Available online at http://www.advancedscientificjournal.com
http://www.krishmapublication.com
IJMASRI, Vol. 2, issue 2, pp. 421- 428, Feb. -2022
https://doi.org/10.53633/ijmasri.2022.2.2.002

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY ADVANCED SCIENTIFIC RESEARCH AND INNOVATION (IJMASRI)

ISSN: 2582-9130 **IBI IMPACT FACTOR 1.5** **DOI: 10.53633/IJMASRI**

RESEARCH ARTICLE

EARLY- STAGE DIABETES PREDICTION USING MACHINE LEARNING CLASSIFICATION ALGORITHMS

[1]Yatharth Kathuria, [1]Vishal Modani, [2]Sachin Garg and [2]Varun Goel

[1]Department of Information Technology Maharaja Agrasen Institute of Technology, New Delhi, India
[2]Assistant Professor, Department of Information Technology Maharaja Agrasen Institute of Technology, New Delhi, India
kathyatharth1999@gmail.com, vishalmodani647@gmail.com, sachingarg@mait.ac.in
varungoel.cs@gmail.com

Abstract

Diabetes has developed as one of the maximum risky threats to the human global. Many have become its sufferers and are not able to pop out of it despite the fact that they're running to keep away from it from developing in addition. Cloud Computing and the Internet of Things (IoT) are equipment that play a completely vital function in today's lifestyles concerning many factors and our poses which includes healthcare tracking of sufferers and aged society. Diabetes Healthcare Monitoring Services may be very vital in recent times due to the fact bodily going to hospitals and status in a queue is a completely useless model of affected person tracking. Maybe if an affected person has very persistent diabetes and it isn't detected at an early level, his/her ought to spend his/her time in lengthy queues for a prognosis which may be risky if left undetected in an extended run. Diabetes also can act as a way for different sicknesses like coronary heart attack, kidney damage, and relatively blindness. This mission uses numerous devices getting to know algorithms consisting of K-nearest neighbours, naive Bayes, assist vector device(Linear Non-Linear), logistic regression, selection tree, and random wooded area with the assist of which we will without problems discover the accuracy of a version predicting that someone has diabetes or not.

Key words: Machine Learning, Classification, Diabetes, Dimensionality Reduction, Analysis Metrics

Introduction

Diabetes is a noxious disorder withinside the global. Diabetes is brought on due to weight problems or excessive blood glucose level, and so forth. It impacts the hormone insulin, ensuing in ordinary metabolism of crabs and enhancing ranges of sugar withinside the blood. Diabetes happens whilst the frame does now no longer make sufficient insulin. According to (WHO) World Health Organization approximately 422 million human beings are afflicted by diabetes mainly from low- or idle-earnings countries. And this will be extended to 490 billion as much as the 12 months 2030. However, the superiority of diabetes is discovered amongst numerous Countries like Canada, China, and India, and so on. The populace of India is now greater than one hundred million so the real range of diabetics in India is forty million. Diabetes is a chief reason of dying withinside the global. Early prediction of sicknesses like diabetes may be managed and shop human lifestyles. To accomplish this, this painting explores the prediction of diabetes through taking numerous attributes associated with diabetes disorder. For this purpose, we use the Dataset taken from the UCI device getting to know Repository, we observe numerous Machine Learning Classification Techniques to are expecting the opportunity of getting diabetes. Machine Learning is a way this is used to teach computer systems or machines explicitly. Various Machine Learning Techniques offer green effects to gather expertise through constructing numerous class and ensemble fashions from the gathered datasets. Such gathered records may be beneficial to are expecting diabetes. Various strategies of Machine Learning can successful to do prediction, however, it's difficult to pick the great method. Thus, for this purpose, we observe famous class and ensemble techniques to the dataset for prediction. Diabetes may be successfully controlled when11 caught early. However, whilst left untreated, it is able to result in capacity headaches that consist of heart8 damage. disorder, stroke, kidney damage, and nerve damage. Normally when you devour or drink, your frame will damage down sugars out of your meals and use them for electricity to your cells. To accomplish this, your pancreas desires to supply a hormone referred to as insulin.2 The insulin is what allows the method of pulling sugar from the blood and setting it withinside the cells for use, or5 electricity. If you have diabetes, your pancreas both produces too little insulin or none at all. The insulin can't be used successfully. This allows4 the blood organization disadvantaged of glucose ranges to upward push at the same time as the relaxation of your cells are9 wished electricity. This can result in an extensive sort of troubles affecting almost each predominant body10 device. Diabetes also can have an effect on your pores and skin, the most important organ of your frame. Along with dehydration, your body's loss of moisture because of excessive blood sugar can reason the pores and skin in your ft to dry and crack. It's vital to absolutely dry your ft after bathing or swimming. You can use petroleum jelly or mild creams, however keep away from letting those regions end up too1 moist. Moist, heat folds withinside the pores and skin are vulnerable to fungal, bacterial, or yeast infections. These have a tendency to develop6 among arms and toes, the groin, armpits, or withinside the corners of your mouth. Symptoms consist of redness, blistering, and itchiness. High-pressure spots beneath Neath your foot can result in calluses. These can end up inflamed or expand ulcers. If you do get an ulcer, see3 your doctor right now to decrease the threat of dropping your foot. You will also be greater inclined to7 boils, inflamed nails.

Literature review

Diabetes is one of the fastest-developing persistent lifestyles-threatening sicknesses which have already affected 422 million human beings international in step with the record of the World Health Organization (WHO), in 2018. Due to the presence of a highly lengthy asymptomatic segment, early detection of diabetes is constantly preferred for a clinically significant final result. Around 50% of everybody stricken by diabetes are undiagnosed due to its lengthy-time period asymptomatic segment. The early prognosis of diabetes is most effective feasible through right evaluation of each not unusual place and much less not unusual place signal signs, which may be discovered in exceptional levels from

disorder initiation as much as prognosis. Data mining class strategies were nicely generic through researchers for the threat prediction version of the disorder. To are expecting the chance of having12 diabetes calls for a dataset, which includes the records of newly diabetic or would-be diabetic sufferers. In these paintings, we've got used this kind of dataset of 520 times, which has been gathered the usage of direct questionnaires from the sufferers of Sylhet Diabetes Hospital in Sylhet, Bangladesh, and authorised through a doctor. We have analysed the dataset with Naive Bayes Algorithm, Logistic Regression Algorithm, and Random Forest Algorithm and after making use of tenfold Cross-Validation and Percentage Split assessment strategies, Random-wooded area has been discovered to have the great accuracy in this dataset. Finally, a generally accessible, person-pleasant device for the end-person to test the threat of getting diabetes from assessing the signs and beneficial hints to govern over the threat elements has been proposed. This mission pro- poses a powerful method for in advance detection of diabetes disorder. An assessment of the exceptional device getting to know strategies used on this examine exhibits which set of rules is great proper for the prediction of diabetes. Diabetes Prediction is turning into the place of hobby for researchers to teach this system to perceive the affected person is diabetic or now no longer through making use of right classifier at the dataset. Based on preceding studies paintings, it's been found that only some class algorithms aren't sufficient. Further, using cross-validation does now no longer enhance the very last accuracy of the bulk of the algorithms. Hence a device is needed for Diabetes Prediction is a vital place in computer systems, wherein all of the class algorithms are carried out and their accuracies are as compared to get pick the great-proper one.

Dataset description

In this research, the early-stage diabetes risk prediction dataset collected from the UCI machine learning repository was used. The dataset was created through a direct questionnaire among diabetic and non-diabetic patients from the diabetes Hospital of Sylhet, Bangladesh. This dataset consists of 520 records and 17 attributes. Among them, 320 records are positive and 200 records are negative. A brief description of attributes is given in the below table.

S.No	Name of the attribute
1	Age
2	Sex
3	Polyuria
4	Polydipsia
5	Sudden Weight Loss
6	Weakness
7	Polyphagia
8	Genital Thrush
9	Visual Blurring
10	Itching
11	Irritability
12	Delayed healing
13	Partial paresis
14	Muscle stiffness
15	Alopecia
16	Obesity
17	Class

This class variable shows the outcome of either 0 or 1 for diabetics which indicates positive or negative for diabetics.

Research Methodology

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

Step1: Import required libraries, Import diabetes dataset.

Step2: Encoding the dependent Variable, Splitting the dataset, and Feature Scaling.

Step3: Perform a percentage split of 75% to divide dataset as Training set and 25% to Test set.

Step4: Select the machine learning algorithm i.e., K Nearest Neighbour, Support Vector Machine(Linear and Non-Linear), Decision Tree, Logistic regression, Random Forest.

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

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

Step7: Perform dimensionality reduction for visualization.

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

A. Data Pre-Processing

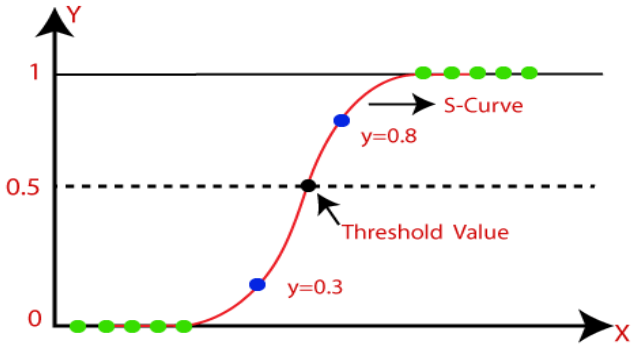
Data pre-processing is the most important process. Mostly healthcare-related data contains missing values, differences in the range of values, etc that can cause the effectiveness of data. To improve quality and effectiveness obtained after the mining process, Data pre-processing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate results and successful prediction. In the data pre-processing, we have first imported sufficient libraries such as NumPy, pandas, matplotlib, and imported datasets as well. We have pre-checked for the null values and missing data in each of the attributes. We have Encoded the Categorical Data after which Splitting of the Data frame into Train and Test Sets is being done Splitting of data- After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train the algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically, the aim of normalization is to bring all the attributes under the same scale. Feature Scaling is further done to improve the training process and thus it will improve the final predictions.

B. Classification Algorithms

***Logistic regression*:**

Logistic regression is a machine learning algorithm for classification. In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function. It is a linear classification model.

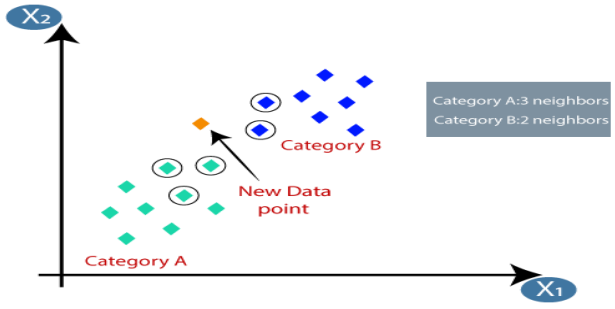
Fig 1: Logistic Regression Classification Model



***K-Nearest Neighbours(K-NN)*:**

Neighbours-based classification is a type of lazy learning as it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the k nearest neighbours of each point. It is a non-linear classification model.

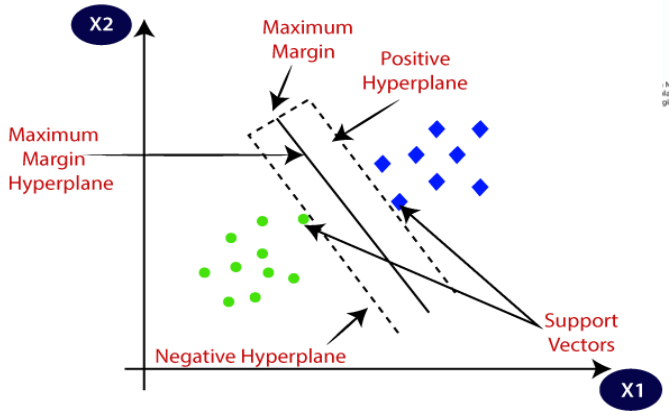
Fig 2: K-Nearest Neighbours Classification Model



***Support Vector Machine(SVM)*:**

Support vector machine is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. It is a linear classification model

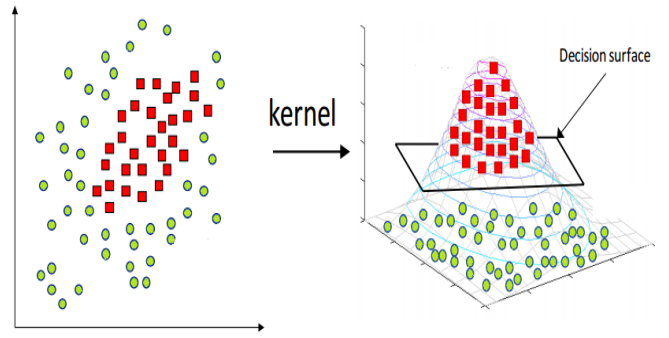
Fig 3: Support Vector Machine Classification Model



Kernel SVM: Kernel Function in Kernel SVM

This method used to take data as input and transform into the required form of processing data. "Kernel" is used due to set of mathematical functions used in Support Vector Machine provides the window to manipulate the data. So, Kernel Function generally transforms the training set of data so that a non-linear decision surface is able to transform to a linear equation in a higher number of dimension spaces. It is a non-linear classification model.

Fig 4: Kernel SVM Classification Model



Naive Bayes:

Naive Bayes algorithm based on Bayes' theorem with the assumption of independence between every pair of features. Naive Bayes classifiers work well in many real-world situations such as document classification and spam filtering. It is a non-linear classification model

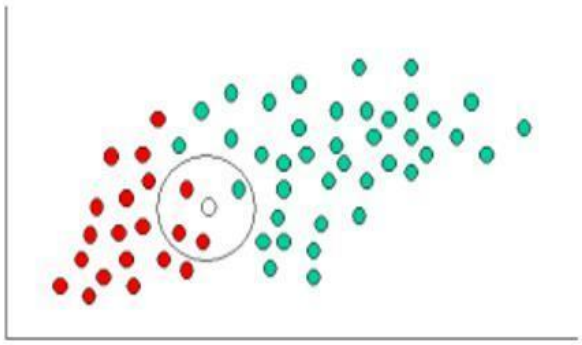
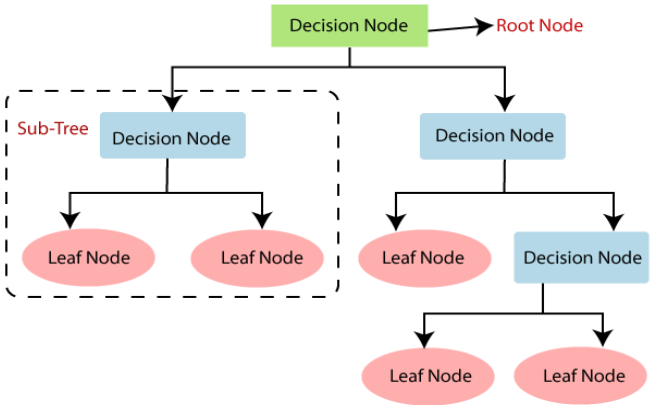


Fig 5: Naïve Bayes SVM Classification Model

Decision Tree Classification:

Given a data of attributes together with its classes, a decision tree produces a sequence of rules (decision node) that can be used to classify the data. They are not power on their own but they are used in other methods that leverage their simplicity and create some very powerful machine learning algorithms which are used to perform facial recognition, etc. It is a non-linear classification model.

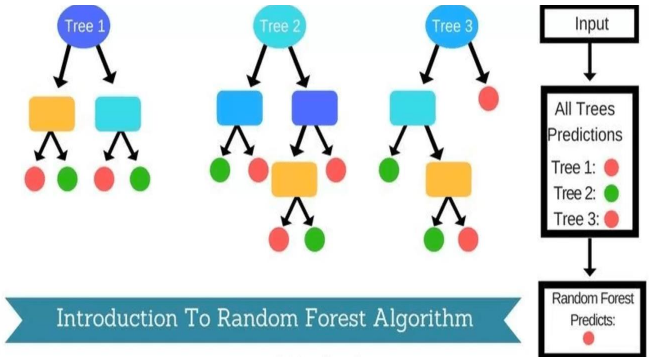
Fig 6: Decision Tree Classification Model



***Random Forest Classification*:**

Random forest classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses an average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement. It is a non-linear classification model and uses ensemble learning.

Fig 7: Random Forest Classification Model



C. Evaluation Criteria

It is the most common evaluation metric for classification problems. It is defined as the number of correct predictions against the number of total predictions. Also finding only the accuracy score sometimes is not enough. So, we also look at other performance metrics like Precision (measuring exactness), Recall (measuring completeness) , and the F1 Score (compromise between Precision and Recall).

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1 Score = 2 * Precision * Recall / (Precision + Recall)

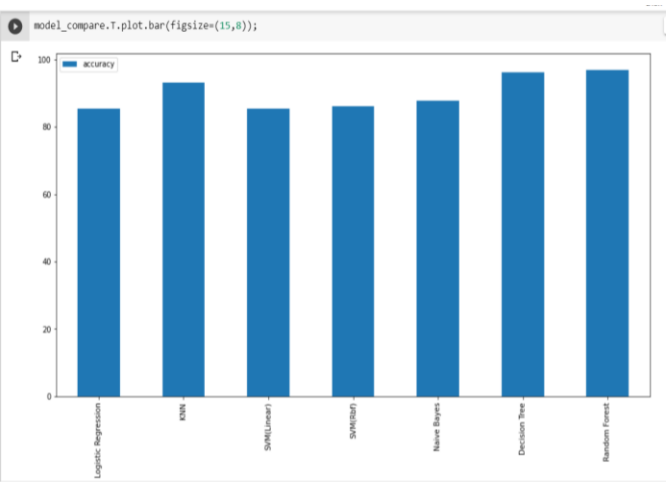
Results

A. *Final Results Analysis*

- Finally, the accuracies of all the models are being compared using a bar graph with the help of pandas library and plot function.

- After a comparison of the accuracies, it is found that **Random Forest Classification Model** having **accuracy=96.9.32** is the best one for "**EARLY STAGE DIABETES PREDICTION**".

Name Of Algorithm	Accuracy	Precision	f1-score	Recal-score
Logistic Regression	85.385%	0.831	0.879	0.932
KNN	93.07%	0.933	0.940	0.946
SVM (Linear)	85.385%	0.831	0.879	0.932
SVM(Rbf)	86.154%	0.833	0.886	0.946
Naive Bayes	87.962%	0.854	0.892	0.946
Decision Tree	96.154%	0.960	0.966	0.973
Random Forest	96.932%	0.973	0.973	0.973



B. Pros and Cons of Each Classification Model

Classification Algorithm	Pros	Cons.
Logistic Regression	Probabilistic	It's assumptions
K-NN	Simple, fast efficient	Need to choose neighbours k
SVM (Linear)	Overcomes overfitting, not biased by outliners	Not for non-linear problems and a large number of features
SVM (Non-Linea)	Overcomes overfitting, not biased by outliners, suitable for non-linear problems	More complex and not for a large number of features
Naive Bayes	Efficient, probabilistic , efficient, suitable for non-linear problems	Assumption of the relevance of features
Decision Tree	No feature scaling, suitable for linear and non-linear problems	Poor results for small datasets, easy overfitting
Random Forest	Powerful, accurate. Suitable for non-linear problems	Need to choose the number of trees, easy overfitting

Future Scope

A particular method to identify diabetes is not a very sophisticated way for initial diabetes detection and it is not fully accurate for predicting diseases. That's why we need a smart hybrid predictive analytics diabetes diagnostic system that can effectively work with accuracy and efficiency. So, we use data mining and classification for exploring and utilizing to support the medical decisions, which improves in diagnosing diabetic patients. Due to the dataset, we have till date is not up to the mark, we cannot predict the type of diabetes, so in the future, we aim to predict the type of diabetes and explore it, which may improve the accuracy of predicting diabetes. Along with that, there are various deep learning methods that could be applied to further improve the efficiency of the diagnosis.

Reference

1. Diabetes, May 2020, [online] Available: https://www.who.int/newsroom/factsheets/detail/diabetes.
2. Maniruzzaman, M., Rahman, M.J and Al-Mehedi Hasan, M. (2018)."Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers", *J Med Syst*, vol. 42, pp. 92.
3. Diabetes Daily, May 2020, [online] Available: https://www.diabetesdaily.com/learn-about-diabetes/what-isdiabetes/how-manypeople-have-diabetes/
4. Safial Ayon and Md. Islam, (2019) "Diabetes Prediction: A Deep Learning Approach", *International Journal of Information Engineering and Electronic Business*, vol. 11, pp. 21-27, 2019.

5. Wang, Q., Cao, W. Guo, J. Ren, J. Cheng, Y and Davis, D. N. (2019) "DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data With Missing Values", *IEEE Access*, vol. 7, pp. 102232-102238.
6. Agrawal, P and Dewangan, A (2015). "A brief survey on the techniques used for the diagnosis of diabetesmellitus", *Int. Res. J. Eng. Technol. (IRJET)*, vol. 02, no. 03.Ahmed, Developing a predicted model for diabetes type 2 treatment plans by using data mining.
7. Islam, M. M. F., Ferdousi, R. Rahman, S and Bushra, H.Y. (2020). Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques", *Computer Vision and Machine Intelligence in Medical Image Analysis. Advances in Intelligent Systems and Computing*, vol. 992. [online]Available: https://doi.org/10.1007/978-981-13-8798-2_12.
8. Early stage diabetes risk prediction dataset. Data Set, Aug 2020, [online] Available: https://archive.ics.uci.edu/ml/datasets/Early+ stage + diabetes + risk + pre diction + dataset.
9. Google Collaboratory, Aug 2020, [online] Available:https://colab.research.google.com/notebooks/intro.ipynb

*****