Application of Data mining methods in the Diagnosis of UCI based Pima Indians Diabetes Database

Ankit Srivastava
Mtech (CSE)
Department of Computer
Science and Engineering
RKDF College of Engineering
Bhopal, India
srivastava.chaman@gmail.com

Prof. (Dr.) Mohit Gangwar
Dean Engineering, Bhabha
University, Bhopal. India
mohitgangwar@gmail.com

Asst. Prof. Manish Rai
Department of Computer
Science and Engineering
RKDF College of Engineering
Bhopal, India
manishrai2587@gmail.com

Abstract:

Data mining is respectively defined as "the non-trivial extraction of implicit, previously unknown, and potentially useful information from data" and "the science of extracting useful information from large data sets or databases". It is in particular used to help people to discover understanding from the set of rough statistics or patterns for analysis. Due to the fact the data mining techniques principally within the subject of knowledge discovery in databases (KDD) several researchers treat data mining and KDD as synonyms. Diabetes mellitus can lead long term damage to human body. The long term damage is mostly known as diabetic complications. Diabetes is directly linked to high blood pressure and it's also contributes to high cholesterol which cause high rate of heart attacks and cardiovascular disease. It's required with the strategy for computationally extricating hidden knowledge structures portrayed in models and patterns from gigantic data vaults.

Keywords: *Data mining, KDD, Mellitus, Naïve Bayes, Decision Trees, k Nearest Neighbors (k-NN)*

I.INTRODUCTION

In the machine learning research community lot of work has been done to solve the classification problem. The Pima Indian Diabetes Dataset is used to test the classification performance of the machine learning methods. This dataset was originally donated by Vincent Sigillito, Applied physics Laboratory, John Hopkins university, Laurel, MD 20707. It was selected from a larger database held by the national Institute of diabetes and digestive and kidney diseases. It is publicly available in the machine learning dataset UCI. This data set was obtained from the UCI Repository of Machine Learning Databases [1]. All patients represented in this dataset are females at least 21 years of Pima Indian heritage living near Phoenix, Arizona, USA. This dataset contains 8 input variables and a single output variable called class. The class value 1 means the patient is tested positive for diabetes and 0 means tested negative for diabetes disease. PIMA Indian Diabetes Dataset from UCI repository contains 768 instances.

II. PIMA INDIAN DIABETS DATASET

Table1.PIMA Dataset Description.

Sno	Attribute	Type
1	Number of times pregnant	Numeric
2	Plasma glucose concentration	Numeric
3	Blood pressure(Diastolic)	Numeric
4	Triceps skin fold thickness(mm)	Numeric
5	2-Hour serum insulin	Numeric
6	Body mass index(kg/m2)	Numeric
7	Diabetes pedigree function	Numeric
8	Age (years)	Numeric
9	Class Variable (True or False)	Nominal

III.LITERATURE REVIEW

In this section different literature survey approach which works towards mining disease parameters towards diabetes using Data mining approaches. In this review describe how the different algorithms and perform upon the dataset.

There have been extensive studies of this dataset in the Machine Learning Literature. Various classification algorithms have been applied to the data set, and no algorithm performs exceptionally well.

In this paper [2] Diabetes mellitus is one of the most serious health challenges facing American Natives in the United States today. The publicly available Pima Indian diabetic database (PIDD) at the UCI Machine Learning Lab has become a standard for testing data mining algorithms to see their accuracy in predicting diabetic status from the 8 variables given. In this study we will try to predict the presence of diabetes based on ensemble of SVM and BP NN. The predictive accuracy was 88.04 which was the best accuracy and it was very promising with regard to the other classification systems in the literature for this problem.

In this paper [3] Diabetes is one of the leading causes of death, disability and economic loss throughout the world. Type 2 diabetes is more common (90-95% worldwide) type of diabetes. However, it can be prevented or delayed by taking the right care and interventions which indeed an early diagnosis. There has been much advancement in the field of various Machines learning algorithms specifically for medical diagnosis. But due to partially complete medical data sets, accuracy often decreases, results in more number of misclassification that can lead to harmful complications. An accurate prediction and diagnosis of a disease becomes a challenging research problem for many researchers. Therefore, aimed to improve the diagnosis accuracy we have proposed a new methodology, based on novel preprocessing techniques, and K-nearest neighbor classifier. The effectiveness of the proposed methodology is validated with the help of various quantitative metrics and a comparative analysis, with previously reported studies using the same UCI dataset focusing on pima-diabetes disease diagnosis. This is the first work of its kind, where

100% classification accuracy is achieved by feature reduction from eight to two that shows the out performance of the proposed methodology over existing methods.

In this paper [4] Parashar A. et al. (2014) have proposed Linear Discriminant Investigation and Support Vector Machine for the conclusion of Pima Indians Diabetes dataset, where LDA diminishes include subsets and SVM is capable to classify the data. They have likewise contrasted SVM and feed forward neural system (FFNN) yet our proposed SVM+LDA gives better order precision as 77.60% with 2 features.

In this paper [5] Healthcare industry contains very large and sensitive data and needs to be handled very carefully. Diabetes Mellitus is one of the growing extremely fatal diseases all over the world. Medical professionals want a reliable prediction system to diagnose Diabetes. Different machine learning techniques are useful for examining the data from diverse perspectives and synopsizing it into valuable information. The accessibility and availability of huge amounts of data will be able to provide us useful knowledge if certain data mining techniques are applied on it. The main goal is to determine new patterns and then to interpret these patterns to deliver significant and useful information for the users. Diabetes contributes to heart disease, kidney disease, nerve damage and blindness. So mining the diabetes data in efficient way is a crucial concern. The data mining techniques and methods will be discovered to find the appropriate approaches and techniques for efficient classification of Diabetes dataset and in extracting valuable patterns. In this study a medical bioinformatics analyses has been accomplished to predict the diabetes. The Pima Indian diabetes database was acquired from UCI repository used for analysis. The dataset was studied and analyzed to build effective model that predict and diagnoses the diabetes disease. In this study we aim to apply the bootstrapping resampling technique to enhance the accuracy and then applying Naïve Bayes, Decision Tree and k Nearest Neighbors (kNN) and compare their performance.

IV.PROPOSED ARCHITECTURE

In view of the problem statement described in the introduction section, we propose a classification model with boosted accuracy to predict the diabetic patient. In this model, we have employed. Different classifiers like Decision Trees, KNN and Naïve Bayes. The major focus is to increase he accuracy by using resample technique on a benchmark well renowned diabetes dataset that was acquired from PIMA Indian Diabetes Dataset from UCI machine learning repository, which consists of eight attributes. To improve the quality of the results obtained after mining and the effectiveness of the complete mining process, data preprocessing is done [6].

The framework is composed of the following important phases:

- ➢ Dataset Selection (PIMA Indian Diabetes Dataset)
- ➢ Data Preprocessing
- ➢ Feature extraction through principle component analysis (PCA)
- ➢ Applying Resample filter
- ➢ Learning by Classifier (Training) i.e. Naïve Bayes, KNN and Decision Trees
- ➢ Achieving trained model with highest accuracy
- ➢ Using trained model for prediction

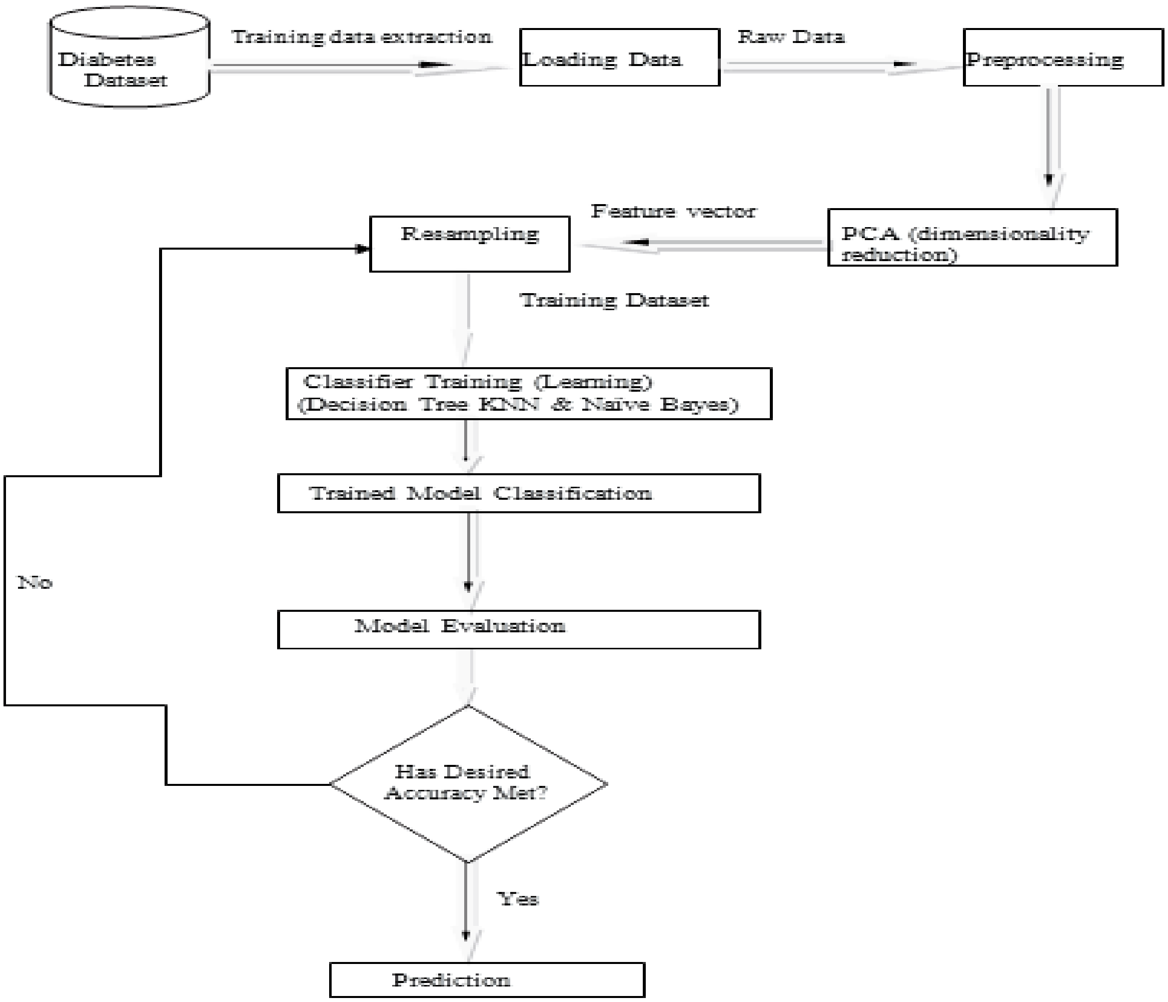

Diabetes Dataset
Training data extraction
Loading Data
Raw Data
Preprocessing
PCA (dimensionality reduction)
Feature vector
Resampling
Training Dataset
Classifier Training (Learning) (Decision Tree KNN & Naïve Bayes)
Trained Model Classification
Model Evaluation
Has Desired Accuracy Met?
No
Yes
Prediction

Figure 1 the proposed framework is shown

Thus in order to proposed a better prediction model using classification and further combine approaches requirement is to further acquire an scheme which contribute on getting better outcome and system, here our proposed methodology a Sensing is utilize scheme in place of traditional hybrid approach.

In Various systems we have observed the different classification algorithms and also we have performed different outlier technique to perform classification, where we have some techniques in which there are deficiency in detection rate and accuracy, also some algorithm having less precision and high recall value which is not suitable as a best outlier practice. Also while dealing with the large categorical dataset a heavy computation and thus complexity is required to make it furnished.

V.PROPOSED METHODOLOGY

A classifier is a tool in machine learning that proceeds a group of data demonstrating the objects we need to classify and tries to forecast which class the new data belongs to. The classification objective set for this study is to achieve enhanced accuracy by using Naïve Bayes, Decision Trees and KNN classifiers and determine which one suits the most for diabetes classification technique. The classifiers we are selected to use in this study are ranked among the top ten best classifiers especially k nearest neighbor and decision trees. The techniques used are Naïve Bayes, J48, J48graft and IBK. These classifiers are selected on the bases of their strengths described below and also due to their frequent use in previous research studies.

1. Naïve Bayes

Naïve Bayes is a data mining classification technique and it is used as a classifier. This classifier is used for probability prediction if a sample belongs to particular class. The quality of Naïve Bayes is high accuracy and fastest to train data. It is usually used on very large datasets. The Naïve Bayes Algorithm is a probabilistic algorithm that is sequential, following steps of execution, classification, estimation and prediction. There are various data mining existing solution for finding relations between the diseases, symptoms and medications, but these algorithms have their own limitations; numerous iterations, high computational time and binning of the continuous arguments etc. Naïve Bayes overcomes various limitations and can be applied on a large dataset in real time. Naive Bayes is very popular in commercial and open-source anti-spam e-mail filters.

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem (from Bayesian statistics) with strong (naive) independence assumptions. [7] An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters necessary for classification.

2. Decision Trees

A "divide-and-conquer" approach to the problem of learning from a set of independent instances lead naturally to a style of representation called as decision trees. The objective of a decision tree classifier is to build a classifier model that classifies the attributes based on the other attribute values. In decision trees, classification process starts from the root node and is sorted using the attribute values. Murthy (1998) gave an outline about decision trees and their efficacy in the field of machine learning. Constructing an optimal binary tree is an NP problem and many researchers have searched for an efficient heuristic technique.

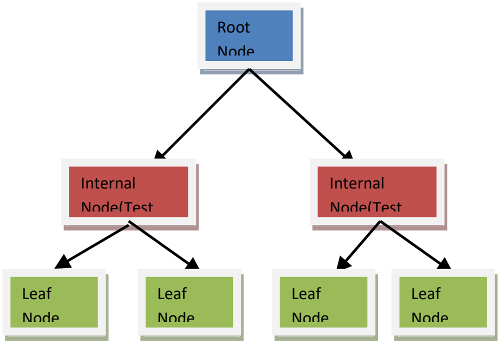


Figure 2 sample decision tree

3. k Nearest Neighbors (k-NN):

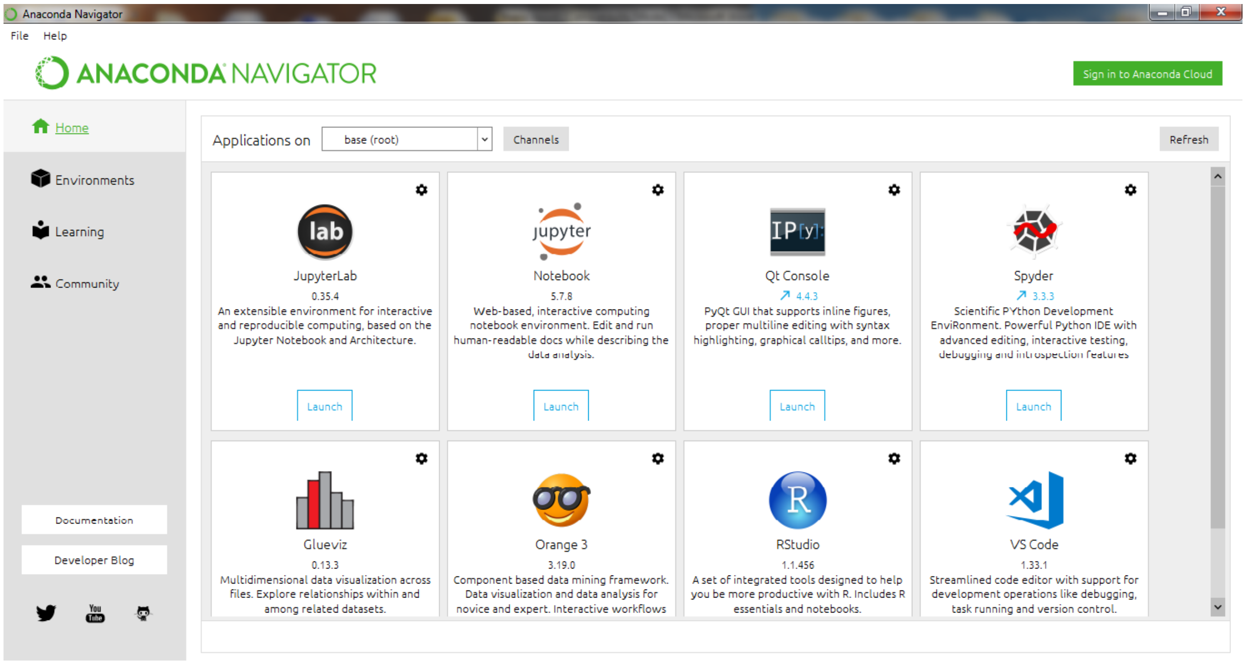
k-NN is a very simple data mining technique and use for classification. k-NN is a sort of instance-based learning, also referred as lazy learning, which basically aims with estimating the function locally and all computation is postponed until classification. It can be beneficial to allocate weight to the contributions of the neighbors, so as to the closer neighbors contribute more to the average than those who are reside more far-away. The distance is mostly measured by using Euclidean distance formula. Here k is static value and mostly it takes an odd value like 1, 3 and 5.

K folds cross validation technique is used for training

data. This technique is mostly used in circumstances wherever the aim is prediction, and we wish to evaluate how a predictive model in practice will perform especially in terms of accuracy. In the prediction problem, a model is generally fed with a dataset that contains known data instances on which training is done (training dataset), as well as a dataset of anonymous data against which the model is being tested so called testing dataset. This technique is used to assess predictive models by dividing the original sample dataset into a training set that is used ahead to train the model, and a test set on which it did testing to evaluate it.

VI.EXPERIMENT & RESULT EVALUATION

In the figure 3 below, a Anaconda 3 initialization is presented. This figure shows the start page of the Anaconda 3, which we have used for analysis. Now next step shows processing done with Anaconda 3 and parameters generator by tool and their analysis is performed.



An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Multidimensional data visualization across files. Explore relationships within and among related datasets.

Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows

A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.

Streamlined code editor with support for development operations like debugging, task running and version control.

Figure 3 Anaconda 3 initialization is presented

Computed Result Comparison:

According to the simulation execution with three major classification techniques, a comparison table is drawn.

Table 2. Comparison of all classifiers performance

Classifier	TP	FN	FP	TN	Accuracy	Precision	Recall	Mean Absolute Error
Naïve Bayes	107	38	39	112	75.90	74.65	75.17	.0249
J48	132	13	4	157	95.54	97.78	92.36	0.045
JGragt	132	13	4	157	95.45	97.50	92.80	.044

k-NN =1	132	13	6	155	94.75	96.45	91.250	.016
k-NN =3	113	32	39	122	77.50	75.16	79.31	.098

Graphical Result Analysis:

An analysis of result graphically is discussed which help in understanding the observe parameter and their graphical monitoring. The comparison of performance of different classifiers is also shown in the graphs below.

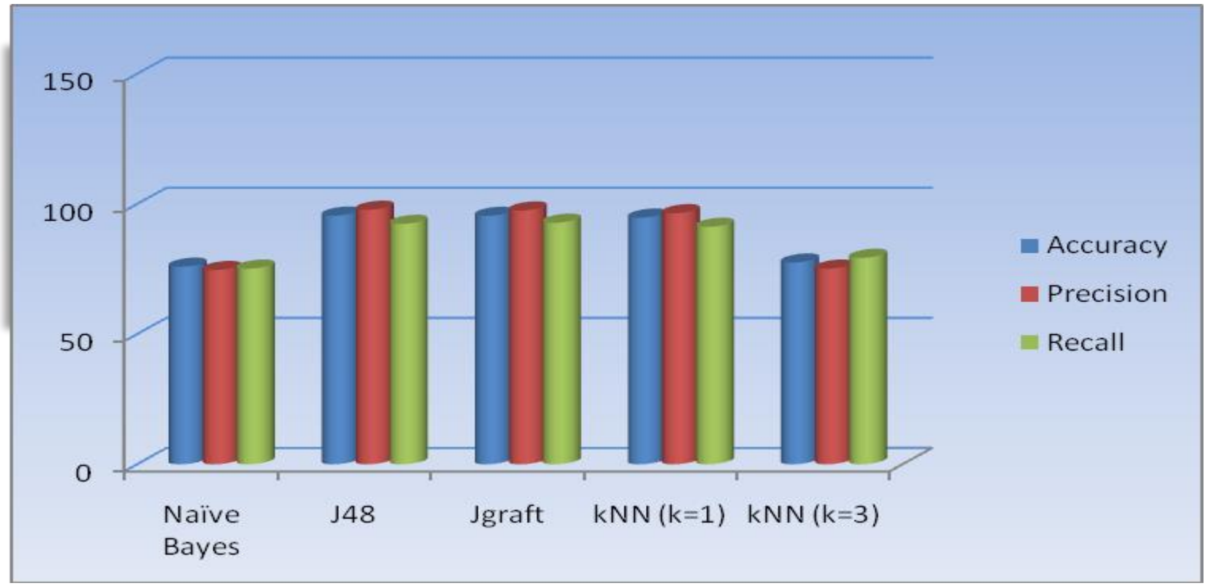


Figure 4 Accuracy comparison graph

VII.CONCLUSION & FUTURE WORK

In our dissertation work an extension of previous work is performed which classify the PIMA Indian Diabetes Dataset . The dataset has been acquired from UCI machine learning repository database. The dataset consists of 768 total instances and nine attributes. A lot of principle definition is takes for classification and apply the standards over dataset. Its classification, defining set of rules and further finding an exact common factor which reveals the diabetes disease. We upgrade the exactness by improving the data in preprocessing stage that truly functions admirably. It is likewise reasoned that the exactness of a model is very reliant on the dataset. In this way, this procedure works very well on PIMA diabetic dataset in any case, may not ensured similar outcomes on an alternate dataset.

In future work incorporates it is plan to utilize further developed classifiers, for example, artificial neural networks (ANN), genetic algorithm (GA) and evolutionary algorithm (EA). A further analysis can be done with real time dataset and finding a usable platform for the user to get utilizing of proposed system for diabetes. A further work will get approach over breaking down more classification algorithm with existing algorithms. In this way, the patient can be cautioned to change their ways of life.

VIII.REFERENCES

[1]*http://ftp.ics.uci.edu/pub/ml-repos/machine-learning databases/pima-Indians-diabetes, 2003.*

[2]*Rahmat Zolfaghari(2012). Diagnosis of Diabetes in Female Population of Pima Indian Heritage with Ensemble of BP Neural Network and SVM. IJCEM International Journal of Computational Engineering & Management, Vol. 15 Issue 4, July 2012 ISSN (Online): 2230-7893 www.IJCEM.org*

[3]*[3]Madhuri Panwar , Amit Acharyya , Rishad A. Shafik and Dwaipayan Biswas(2016). K-Nearest Neighbor Based Methodology for Accurate Diagnosis of Diabetes Mellitus. 2016 Sixth International Symposium on Embedded Computing and System Design (ISED) 978-1-5090-2541-1/16/$31.00 © 2016 IEEE*

[4] *Parashar A., Burse K., Rawat K. (2014). A Comparative Approach for Pima Indians Diabetes Diagnosis using LDA-Support Vector Machine and Feed Forward Neural Network. International Journal of Advanced Research in Computer Science and Software Engineering. Vol. 4, pp. 378-383, ISSN: 2277 128X.*

[5] *Uswa Ali Zia, Dr. Naeem Khan (2017). Predicting Diabetes in Medical Datasets Using Machine Learning Techniques. International Journal of Scientific & Engineering Research Volume 8, Issue 5, May-2017 ISSN 2229-5518.*

[6] *Farahmandian M., Lotfi Y., Maleki I. (2015). Data Mining Algorithms Application in Diabetes Diseases Diagnosis: A Case Study. MAGNT Research Report. Vol. 3, PP. 989-997, ISSN. 1444-8939.*

[7]*Kevin P. Murphy, " Naïve Bayes classifier", Department of Computer Science, University of British Columbia,2006.*