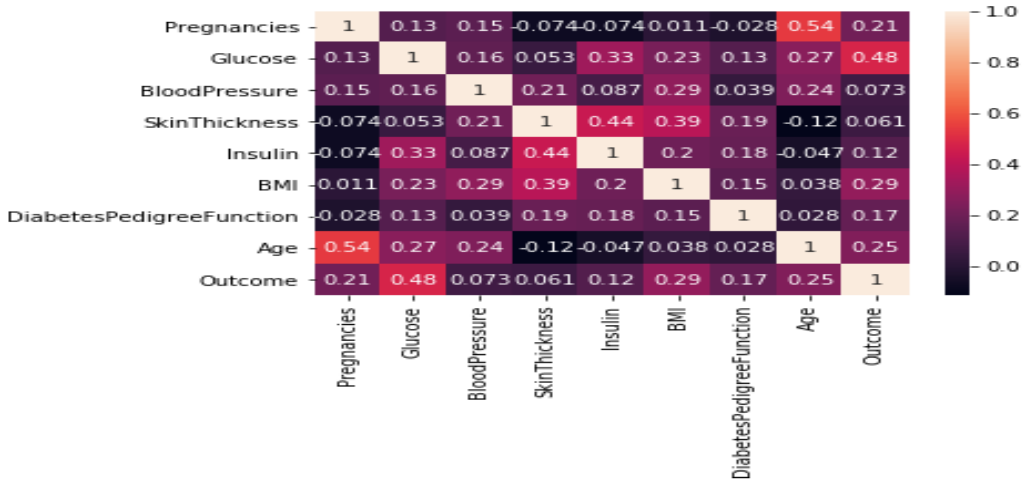


Fig 4 Parameters measurement scale

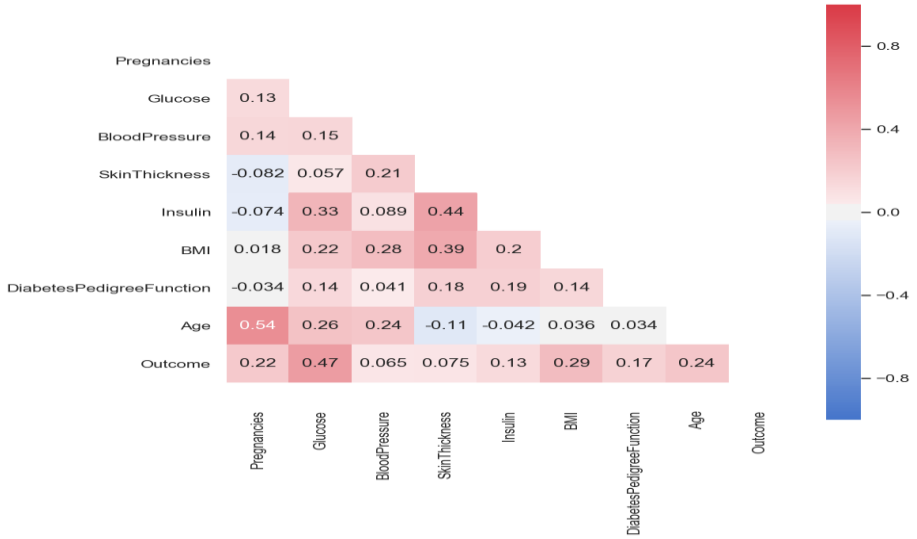


Fig 5 Contribution towards Classification

The above figure is the outcome of the logistic regression, it establishes the relationship between the attributes and how they contribute to the classification of the problem. It is clear from the outcomes that glucose, pregnancies, BMI and age attributes contribute more in the classification. It is to be noted that pregnancies also contribute to a major extent in the classification which was not identified in the decision tree classifier.

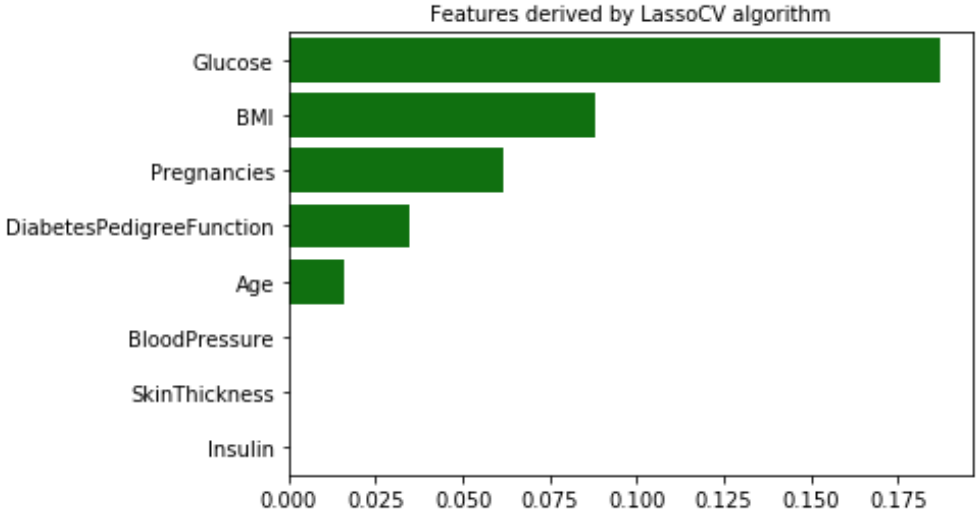


Fig 5 Features derived for prediction

Table 1 Misclassification rate of algorithms

Classifier	Acc. Of Training Set	Acc. Of Test Set	True -ve	True +ve	False –ve	False +ve	Accuracy	Misclassification
Decision Tree	0.852	0.729	92	31	21	48	.72	0.27
KNN	0.79	0.17	108	15	40	29	.71	0.28
SVM	.74	.74	97	44	25	26	.74	0.26
Random Forest	-	-	126	4	5	83	.90	0.10
Naïve Bayes	-	-	107	23	26	46	.76	0.24
Logistic Regression	-	-	163	17	26	46	.79	0.21
Linear Regression	-	-	-	-	-	-	.90	0.10

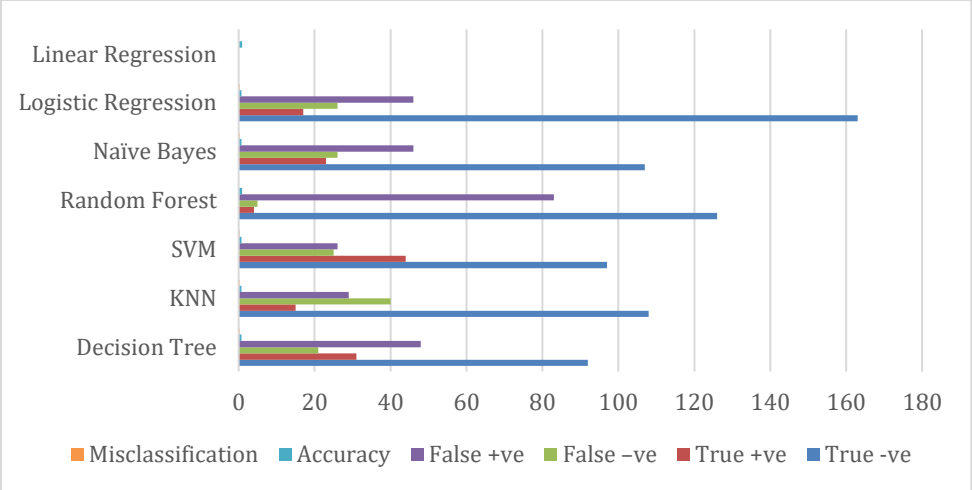


Fig 6 Graph of Misclassification of algorithms

4 results

The classifiers utilized for contrasting the boundaries exactness, accuracy, beneficiary working trademark bend and time taken by the calculations are Gaussian Naïve Bayes, Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Logistic relapse, Linear SVC. The Bayes techniques are a bunch of administered learning calculations dependent on applying Bayes' hypothesis with the "innocent" supposition of contingent freedom between each pair of highlights given the estimation of the class variable.

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} \quad P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Since P(x1,…, xn) is constant given the input, we can use the following classification rule:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} \quad P(x_i \mid y) \Downarrow \hat{y} = arg \max_{y} P(y) \prod_{i=1}^{n} \quad P(x_i \mid y),$$

Notwithstanding their plainly distorted speculations, unsuspecting Bayes classifiers have worked outstandingly in some certifiable conditions, extensively report gathering and spam isolating. They require an unassuming

amount of getting ready data to survey the indispensable limits. Straightforward Bayes understudies and classifiers can be fast diverged from more refined procedures. The decoupling of the class unforeseen component spreads suggests that each scattering can be independently evaluated as a one dimensional apportionment. This accordingly helps with helping issues beginning from the scourge of dimensionality. GaussianNB realizes the Gaussian Naive Bayes count for portrayal.

The limits σy and μy are surveyed using most prominent likelihood. BernoulliNB realizes the honest Bayes planning and gathering computations for data that is passed on by multivariate Bernoulli scatterings; i.e., there may be various features anyway everybody is believed to be a twofold regarded (Bernoulli, Boolean) variable. Thus, this class anticipates that tests should be addressed as twofold regarded component vectors; at whatever point gave some other kind of data, a BernoulliNB case may binarize its data (dependent upon the binarize parameter).The decision rule for Bernoulli guileless Bayes relies upon:

$$P(x_i|y)=P(i|y)\, x_i+(1-P(i|y))(1-x_i)$$

which changes from multinomial NB's norm in that it explicitly rebuffs the non-occasion of a component I that is a marker for class y, where the multinomial variety would essentially dismiss a non-happening feature. Because of text request, word occasion vectors (rather than word check vectors) may be used to get ready and use this classifier. BernoulliNB may perform better on some datasets, especially those with more restricted records. It is fitting to survey the two models, if time awards.

Multinomial Naïve Bayes checks the prohibitive probability of a particular word given a class as the overall repeat of term t in records having a spot with class(c). The assortment thinks about the amount of occasions of term t in getting ready chronicles from class (c), including different occasions. Vital backslide is a game plan estimation used to apportion observations to a discrete course of action of classes. A part of the cases of collection issues are Email spam or not spam, online trades Fraud or not Fraud, Tumor Malignant or Benign. Determined backslide changes its yield using the key sigmoid ability to reestablish a probability regard.

Direct SVM is used for straightly unmistakable data, which suggests if a dataset can be assembled into two classes by using a lone straight line, by then such data is named as straightforwardly separable data, and classifier is used called as Linear SVM classifier.

Fig 7 Performance measure of classifiers

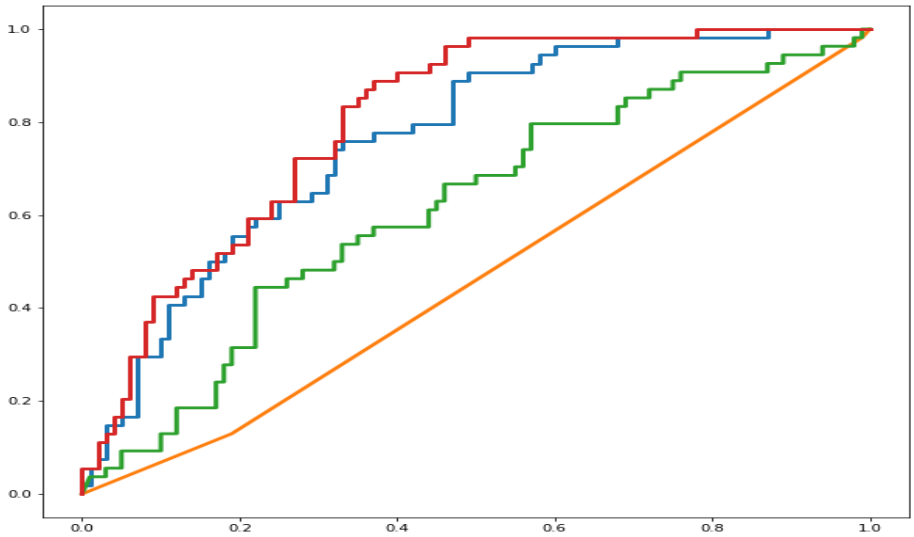
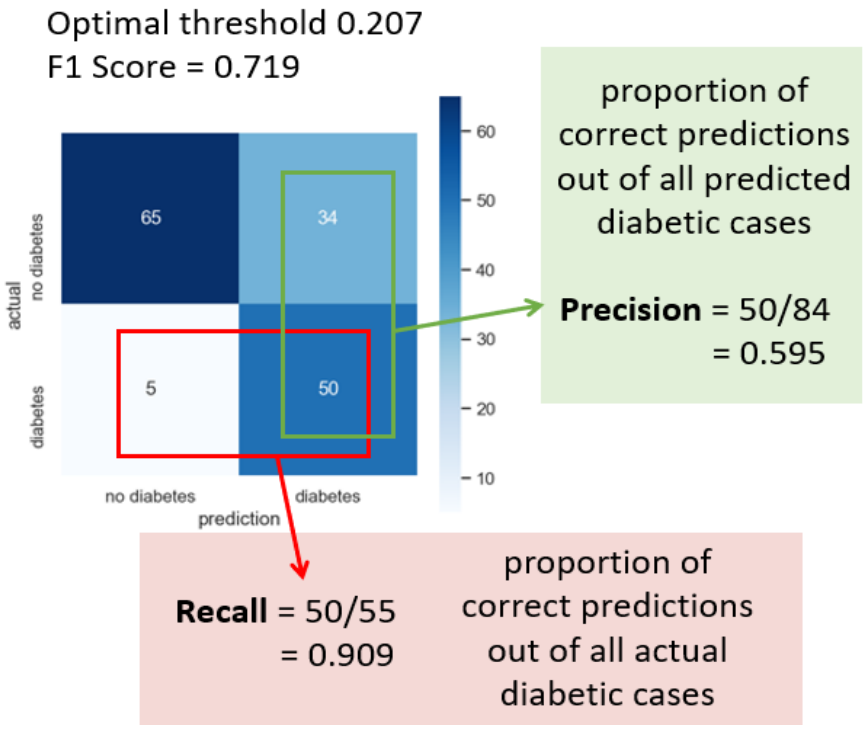


Fig 7 Result comparison of classifier



5 conclusion

We applied numerous calculations and did a great deal of highlight control and extraction. We got the best precision of 90% utilizing arbitrary woodland and straight relapse. The pertinence of this order venture isn't to erroneously arrange the patient which would be expected, so the emphasis should be on "Review" metric. Gaussian Naive Bayes model has done well overall (0.909). In this task, the Gaussian Naive Bayes model has accomplished an expectation score of 0.909, i.e., out of every single diabetic patient, 90.9% of them will be effectively grouped utilizing clinical indicative estimations.

References

1. Zarkogianni, K., Athanasiou, M., Thanopoulou, A. C., & Nikita, K. S. (2017). Comparison of machine learning approaches toward assessing the risk of developing cardiovascular disease as a long-term diabetes complication. IEEE journal of biomedical and health informatics, 22(5), 1637-1647.

2. Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., & Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PloS one, 12(7), e0179805.
3. Farzi, S., Kianian, S., & Rastkhadive, I. (2017, December). Predicting serious diabetic complications using hidden pattern detection. In 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI) (pp. 0063-0068). IEEE.
4. Huzooree, G., Khedo, K. K., & Joonas, N. (2017, July). Glucose prediction data analytics for diabetic patients monitoring. In 2017 1st International Conference on Next Generation Computing Applications (NextComp) (pp. 188-195). IEEE.
5. Chen, P., & Pan, C. (2017, February). Evaluation of the relationship between diabetes and large blood vessel disease. In 2017 13th IASTED International Conference on Biomedical Engineering (BioMed) (pp. 200-207). IEEE.
6. Liu, B., Li, Y., Sun, Z., Ghosh, S., & Ng, K. (2018, April). Early prediction of diabetes complications from electronic health records: A multi-task survival analysis approach. In Thirty-Second AAAI Conference on Artificial Intelligence.
7. Ji, X., Chun, S. A., & Geller, J. (2016). Predicting comorbid conditions and trajectories using social health records. IEEE transactions on nanobioscience, 15(4), 371-379.
8. Burda, V., Novák, D., & Schneider, J. (2016, August). Evaluation of diabetes mellitus compensation after one year of using Mobiab system. In 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 6002-6005). IEEE.
9. Anand, A., & Shakti, D. (2015, September). Prediction of diabetes based on personal lifestyle indicators. In 2015 1st International Conference on Next Generation Computing Technologies (NGCT) (pp. 673-676). IEEE.
10. Joshi, S., & Borse, M. (2016, September). Detection and prediction of diabetes mellitus using Back-propagation neural network. In 2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE) (pp. 110-113). IEEE.
11. Chee YJ, Ng SJH, Yeoh E. Diabetic ketoacidosis precipitated by Covid-19 in a patient with newly diagnosed diabetes mellitus. Diabetes Res Clin Pract 2020 April 24
12. Li J, Wang X, Chen J, Zuo X, Zhang H, Deng A. COVID-19 infection may cause ketosis and ketoacidosis. Diabetes Obes Metab 2020 April 20
13. Ren H, Yang Y, Wang F, et al. Association of the insulin resistance marker TyG index with the severity and mortality of COVID-19. Cardiovasc Diabetol 2020;19:58-58.
14. J. Omana and M. Moorthi, "Naïve Bayes based Summarizing Ruleset in Prediction of Diabetes Mellitus using Magnum Opus," 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2019, pp. 255-259, doi: 10.1109/I-SMAC47947.2019.9032528.
15. M .Baskar, J. Ramkumar, V.Venkateswara Reddy, G.Naveen Reddy, "Cricket Match Outcome Prediction using Machine Learning Techniques", International Journal of Advanced Science and Technology, Vol. 29, No. 4, pp: 1863-1871, ISSN: 2005-4238, April 2020.
16. M .Baskar, J. Ramkumar,Ritik Rathore, Raghav Kabra, "A Deep Learning Based Approach for Automatic Detection of Bike Riders with No Helmet and Number Plate Recognition", International Journal of Advanced Science and Technology, Vol. 29, No. 4, pp: 1844-1854, ISSN: 2005-4238, April 2020.
17. Baskar, M., Renuka Devi, R., Ramkumar, J. *et al.* Region Centric Minutiae Propagation Measure Orient Forgery Detection with Finger Print Analysis in Health Care Systems. *Neural Process Lett* (2021). Springer, January 2021. https://doi.org/10.1007/s11063-020-10407-4.