Predicting Diabetes in Medical Datasets Using Machine Learning Techniques

Uswa Ali Zia, Dr. Naeem Khan

Abstract-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 WEKA software was employed as mining tool for diagnosing 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 Trees and k Nearest Neighbors (kNN) and compare their performance.

Index Terms- Healthcare, Diabetes, Classification, K-nearest neighbours, Decision Trees, Naive Bayes.

1. INTRODUCTION

Computers have brought substantial improvements to technology that lead to the production of massive volumes of data. Additionally, the advancements and innovations in the healthcare database management systems generate a huge number of medical databases. Healthcare industry contains very large and sensitive data. This data needs to be treated very carefully to get benefitted from it. There is need to develop some more accurate and efficient predictive models that helps in diagnosing a disease although it was revealed that diabetes mellitus is the diseases which becomes one of the global hazard. Diabetic Mellitus is a set of associated diseases in which the human body is unable to control the quantity of sugar in the blood. It is a group of metabolic diseases which results in high blood sugar level, may be as the body does not produce sufficient insulin, or may because cells do not react to the produced insulin. This disease becomes a global hazard and will increasing rapidly so it is estimated that almost sixty million people from all over the world will be effected by diabetics in 2025. Hence there it is needed to analyses the already available huge diabetic data sets to discover some incredible facts which may help in producing some prediction model.

The focus is to develop the prediction models by using certain machine learning algorithms. The machine learning is a sort of artificial intelligence that enables the computers to learn without being explicitly programmed. Machine learning emphases on the development of computer programs that can teach themselves to change and grow when disclosed to new or unseen data. Machine learning algorithms are mostly categorized as being supervised or unsupervised. A supervised learning algorithm uses the past experience to make predictions on new or unseen data while unsupervised algorithms can draw inferences from datasets. The supervised learning is also called classification.This study uses classification technique to produce a more accurate predictive model as it is one of themost commonly applied machine learning technique that examines the training data and creates an inferred function, which can be used for mapping new or unseen examples. The major goal of the classification technique is to forecast the target class accurately for each case in the data. Classification Algorithms generally require that the classes be defined grounded on the data attribute values. They often define these classes by looking at the characteristics of data already known to belong to class. This process of finding useful information and patterns in data is also called Knowledge Discovery in Databases (KDD) which involves certain phases like Data selection, Pre-processing, Transformation, Classification and Evaluation.

Before applying any classification algorithm it is necessary to prepare or preprocess the acquired original dataset to enhance the performance of a classifier. Besides managing the noise and dealing with the missing value, there is a common issue in the real environment datasets that the target class values are not equal or are not balanced. Several real world application for example medical diagnoses, fraud detection, network interruption detection,

fault monitoring,detection of pollution, biomedical, bioinformatics and remote sensing suffer from these phenomena. This disorder is known as class imbalance. Class imbalance problem recently becoming a hot issueand being examinedby machine learning and data mining researchers. Besides other major challenges faced by machine learning and data mining fields, class imbalance is also among one of these challenges. Imbalance data sets reduces the performance of data mining and machine learning techniques and also affect on the total accuracy and decision making as beinginclined to the majority class, which lead to misclassifying the minority class samples or may handle them as noise.This affects prediction accuracy of the classifier. The prediction accuracy in medical datasets is generally low while using conventional classification techniques without applying additional preprocessing or data preparation techniques. One of the solutions is resample for dealing with class imbalance problem. It is a preprocessing method that handles the imbalance problem by creatingalmost balanced training data set and adjusting the preceding distribution for both minority and majority class. Sampling methods compriseof under sampling, over sampling and sometimes hybrid techniques. Under sampling approach will balance the data by eliminatingsamples from majority class whereas the over sampling method will balance the data by creating theduplicates of the present samples or by adding new samples to the minority class.Resample is one such technique which ensures selection of same sizes of class instances for each type of class labels.Therefore we consider resample as one approach to enhance classification accuracy.

In this study we have applied bootstrapping method which is a statistical re-sampling technique that allows to randomly replacing different set of data points within a dataset, and hence results in higher accuracy. Resampling methods useby computer to produce a huge amount of simulated samples. Patterns in these samples are then summarized and evaluated. The strengths of using bootstrap resampling technique are that each sample must have an equal probability of being selected. The simulated samples take full advantage of the information in the sample. Resampling is suggested to be done with replacement. This technique will be simpler and more accurate, needs less assumption, and have better generalizability. Resampling gives particularly rich advantages where expectations of traditional parametric tests are not met, as with minor samples from non-normal distributions.

Therefore this technique will help equalizing the minority classes as it aims at obtaining the same size of data points for each class. The efficiency of different classification techniques would be then evaluated to suggest the suitable choice. The classification algorithms have been applied to the PIMA Indians Diabetes Dataset of National Institute of Diabetes and Digestive and Kidney Diseases that contains the data of female diabetic patients.

2. LITERATURE REVIEW

Yasodha*et al*.[1] uses the classification on diverse types of datasets that can be accomplished to decide if a person is diabetic or not. The diabetic patient's data set is established by gathering data from hospital warehouse which contains two hundred and forty nine instances with seven attributes. These instances of this dataset are referring to two groups i.e. blood tests and urine tests. In this study the implementation can be done by using WEKA to classify the data and the data is assessed by means of 10-fold cross validation approach, as it performs very well on small datasets, and the outcomes are compared. The naïve Bayes, J48, REP Tree and Random Tree are used. It was concluded that J48 works best showing an accuracy of 60.2% among others.

Aiswarya*et al*. [2] aims to discover solutions to detect the diabetes by investigating and examining the patterns originate in the data via classification analysis by using Decision Tree and Naïve Bayes algorithms. The research hopes to propose a faster and more efficient method of identifying the disease that will help in well-timed cure of the patients. Using PIMA dataset and cross validation approach the study concluded that J48 algorithm gives an accuracy rate of 74.8% while the naïve Bayes gives an accuracy of 79.5% by using 70:30 split.

Gupta *et al.* [3] aims to find and calculate the accuracy, sensitivity and specificity percentage of numerous classification methods and also tried to compare and analyse the results of several classification methods in WEKA, the study compares the performance of same classifiers when implemented on some other tools which includes Rapidminer and Matlabusing the same parameters (i.e. accuracy, sensitivity and specificity). They applied JRIP, Jgraft and BayesNet algorithms. The result shows that Jgraft shows highest accuracy i.e 81.3%, sensitivity is 59.7% and specificity is 81.4%. It was also concluded that WEKA works best than Matlab and Rapidminner.

Lee *et al.* [4] focus on applying a decision tree algorithm named as CART on the diabetes dataset after applying the resample filter over the data. The author emphasis on the class imbalance problem and the need to handle this problem before applying any algorithm to achieve better accuracy rates. The class imbalance is a mostly occur in a

dataset having dichotomous values, which means that the class variable have two possible outcomes and can be handled easily if observed earlier in data preprocessing stage and will help in boosting the accuracy of the predictive model.The study illustrates the effect of resampling means in field of medical the dataset used inthis study was acquired from the National Health and Nutrition Examination Survey (NHANES) 2009–2010. The attributes of the dataset includes glucose (fasting and non-fasting) and body mass index.On this data the researcher built some decision tree models to forecast undiagnosed diabetes among adults. The Centers for Disease Control and Prevention declared that the occurrence of diagnosed and undiagnosed diabetes are about6.0% and 2.3%, respectively and results in its large burden to the social order, to identify undiagnosed diabetes for improved decision-making of health care providersefforts were dedicated.Classification and Regression Tree (CART) being a recursive partitioning method aim at excruciating the data into different parts based on the maximum significant exposure variables carried out by this method. The tool used for experimentation is R software. The data was splinted into ratio of 70:30. Finally the maximum accuracy achieved by this study is 67%.

Chikh*et al.* [5] used enhanced AIRS2 called MAIRS2 to increase the diagnostic accuracy of diabetes diseases. K-nearest neighbors algorithm swap with the fuzzy K-nearest neighbors to enhance the diagnostic accuracy of diabetes diseases. The diabetes dataset acquired from UCI machine learning repository. The authors attained a good tradeoff between classification accuracy and data reduction. The propose system (MAIRS2) that performed better than classical AIRS2. The authors achieved highest classification accuracy by MAIRS2 is 89.10%.

Sharmila*et al.* [6] aims to analyze the data in predicting thediabetes from medical record of the patients. The study states that approximately 40 million Indianssuffer from diabetes till now. his. This study is analysing the diabetes from huge medical records by using decision trees with statisticalimplication using R tool .R is a sequential programming language for the analysis, graphics and software development activities for datamining and in various fields. The datasets were collected from Chennai to analyze having ten attributes (i.e. pregnant, LDL, post prandial HDL, BMI, HBAIC, age, creatinine, family) and a class variable. There are four possible outcomes i.e. either the patient is positive fordiabetes, pre-diabetes, gestationaldiabetesand non-diabetic. The csv file are loaded into R. after the prepressing the decision tree algorithm is applied to predict all the four possible diabetics outcomes as defined above and produces the results. The R tool analyses datasets in 748.54 seconds. This study uses R tool which is quite effective, extensible and having comprehensiveenvironment for statistical computing and graphics. Another important feature of R is that it supports a variety of file formats (XML, binary files, CSV) and also user created R packages. The study also uses decision trees for the reason that they are easy to understand, economical to construct, easy to incorporate with database system and is relatively accurate in several applications. In this study a thorough analysis of the diabetic datasets was done efficiently with the help of R. this information which was discovered from this study can also be used to build efficient prediction models.

Sadhana*et al.* [7] emphasis on the need to analyse the already available huge diabetic data sets to analyzedso to discover some vital facts which may help in producing some prediction model. Besides using the data mining techniques (as previously used) this study is going to uses Hadoop, hive and R for analysing the datasets. The datasets were taken from Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases. Total eight attributes (no. of pregnancies, glucose plasma concentration, blood pressure, serum insulin, body mass index, age, diabetes pedigree and skin fold depth) were used to produce the result as a patient being effected by diabetes when output is 1 and not whenindicated 0. The raw csv file is injected to hive as input where these datasets were analysed on the basis of these attributes. The output of hive is given to R as input which performs statistical analyses along with producing the graphs. The basic benefit of Hive is that it acts as a data warehouse solution that constructed at top of Hadoop. The results produced are highly efficient as hive has analysed seven hundred and sixty eight datasets in just 19 seconds. The graphs

generated by R can help to understand the outcomes in a simpler manner. The study claims that a prediction model should be developed by using such graphs or information.

Gowsalya*et al.*[8] aims to propose a system with the ability of forecasting the risk of readmission of diabetic patients within coming 30 days and can accomplish this take with help of MapReduce technique. This risk factor obtained will aids the physicians in suggesting suitable care for the patients.The study presents solution which usesHadoopMapReduce to analyse huge datasets and mine valuable observations from the dataset that aids in assigning the resourceseffectively. For new patients, this system makes use of the information of the prior patients with similar illnesses and reuses those suggestions. The system collects the data straight from the patients (body sensors) and their associated doctors. This data is then stored on Hadoop Distributed File System (HDFS) and MapReduce technique is applied by HDFS. Analysis is performed on the datasets with information of hospital admission, diabetic encounter, laboratory tests, medications, time to stay in the hospital. The rate of the readmission is calculated on the features like age, HbAIC result and modification in prescription. Haemoglobin A I C (HbAIC) is an considered important factor as a measure of glucose control, which is mostly results to measure of diabetes. The likelihood of getting readmitted is high if the value is greater than 8%.The use of distributed file system for the development of this proposed systemuses inexpensive present hardware and stores data across nodes. This predictive system helps hospitals and other health care organizations to assign clinicians, nurses, machinery, and other resources in a better way.

Eswari*et al.*[9]focuses on a prediction model by analyzingthe algorithm in Hadoop/Map Reduce environment to predict the widespreadtypes of diabetes and the related problems and also the treatment. The suggesteddesign of predictive analysis system is constructed on numerous levels e.g. data collection, warehousing, predictive analysis, processing analysed reports. The system analysis by working in Hadoop/Map Reduce setting to categorize the type of diabetics, its problems and the type of treatment suggested for such patients. The suggested system uses Hadoop as the open-source distributed data processing platform. Hadoophas the ability toplay both functionsofa data manager as well as analytics tool.Big Data Analytics in Hadoop'sapplication gives an organized approach for attaining improved results such as availability and affordability of healthcare facility to population as this research ambitions to deals with the study of curing diabetes in medical industry via the big data analytics.

Salian*et al.* [10] explains that analysing the big data will help in predicting the risk of diabetic patient's readmission efficiently by determining the risk predictors that can be a reason of readmission of diabetic patients. The study suggested a predictive model that can find the patients with chronic diabetes diseases and are most likely to be get admitted again and again. In the suggested system works by loading the raw data is loaded into the Hadoop File System (HDFS) firstly and then by using Hive queries, all the nominated predictive variables are recovered into a comprehensible dataset to

use formodeling. And then model works by selecting and applying various classifications, prediction method using Hadoop. The accurateness of the results was checked by confusion matrix. The top five readmission predictors in diabetic dataset are body mass index,plasma glucose, age, pregnant, pedigree function are top predictors in the proposed model. This study shows thatthe risk of readmission for diabetes patients can be evaluated by big data analytics. Predictive modeling has been worked by applying decision tree classification method. The chance of readmission in diabetic patient is successfully predicted by this proposed model.

Raghupathi*et al.*[11] defines the potential and possibilities of big data analytics in healthcare.Along through the potential of big data analytics the study also highlighted several challenges to address.The analysis of big data in the health care sector results in cost reduction and quality treatment to the patients, further benefits includes to identify those individuals who would be benefitted from anticipatory care or by changing their routine in a proactive manner; outlining the broad scale disease to support prevention initiatives; gathering and issuing data on medical actions, identifying, predicting and dropping fraud by applying advanced analytic systems for fraud recognition and checking the correctness and stability of claims. Several challenges are also highlighted which includes governance issues including ownership, security, privacy have however to be addressed. By overcoming the existing limitation as defined above will help in more fast progress in analysing the big data in healthcare.

Hay *et al.* [12] tries to maps the geographical areas where there is a greater chance of an infectious disease to be occurred and those areas where the chances are relatively low. The analysis is based upon the environmental factors such as temperature and rain fall. The source data will be gathered from various sources and in various formats required to be treated in real time and thus make use of big data techniques to map the surveillance of disease in real time. The input data from various sources is to be processed in real time and uses techniques (such as data mining or machine learning) to map the surveillance of disease in real time. Using data mining techniques such as machine learning and the use of multitude sourcing provides an opportunity of creating a continually or frequently updated atlas of infectious diseases. Though using big data analytics techniques it is possible to provide the risk map in real time.

Weber *et al.* [13] emphasis to identify all the diverse but useful data sourceslike social media, census records, and numerous other types of data and then link them together while taking care of the privacy and security, so as to get fully benefitted from big data.The biomedical data is distributed across different isolated areas so it is necessary to link them all to get better insights from this available data by analysing it. Although before linking data from all sources for analysis it is also necessary to distinguished between the useful sources and the irrelevant data sources. The study applies the probabilistic linkage algorithm for linking the diverse sources. This algorithm's main advantage is that the same technique is used to match the patient's crossways different electronic health records can be stretched to the data sources outside the health care.

Meredith*et al.*[14] defines the importance of big data in prevention of certain disease by continually measuring and analysing the data in real time from different sources and suggest precautions to particular individual about his/her disease while lowering the cost. Big data can assist action on the risk factors such as physical activity, nutrition, use of tobacco, andexposure to pollution. The study describes two case studies to show how big data is helpful in disease prevention. Disease prevention is based upon to identify modifiable risk factors for disease like exercise, diet, alcohol consumption, smoking and pollution get insights then lead to interventions to improve therisk factors and improve health.

Rao*et al.* [15] enlightens the security challenges related to big data with particular reference to healthcare sector. The study aimed to propose feasible security solutions to get fully benefitted from big data relating to healthcare. The study explains the necessity of big data analysis in healthcare sector to do proactive and reactive analysis of the information which will results in providing chances for

forecasting, realizing uncertain needs, and decreasing risks as along with providing tailored services. The study also proposed four security models which are data de-identification model, data centric approach to security, walled garden model and jujutsu security.Security solutions should be implemented in such a way that they should guarantee safe analytics and securing big data frameworks.

Augustine *et al.* [16] focuses on the benefits of using Hadoopas being more flexible, scalable and as a more economical solution for the analysis of big medical data (images) produce in healthcare sectors. Hadoop provides solution to analyse the medical images by combining these medical images from numerous sources and extracts the important data for accurate diagnosis. The study emphasis on the use of an interface called Hadoop Image Processing Interface (HIPI) supports the image processing as accomplished in Hadoop.

3.PROPOSED FRAMEWORK

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 the 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. The proposed framework is shown in Figure 1.

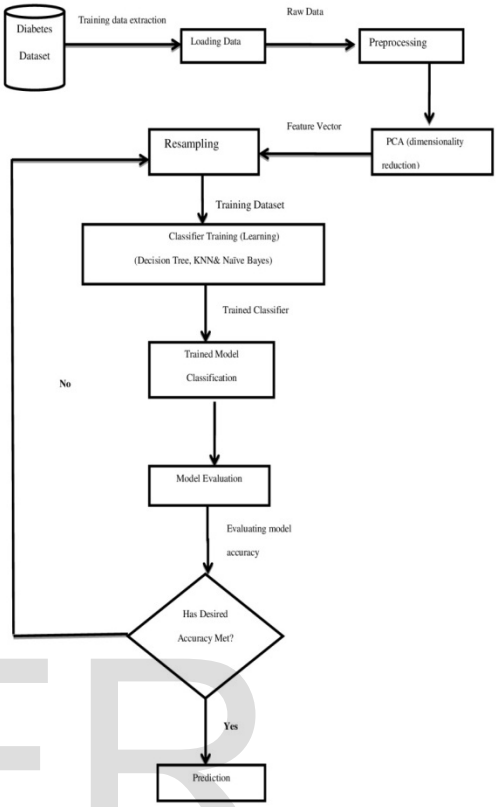

Figure 1. The proposed Classification Model for Diabetes Datasets.

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

The detail description of the components and the activities performed against each component is mentioned below.

3.1 Dataset Selection (Diabetes Dataset)

In data mining and machine learning, the data selection is a process in which the most relevant data is selected from a specific domain to derive values that are informative and facilitate learning within that domain. In the study, we have used diabetes dataset having eight attributes that are used to predict the symptom of gestational diabetes in a female patient. This dataset was obtained from UCI repository and is a benchmark dataset.On the basis of historical information stored in the dataset such as age, body mass index, blood pressure and number of times pregnant the classifiers are trained for making decision whether diabetes test for an individual is positive or negative. The PIMA diabetes dataset only represents the Indian national females who are at least 21 years old.All of the attributes are of numeric-valued continuous data type. The attribute for class label is dichotomous variable(i.e., the binary response variable) within the PIMA dataset follows each tuple of the dataset. PIMA Indian Diabetes Dataset from UCI repository contains 768 instances.The PIMA dataset is converted from CSV to ".ARFF" format accepted by WEKA 3.6.13. The complete details of all the eight attributes are listed in Table below.

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-Hourseruminsulin	Numeric
6	Body mass index(kg/m^2)	Numeric
7	Diabetes pedigree function	Numeric
8	Age (years)	Numeric
9	Class Variable (True or False)	Nominal

3.2 WEKA Tool

WEKA 3.6.13is used in this study. WEKA stands for the Waikato Environment for Knowledge Analysis and this tool is developed and distributed freely by the University of Waikato, New Zealand .WEKA is one of the most famous tool for data processing and data analysis. Since WEKA software has been written in Java language, therefore, it runs on almost every platform. It consists of variety of machine learning algorithms and is capable to solve a multitude of data mining and machine leaning problems. WEKA supports many machine learning and data mining tasks such that regression, classification, prediction, feature selection and visualization. WEKA provides a database connection to access data and manipulates it. WEKA allows us to create, run, modify and analyze experiments in more way that is suitable.The most prominent advantages of WEKA include its free availability, portability, a broad collection of data preprocessing and modeling techniques and the friendly graphical user interface makes it easy to use. WEKA performance is comparatively better than other data mining tools named TANAGRA and MATLAB. Different classification techniques show much better results on WEKA than other tools [18].

3.3 Data preprocessing

Data preprocessing is a technique of machine learning that comprises of converting raw data into an logical or comprehensible format. The real world data is mostly incomplete, inconsistent, unreliable, redundant and having missing values etc. Data preprocessing is an conventional technique of eliminating such problems which are also known as noise. Preprocessing involves certain activities like data cleaning, integrating the data, transformation of data, data reduction, data discretization and data cleaning. Here the dataset is checked for duplicate values, missing values and type miss-matches etc. All these inconsistencies are eliminated from this dataset, in the phase called data preprocessing phase. It is important to clean the dataset before training it on a classifier in order to better learn the hidden patterns in the dataset. The set of pertinent feature vector fed to the classifier help it learn more accurately in a shorter span of time.

3.4 Feature Extraction through Principle Component Analysis (PCA)

After setting the classification objectives, we apply principle component analysis (PCA) on the dataset to determine the most suitable set of attributes that can help achieve better classification. The set of attribute suggested by the PCA are termed as feature vector in this study.

Feature reduction or dimensionality reduction will benefitted us by reducing the computation and space complexity. Simple and more robust models should be developed, which are easier to understand and also saves the cost. Therefore, we applied PCA on the entire PIMA dataset within the WEKA tool. A threshold value of 0.21 is selected and all the attributes having range of greater than and equal to 0.21 is selected for further experimentation.

3.5 Resample Filter

The supervised Resample filter is applied to the preprocessed dataset. As the class attribute is of nominal data type therefore we are using supervised resample filter in WEKA, which produces a random subsample of a dataset using either by doing sampling with replacement or sampling without replacement. Re-sampling is a series of methods used to reconstruct your sample data sets, including training sets and validation sets. The original dataset must fit completely in memory. The amount of instances in the generated dataset may be identified. This filter helps to preserve the class distribution in the subsample, or to bias the class distribution to a near balanced distribution. It can provide more "useful" different sample sets for learning process. This approachis very easy to implement and fast to run.The unbalanced classes do not have the same number of instances, this is true for the experimental database .When the distribution of instances is not uniform, the resampling of the experimental database is necessary. In this study, we adopted bootstrap method of resample on the datasetwhich obtains a random sample with replacement from a sample. In order to achieve balanced classes, WEKA can use a resampling with replacement which replicates some instances within classes,whenever the classes have just a few instances. The parameters defined are set according to our requirements. This approach helps in balancing the imbalanced datasets and also gives us an enhancement in our preferred accuracy measures.

3.6 Classifiers

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 neighhour 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.

3.6.1 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 usuallyused 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.

3.6.2 Decision Trees

Decision tree is a classification technique. This technique is mostly use for prediction and classification. A tree comprises of paths, branches and leave nodes. Collection of branches is called path and represents the attribute value. Leaves represented Class value. Each path in decision tree symbolizes a rule which is used for classification or prediction. Decision tree divides the data into subsets or nodes. Root node represents the complete dataset. Tree pruning is preformed after tree is built completely. Pruning is started from the lead node.

Being particular to J48 Decision tree classifier, it works on the following simple algorithm. While classify a new item, firstlyit generate a decision tree grounded on the attribute values of the existing training data. So, each and every time it encounters a set of items (training set) it recognizes the attribute that distinguishes the numerous instances utmost

clear. Among the likely values of this feature, if there is some value for which the data instances belonginginside its category have the similar value for the target variable, then we terminate that branch and ascribe to it the target value that we have obtained.

3.6.3 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. In k-fold cross-validation, the original sample is divided at random into k equivalent size subsamples. Of these k subsamples, a particular subsample is reserved as the validation data and used for testing the model, while the k-1 remaining subsamples are utilized as training data. After that this cross-validation process is recurring k times (called the folds), with every of the k subsamples just used one time as the validation data. It works in loop manner. One benefit to use this technique is that every observation is used for both training and validation, and every single observation is utilized for validation justone time. In this study we set the value of k=10.

WEKA tool is used for training and testing (learning) model. Learning model accuracy is checked through MSE (Mean Squared Error). While training the classifier, it is important to determine that classifier has efficiently learnt from the dataset. For this purpose, mean square error (MSE) technique is widely used. The aim is to train the classifier until mean square error becomes negligible.If desired data accuracy is met then trained model will be saved otherwise preprocessing step will be performed again.

4. Experimentation and Results

For experimentation PIMA Indian diabetes dataset is used in this study. In the PIMA dataset, we have two class problems of diabetes in individual patient having tests either positive or not. The dataset has been acquired from UCI machine learning repository database. The dataset consists of 768 total instances and nine attributes, namely, Diastolic blood pressure (mm Hg), Plasma glucose concentration, Number of times pregnant, Body mass index, 2-Hour serum insulin, Triceps skin fold thickness (mm), Age (years) and Diabetes pedigree function. After preprocessing the data instances are reduced. We also applied PCA to reduce the dimensionality of dataset. By applying PCA on all the attributes, PCA returned six attributes to be used for training the classifiers. Then applying resample filter with no replacement that disables the data to be replicated. The classifiers are applied. The naïve Bayes, Decision Trees and Lazy classifiers are applied one by one on the same data. We applied these classifiers on PIMA Indian diabetes dataset. The classification results are evaluated by comparing them in terms of correctly classified and incorrectly classified instances. There are certain performs measures produced by WEKA other than accuracy, precision and recall. They include F measures and ROC area. The F measure is actually the weighted average of Precision and Recall. Hence, this measure get both false positives and false negatives into account. Naturally F measure is usually more useful than accuracy, when class distribution is uneven. It works very well when false positives and false negatives have almost same cost. If the cost of false positives and

false negatives are very dissimilar, then a healthier choice is to consider Precision and Recall rather than Accuracy. The formula for F1 Score is 2*(Recall * Precision)/(Recall + Precision). Similarly the ROC (receiver operating characteristic**)** curvere presents a graphical plot which shows the performance of a binary classifier system because its discrimination threshold is varied. The ROC curve is generated by plotting the true positive rate (TPR) and the false positive rate (FPR) at several threshold values. The term sensitivity, recall or probability of detection also indicates the true-positive rate in machine learning.

This study is limited to three performance measure that includes accuracy, precision and recall.

Accuracy is the utmost spontaneous performance measure. Itbasically deals with ratio of correctly predicted observations. It is best to measure the accuracy when the class is balanced; therefore our focus is to enhance the accuracy. The formula used to calculate the Accuracy is mentioned below:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+FP} \quad \ldots\ldots (1)$$

Precision indicates the number of True Positives divided by the number of True Positives and False Positives. Hence, it shows the number of positive predictions divided by the total number of positive class values predicted. Precision is also termed as the Positive Predictive Value (PPV). The formula is mentioned below:

$$Precision = \frac{TP}{TP+FP} \ldots\ldots (2)$$

Whereas recall indicates the number of True Positives divided by the number of True Positives and the number of False Negatives. Hence it is the number of positive predictions divided by the number of positive class values in the test data. Recall also sometimes titledas Sensitivity or the True Positive Rate.The formula is mentioned below:

$$Recall = \frac{TP}{TP+FN} \ldots\ldots (3)$$

Performances of each classifier are measured in these terms by using equation 1, 2 and 3.

The final results are shown below:

Table 2. Confusion Matrix for Naïve Bayes

		Predicted	
		Negative	Positive
Actual	Negative	True positive(TP)=107	False negative(FN)=38
	Positive	False positive(FP)=39	True negative(TN)=112

Accuracy =74.8%

Precision= TP/TP+FP*100 = 73.28%

Recall =TP/TP+FN*100 = 73.79%

Table 3. Confusion Matrix for Decision tree (J48)

		Predicted	
		Negative	Positive
Actual	Negative	True positive(TP)=132	False negative(FN)=13
	Positive	False positive(FP)=4	True negative(TN)=157

Accuracy =94.44%

Precision= TP/TP+FP*100 =97.05%

Recall =TP/TP+FN*100 = 91.03%

Table 4.Confusion Matrix for Decision tree (J48Graft)

		Predicted	
		Negative	Positive
Actual	Negative	True positive(TP)=132	False negative(FN)=13
	Positive	False positive(FP)=4	True negative(TN)=157

Accuracy =94.4% , precision =97%, Recall =91.3%

Table 5. Confusion Matrix for KNN (k=1)

		Predicted	
		Negative	Positive
Actual	Negative	True positive(TP)=132	False negative(FN)=13
	Positive	False positive(FP)=6	True negative(TN)=155

Accuracy =93.79%, Precision=95.65%, Recall = 91.03%

Table 6. Confusion Matrix for KNN (k=3)

		Predicted	
		Negative	Positive
Actual	Negative	True positive(TP)=113	False negative(FN)=32
	Positive	False positive(FP)=39	True negative(TN)=122

Accuracy =76.79%

Precision= TP/TP+FP*100 =74.34%

Recall =TP/TP+FN*100 = 77.93%

Table 7. Comparison of all classifiers performance

Classifier	TP	FN	FP	TN	Accuracy %	Precision	Recall	Mean Absolute Error
Naïve Bayes	107	38	39	112	74.84	73.28	73.79	0.249
J48	132	13	4	157	94.44	97.05	91.03	0.045
JGragt	132	13	4	157	94.44	97.05	91.03	0.044
k-NN =1	132	13	6	155	93.79	95.65	91.03	0.016
k-NN =3	113	32	39	122	76.79	74.34	77.93	0.098

The comparison of performance of different classifiers is also shown in the graphs below.

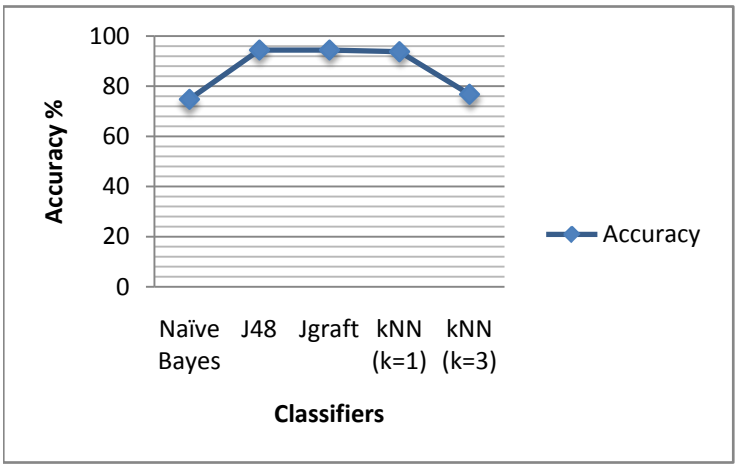

Figure 2. Accuracy comparison graph

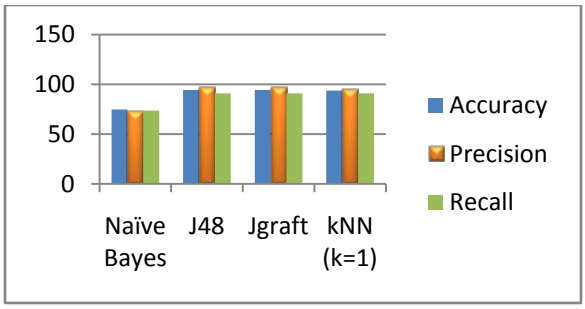

Figure 3. Comparison of all performance measures

5. COMPARISON OF RESULTS

We compared the results achieved in this study with the results reported by other researcher in the existing literature. We mainly focused on the method used and the accuracy achieved by the other studies. A comparison of our framework with other studies is provided in Table 8.

Table 8: Results Comparison Table

Reference	**Proposed model / Method**	**Dataset Used**	**Purpose**	**Accuracy Achieved (%)**
N. Gupta *et al*. (2013)	Decision Tree	PIMA Indian Diabetes Dataset	To predict diabetes	81.33%
P.Yasodha , M. Kannan (2011)	Bayes Net	A hospital repository	To predict diabetes	66.2%
A. Iyer*et al*. (2015)	Decision Tree	PIMA Indian Diabetes Dataset	To predict diabetes	74.8%
K. Rajesh, V. Sangeetha (2012)	Decision Tree	PIMA Indian Diabetes Dataset	To predict diabetes	87%
Lee (2014)	Decision Tree	National Health and Nutrition Examination Survey	To predict diabetes	67%
Chick *et al.* (2012)	k-NN	PIMA Indian Diabetes Dataset	To predict diabetes	89.10%
Our proposed framework	Decision Trees Naïve Bayes kNN(k=1) kNN(k=3)	PIMA Indian Diabetes Dataset	To improve diabetes prediction	94.44% 74.89% 93.79% 76.79%

In this study, we are used classification algorithms Naïve Bayes, Decision Trees and kNN for prediction diabetes. The result obtained from this study is compare with the similar study of other authors. From the comparison table we have notice the decision trees work better than others. The decision tree algorithms i.e. J48 and Jgraft outperforms over other classifiers and previous studies. It achieves the highest accuracy rate of 94.44%. The decision tree is simple and good classifier for prediction diabetes. A comparison of the accuracy produced by all the classifiers before applying resampling and theaccuracy produced by them after applying resampling is given below:

Table 9: comparison before and after applying Resampling

Classifiers	**Without Bootstrapping (Accuracy Rates%)**	**After Bootstrapping (Accuracy Rates%)**
Naïve Bayes	71.45%	74.89%
Decision Tree (J48)	78.43%	94.44%
Decision Tree (J48graft)	78.43%	94.44%
k-NN (k=1)	69.93%	93.79%
k-NN (k=3)	72.22%	76.79%

6.CONCLUSION AND FUTURE WORK

Data mining plays an important role in various fields such as artificial intelligence (AI) and machine learning (ML), statistics and database systems. The core objective of this study is to enhance the accuracy of predictive model. The accuracy can be increase by improving the performance of the data, the algorithms or even by algorithm tuning. We enhance the accuracy by improving the data in preprocessing phase that really works well. Applying bootstrapping resampling technique on this PIMA dataset will increases the accuracy of almost all classifiers but the decision trees leads over others. It is also concluded that the accuracy of a model is highly dependent on the dataset. So, this technique works very well on PIMA diabetic dataset but may not guaranteed the same results on a different dataset. In future work includes it is plan to use further

more advanced classifiers such as artificial neural networks (ANN), genetic algorithm (GA) and evolutionary algorithm (EA).

The diabetes dataset considered in this study might not consider some other important factors that are related to gestational diabetes, like metabolic syndrome, family history, habit of smoking, lazy routines, some dietary patterns etc. The appropriate prediction model would want additional relevant data to make it more accurate. This would be accomplished by gathering diabetic patient's datasets from various sources, to generate a better relevant prototype. This is a limitation of this research.

ACKNOWLEDGMENT

I would like to thank Dr. Naeem Khan for his guidance and cooperation in completing my research work. I would also be thankful for my family for their support throughout my life.

- *Uswa Ali Ziais a student of MSCS at SZABIST Islamabad. E-mail: malka_09@yahoo.com*
- *Dr. Naeem Khan is working as assistantprofessor (department of Computer Science) at SZABIST Islamabad. Email: dr.naeem@szabist-isb.pk*

References

[1]P.Yasodha and M. Kannan, "Analysis of a Population of Diabetic Patients Databases in WekaTool", *International Journal of Scientific & Engineering Research*, vol. 2, no. 5, 2011.

[2]A. Iyer, J. S and R. Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques", *IJDKP*, vol. 5, no. 1, pp. 01-14, 2015.

[3]N. NiyatiGupta,A.Rawal, and V.Narasimhan,"Accuracy, Sensitivity and Specificity Measurement of Various Classification Techniques on Healthcare Data", *IOSR Journal of Computer Engineering*, vol. 11, no. 5, pp. 70-73, 2013.

[4]P. Lee, "Resampling Methods Improve the Predictive Power of Modeling in Class-Imbalanced Datasets", *International Journal of Environmental Research and Public Health*, vol. 11, no. 9, pp. 9776-9789, 2014.

[5] M. Chikh,M. Saidi, and N. Settouti, "Diagnosis of diabetes diseases using an Artificial Immune Recognition System2 (AIRS2) with fuzzy K-nearest neighbor," *Journal of medical systems*,vol.36, no.5, pp. 2721-2729, 2012.

[6] K. Sharmila and S. Manickam, "Efficient Prediction and Classification of Diabetic Patients from bigdata using R,"*International Journal of Advanced Engineering Research and Science*, vol. 2, Sep 2015.

[7] S. Sadhana and S. Savitha, "Analysis of Diabetic Data Set Using Hive and R," *International Journal of Emerging Technology and Advanced Engineering*, vol. 4, July 2014.

[8] M. Gowsalya, K. Krushitha, and C. Valliyammai, "Predicting the risk of readmission of diabeticpatientsusing MapReduce," pp. 297–301,2014.

[9] N. M. S. kumar, T. Eswari, P. Sampath, and S. Lavanya, "Predictive methodology for diabetic dataanalysis in big data," *Procedia Computer Science*, vol. 50, pp. 203–208, 2015.

[10] S.Salian and G. Harisekaran, "Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients," *International Journal of Science and Research*, vol. 4, April 2015.

[11] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential", *Health Information Science and Systems*, vol. 2, no. 1, p. 3, 2014.

[12] S. Hay, D. George, C. Moyes and J. Brownstein, "Big Data Opportunities for Global Infectious Disease Surveillance", *PLoS Med*, vol. 10, no. 4, p. e1001413, 2013.

[13] G. Weber, K. Mandl and I. Kohane, "Finding the Missing Link for Big Biomedical Data", *JAMA*, 2014.

[14]M. Barrett, O. Humblet, R. Hiatt and N. Adler, "Big Data and Disease Prevention: From Quantified Self to
Quantified Communities", *Big Data*, vol. 1, no. 3, pp. 168-175, 2013.

[15] S. Rao, S. Suma and M. Sunitha, "Security Solutions for Big Data Analytics in Healthcare", *2015 Second*

International Conference on Advances in Computing and Communication Engineering, 2015.

[16]D. Peter Augustine, "Leveraging big data Analytics and Hadoop in developing India's

healthcare services," *International Journal of Computer Applications*, vol. 89, no. 16, pp. 44–50, 2014.

[17] S.David and A. Saeb, "Comparative Analysis of Data Mining Tools and Classification Techniques usingWEKA in Medical Bioinformatics", *Computer Engineering and Intelligent Systems*, vol.4, no.13, 2013.