International Journal of Advanced Scientific Technologies, Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) Volume.3,Special Issue.1,March.2017

A NOVEL TECHNIQUE TO PREDICT DIABETIC DISEASE USING DATA MINING – CLASSIFICATION TECHNIQUES

G. Krishnaveni*, Prof. T.Sudha@

**** Research scholar, Dept. of Computer Science, SPMVV University, Tirupathi.***

@ Professor, Dept. of Computer Science, SPMVV University, Tirupathi.

***ABSTRACT:** Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. The software of data mining is an analytical tool for analyzing data. Data mining has become a main strategy in many industries to improve outputs and decrease costs. Now days in healthcare management this field will become very useful. Data mining techniques has became great potential for the healthcare industry to predict health deceases by using systematic data and analytics to identify inefficiencies and best practices that improve care and reduce costs. These techniques are fast in nature and take less time for the prediction system to improve the diabetic decease with more accuracy. In this paper we are applying the various classification techniques over diabetic mellitus decease dataset for the prediction of decease and non decease person. The diabetic database is preprocessed to make the mining process more efficient. The preprocessed data is used to predict using classification algorithms like Discriminent analysis, KNN, Naïve Bayes and Support vector machine. These classifiers can be efficiently used in bioinformatics problem. We are analyzing the various classification techniques like Discriminent analysis, KNN, Naïve Bayes and Support vector machine with linear and RBF kernel function and showing their accuracy.*

***Keywords:** Diabetic data, Discriminent analysis, KNN, Naïve Bayes and Support Vector Machine.*

I. INTRODUCTION

Data mining is process of extracting knowledge from huge amount of databases. Data mining is useful mostly in exploratory analysis because of nontrivial information in huge amounts of data. The techniques of data mining are useful for predicting the various deceases in healthcare departments. Now a days decease prediction plays an important role in data mining. There are different types of decease predicting in data mining namely heart disease, lung cancer, breast cancer and diabetic. This paper analyzes the diabetic decease predictions using classification algorithms. Diabetes mellitus now a day's became a major health problem. Diabetes is a decease where the body could not produce insulin or not using produced insulin properly. The insulin dependent diabetes mellitus (IDDM) is a chronic decease that appears when hormone insulin has not been produced enough in patient's a body. When not enough insulin is produced pancreas cannot control blood glucose level and causes the glucose level increase in blood which causes various deceases like retinopathies and nephropathies. Normally the diabetes decease can be divided into two classes, type 1 which is called insulin dependent and another one is type 2 which is called insulin independent. Type 1 can be seen in children of less age group whereas type 2 normally seen in the adult people. This decease id increasing day by day according to International diabetes Federation, there currently 246 million diabetic people worldwide, and this number is expected to rise to 380 million by 2025. Introducing data mining into medical analysis are to increase diagnostic accuracy, to reduce costs and to save human resources. In this context, various classification techniques are used to predict the onset of diabetes mellitus in Pima Diabetic dataset and showed that which technique is better approach by using their accuracy. In this paper we obtained more than 75% accuracy. In this paper , various techniques are implemented for the forecasting of Diabetes and concluded with best forecasting techniques which has a maximum accuracy. Implemented techniques are listed below :

1. Discriminant analysis.
2. KNN Algorithm
3. Naïve Bayes
4. Support Vector Machine with
5. SVM with Linear Kernel function.
6. SVM with RBF Kernel function.

II.THE DATASET

The dataset which we used for our work is Pima Indian Diabetic Dataset. This is a popular dataset from National Institute of Diabetes and Digestive and Kidney Deceases. 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.

Number of Instances: 768

Number of Attributes: 8 plus class

For Each Attribute: (all numeric-valued)

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 (No or Yes)

Missing Attribute Values: Yes

Class Distribution: (Class value 'yes' is interpreted as "tested positive for diabetes" and class value 'no' is interpreted as "tested negative for diabetes")

Table 1: Class Distribution

Class Value	Number of instances
No	500
Yes	268

Table 2 : Statistical Analysis of Dataset

Para meter No	Mean	Standard Deviation	Min value	Max Value
1	3.845052	3.369578063	0	17
2	120.8945	31.9726182	0	199
3	69.10547	19.35580717	0	122
4	20.53646	15.95221757	0	99
5	79.79948	115.2440024	0	846
6	31.99258	7.88416032	0	67.1
7	0.471876	0.331328595	0.78	2.42
8	33.24089	11.76023154	21	81

III. MATERIALS AND METHODS

A. *Related Work*

1. Smith et al, investigated the diagnostic, binary- valued variable whether the patient shows signs of diabetes according to World Health Organization criteria(i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care). The population lives near Phoenix, Arizona, USA.
2. Han et al. implemented a classifier on PIMA dataset by decision tree that was formed with Rapid Miner.
3. Jayalakshmi et al. designed a system that was applied to PIMA dataset for classification aim. The system made us of Artificial Neural Network for classification.
4. Patilet et al. produced association rules for PIMA dataset .
5. Arora et al. used UCI database for both classification and comparison of the classification methods they used. They made use of 5 different dataset (including PIMA) from UCI and applied j48 and Multilayer Perception (MLP) for classification and comparison aims.

B. *MATLAB*

Matlab is matrix laboratory. It is a multi paradigm numerical computing environment & fourth generation programming language. A proprietary programming language developed by Math Works, Matlab allows matrix manipulations, plotting of functions & data, implementation of algorithm, creation of user interfaces, and interfacing with programs written in other languages c, c++, java, Fortran & Python etc.

IV. DATA PREPROCESSING

Data preprocessing is the first step in data mining. In Data preprocessing the misclassified data is removed. In this process Cleaning and filtering of the data is carried out with respect to the data and data mining algorithm employed so as to avoid the creation of deceptive or inappropriate rules or patterns. In preprocessing first it selects an attribute for selecting a subset of attributes with good predicting capability. It handles all missing values and investigates each possibility. If an attribute has more than 5% missing values then the records should not be

deleted and it is advisable to impute values where data is missing, using a suitable method.

V. TRAINING AND TEST DATA SELECTION

In this paper, we used cross validation for dividing dataset into training and testing data. It is an inherent part of machine learning. In this context we use holdout validation. We partition the data into training set and test set. The training set will be used to train the model parameters. Then the trained model is used to make prediction on the test set. Predicted values will be compared with actual data to compute the confusion matrix. Confusion matrix is one way to visualize the performance of a machine learning technique. In this paper, we will hold 40% of the data, selected randomly, for test phase.

VI. K NEAREST NEIGHBOR ALGORITHM

In K Nearest Neighbor algorithm object is assigned to the class which is most common in its neighbors. This algorithm is mathematical computational algorithm and is used for binary classification i.e. 0 & 1. It works best where the data has exactly two output classes. The input consists of the k closest training examples in the feature space. In KNN classification, the output is a class member. An object is classified by the majority of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

Confusion matrix and Accuracy: A confusion matrix is a best tool to analyzing the accuracy of classifier. Structure of confusion matrix is given below:

	C1	C2
C1	True positive	False negative
C2	False positive	True negative

True Positive (TP) :- Refers to positive tuples that were correctly labeled by classifier.

True Negative (TN) :- Refers to negative tuples that were correctly labeled by classifier.

False Positive (FP) :- Refers to the negative tuples that were incorrectly labeled by classifier.

False Negative (FN) :- Refers to the positive tuples that were incorrectly labeled by classifier.

Accuracy :- (TP + TN) / (TP+TN+FP+FN)

Confusion matrix of KNN algorithm :-

TP	FP	FN	TN
81.6092	18.3908	48.1481	51.8519

Accuracy:- 100 * (81.6092 + 51.8519) / (81.6092 + 51.8519+18.3908 + 48.1481) = 71.1359

VII. DISCRIMINANT ANALYSIS

Discriminant analysis is a classification problem, where two or more groups or clusters or populations are known a priori and one or more new observations are classified into one of the known populations based on the measured characteristics. Discriminant analysis is carried out in 7 steps:

1. Collect training data: The data with known labels. Here, we actually known which tuple belong to which category. For example in diabetic data, we actually know who are diabetic and who non diabetic are.
2. Prior probabilities: The prior probability p_i represents the expected portion of the community that belongs to population π_i. There are three common choices:
   a. Equal priors
   b. Arbitrary priors
   c. Estimated priors
3. Determine variance-covariance matrices are homogeneous for the two or more populations involved by using Bartlett's test. The result of this test will determine whether to use Linear or quadratic discriminate analysis.
4. Estimate the parameters of the conditional probability density functions $f(\mathbf{X} \mid \pi_i)$. Here we shall make the following standard assumptions :
   a. The data from group *i* has common mean vector μ_i
   b. The data from group *i* has common variance-covariance matrix Σ.
   c. Independence: The subjects are independently sampled.
   d. Normality: The data are multivariate normally distributed.
5. Compute discriminate functions. This is the rule for classification of the new object into one of the known populations.
6. Use cross validation to estimate misclassification probabilities.
7. Classify observation with known group memberships.

Confusion matrix and Accuracy:

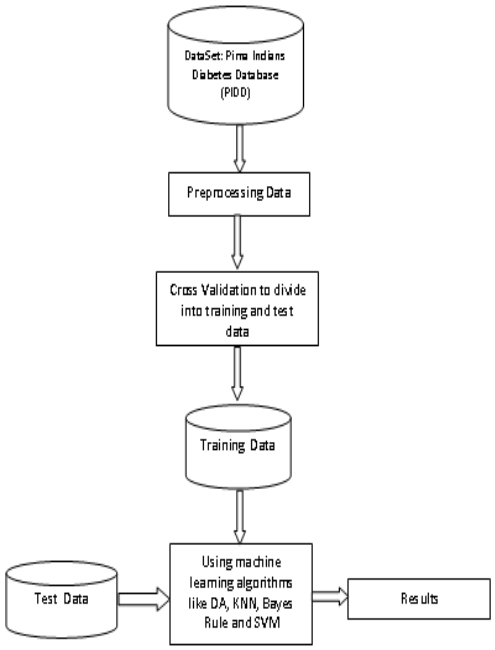

Figure 1: Architecture of Proposed System

TP	FP	FN	TN
88.5057	11.4943	46.0317	53.9683

Accuracy: 100 * (88.5057 +53.9683) / (88.5057 +53.9683+11.4943 +46.0317) = 76.3501

VIII. NAIVE BAYES ALGORITHM

Naïve Bayes classifier is mainly suitable when the dimensionality of the inputs is high. Due to its simplicity, Bayes can offer outclass more refined classification methods. This model recognizes the characteristics of patients with heart decease. It is the foundation for many machine-learning and data mining methods. Naive Bayes algorithm considers each of the feature to contribute independently to the probability that the person has a heart disease. Naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature. This algorithm is used to create models with predictive capabilities.

Confusion matrix and Accuracy:

TP	FP	FN	TN
87.0690	12.9310	43.9153	56.0847

Accuracy: 100 * (87.0690 +56.0847) / (87.0690 +56.0847+12.9310 + 43.9153) = 76.1639

IX. SUPPORT VECTOR MACHINE ALGORITHM

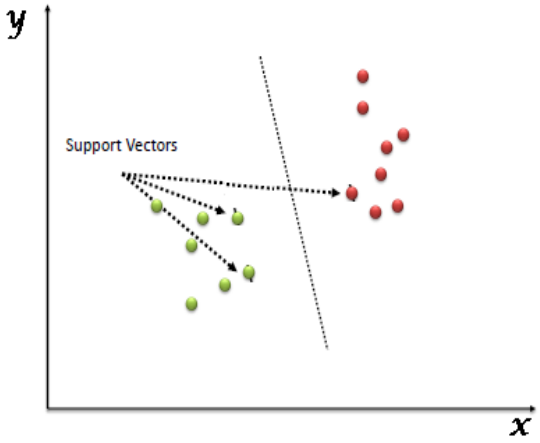

Figure 2: Support Vector Machine

SVM is related to statistical learning theory. SVM was first introduced in 1992. "Support Vector Machine" (SVM) is a supervised machine learning algorithm. SVMs are mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper- plane that differentiates the two classes very well. SVM is kernel-based technique which is major development in machine learning algorithms. These are extension to nonlinear model of the generalized algorithm developed by Valdimir Vapnik.

Support Vectors: Support Vectors are simply the co-ordinates of individual observation. Support vectors are data points that lie on the hyper plane. They are the most difficult to classify.

SVM: A new generation of learning algorithm

Pre 1980:- In this era almost all machine learning methods are linear models

1980's:- Decision trees and NNs are efficient learning of nonlinear decision surfaces. But there is a problem that little theoretical basis and all suffer from local minima.

1990's:- Efficient learning algorithm for non-linear functions based on computational learning theory. Efficiency in this era is using kernel functions for separating non-linear data into linear data.

Identify the right hyper plane:- In figure we have three hyper-planes (H1, H2 and H3), The goal is to identify the

right hyper plane to classify two classes in this case stars and circles.

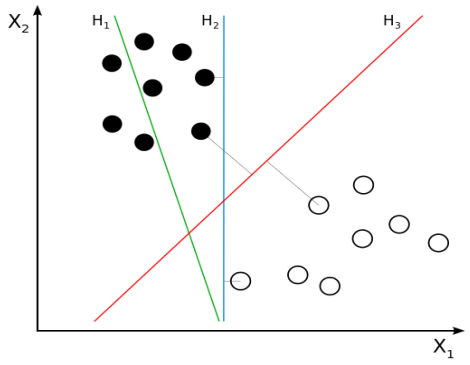

Figure 3 : H1 does not separate the classes.
H2 does but only with small margin.
H3 separate them with the maximum margin

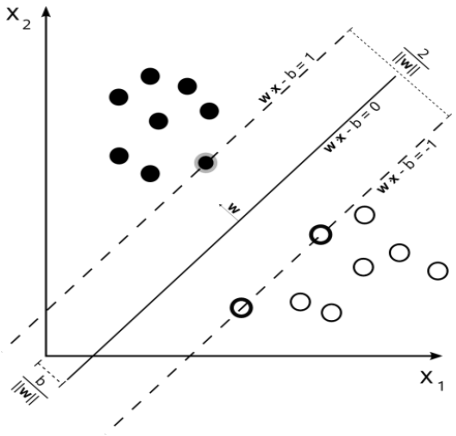

Figure 4: Maximum hyperplane and margins
for an SVMtrained with samples from two classes

Maximizing the distances between nearest data points (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as Margin.

Linearly Separable Binary Classification:

We have L training points, where each input xi has D attributes (i.e. is of dimensionality D) and is in one of two classes yi = -1 or +1, i.e our training data is of the form:

$$\{x_i, y_i\} where\ i = 1 \dots l, y_i \in \{-1, 1\}, x \in RD$$

Here we assume the data is linearly separable, meaning that we can draw a line on a graph of x1 vs x_2 separating the two classes when D=2 and a hyperplane on the graphs $x_1, x_2, \dots x_D$ for D>2

$$x_i . w + b \geq +1\ for\ y_i = +1 \qquad (1)$$

$$x_i . w + b \leq +1\ for\ y_i = -1 \qquad (2)$$

These equations can be combined into:

$$y_i(x_i . w + b) - 1 \geq 0\ \forall i$$

Kernel Functions:- A kernel is a similarity function. Kernel function is the domain expert, provide to a machine learning algorithm. It takes two inputs and spits out how similar they are. The main important thing is computing the feature vector corresponding to the kernel. Many of the machines learning algorithms are written to only use dot products and then replace the dot products with kernels. If not for ability to use kernel functions directly, we stuck with low dimensional and low performance feature vectors. This trick is called kernel trick.

1. Kernel function is not meant for to transforms single feature vector into a higher dimensional feature vector. Thus f(x) = [x, x^2] is not a kernel. It is simply a new feature vector. You do not need kernels to do this. You need kernels if you want to do this, or more complicated feature transformations without blowing up dimensionality.
2. A kernel is not only restricted to SVMs. Other learning algorithm which works with dot products can be written down using kernels. The idea of SVMs is beautiful, the kernel trick is beautiful, and convex optimization is beautiful, and they stand quite independent.

Some kernel functions available from the existing literature.

Linear Kernel :- The Linear kernel is the simplest kernel function. It is given by the inner product <x,y> plus an optional constant C. Kernel algorithms using a linear kernel are often equivalent to their non-kernel counterparts,

$$\mathbf{k(x, y) = x^T y + c}$$

Polynomial Kernel:- The Polynomial kernel is a non-stationary kernel and these kernels are well suited for problems where all the training data is normalized.

$$\mathbf{k(x, y) = (\alpha x^T y + c)^d}$$

Adjustable parameters are the slope alpha, the constant term c and the polynomial degree d.

Gaussian Kernel:- The Gaussian kernel is an example of radial basis function kernel.

$$k(x, y) = \exp(-\frac{||x - y||^2}{2\sigma^2})$$

Alternatively, it could also be implemented using

$$k(x, y) = \exp(-\gamma||x - y||^2)$$

The adjustable parameter sigma plays a major role in the performance of the kernel, and should be carefully tuned to the problem at hand. If overestimated, the exponential will behave almost linearly and the higher-dimensional projection will start to lose its non-linear power. In the other hand, if underestimated, the function will lack regularization and the decision boundary will be highly sensitive to noise in training data.

Exponential Kernel: The exponential kernel is closely related to the Gaussian kernel, with only the square of the norm left out. It is also a radial basis function kernel.

$$k(x, y) = \exp(-\frac{||x - y||}{2\sigma^2})$$

Confusion Matrix and Accuracy for Linear Kernel Function:-

TP	FP	FN	TN
85.6322	14.3678	47.0899	52.9101

Accuracy:- 100 * (85.6322 +52.9101) / (85.6322 +52.9101+14.3678 + 47.0899) = 74.1155

Confusion Matrix and Accuracy for RBF Kernel Function:-

TP	FP	FN	TN
77.5862	22.4138	32.2751	67.7249

Accuracy:- 100 * (77.8562 +67.7249) / (77.5862 +67.7249+22.4138 + 32.2751) = 74.1155

Visualization and Accuracy of Classifiers:-

X. PERFORMANCE AND EVALUATION

Performance of model can be evaluated various performance measures: classification accuracy, sensitivity and specificity. These measures are evaluated using true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

Actual Vs. Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Where TP=True Positive, FP=False Positive, TN=True Negative, FN=False Negative

These measures showed that The Discriminate analysis and Naïve Bayes can give accuracy of 76.35% an 76.16 80% , Which is the best accuracy. Mostly the accuracy result of the Bayes classifier is placed between 71% to 74% depending upon the number of cross validation applied on the dataset when performing the test.

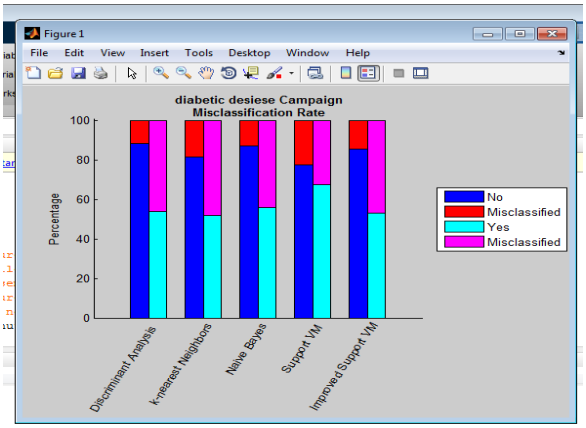

Figure 5: Visualization and Accuracy of Classifiers

XI. CONCLUSION

Naives Bayes is more productive than other classifiers. Thus this article introduces a successful Diabetic mellitus Diagnosing technique which helps to predict the disease that can finally decrease the manual work. We began with observing the symptoms as it are very difficult to predict diabetes mellitus decease finding symptoms. In the second step we preprocess the diabetic database to make the mining process more efficient. Finally, the results are compared with the help of different prediction classifiers Discriminent analysis, KNN, Naïve Bayes and Support vector machine. The last results were compared using different performance measures. These measure used true positive (TP), true negative (TN), false positive (FP) and false negative (FN) to calculate results. Performances of our technique were measured by Accuracy: 74.1155%. The proposed approach has demonstrated that mining helps to retrieve useful correlation even from attributes which are not direct indicators of the class we are trying to predict. Besides these information analysis results can be utilized for further research as a part of upgrading the accuracy of the prediction system in future.

REFERENCES

[1] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," International Journal of Medical Informatics, vol.77, pp. 81-97, 2008.

[2] Abidsarvwar, Vinod Sharma, "Intelligent Naïve Bayes Approach to Diagnose Diabetes Type-2" Special Issue of International Journal of Computer Applications on Issues and Challenges in Networking, Intelligence and Computing Technologies ,November 2012.

[3] SonuKumari, Archana Singh, "A data mining approach for the diagnosis of diabetes mellitus" ,IEEE conference on intelligent systems and control ,pp-373-375,2013.

[4] RaksehMotka,ViralParmar, "Diabetes Mellitus Forecast Using Different Data mining Techniques IEEE International Conference on Computer and Communication Technology (ICCCT),2013.

[5] VeenaVijayan V., AswathyRavikumar, "Study of Data Mining algorithms for Prediction and Diagnosis of Diabetes Mellitus", International Journal of Computer Application, Vol 94,pp .12-16,June 2014

[6] GunasekarThangarasu, P.D.D. Dominic, "Prediction of HiddenKnowledge from Clinical Database using Data mining Techniques" International Conference on Computer and Information Sciences (ICCOINS), .1-5,2014.

[7] JefriJuniferPangaribuan ,Suharjito "Diagnosis of Diabetes MellitusUsing Extreme Learning Machine ".pp-33-38,2014