DIABETES PREDICTION USING DATA MINING

[1]Prof. Dr. ML Sharma, Prof. Sunil Maggu, [3]Akshita Goel, [4]Deepali Guleria, [5]Srishti Bansal
[1]HOD, [2]Assistant Professor, [3]Student, [4]Student, [5]Student
[1]Information Technology
[1]Maharaja Agrasen Institute of Technology, New Delhi, India

Abstract: Data mining is use to extract those patterns that were previously hidden and unknown. One of the tasks of data mining is Classification. Health care data is often large, diverse and multiplex because different variable types are present. These days, it is necessary to get knowledge from such data. Data mining is utilized to extract knowledge by constructing models from healthcare data such as sugar patient sets. Diabetes is a very serious disease and it is a major health challenge worldwide. Using data mining methods for early prediction of diabetes has gained major popularity.

In this paper, we are going to predict whether a person is diabetic or not using data mining. K-Nearest Neighbour algorithm is used for diabetes data and classification. We have compared the accuracy for classifying the data.

1. INTRODUCTION

Diabetes is the most chronic disease in society and in all age groups. In this disease either body does not produce or properly use the insulin. Diabetes is a condition that weakens the body's ability to process blood glucose (blood sugar). Insulin is important as our body cells need glucose for growth. When someone has diabetes, little or no insulin is discharged. In this state, the body is unable to use insulin even when plenty of glucose is available in the bloodstream.

Different kinds of diabetes can occur, and controlling them depends upon the type. Diabetes may not certainly occur in a person leading an inactive lifestyle or being overweight. In some cases it is present from childhood.

Some people are having prediabetes meaning when their blood sugar is in the range of 100 to 125 milligrams per decilitre (mg/dL). Normal blood sugar level range is 70 and 99 mg/dL, whereas diabetic person will have a blood sugar higher than 126 mg/dL. The prediabetes level means that blood sugar level is higher than normal but not so high to cause diabetic condition. But people having prediabetes are at probability of developing type 2 diabetes, although they don't experience the symptoms of full diabetes.

Without early detection and proper management, blood sugar levels can increase. Diabetes should not be left untreated or improperly managed, as it can result in a variety of complications, including kidney failure, amputation, heart attack, blindness, stroke, etc.

Diabetes complications can be avoided by keeping the blood pressure and blood sugar at proper level. For this, proper diagnoses should be done as early as possible to provide suitable treatment. The main advantage of information technology is that a large data storage of past patients' records is maintained and monitored by hospitals continuously for various references. These data records help the doctors to study different patterns in the data set. The design pattern found in data sets is used for prediction, diagnosis and collation for the various chronic diseases. Here, we use such dataset to help us predict whether a person is diabetic or not by training our machine using previous datasets.

1.1 TYPES OF DIABETES

Four types of Diabetes are there. These are Type1, Type2, Gestational diabetes, congenital diabetes.

1. Type 1: In this the body stops producing insulin. It is usually diagnosed in young adults and children, called juvenile diabetes. Only 5% of diabetic people have this form of the disease. Here the pancreas does not produce insulin. Type 1 diabetic must get a synthetic structure of insulin. They either receive it from an insulin pump or from a shot.
2. Type 2: It is the most common form of diabetes. In this, the insulin is not properly used by the body, called insulin resistance. Type 2 diabetic patient may need to take diabetes pills or insulin. In some cases, exercise and a proper meal plan can help a lot.
3. Gestational Diabetes: This type of diabetes occurs during pregnancy. It occurs when the body cannot produce sufficient insulin to handle the consequences of a growing baby and the changing hormone levels.
4. Congenital Diabetes: Genetic defects of insulin secretion, cystic fibrosis-related diabetes, steroid diabetes induced by high doses of glucocorticoids are the main cause of this type of diabetes.

1.2 SYMPTOMS OF DIABETES

The general symptoms of diabetes:

1. Frequent vomiting
2. Blurred vision
3. Increased thirst
4. Frequent hunger
5. Slow healing infection
6. Loss of body weight
7. Frequent urination

2. METHODOLOGY

Several techniques are used in data mining to describe the type of mining. These are:

(1) Association

Association is a mining function that find the probability of the relationship between data items in the collection of data set in which relationships are expressed as association rules.

(2) Classification

Classification is a data mining technique that it is used to assign each data items in a collection is data items into a target category or class. Mathematical procedures are used here such as linear programming, statistics decision trees and neural network. Pattern recognition is a type of classification where an input data is categorized to likeness among one of many already defined classes.

(3) Clustering

Clustering is a technique of data mining which uses automatic method and undertakes a significant group of objects which have same characteristics. Classes and places objects are described in each class. It differs from the other classification techniques as the objects are allocated into predefined classes there.

(4) Decision trees

A decision tree is a structure that has a root node, branches, and the leaf nodes. This is one of the most commonly used procedures in data mining as its prototype is easy to find for the users.

(5) Prediction

The prediction is a technique that determines the association amongst self-determining variables (as its name suggests) and association amongst both dependent and self-determining variables.

(6) Sequential Patterns

Sequential patterns is a study or a technique in data mining that determines or identifies the comparable patterns, trends or consistent events over a professional period in the transaction data.

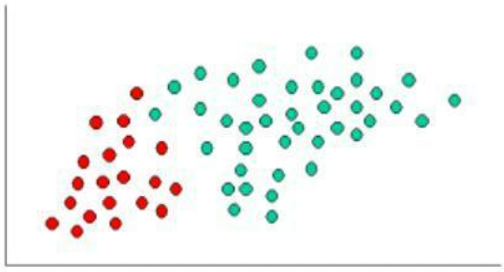
(7) Regression analysis

Regression analysis is a statistical method in data mining that is used amongst variables for approximating the associations.

3. DIFFERENT DATA MINING ALGORITHMS HAVE BEEN PROPOSED TO CLASSIFY, PREDICT AND DIAGNOSE DIABETES

3.1 GAUSSIAN NAIVE BAYES

Gaussian Naive Bayes is an algorithm used for classification amongst two-class (binary) and multi-class classification problems and when described using binary or categorical input values (respectively), it is easiest to understand. For numerical variables, normal distribution is assumed. Bayes Theorem of conditional probability is the basis of Gaussian Naive Bayes. It is called *naive Bayes* or *idiot Bayes* because the probability calculation done for each value in the dataset is



found using Bayes Theorem. This is a strong assumption that the attributes do not interact and this is most unlikely in real data. But this algorithm's approach performs quite where the above assumption does not hold on data.

Naive Bayes can be extended to real-valued attributes, commonly by assuming a Gaussian distribution and this extension is called as the Gaussian Naive Bayes. Although to estimate the distribution of the data other functions can also be used, but the

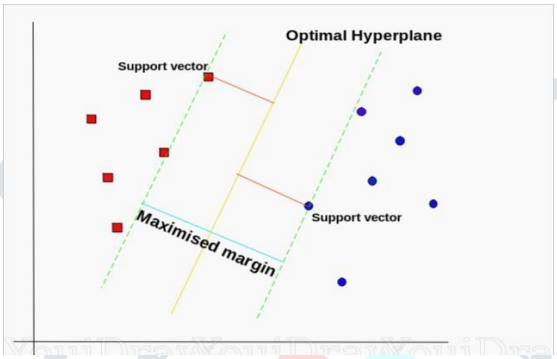
Gaussian (or Normal) distribution is the easiest to deal with because the mean and the standard deviation from our training data is needed to be estimated only.

Pros of Gaussian Naive Bayes:
1. To predict class of test data set is fast and easy and performs in multi class predictions very well.
2. A Naive Bayes classifier performs better when compared to other models of classification like logistic regression etc. When the assumption of independence holds, and we need less training data.

Cons of Gaussian Naive Bayes model:
1. Model will be assigned a zero (0) probability if, categorical variable has a category in the test dataset which wasn't observed in the training dataset and will be unable to make a prediction. This is often known as "Zero Frequency".

3.2 SUPPORT VECTOR MACHINE

SVM (Support Vector Machine) is an algorithm that falls under supervised learning and is used for classification and regression problems. It can solve linear and non-linear problems and works well for many practical problems. According to SVM, the algorithm creates a line or a hyperplane which separates the data into classes. The input data is then classified according to the side of the line it falls on.

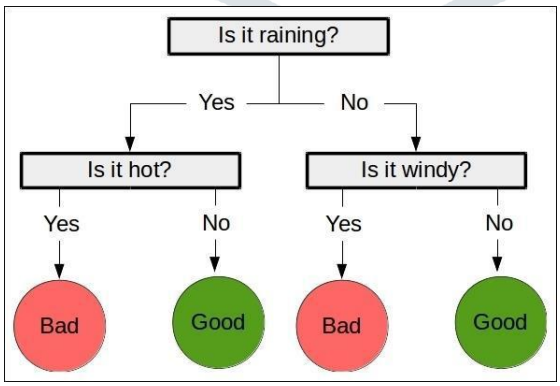


Pros of SVM:
1. When the number of dimensions is larger than the number of models, SVM is pretty effective.
2. A subgroup of training points are used in decision function (called the support vectors), making its memory very efficient.

Cons of SVM:
1. When the dataset is huge, it doesn't perform well because the training time required is very high.
2. It doesn't accomplish very well when the target classes are overlapping.

3.3 DECISION TREES

A decision tree is like graph in whose nodes represent a place from where we can pick up an attribute and ask one question and edges represent the answers to those questions; and leaves represent the class label or the actual output. It is used in non-linear decision making with simple linear decision surface. The examples are classified in decision trees by sorting them down from the root to some leaf node and the classification is provided to the example by the



leaf node. Every node for some attribute acts as a test case and every edge descending from that node corresponds to one of the possible answers to the test case. This process is recursive in nature and for every subtree rooted at new nodes, it is repeated.

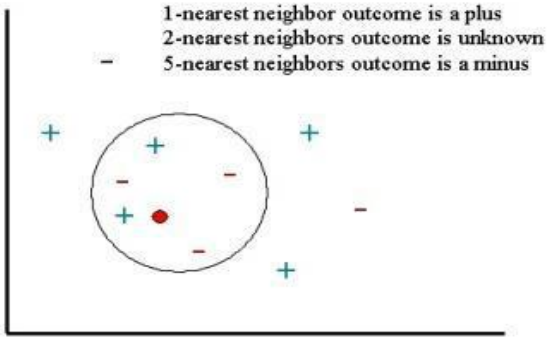
Pros of Decision Tree:
1. Ability of selecting the most discriminatory features.
2. Data classification without many calculations.

Cons of Decision Tree:
1. Overfitting**:** This occurs when the algorithm captures noise in the dataset.

2. High Variance**:** The prediction model becomes unstable with even a small variance in data.

3.4 K- NEAREST NEIGHBOURS (KNN)

In Machine Learning, K-Nearest Neighbours is one of the most basic but important algorithm of classification. It belongs to the domain of supervised learning and is intensively used in intrusion detection, pattern recognition and data mining. It is non-parametric which means it doesn't make any underlying assumptions about distribution of data & hence, is widely used in real-life scenarios. From the data, the model structure is determined. K-Nearest Neighbours Algorithm is based on **feature similarity** and doesn't use training data points for any *generalization* i.e. *no explicit training phase* is present or even its there, it is very minimal. There is lack of generalization which means that KNN keeps all the training data.



KNN is used for classification where the output is shown as a class membership. By a majority vote of the neighbours, an object is classified, then the object is assigned to the class which is most common amongst its k nearest neighbours. When it gives output as the value for the object and predicts continuous values, it can be used for regression. The value so obtained is the average of values of all its k nearest neighbours.

Pros of KNN:

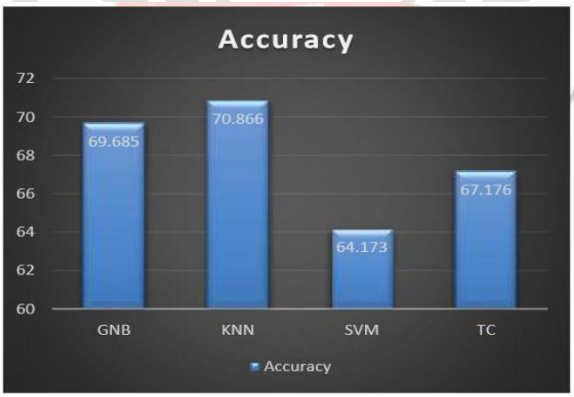
1. Implementation is simple.
2. Featuring the choices is flexible.
3. Multi-class cases are handled naturally.
4. In practicing with enough representative data, it can do well.

Cons of KNN:

1. While finding nearest neighbours, there occur great search problems.
2. Data loading is a problem.
3. We have a significant distance function.

3.5 PERFORMANCE COMPARISON

We tried to optimize every algorithm and found that KNN algorithm is best suitable for our application.



ALGORITHMS	ACCURACY	ERROR RATE
Naive Bayes	69.685	0.33
KNN	70.866	0.34
SVM	64.173	0.29
Decision Tree	67.176	0.28

4. OBJECTIVE AND SCOPE

Data extracting methods aim to extract the data from dataset and then transform it into a clear construction for additional uses. This diagnostic method plans to examine the information and fetches organized associations connecting variables or reliable patterns, and then applying the detected patterns to confirm the findings.

The objective of the project is to build up a framework by utilizing the data mining algorithms that can predict whether the patient has diabetes or not. Also, predicting the illness early prompts the treatment to the patients. Hidden learning from a colossal measure of diabetes-related information can be removed by data mining. That is why, in diabetes examination, it has a critical part. The point drawn from this is to build up a framework which can anticipate the diabetic hazard level in a patient with high accuracy.

Most of the people don't have enough money to spend on the diagnosis test for diabetes and because of that they suffer a lot. This project will help these people to easily detect diabetes without spending any money for the expensive test. It will not only help them to easily detect diabetes but also do it in a very simple way. So, any person with less knowledge can also access it easily for their health.

Such framework can be used to predict and treat various other chronic diseases. It will be a great boon to the medical field and can be used to improve the health conditions of people.

5. DATASET

Collection of data is called a dataset. A data set, mostly, corresponds to the contents of one database table, or statistical data matrix whether single or multiple, where row and column represents different variables. On the selection of these instances from larger database, several constraints are placed.

The diabetes data set was originated from UCI Machine Learning Repository and it contains 10 attributes of 768 number of instances.

Here, patients of Pima Indian heritage are at least 21 years old. From the studies, it has been observed that percentage of Pima Indian population suffering from the diabetes is more than 70.

Name of Data	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration 2 hours in an oral glucose tolerance test
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
Age	Age(years)
Sex	Male or Female
Outcome	Class variable (0 or 1)

Attribute Information Table

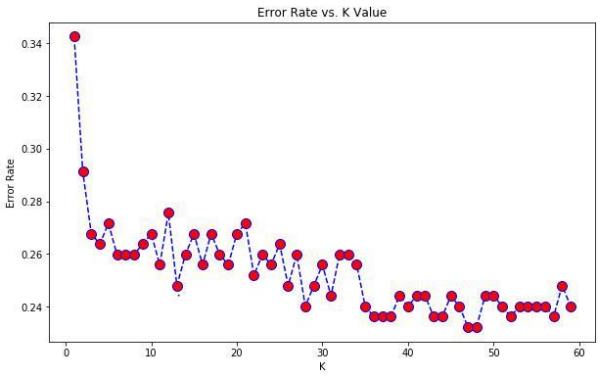
6. IMPLEMENTATION RESULTS

The dataset of 768 instances was implemented for the prediction of diabetes in Python 3 using the Jupyter Notebook with Anaconda Navigator. The machine was trained with the available dataset and then the KNN algorithm was applied to it.

When the different attribute values were not standardized, the accuracy score obtained was 0.7125 i.e. 71.25%.

On standardizing the attribute values, the accuracy score obtained was 0.7322 i.e. 73.22%.

Finding the Appropriate K Value
A graph was plotted between the error rate and the K value as shown below:



According to this, we find that at K=48 the error rate in KNN are minimum.

For K=48, the accuracy score obtained was 0.7677 i.e. 76.77%.

6.1 ACCURACY IMPROVEMENT BY USING BAGGING AND BOOSTING ALGORITHMS

6.1.1 BAGGING AND BOOSTING:

Bagging (Bootstrap Aggregating) **is** an ensemble method which creates discrete samples of the training dataset and then assigns a classifier for each sample. Multiple classifiers results are then combined to give the final outcome. The point here is that each sample of the training dataset has some different thing in it, to give each trained, a precisely different view and perspective on the problem. Bagging is used to decrease the variance in the predicted value by generating additional information for training from the dataset using combinations to produce complex sets of the original data.

Boosting is an ensemble method in which the base classifier is prepared on the training data for accurate results. A second classifier is then afterwards created to focus on those instances in the training data that the first classifier got wrong. This process is continued until a limit is reached in the number of accuracy. Boosting is an iterative method which basically adjusts the weight of an observation based on the previous classification. If an observation was classified wrongly, it then tries to adjust the weight of this observation by either increasing or decreasing it. Thus through boosting strong predictive models are bulid.

6.1.2 BAGGING AND BOOSTING RESULTS:

By applying the bagging algorithm with the KNN, the accuracy was increased and obtained to be 0.7637 i.e. 76.37%

By applying the gradient boosting algorithm with the KNN, the accuracy was increased and obtained to be 0.7598 i.e. 75.98%.

6.2 DIFFERENT ACCURACIES OBTAINED

Different accuracies that were obtained by applying the KNN algorithm on the dataset in different ways are given as:

ALGORITHM	ACCURACY RATE
Without standardization	71.25
With standardization	73.22
With bagging	76.37
With boosting	75.98

7. CONCLUSION

In the medical field, different hidden patterns from the data can be extracted using data mining and machine learning algorithms. Using these, important clinical parameters can be analysed ,various diseases can be predicted, forecasting tasks in medicine can be done, patients can be managed, etc. 768 records diabetes data set were obtained from UCI. Among the number of algorithms that were proposed for prediction, KNN algorithm was found to be the most suitable for this application.

So, K-nearest neighbour algorithm was used for the diagnosis and prediction both. We calculated the accuracy rate. The accuracy rate basically shows that in the test dataset how many outputs of the data are same as output of the data of different features of training dataset. K-Nearest Neighbour which is widely used for diagnostic purposes, is one of the most effective AI algorithms. It provides very accurate and efficient results compared to others.

The ensemble methods like Bagging and Boosting when used with the KNN algorithm provide better prediction performance or accuracy rather than single one.

8. REFERENCES

[1] Kavakiotis, Ioannis, Olga Tsave, AthanasiosSalifoglou, NicosMaglaveras, IoannisVlahavas, and IoannaChouvarda. "Machine learning and data mining methods in diabetes research.Computational and structural biotechnology journal (2017).

[2] Iyer Aiswarya, Jeyalatha S and Sumbaly Ronak. Diagnosis of diabetes using classification mining techniques. International Journal of Data Mining & Knowledge Management Process. 2015; 5:1-14.

[3] Velide Phani Kumar and Velide Lakshmi. A Data Mining Approach for Prediction and Treatment of diabetes Disease.International Journal of Science Inventions Today. 2014; 3:73-9.

[4] Zheng, Tao, Wei Xie, LilingXu, Xiaoying He, Ya Zhang, Mingrong You, Gong Yang, and You Chen. "A machine learning-based framework to identify type 2 diabetes through electronic health records." International journal of medical informatics 97 (2017): 120-127.

[5] Komi, Messan, Jun Li, YongxinZhai, and Xianguo Zhang. "Application of data mining methods in diabetes prediction." In Image, Vision and Computing (ICIVC), 2017 2nd International Conference on, pp. 1006-1010. IEEE, 2017.

[6] American Diabetes Association. Diagnosis and Classification of Diabetes Mellitus. Diabetes Care. 2009;32(Suppl 1): S62-S67.

[7] Pradeep, K. R., and N. C. Naveen. "Predictive analysis of diabetes using J48 algorithm of classification techniques." In Contemporary Computing and Informatics (IC3I), 2016 2nd International Conference on, pp. 347-352. IEEE, 2016.

[8] Perveen, Sajida, Muhammad Shahbaz, Aziz Guergachi, and Karim Keshavjee. "Performance analysis of data mining classification techniques to predict diabetes." Procedia Computer Science 82 (2016): 115-121.

[9] Saravananathan, K., and T. Velmurugan. "Analyzing Diabetic Data using Classification Algorithms in Data Mining." Indian Journal of Science and Technology 9, no. 43 (2016).

[10] Thirumal, P. C., and N. Nagarajan. "Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study." ARPN Journal of Engineering and Applied Science 10, no. 1 (2015): 8-13.

[11] kumarDewangan, A., & Agrawal, P. (2015). Classification of Diabetes Mellitus Using Machine Learning Techniques. International Journal of Engineering and Apllied Sciences, 2(5), 145-148.

[12] "Diagnosis of Diabetes Mellitus based on Risk Factors" International Journal of Computer Applications (0975 – 8887) Volume 10-No.4, November 2010.

[13] Mekruksavanich, Sakorn. "Medical expert system based ontology for diabetes disease diagnosis." In Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference on, pp.383-389. IEEE, 2016.

[14] http://www.kaggle.com/uciml/pima-indians-diabetes-database