Analyze Data Mining Algorithms For Prediction Of Diabetes

[1] Priya B. Patel, [2] Parth P. Shah, [3] Himanshu D. Patel
[1,2,3]Student
[1] Computer Engineering Department,
[1] BVM Engineering College, Vallabh Vidyanagar, India.

***Abstract*— Purpose of data mining is to extract useful information from large collection of data. Before understanding what is Diabetes we need to understand role of insulin in our body. Insulin serve "Gateway" to open body cells, it allows our body to use the glucose for energy. Insulin controls glucose level in our body. Diabetes is a disease in which level of glucose in blood is increase. Traditionally diabetes diagnosed by physical and comical test, But it not give accurate result. To overcome this limitation we make prediction of disease using different Data Mining algorithm for prediction and diagnosis of diabetes mellitus. The main data mining algorithms discussed in this paper are Gaussian Naive Bayes, KNN, SVM and Decision Tree. The data set chosen for experimental simulation is based on Pima Indian Diabetic Set from University of California, Irvine (UCI) Repository of Machine Learning databases.**

***Index Terms*— Data mining, Diabetes, GNB algorithm, KNN algorithm, SVM algorithm, Decision tree algorithm..**

I. INTRODUCTION

Data mining, also entitled knowledge discovery in databanks, in computer science, the development of discovering stimulating and valuable patterns and associations in huge volumes of data. He field combines tools from statistics and artificial intelligence (such as neural networks and machine learning) with database management to analyze large digital collections, known as data sets.

Data mining is widely used in business (insurance, banking, retail), science research (astronomy, medicine), and government security (detection of criminals and terrorists).[1]

The health sector has more need for data mining today than ever. There are quite a lot of arguments that could be sophisticated to support the use of data mining in the health sector (Data overload, early detection and/or avoidance of diseases, Evidence-based medicine and prevention of hospital errors. Non-invasive finding and decision support, Policy-making in public health and additional value for money and price savings) [2]

The Disease Prediction plays an important role in data mining. There are different types of diseases predicted in data mining namely Hepatitis, Lung Cancer, Liver disorder, Breast cancer, Thyroid disease, Diabetes etc… This paper analyzes the Diabetes predictions.There are mainly four types of Diabetes Mellitus. They are Type1, Type2, Gestational diabetes, congenital diabetes. [2]

In type 1 diabetes, the body does not produce insulin. It is usually diagnosed in children and young adults, and was previously known as juvenile diabetes. Only 5% of people with diabetes have this form of the disease[3].Type 2 diabetes is the most common form of diabetes. In this type the body does not use insulin properly. This is called insulin resistance. Gestational diabetes mellitus is a type of diabetes that occurs during pregnancy. It occurs when body cannot produce enough insulin to handle the effects of a growing baby and changing hormone levels.[4]

The Congenital diabetes is caused due to genetic defects of insulin secretion, cystic fibrosis-related diabetes, steroid diabetes induced by high doses of glucocorticoids [2].If left untreated or improperly managed, diabetes can result in a variety of complications, including heart attack, stroke, kidney failure, blindness, problems with erection impotence) and amputation[4].

Keeping your blood pressure and blood glucose (sugar) at target will help to avoid diabetes complications. For this it should be diagnosed as early as possible to provide suitable therapy. The main advantage of information technology is that a huge data storage of past patient's records are maintained and monitored by hospitals continuously for various references. These medical data helps the doctors to examine different patterns in the data set. The designs found in data sets may be used for collation, prediction and Diagnosis of the diseases.

II. Background

Diabetes is the most recurrent disease nowadays in all society and in all age groups. It is a disease in which the body does not produce or properly use insulin. The cells in our body require glucose for growth for which insulin is quite crucial. When someone has diabetes, little or no insulin is discharged. In this situation, plenty of glucose is available in the blood stream but the body is unable to use it. The types of diabetes are Type-1 Diabetes, Type-2 Diabetes, Gestational Diabetes.[6]

- ***Types of Diabetes:***

Type 1 diabetes is the structure where the pancreas does not produce insulin. It was formerly known as insulin dependent diabetes mellitus or juvenile-onset diabetes. Ten percent of sufferers have this structure. People with this structure must obtain a synthetic structure of insulin. They either receive it from a shot or from an insulin pump. Whereas, in type 2 diabetes the pancreas does make insulin. This form was previously named non-insulin-dependent diabetes mellitus or maturity-onset diabetes. However, it may not produce enough. In other cases, the body does not use it accurately. This is known as insulin resistance. People with type 2 diabetes may need to take diabetes pills or insulin. In some cases it can be managed with exercise and a meal plan as well [7].

- ***General symptoms of diabetes:***

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

- ***Diagnose test***

1. Urine test
2. Fasting blood glucose level
3. Random blood glucose level
4. Oral glucose tolerance test
5. Glycosylated hemoglobin.[2]

The data study consists of diabetes dataset. It includes name of the attribute as well as the Explanation of the attributes. Pima Indian Diabetes Dataset and Indian Council of Medical Research–Indian Diabetes (ICMR-INDIAB) study provides data about the Diabetes. World Health Organization (WHO) report had shown a marked increase in the number of diabetics and this trend is expected to grow in the next couple of decades. In the International Diabetes Federation Conference 2003 held in Paris, India was labeled as "Diabetes Capital of the World". Of about 190 million diabetics worldwide, more than 33 million are Indians. According to the diabetes Atlas of 2009, there were 50.8 million people with diabetes in India. The Worldwide figure is expected to rise to 330 million, 52 million of them Indians by 2025.

III. DATASET

Overview:

The dataset is a collection of data. Mostly a data set corresponds to contents of single database table, or a single/multiple statistical data matrix, where column and row represents the different variable. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. From the studies over the years More than 70% Pima Indian population is suffering from the diabetes. The dataset mainly contain 9 attributes of 768 number of instance.[1]

Attribute Information:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

IV. METHODOLOGY

Several core techniques that are used in data mining describe the type of mining.

(1) Association

Association is the popular data mining technique. In connotation, a pattern is exposed centered on a association amongst things in the equal transaction.

(2) Classification

Classification is a typical data mining technique constructed on machine learning. Normally, classification is used to organize each thing in a set of data into one of a predefined set of sets. Classification technique uses mathematical procedures such as decision trees,neural network, statistics and linear programming,. In classification, we define the software that can absorb how to classify the data things into sets

(3) Clustering
Clustering is a data mining technique that marks a significant or beneficial cluster of objects which have alike characteristics using the automatic method. The clustering technique describes the classes and places objects in each class, while in the classification techniques, objects are allocated into predefined classes.

(4) Decision trees
The A decision tree is one of the furthermost common used data mining procedures as its prototypical is easy to recognize for users.

(5) Prediction
The prediction, as its name indirects, is one of a data mining techniques that determines the association amongst self-determining variables and association amongst dependent and self-determining variables.

(6) Sequential Patterns
Sequential patterns study is one of data mining technique that search for to determine or identify comparable patterns, consistent events or trends in transaction data over a professional period.

(7) Regression analysis
It is a statistical method for approximating the associations amongst variables.[8]

(8) Classification algorithms
They involve that the classes be defined centered on data attribute values. Pattern recognition is a kind of classification where an input pattern is categorized into one of numerous classes centered on its likeness to these predefined classes.

Data classification is a two-step process.

1) Model assembly: describing a set of predetermined classes
- Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
- The set of tuples used for model construction is training set
- The model is represented as classification rules, decision trees, or mathematical formulae

2) *Model convention:* for classifying future or unknown objects
- Estimate accuracy of the model
- The known label of test sample is compared with the classified result from the model
- Accuracy rate is the percentage of test set samples that are correctly classified by the model
- Test set is independent of training set (otherwise over fitting)
- If the accuracy is acceptable, use the model to classify new data[1]

V. MACHINE LEARNING METHODS

Machine Learning provides "computers the facility to learn deprived of being explicitly programmed Machine learning algorithms are characteristically classified into three wide groups, depending on the nature of the learning "signal" or "feedback" accessible to a learning system. These are

(1) *Supervised learning:*
The computer is offered with instance inputs and their preferred productions, given by a "teacher", and the objective is to learn a common rule that draws inputs to outputs.

(2) Unsupervised learning:
Not any labels are specified to the learning algorithm, leaving it on its individual to discover structure in its input. Unsupervised learning can be a aim in itself (discovering secrete patterns in data) or a means towards an end (feature learning).

(3) *Reinforcement learning:*
A computer program cooperates with a dynamic environment in which it need to accomplish a certain .

1

Different data mining algorithm has been proposed for classify, predict and diagnose diabetes.

GAUSSIAN NAIVE BAYES

Naive Bayes Classifier Overview

The Naive Bayes Classifier method is founded Bayesian theorem .GNB mainly used when the dimensionality of the inputs is high. In spite of its simplicity, GNB can often outperform extra sophisticated classification approache.

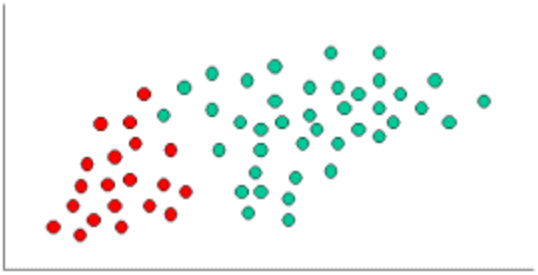


Figure 1 GNB Classification

[10]

To exhibit the idea of Naïve Bayes Classification, contemplate the example displayed in the illustration above. As specified, the objects can be classified as either GREEN or RED. Our task is to categorize new cases as they reach, i.e., select to which class label they fit, based on the presently escaping objects.

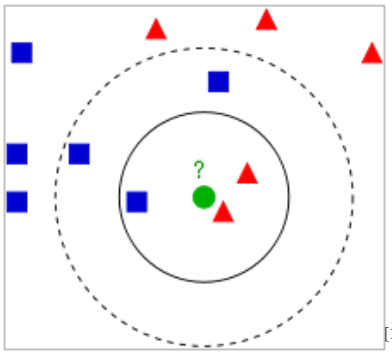
Since there are double as many GREEN objects as RED, it is sensible to consider that a new case (which hasn't been observed yet) is twofold as likely to have attachment GREEN rather than RED. In the Bayesian analysis, this certainty is known as the prior probability. Prior probabilities are based on prior experience, in this case the percentage of GREEN and RED objects, and often used to forecast consequences before they actually occur.

KNN

KNN is known as a instance-based learning, or also popoular as a lazy learning, where the role is only estimated locally and all calculation is postponed until classification.The k-NN algorithm is the most simplest among all machine learning algorithms. It studies all existing cases and classifies into new cases based on common factor. (e.g., distance functions). Features for which it is very popular is its simplicity of interpretation and short calculation time even with such ease, it can provide greatly feasible results. The neighbors are taken from a set of objects for which the object property value (for k-NN regression) or the class (for k-NN classification) are recognized. This can be supposed of as the training set for the algorithm, though no explicit training step is essential. Significance of k is constantly a positive integer. Here neighbors are selected from set of objects so constantly accurate classification is recognized. Categorizing objects based on the following exercise data in the feature space.

Following are steps used to perform prediction through KNN algorithm.

Step1: Define k means the magnitude of nearest neighbors.
Step2: Calculate the distance among the query instance and all the training samples.
Step3: The distance of entire the training samples are organized and nearest neighbor constructed on the k minimum distance is resolute.
Step4: Acquire entire the classes of the training data for the structured value which falls under k.
Step5: Utilize simple majority of group of nearest neighbors as the forecast value of the query instance.



[11]

Figure 2 KNN classification

Example of k-NN classification. The test example (green circle) should be characterized either to the first class of blue squares or to the another class of red triangles. If k = 3 (solid line circle) it is assigned to the another class as there are 2 triangles and only 1 square exclusive the inner circle. If k = 5 (dashed line circle) it is allotted to the first class (3 squares vs. 2 triangles inside the outer circle).

There will a number of examples for drill. These samples are stored in an n –dimensional space. When an unfamiliar test label is given, the k-nearest neighbor classifier quests these samples which are nearby to the unknown sample. Closeness is usually defined in terms of Euclidean distance .The Euclidean distance is between two points P (p1, p2… Pn) and Q (q1, q2… qn) given by equation.

How does the KNN algorithm work?

The training cases are vectors in a multidimensional feature space.each vector has its class label. In KNN algorithm training phase involves storing the class labels of the training samples and the feature vectors.

In the classification phase, k is defined as a user-defined constant, and an vector without label is classified by allocating the label. These which is most frequent amongst the k training samples nearest to that query point.

A normally used distance metric for continuous variables is Euclidean distance. For distinct variables, such as for text classification, additional metric can be used, such as the overlap metric (or Hamming distance). In the perspective of gene expression microarray data, for example, k-NN has also been involved with association coefficients such as Spearman.Often, Pearson, the classification exactness of k-NN can be enhanced expressively if the distance metric is learned with dedicated algorithms such as neighborhood components analysis orLarge Margin Nearest Neighbor.

KNN Pros

- Simple to implement
- Flexible to feature / distance choices
- Naturally handles multi-class cases
- Can do well in practice with enough representative data

KNN Cons

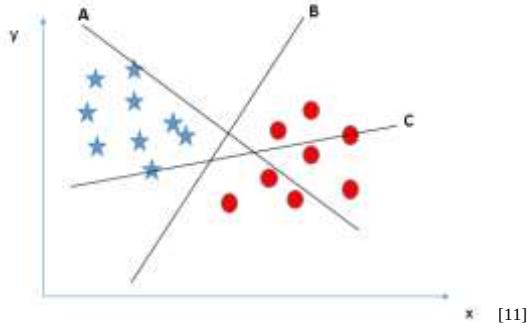
- Great search problem to find nearest neighbours
- Loading of data
- Must know we have a significant distance function

SVM [SUPPORT VECTOR MACHINE]

A Support Vector Machine (SVM) is a discriminative classifier. SVM can be defined by a separating hyperplane. In other words,labeled training data has been given (known as supervised learning), output of these algorithm is an optimal hyperplane.

How SVM algorithm work?

Identify the right hyper-plane (Scenario-1):



[11]

Figure 3 SVM Classification Scenario-1

You need to remember a thumb rule to identify the right hyper-plane: “First-rate the hyper-plane which ghettoizes the two classes better”. In this scenario, hyper-plane “B” has brilliantly achieved this job.

Recognize the right hyper-plane (Scenario-2): Here, we have three hyper-planes (A, B and C) and all are ghettoizing the classes well. Now, How can we identify the right hyper-plane?

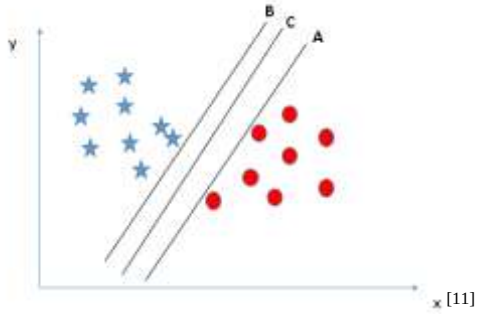


Figure 4 SVM Classification Scenario-2

Here, maximizing the distances among nearest data point (either class) and hyper-plane will help us to choose the right hyper-plane. This distance is called as Margin. Let's look at the below snapshot:

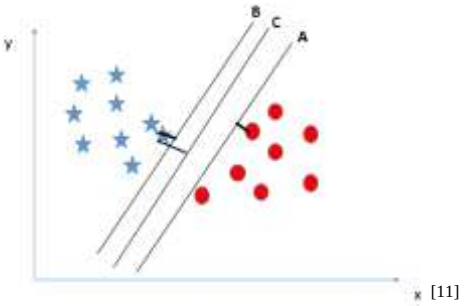


Figure 5 SVM Classification Scenario-2:Margin

Above, you can see that the margin for hyper-plane C is high as equated to both A and B. Hence, we name the right hyper-plane as C. Additional lightning purpose for choosing the hyper-plane with higher margin is robustness. If we choice a hyper-plane having squat margin then there is great chance of miss-classification.

SVM Pros

- It works really fine with perfect margin of parting
- It works effectively in high dimensional spaces.
- It is effective in cases where number of dimensions is larger than the number of models.
- It uses a subgroup of training points in the decision function (called support vectors), so it is also memory efficient.

SVM Cons

- It doesn't perform well, when we have huge data set because higher training time is required.
- It also doesn't accomplish very well, when the data set has further noise i.e. target classes are overlapping
- SVM does not offer probability estimates directly,In these method five-fold cross-validation has performed. It is related to SVC methodology.

DECISION TREE

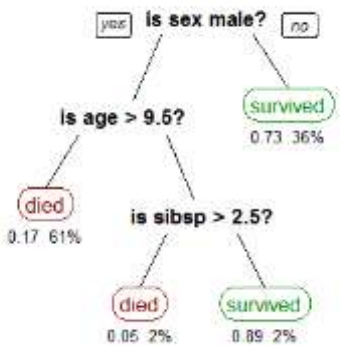


Figure 6 A tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf.

A decision tree is a simple depiction for classifying samples. Here Solitary aim is classification.In this methodology, assume that all of the input features have finite distinct domains. Each component of the domain of the classification is called a class. In decision tree which also known as classification tree, every interior (non-leaf) node is categorized with an input feature. The arcs are labeled with all of the possible values of the output feature or the arc leads to a subordinate decision node on a diverse input feature. Each leaf of the tree is labeled with a class or a likelihood scattering over the classes [12].

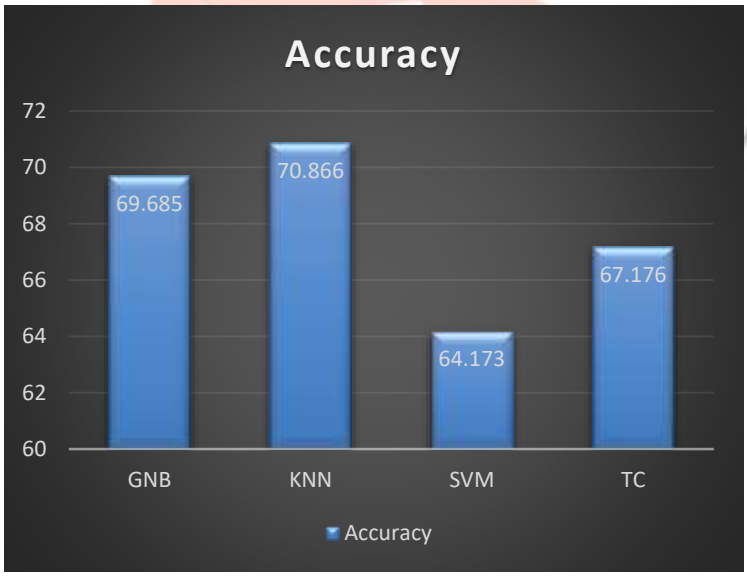
Algorithm for Decision Tree Induction

The decision tree induction algorithm works on the concept of recursively, by choosing the finest quality to divide the data and expanding the leaf nodes of the tree until the ending condition is met. The choice of finest split test condition is determined by equating the impurity of child nodes. Afterward construction of the decision tree, a tree-prunning phase can be accomplished to reduce the size of decision tree. Decision trees that are too huge are inclined to a phenomenon identified as overfitting. Pruning supports by trimming the branches of the initial tree in a way that increases the interpretation capability of the decision tree. [13]

VI. PERFORMANCE STUDY OF ALGORITHM

Algorithms	Accuracy	Error rate
Gaussian Naive Bayes	69.685	0.33
KNN	70.866	0.34
SVM	64.173	0.29
Decision tree :	67.176	0.28

Table 1 Accuracy of different Algorithms



Graph 1 camparision of Accuracy of different Algorithms

VII. CONCLUSION

Data mining and machine learning algorithms in the medical field extracts different hidden patterns from the medical data. They can be used for the analysis of important clinical parameters, prediction of various diseases, forecasting tasks in medicine, extraction of medical knowledge, therapy planning support and patient management. A number of algorithms were proposed for the prediction and diagnosis of diabetes. These algorithms provide more accuracy than the available traditional systems.
We tried and optimize every algorithm and we found out KNN algorithm best suitable for over application

VIII. REFERENCES

[1] https://www.britannica.com/technology/data-mining
[2] http://www.diabetes.ca/about-diabetes/types-of-diabetes
[3] http://www.diabetes.org/diabetes-basics/type

[4] http://www.diabetes.ca/diabetes-and-you/living-with-gestational-diabetes]
[5] http://www.diabetes.ca/diabetes-and-you/complications
[6] P. Radha, Dr. B. Srinivasan "Predicting Diabetes by cosequencing the various Data Mining
[7] Classification Techniques" IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014.
[8] https://www.papermasters.com/diabetes.htm
[9] http://www.zentut.com/data-mining/data-mining-techniques
[10] http://www.zentut.com/data-mining/data-mining-techniques
[11] http://www.statsoft.com/textbook/naive-bayes-classifier
[12] https://www.analyticsvidhya.com/blog/2015
[13] https://en.wikipedia.org/wiki/Decision_tree_learning
[14] http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/lguo/decisionTree.html