A Critical Study on the Performance Evaluation of various Feature Selection and Classification algorithms using Diabetes Disease Dataset

Varun Dhanalakota[1*], Vikas B[2]

[1*]Computer Science and Engineering, GITAM Institute of Technology, GITAM, Visakhapatnam
[2] Computer Science and Engineering, GITAM Institute of Technology, GITAM, Visakhapatnam
(e-mail: varun.dhanalakota@gmail.com)

Abstract— Health Informatics has been emerging as a significant field by the application of technology towards the medical health of the patients. Diabetes mellitus has become a major cause of death prevailing over the past few decades. Applying the data mining techniques against the diabetes dataset appears to predict whether the patient is affected with diabetes or not. Hence, reduces the effort of the doctors. This research is based on the idea of comparing certain classification algorithms, considering their rate of accuracy and performance. The classification algorithms are evaluated on a diabetes data. This comparative study explores the various commonly used algorithms in order to get the best possible outcomes. To achieve the above, we perform a preprocessing step on the data using a supervised attribute filter that selects attributes and the performance of the each classification algorithm is measured with the help of an evaluation function.

Keywords— *Health informatics, diabetes mellitus, feature selection, classification*

I. INTRODUCTION

Health informatics or medical informatics can be described as the management and use of the patient's health care information in the field of technology. Diagnosis of diseases in the earlier days, say using traditional methods involved a lot of work to be done manually by the physician. As the world is moving towards automation, this work can be utilized in the modern diagnosis of diseases.

The number of people affected with diabetes has quadrupled since 1980 [1]. Generally, diabetes is of two types- Type 1 and Type 2. Type 1 diabetes is caused mainly due to the decreased amount of insulin produced inside our body which results in high blood sugar levels. Whereas, Type 2 diabetes is caused mainly due to obesity and when the body resists the effects of insulin [2].

According to the International Diabetes Federation, 415 million people are affected with diabetes across the world and out of those 90% are suffering from Type 2 diabetes [3]. High blood sugar level in pregnancy also causes diabetes to the delivered baby. These statistics show us the need to diagnose diabetes in the early stage before it starts affecting the people.

Health informatics has been improving with the advent of data mining and other soft computing techniques. One such application is basically classifying the problem based on its parameters into either of the two classes, presence or absence of diabetes. Data mining is a process that lets us discover unknown patterns and trends in the dataset, and to do so it is essential to preprocess the data which simply means removing missing values, outliers and adding class labels if not present.

The prediction of the diabetic state can be analyzed using various classification techniques. The data can be classified using neural networks [4].The dataset can be classified using hybrid systems that are intelligent enough to diagnose the patients accurately [5]. The analyzed data can be visualized using various visualization techniques [6].

Rest of the paper is organized as follows, Section II contains the related work of diagnosing the diabetes dataset using classification, Section III explains the various steps involved to diagnose the diabetes disease dataset, Section IV discusses the performance results calculated by evaluating the classification algorithms, and Section V concludes the overall research work done in this paper with the future directions.

II. RELATED WORK

Many researchers have tried to predict the diabetic state of the patients using various approaches. Few of the related works are mentioned below.

Panigrahi Srikanth and Dharmaiah Devarapalli [7] classified the diabetes dataset using two most popular classification algorithms Naive Bayes and Decision Tree algorithms. In their study, they constructed the evolution function using the two approaches and calculated the performace using the error rate found by each algorithm.

NirmalaDevi M, Balamurugan S and Swathi U.V [8] proposed a model that combines K-means algorithm and KNN algorithm to improve the performance of simple KNN. The proposed model consists of multi preprocessing steps. The kmeans algorithm removes the outliers and replaces missing values with the mean value of each attribute. The KNN algorithm classifies the improved data with a better accuracy upto 97.4%.

Aakanksha Mahajan, Sushil Kumar and Rohit Bansal [9] proposed a model to classify the dataset. The proposed model uses Particle Swarm Optimization algorithm to perform attribute selection and the KNN algorithm is used to classify the data. This model can decrease the preprocessing time and performs well with an accuracy upto 77%.

Vikas B, B.S.Anuhya, K Santosh Bhargav, Sipra Sarangi, Manaswini Chilla [10] performed a rule based mining technique, called Apriori algorithm on their dataset to find strong association rules and extract frequent item sets from the data.

In order to get high performance results while diagnosing diabetes, the data must be preprocessed before applying the classification techniques on it.

III. METHODOLODY

In this study we perform four classification algorithms and compare the performance of the classifiers with the others to see which performs more accurately. To diagnose any disease using data mining techniques, we must follow three steps.

They are as follows:

- Collecting the Data
- Preprocessing the Data
- Applying a classification algorithm to the preprocessed data to get expected results.

The above three steps have been followed to get our expected results.

A. Data Collection

The first step is the collection of data. In order to lead the research, the data selected for this study is Pima Indian diabetes dataset from the Kaggle website [11]. It is available open to everyone.

It contains the data of pregnant women with age of atleast 21 years, diagnosed with presence or absence of diabetes. It consists of 768 tuples each with 8 attributes along with the class label at the end. Out of 768 tuples, 500 tuples are diagnosed negative and the rest are diagnosed positive for diabetes.

The attributes in the diabetes dataset are as presented in the table1 below.

Table1. Attribute list present in the diabetes dataset

Pregnancy count	numeric
Glucose level	numeric
Blood pressure (mmHg)	numeric
Thickness of skin (mm)	numeric
2-hour serum insulin (mu U/ml)	numeric
BMI (kg/m)^2	numeric
Diabetes pedigree function	numeric
Age	numeric
Class label (positive -1 negative -0)	numeric

B. Preprocessing of data

After collecting the data, the second step is to preprocess the data for better performance results. Data preprocessing involves cleaning and reducing of the dataset.

A. Data Cleaning

Prediction and analysis of diagnosis requires the data to be clean and preprocessed. The quality greatly influences the performance of classification techniques. The diabetes dataset contains 376 tuples with missing values. The missing attributes in the dataset are manually filled with 0s.

B. Data Reduction

To reduce the time taken analyze the data, we have performed data reduction. Data reduction techniques reduce the dataset to a smaller set in terms of attributes, while maintaining the integrity of data. We have applied three such techniques to perform reduction.

The applied feature selection techniques are as follows:

- Information Gain Attribute Evaluation
- Principal Components Analysis
- Correlation-based Feature Subset Evaluation

The Information Gain and Principal Components techniques use the Ranker search method. While the Greedy search method is used for the Correlation-based Feature Subset Evaluation technique.

C. Classification of data

Classifying the data is the final step that is to be done in diagnosing the diabetes dataset. Classification is a data mining technique that analyzes the data and extracts models that predicts a categorical class label for each tuple. Classification works only for categorical data [12]. After data reduction, different classification techniques are applied on the dataset to get their respective performance results.

The applied classification techniques in this study are:

- Naive Bayes classifier
- Decision Tree classifier(J48)
- SMO classifier
- Voted Perception classifier

All the classification techniques have been performed and evaluated using WEKA, a data mining tool. WEKA was originally developed at the University of Waikato in New Zealand. WEKA is a java based program that contains a large collection of modern data mining techniques and machine learning algorithms. It also provides standard techniques to preprocess and classify the data [13].

IV. RESULTS AND DISCUSSION

A. Information Gain Attribute Evaluation

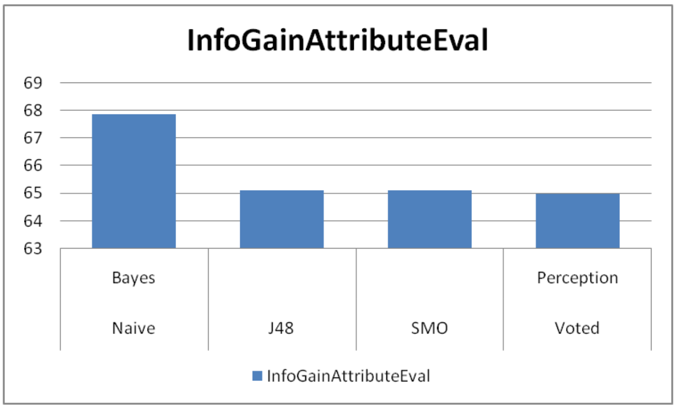
It is a feature selection technique that measures how each attribute contributes in decreasing the overall entropy. The attributes that do not reduce the entropy greatly are not selected for the classification and only the attributes that reduce the entropy of the data are selected for classification.

Table 1. Performance evaluation of Information Gain Attribute Evaluation

	Naive Bayes	J48	SMO	Voted Perception
InfoGain AttributeEval	67.8385	65.1042	65.1042	64.974

The performance of each classification algorithm for the Information Gain Attribute Evaluation technique is tabulated in the table 1(in terms of the rate of accuracy). These results are visualized in the figure1 presented below.

Figure 1. Graphical Representation of Performance evaluation of Information Gain Attribute Evaluation



By observing the obtained results, among the four classification techniques applied on the diabetes disease dataset, reduced by the Information Gain Attribute Evaluation filter, Naive Bayes classifier appears to perform diagnosis with the highest accuracy of 67.8385%.

B. Principal Components Analysis

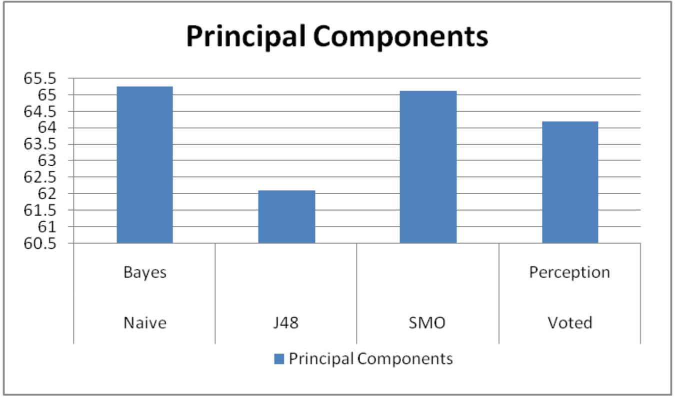
PCA (also called as K-L method) searches for n-dimentional orthogonal vectors that are used to represent the data, where k is less than the total number of vectors. In this technique smaller set of attributes are created and then the original data can be projected onto this smaller set. Thus, dimensionality reduction is performed.

Table 2. Performance evaluation of Principal Components Analysis

	Naive Bayes	J48	SMO	Voted Perception
Principal Components	65.2344	62.1094	65.1042	64.1927

The performance of each classification algorithm for the Principal Components technique is tabulated in the table2. These results are visualized in the figure 2 presented below.

Figure 2. Graphical Representation of Performance evaluation Principal Components Analysis



By observing the obtained results, among the four classification techniques applied on the dataset, reduced by the Principal Components filter, Naive Bayes classifier appears to perform diagnosis again with the highest accuracy of 65.2344%.

C. Correlation-based Feature Subset Selection

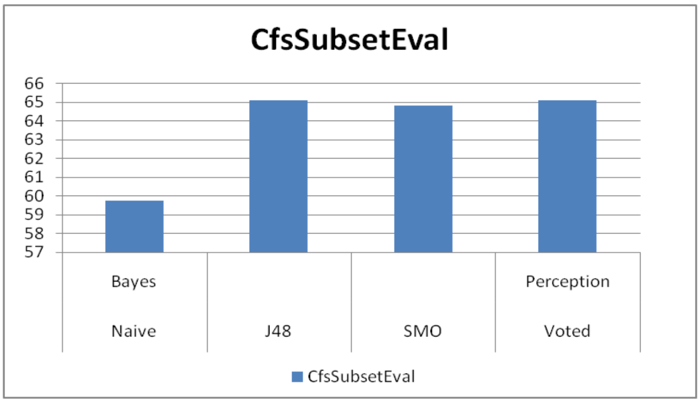
This feature selection technique selects a set of attributes through a correlation based approach. The selected feature subsets contain that have the highest correlation with the class are selected while the other sets are removed from the dataset.

Table 3. Performance Evaluation of Correlation-based Feature Subset Evaluation

	Naive Bayes	J48	SMO	Voted Perception
Cfs SubsetEval	59.7656	65.1042	64.8438	65.1042

The performance of each classification algorithm for the Correlation based Feature Subset Evaluation technique is tabulated in the table 3. These results are visualized in the figure3 presented below.

Figure 3. Graphical Representation of Performance evaluation Correlation-based Feature Subset Evaluation



Among the four classification techniques applied on the dataset, reduced by the Correlation-based Feature Subset Evaluation filter, both J48 and Voted Perception classifiers perform diagnosis equally with the most accuracy of 65.1042%.

V. CONCLUSION AND FUTURE SCOPE

Classification is one of the most renowned forms of analysis in data mining. Applying classification techniques in the medical field helps us diagnose various diseases such as diabetes. In this study, we came across evaluation of four different classification techniques namely Naive Bayes, J48, Sequential Minimal Optimization and Voted Perception used for analyzing the Pima Indian diabetes dataset that is reduced by three different feature selection techniques one at a time. By observing the results, we can find the best classification algorithm suited for a particular feature selection technique. On applying Information Gain Attribute Evaluation technique, Naive Bayes classifier performs diabetes diagnosis most accurately with an accuracy of 67.8385%. On applying Principal Components Analysis technique, Naive Bayes classifier performs diabetes diagnosis most accurately with an accuracy of 67.2344%. On applying Correlation-based Feature Subset Evaluation technique, both J48 and Voted Perception classifiers performs diabetes diagnosis most accurately with an accuracy of 65.1042%.

There can be further enhancement in the evaluation and performance of the classification techniques by the application of ensemble methods.

REFERENCES

[1] CNN-https://edition.cnn.com/2016/04/06/health/diabetes-quadruples-who-report/index.html

[2] Malchoff C D, “Diagnosis and classification of diabetes mellitus”, Diabetes Care, 2011.

[3] Madhuri Panwar, Amit Acharyya , Rishad A Shafik, Dwaipayan Biswas, “K-Nearest Neighbour Based Methodology for Accurate Diagnosis of Diabetes Mellitus”, 2016 6th International Symposium on Embedded Computing and System Design.

[4] Ihsan Salam Jasim, Adil Deniz Duru, Khalid Shaker, Baraa M Abed, Hadeel M Saleh, “Evaluation and Measuring Classifiers of Diabetes Diseases”, ICET 2017.

[5] Rahul Kala, Anupam Shukla, Ritu Tiwari, “Comparitive Analysis of Intellegent Hybrid Systems for detection of PIMA Indian Diabetes”, 2009 World Congress on Nature & Biologically Inspired Computing(NaBIC 2009).

[6] C M Velu, K R Kashwan, “Visual Data Mining Techniques for Classification of Diabetic Patients”, 2013 3rd IEEE International Advance Computing Conference (IACC 2013).

[7] Panigrahi Srikanth, Dharmaiah Deverapalli, “A Critical Study of Classification Algorithms Using Diabetes Diagnosis”, In the Proceedings of the 2016 IEEE 6th International Conference on Advanced Computing.

[8] NirmalaDevi M, Balamurugan S, Swathi U V, “An amalgam KNN to predict Diabetes Mellitus”, In the Proceedings of the 2013 IEEE International Conference on Emerging Trends in Computing,Communication and Nanotechnology(ICECCN 2013).

[9] Aakanksha Mahajan, Sushil Kumar, Rohit Bansal, “Diagnosis of Diabetes Mellitus Using PSO and KNN Classifier”, In the Proceedings of the 2017 International Conference on Computing and Communication Technologies for Smart Nation(IC3TSN).

[10] Vikas B., Anuhya B.S., Bhargav K.S., Sarangi S., Chilla M., “Application of the Apriori Algorithm for Prediction of Polycystic Ovarian Syndrome (PCOS)” In Information Systems Design and Intelligent Applications. Advances in Intelligent Systems and Computing vol 672, Bhateja V., Nguyen B., Nguyen N., Satapathy S., Le DN, Eds, Springer, Singapore.

[11] PimaIndians Diabetes Database| Kaggle-https://www.kaggle.com/uciml/pima-indians-diabetes-database

[12] Vikas B, Sipra Sarangi, Manaswini Chilla, K Santosh Bhargav, B S Anuhya, “A Literature Review on the Rising Phenomenon PCOS”, International Journal of Advances in Engineering & Technology, 2(10).

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, “The WEKA data mining software: an update”.

Mr. Varun Dhanalakota is currently pursuing his Bachelor's degree from Department of CSE, GITAM since 2015. His research interests include Data Analytics, Data Mining, Neural Networks.

Mr Vikas B pursued Bachelor of Technology and Master of Technology from JNTUH, Hyderabad. He is currently pursuing Ph.D. in the Department of CSE, GITAM. His main research work focuses on Deep Learning, Cryptography Algorithms, Machine Learning and Data Mining. He has years of teaching experience and 2 years of research experience