International Journal of Emerging Technology and Innovative Engineering
Volume 5, Issue 4, April 2019 (ISSN: 2394 – 6598)

ANALYSIS AND PREDICTION OF DIABETES USING MACHINE LEARNING

S.Saru
UG student
Sri Krishna College of Technology,
Coimbatore

S.Subashree
UG student
Sri Krishna College of Technology,
Coimbatore

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. 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 have 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 (KNN) and compare their performance.

Keywords—Healthcare, Diabetes, Classification, K-nearest neighbors, Decision Trees, Naive Bayes.

1.INTRODUCTION

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 the most 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,Transformation, Classification and Evaluation. Several real -world application for example medical diagnoses, fraud detection, network interruption detection, fault monitoring and detection of pollution, biomedical, bioinformatics and remote sensing suffer from these phenomena. 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 SURVEY

The objective of the research is" Prediction of Diabetes by consequence the various Data Mining Classification Techniques" describes the various Data Mining Classification Techniques. There are many classification techniques used in this paper for predicting diabetes [2]. Research paper, "Disease Prediction in Data Mining Technique"– A Survey. The disease prediction plays an important role in data mining. This paper analyses about various diseases like Heart disease prediction, Breast cancer prediction, Diabetes by using many techniques like Classification, Clustering, Decision Tree, Naive Bayes methods in order to predict the diabetes disease. This paper also tells about predictive and descriptive type about the data. Prediction involves some fields in the data set to predict the values of other variables. On the other hand, Description focuses on finding patterns of the data that can be interpreted by humans. The different algorithm of data mining is used in the field of medical prediction are discussed in this paper [1]. The Research paper, "Analysis of various Data Mining Techniques to Predict Diabetes Mellitus", concentrates about overall population affected by diabetes worldwide. This paper also predicts about the overall population affected by diabetes will also double the rate of diabetes of the population by the upcoming years. This Paper aims about the early prediction of the diabetes will save the life of the human. The paper analyses about the three types of diabetes and their causes. It also uses the prediction, classification technique. This provides the higher accuracy for the disease prediction [5]. The research paper, "Review on Prediction of Diabetes using Data Mining Technique", elaborates about detailed review of existing data mining methods used for prediction of diabetes. It also gives about the types of diabetes disease Type1, type2, and type3.The aim of the diabetes is to predict the diabetes with the help of Data mining methods such as the K-Nearest Neighbor Algorithm, Bayesian Classifier, Naive Bayesian Classifier, Bayesian Network, all the methods are used for prediction of diabetes. This paper also mentions about the effects of diabetes on patients [7]. The research paper, "A Survey on Naive Bayes Algorithm for Diabetes Data Set Problems", explores about various Data mining algorithm approaches of data mining that have been utilized for diabetic disease prediction. In this paper Classification and Naive Bayes is one of the most used algorithms for the prediction of disease [3]. The research paper," Prediction of Diabetes Mellitus Techniques", describes about the Decision Tree, Naive Bayes, K nearest neighbors algorithm (k-NN), Classification and Clustering. By using this effective algorithm methods diabetes prediction can be done [4]. The research paper, "A Comparative Study of Classification Algorithms for Disease Prediction in Health Care. This paper describes about Diseases Prediction; Classification algorithm; Data Mining, Decision tree. The main aim of this paper is to find out best classifier from different classification algorithm that can be used to predict disease on applying data set of the patients [6].

3.PROPOSED METHODOLOGY

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 framework is composed of the following important phases: • Dataset Selection (PIMA Indian Diabetes Dataset) • Data Pre -processing • Feature extraction through principal component analysis

(PCA) • Applying Resample filter • Learning by Classifier (Training) i.e. Naïve Bayes, KNN and Decision Trees • Achieving trained model with highest accuracy.

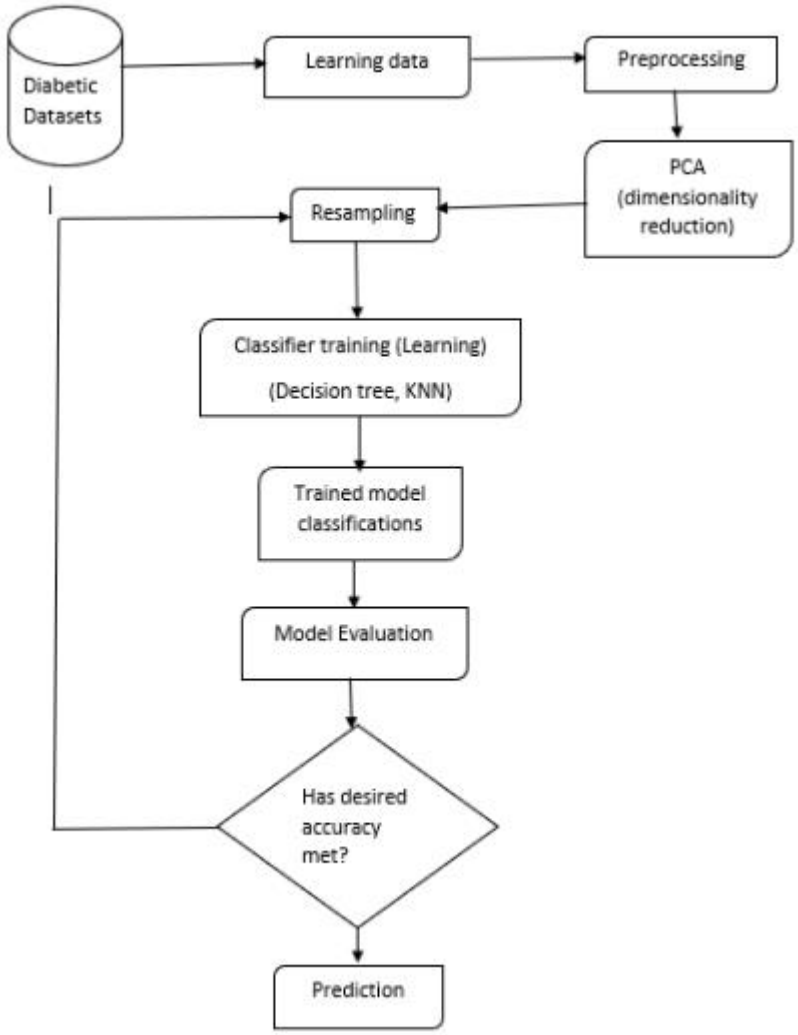

Fig 1. Process Flow

4.CLASSIFICATION OF ALGORITHM

RANDOM FOREST

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees habit of overfitting to their training set.

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set X = x1, …., and with responses Y = y1, …., in bagging, repeatedly (B times) selects a random sample with replacement o the training set and fits trees to these samples:

For b = 1, …., B:

1.Sample, with replacement, n training examples from X, Y, call these Xu, By.

2.Train a classification or regression tree fb on Xu, By.

After training, predictions for unseen samples *x′* can be made by averaging the predictions from all the individual regression trees on *x′*:

f ^ = 1 B ∑ b = 1 B f b (x ′) {\displaystyle {\hat {f}}={\frac {1}{B}}\sum _{b=1}^{B}f_{b}(x')}

or by taking the majority vote in the case of classification tree.

K- NEAREST NEIGHBORS

In pattern recognition, the k- nearest neighbours' algorithm (k-nan) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression.

- · In KNN classification, the output is a class membership. An object is classified by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small). If k=1, then the object is simply assigned to the class of that single nearest neighbour.
- · In KNN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbours.
- · KNN is a type of instance-based learning or lazy learning, where the function is only approximated locally and computation is deferred until classification. The KNN algorithm is among the simplest of all machine learning algorithms.

DECISION TREE

A decision tree is a decision support tool that uses a tree-like model of their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Decision trees are commonly used in operation research, specifically in decision analysis, to help identify a strategy most likely to reach goal, but are also a popular tool in machine learning

A decision tree consists of three types of nodes:

1. Decision nodes – typically represented by squares
2. Chance nodes – typically represented by circles
3. End nodes – typically represented by triangles

SUPPORT VECTOR MACHINE

In machine learning **support-vector machines** (**SVMs**, also **vector networks**) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model linear that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

LOGISTIC REGRESSION

In statistics, the **logistic model** (or **logit model**) is a widely used statistical model that, in its basic form, uses a logistic function to model a binary dependent variable; many more complex extensions exist. In regression analysis, **logistic regression** (or **logit regression**) is estimating the parameters of a logistic model; it is a form of binomial regression. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail, win/lose, alive/dead or healthy/sick; these are represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from ***log****istic un****it***, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probity model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a *constant* rate, with each dependent variable having its own parameter.

5. DATASET SELECTION (DIABETES DATASET)

S No	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/m^2)	Numeric

7.	Diabetes pedigree function	Numeric
8.	Age(years)	Numeric
9.	Class Variable (True or False)	Nominal

Table 1. Dataset Information

6.COMPARISON AND RESULTS

Reference	Proposed model / Method	Dataset Used	Purpose	Accuracy Achieved (%)
N. Gupta et.al (2013)	Decision Tree	PIMA Indian Diabetes Data	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 Data	To predict diabetes	74.8%
K.Rajesh, V.Sangeetha (2015)	Decision Tree	PIMA Indian Diabetes Data	To predict diabetes	87%
Lee (2014)	Decision Tree	National Health and Nutrition Examination Survey	To predict diabetes	67%

Chick et al (2012)	KNN	PIMA Indian Diabetes Data	To predict diabetes	89.10%
Our Proposed Framework	Decision Trees Naïve Bayes KNN(K=1) KNN(K=3)	PIMA Indian Diabetes Data	To model diabetic prediction	94.44% 79.84% 93.79% 76.79%

Table 2. Comparison of algorithms and its accuracy

Classifiers	Without Bootstrapping (Accuracy rate %)	After bootstrapping (Accuracy rate %)
Logistic regression with SVM	71.45%	74.89%
Decision tree(J48)	78.43%	94.4%
k-NN(k=1)	69.93%	93.79%
k-NN(k=3)	72.2%	76.69%

Table 3.Bootstrapping Accuracy Rate

7. CONCLUSION

There are Various data mining method and its application were studied or reviewed. Application of machine learning algorithm were applied in different medical data sets including machine Diabetes dataset. Machine learning methods have different power in different data set. We obtained 768record diabetes data set from UCI. the comparison of individual algorithm and the proposed method is done on this study. We applying 10 cross validation us for evaluation of the performance of these machine learning classification methods purpose. In this study the proposed method provides high accuracy with accuracy value of 90.36% and decision Stump provided less accuracy than other by providing 83.72% accuracy. Therefore, using ensemble method used to provide better prediction performance or accuracy than single one.

8. FUTURE WORK

In this study we concentrated only Diabetes disease for future it can be extended to apply this method in another diseases Small amount sample data used on this study.it can be apply in large amount of data for future extension .on this study also only a single data set used therefore for future multiple data set can be used for prediction .in this study only limited base classifier used .for future it is possible to use another base classifier like ANN, Naive Bayes, KNN, Random tree ,and other.

REFERENCES

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

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

[3] NIyati Gupta, 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] M. Chicheme. Said, and N. Setout, "Diagnosis of diabetes diseases using an Artificial Immune Recognition System 2 (AIRS2) with fuzzy K-nearest neighbour," Journal of medical systems, vol.36, no.5, pp. 27212729, 2012.

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

[6] 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.

[7] Sassanian and G. Hari Sekaran, "Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients," International Journal of Science and Research, vol. 4, April 2015.

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

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

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

[11] M. Barrett, O. Humbled, 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.

[12] 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.

[13] Sivaram, M., B. DurgaDevi, and J. Anne Steffi. "Steganography of two lsb bits." International Journal of Communications and Engineering 1.1 (2012): 2231-2307.

[14] Sivaram, M., et al. "Exploiting the Local Optima in Genetic Algorithm using Tabu Search." Indian Journal of Science and Technology 12 (2019): 1.

[15] Mohammed, Amin Salih, et al. "DETECTION AND REMOVAL OF BLACK HOLE ATTACK IN MOBILE AD HOC NETWORKS USING GRP PROTOCOL." International Journal of Advanced Research in Computer Science 10.6 (2018).

[16] Viswanathan, M., et al. "Security and privacy protection in cloud computing." Journal of Advanced Research in Dynamical and Control Systems (2018): 1704-1710.

[17] Nithya, S., et al. "Intelligent based IoT smart city on traffic control system using raspberry Pi and robust waste management." Journal of Advanced Research in Dynamical and Control Systems, Pages (2018): 765-770.

[18] Dhivakar, B., et al. "Statistical Score Calculation of Information Retrieval Systems using Data Fusion Technique." Computer Science and Engineering 2.5 (2012): 43-5.

[19] Mohammed, Amin Salih, Shahab Wahhab Kareem, and M. Sivaram. "Time series prediction using SRE-NAR and SRE-ADALINE." (2018): 1716-1726.

[20] Abraham, Steffin, Tana Luciya Joji, and D. Yuvaraj. "Enhancing Vehicle Safety with Drowsiness Detection and Collision Avoidance."

International Journal of Pure and Applied Mathematics 120.6 (2018): 2295-2310.

[21] Porkodi, V., et al. "Survey on White-Box Attacks and Solutions." Asian Journal of Computer Science and Technology 7.3 (2018): 28-32.

[22] Malathi, N., and M. Sivaram. "An Enhanced Scheme to Pinpoint Malicious Behavior of Nodes In Manet's." (2015).

[23] Sivaram, M. "Odd and even point crossover based Tabu ga for data fusion in Information retrieval." (2014).

[24] Sivaram, M., et al. "Emergent News Event Detection from Facebook Using Clustering."

[25] Punidha, R. "avithra K, Swathika R, and Sivaram M,"Preserving DDoS Attacks sing Node Blocking Algorithm." International Journalof Pure and Applied Mathematics, Vol. 119, o. 15, 2018." 633-640.

[26] Batri, K., and M. Sivaram. "Testing the impact of odd and even point crossover of genetic algorithm over the data fusion in information retrieval." European Journal of Scientific Research (2012).

[27] Mohamme, Sivaram Yuvaraj Amin Salih, and V. Porkodi. "Estimating the Secret Message in the Digital Image." International Journal of Computer Applications 181.36 (2019): 26-28.

[28] Manikandan, V., et al. "PRIVACY PRESERVING DATA MINING USING THRESHOLD BASED FUZZY CMEANS CLUSTERING." ICTACT Journal on Soft Computing 9.1 (2018).

[29] Obulatha-II-ME-CSE, Miss O. "Position Privacy Using LocX."

[30] Sivaram, M., et al. "The Real Problem Through a Selection Making an Algorithm that Minimizes the Computational Complexity."

[31] Sivaram, M., et al. "DETECTION OF ACCURATE FACIAL DETECTION USING HYBRID DEEP CONVOLUTIONAL RECURRENT NEURAL NETWORK."

[32] V. Porkodi, Dr.D. Yuvaraj, Dr. Amin Salih Mohammed, V. Manikandan and Dr.M. Sivaram. "Prolong the Network Lifespan of WirelessSensor Network" (2018): 2034-2038.

[33] M, Sivaram, ENABLING ANONYMOUS ENDORSEMENT IN CLOUDS WITH DECENTRALIZED ACCESS CONTROL (March 13, 2019). Available at SSRN: https://ssrn.com/abstract=

[34] M, Sivaram, INTEGER WAVELET TRANSFORM BASED APPROACH FOR HIGH ROBUSTNESS OF AUDIO SIGNAL TRANSMISSION (March 13, 2019). Available at SSRN: https://ssrn.com/abstract=

[35] M, Sivaram, HEALTHCARE VISIBLE LIGHT COMMUNICATION (March 13, 2019). Available at SSRN: https://ssrn.com/abstract=

[36] M, Sivaram, Preserving DDoS Attacks Using Node Blocking Algorithm (March 13, 2019). Available at SSRN: https://ssrn.com/abstract=