Diabetic Prediction using Machine Learning Techniques

[1] Shreya C, [2] Dr. Manjula S

[1]MTech student, [2]Assistant Professor
[1]Department of Computer Science and Engineering, JSSS &TU Mysuru,
[2]Department of Computer Science and Engineering, JSSS&TU, Mysuru.

Abstract : Diabetes mellitus is a common disease of human body caused by a group of metabolic disorders where the sugar levels over a prolonged period is very high. It affects different organs of the human body which thus harm a large number of the body's system, in particular the blood veins and nerves. Early prediction in such disease can be controlled and save human life. To achieve the goal, this research work mainly explores various risk factors related to this disease using machine learning techniques. 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 three popular machine learning algorithms, Naive Bayes (NB), K-Nearest Neighbor (KNN) and Decision Tree (DT), on adult population data to predict diabetic mellitus. Naïve bayes, KNN for diabetic disease prediction and Decision tree is used for time prediction i.e., for how long patient will not be affected by the diseases. Our experimental results show that Naive Bayes achieved higher accuracy compared to KNN.

IndexTerms - Diabetes, KNN, Naïve Bayes, Decision Tree, Machine Learning Algorithms.

I. INTRODUCTION

There are many databases in the medical sector. These databases may include data that is organised, semi-structured, or unstructured. Diabetes has developed into a highly severe disease in developing nations such as India as a result of the need to preserve the status quo. Diabetes mellitus (DM) is a chronic non-communicable illness. Diabetes affects millions of individuals. Each year, about 25 million individuals die of diabetes. This number is projected to rise to 629 million by 2045. [1]

Early illness prediction can be managed and save human lives. In order to accomplish this objective, this study focuses on the early prediction of diabetes by taking into consideration different risk considerations associated with this illness. For the aim of the research, diagnostic data sets with 21 diabetes characteristics of 2801 data records are collected. These characteristics include age, history of smoking, eye issue, genetic problems, etc. We develop a prediction model based on these characteristics using different machine learning methods to predict diabetes mellitus.

Machine learning technology provides an efficient outcome by creating models from gathered information to extract knowledge. Knowledge extracted from such data may be helpful in predicting diabetes individuals. Different machine learning methods are capable of predicting diabetes mellitus. But choosing the optimal method to forecast based on these characteristics is extremely challenging. For purposes of the research, we utilise machine learning methods such as Naïve Bayes, KNN for prediction for diabetic diseases, and Decision Tree for the prediction of time - that is, how long patients on adult population data will not be impacted by diabetic mellitus.

.

II. LITERATURE SURVEY

The main purpose of the literature survey is that to get a idea about the related works and research done in the domain of our study/Project. This will enhance our knowledge in the field of study and also helps to improve the overall structure of our project.

1.N. Sneha and Tarun Gangil proposed "Analysis of Diabetes Mellitus for Early Prediction Utilizing Optimal Function Selection". Diabetes is a variety of illnesses. They try to use predictive analytics to select properties that help early detection of diabetes. Analyze the capabilities of the dataset and select the best one based on the correlation values. For diabetes data analysis, the decision tree algorithm and random forest, which have the highest specificity of 98.20% and 98.00%, respectively, are the best. The accuracy of the SVM is 77% and the accuracy of the NB is 82.30%, which allows the function to be successfully mapped in the lower and higher dimensions.

2. Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid and Munam Ali Shah proposed "Prediction of Diabetes Using Medical Machine Learning Algorithms".
Predictive analytics in the medical field has the potential to revolutionize the way medical researchers and physicians analyze and select data. They applied machine learning techniques to predictive analytics in this work. SVM, KNN, LR, DT, RF, and NB are examples of such algorithms. Predictions of diabetes were made using the 768 record PIMAIndian data set. The predictive model was trained and tested using eight features. From the experimental results, it is clear that SVM and KNN have the highest precision for predicting diabetes.

3. Lakshmi K.S proposed "Extracting relevant rules from medical records of diabetics". The medical database provides a wealth of information. The widespread use of electronic medical record systems and recent advances in medical technology have allowed hospitals and other medical facilities to generate vast amounts of medical text data.This treatise describes a new way to identify relevant rules available in medical records. The extracted rules show the relationship between the illness, the symptoms of a particular illness, the drugs used to treat the illness, and the most common age range of the individual who develops the particular illness. To extract rules, NLP (Natural Language Processing) tools are integrated with data mining techniques (Apriori algorithm and FPGrowth algorithm).

III.EXISTING SYSTEM

Diabetes identification is very important with regard to its serious consequences. The present process is a manual one in which the patient's report is manually evaluated by the relevant doctor, and the patient additionally does home glucose monitoring using a diabetes monitor. Most of the previous research investigations concentrated only on 4-5 factors and the findings are less reliable. Earlier test reports done manually, which takes time.

IV. PROPOSED WORK

We are attempted at developing a model in which the system predicts whether the patient is diabetic or not. If the patient is not diabetic, we predict the time duration ie for how long he won't be affected by the disease . To also evaluate the performance and accuracy of various techniques. This is created as an application in real time to assist control early predictions and save lives.

.

4.1 Data-set

In this work, we collect diabetes data from the opens source https://archive.ics.uci.edu/ml/datasets/Diabetes. The dataset consists of various attributes or risk factors of diabetes mellitus of 2801 patients. We have summarized the attributes

SerialNo	Parameter
1.	Age
2.	Gender
3.	Relation
4.	DOD
5.	DD
6.	SugarTested
7.	Value
8.	LEyeE
9.	Symptoms
10.	FamilyHistory
11.	DiabetesTreatment
12.	diabetestreatedfrom
13.	PastSmoked
14.	SmokePerDay
15.	StartedSmoking
16.	Drinking
17.	Weight
18.	Height
19.	AC
20.	BPSystolic
21.	BPDiastolic

V. METHODOLOGY

5.1 Machine Learning

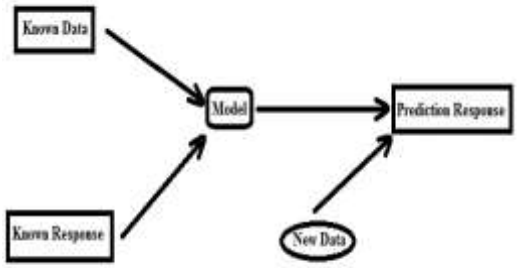
Machine learning is a process of studying a system based on data. Machine learning is a part of data science where we use machine learning algorithms to process data.

5.1.1 Supervised Learning Technique

It's a predictive model used for the tasks where it involves prediction of one value using other values in the data-set. Supervised learning will have predefined labels. It classifies an object based on the parameters to one of the predefined set of labels.

We have many algorithms to build model in supervised learning such as KNN, Naive bayes, Decision Tree, ID3, Random Forest, SVM, Regression techniques etc…. Depending of the requirement, labels, parameters and data-set we select the appropriate algorithm for predictions. Algorithm is used to build a model that makes predictions based on evidence in the presence of uncertainty.

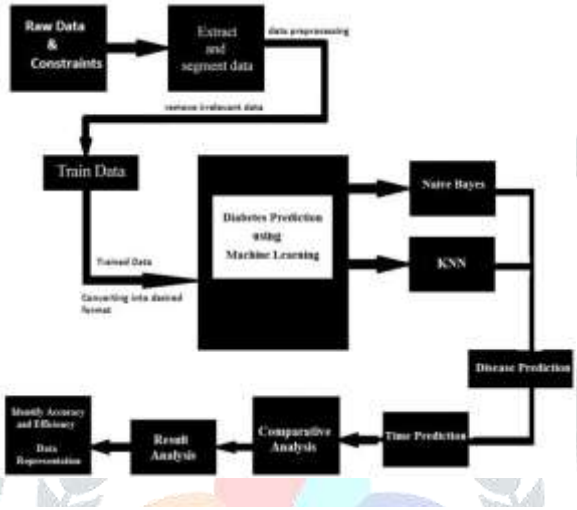
In this project for prediction we make use to "*Bayesian Classifier or KNN algorithm*" which is an efficient and works fine for all different sets of parameters. It also generates accurate results.



Proposed Model

5.2 Classification Rules

Basically classification is used to classify each item in a set of data into one of the predefined set of classes or groups.



Once the data has been ready for modelling, we employ machine learning classification techniques to predict diabetes mellitus. Hence we give an overview of these techniques.

- **Naive Bayes**: Naive Bayes is a popular probabilistic classification technique proposed by John et.al. Naive Bayes also called Bayesian theorem is a simple, effective and commonly used machine learning classifier. The algorithm calculates probabilistic results by counting the frequency and combines the value given in data set. By using Bayesian theorem, it assumes that all attributes are independent and based on variable values of classes. In real world application, the conditional independence assumption rarely holds true and gives well and more sophisticate classifier results.

- **Naive Bayes Algorithm Steps**

    - We scan the retrieval of data from servers for mining, such as the database, cloud, excel sheet etc.

    - We compute the likelihood of occurrence using the following formula for each attribute We should apply the formulas for each class.

    - P=(n_c + m*p)/(n+m)

    - We multiply the output of each attribute with P for each class and utilise the final outcomes for classification.

    - Compare and categorise the values of the attributes to one of the preset class set.

- **K-Nearest Neighbour Algorithm**: K-nearest neighbour is simple classification and regression algorithm that used non parametric method proposed by Aha et.al. . The algorithm records all valid attributes and classifies new attributes based on their resemblance measure. To determine the distance from point of interest to points in training data set it uses tree like data structure. The attribute is classified by its neighbours. In a classification technique, the value of k is always a positive integer of nearest neighbour. The nearest neighbours are chosen from a set of class or object property value

- **K-Nearest Neighbour Algorithm Steps**

  - Retrieve a sample record with the column and row names PimaIndianDiabetes
  - Retrieve records containing attribute and string tests.
  - Calculate the Euclidean distance using the equation.
  - The following tests to ensure that no two values of K are equivalent. Previous Next
  - Using the following minimum and Euclidean distances, determine the nth column
  - Identical output values were discovered.
  - If the patient's results match, he or she has diabetes. Unless that is the case, it is not.

- **Decision Tree**: A decision tree is a tree that provides powerful classification techniques to predict diabetes mellitus. The majority of the information highlights limited discrete areas and feature called the "classification". Every discrete area and feature of the domain is called a class. An input feature of the class attribute is labelled with the internal node in a decision tree. The leaf node of the tree is labelled by attribute and each attribute associated with a target value. The highest information gain for all the attribute is calculated in each node of the tree.

  - **Decision Tree Algorithm Steps**

    - Build a node tree as an input function.
    - To forecast the output of the fea-ture with the greatest information gain.
    - For each characteristic of each tree node, the maximum information gain is computed.
    - Repeat step 2 with the feature not used in the above node to create a subtree

VI. RESULTS AND DISCUSSION

6.1 Data Visualization

6.1.1 Naive Bayes Algorithm Results

Result Analysis

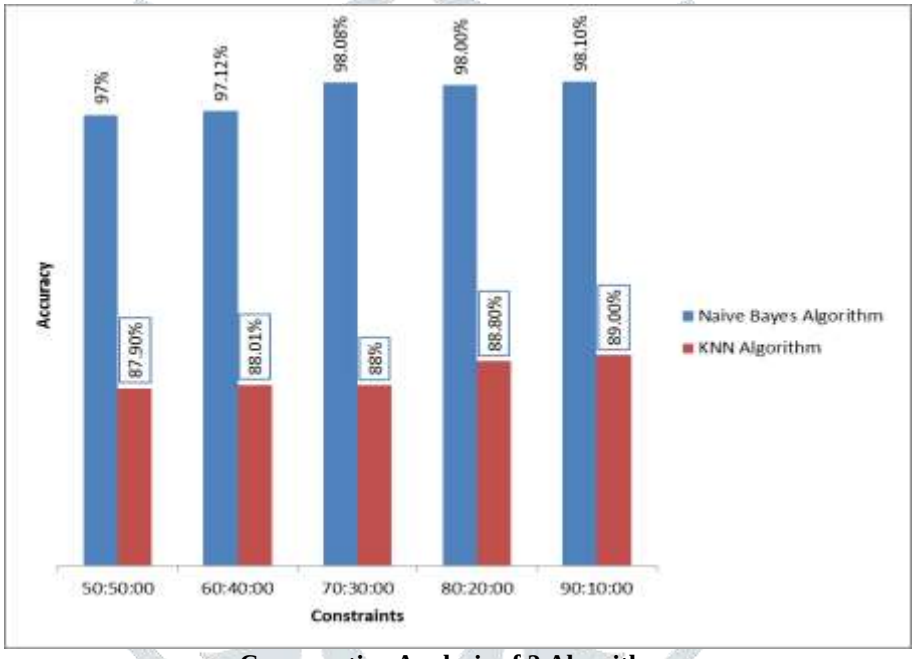
Naive Bayes	Constraint
Accuracy	98.1818181818182 %
Time (milli secs)	569
Correctly Classified	98.1818181818182 %
InCorrectly Classified	1.81818181818181 %

6.1.2 KNN Algorithm Results

Result Analysis

KNN	Constraint
Accuracy	89.0909090909091 %
Time (milli secs)	1075
Correctly Classified	89.0909090909091 %
InCorrectly Classified	10.9090909090909 %

6.1.3 Comparative Analysis



Comparative Analysis of 2 Algorithms

VII. CONCLUSION

In this work, we have analyzed the early prediction of diabetes by taking into account various parameters using machine learning techniques. To predict diabetes effectively, we have done our experiments using three popular machine learning algorithms, Naive Bayes (NB), K-Nearest Neighbor (KNN) and Decision tree, on adult population data to predict diabetes. The performance of Naive Bayes is significantly superior to other machine learning technique for the classification of diabetic data. The experimental results could assist health care to take early prevention and make better clinical decisions to control diabetes and thus save human life.

REFERENCES

[1] World Health Organization, "Report of a study group: Diabetes Mellitus," World Health Organization Technical Report Series, Geneva, 727, 1985.

[2] Kemal Polat, Salih Gunes, and Ahmet Arslan, "A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine," Expert Systems with Applications, vol. 34. 1, January. 2008, pp. 482-487.

[3] Kayaer K and Yildirim T, "Medical diagnosis on Pima Indian diabetes using general regression neural networks," Proceedings of the international conference on artificial neural networks and neural information processing, 2003, pp. 181-184.

[4] Jack W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," Proc. Annu. Symp. Comput. Appl. Med. Care, November 9. 1988, pp. 261-265.

[5] Karegowda A. G., Manjunath A. S. and Jayaram M. A., "Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes," International Journal on Soft Computing, vol. 2. 2, 2011, pp. 15-23.
[6] Carpenter G. A. and Markuzon N., "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases," Neural Networks, vol. 11. 2, 1998, pp. 323-336.
[7] Wold S., Esbensen K. and Geladi P., "Principal component analysis," Chemometrics and intelligent laboratory systems, vol. 2. 1-3, 1987, pp. 37-52.
[8] Balakrishnama S. and Ganapathiraju A., "Linear discriminant analysis-a brief tutorial," Institute for Signal and information Processing, vol. 18, 1998.
[9] Deng L. and Yu D., "Deep learning: methods and applications," Foundations and Trends in Signal Processing, vol. 7. 3-4, 2014, pp. 197-387.
[10] Lee H., "Tutorial on deep learning and applications," NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning, 2010.
[11] Safavian S. R. and Landgrebe D., "A survey of decision tree classifier methodology," IEEE transactions on systems, man, and cybernetics, vol. 21. 3, 1991, pp. 660-674.
[12] Suykens J. A. K. and Vandewalle J., "Least squares support vector machine classifiers," Neural processing letters, vol. 9. 3, 1999, pp. 293-300.
[13] Hosmer Jr. D. W., Lemeshow S. and Sturdivant R. X., "Applied logistic regression," John Wiley & Sons, 2013.
[14] Lin Y., "Support vector machines and the Bayes rule in classification," Data Mining and Knowledge Discovery, vol. 6. 3, 2002, pp. 259-275