KnE Social Sciences

PwR Symposium
Annual PwR Doctral Symposium 2018–2019
Volume 2019

Knowledge E
enriching | engaging | empowering

Conference Paper

Indepth Analysis of Medical Dataset Mining: A Comparitive Analysis on a Diabetes Dataset Before and After Preprocessing

Latifa Nass[1], Stephen Swift[2], and Ammar Al Dallal[3]

[1]College of Business, Arts and Social Sciences, Brunel University London, UK
[2]College of Engineering, Design and Physical Sciences, Brunel University London, UK
[3]College of Engineering, Ahlia University, Bahrain

Abstract

Most of the healthcare organizations and medical research institutions store their patient's data digitally for future references and for planning their future treatments. This heterogeneous medical dataset is very difficult to analyze due to its complexity and volume of data, in addition to having missing values and noise which makes this mining a tedious task. Efficient classification of medical dataset is a major data mining problem then and now. Diagnosis, prediction of diseases and the precision of results can be improved if relationships and patterns from these complex medical datasets are extracted efficiently. This paper analyses some of the major classification algorithms such as C4.5 (J48), SMO, Naïve Bayes, KNN Classification algorithms and Random Forest and the performance of these algorithms are compared using WEKA. Performance evaluation of these algorithms is based on Accuracy, Sensitivity and Specificity and Error rate. The medical data set used in this study are Heart-Statlog Medical Data Set which holds medical data related to heart disease and Pima Diabetes Dataset which holds data related to Diabetics. This study contributes in finding the most suitable algorithm for classifying medical data and also reveals the importance of preprocessing in improving the classification performance. Comparative study of various performances of machine learning algorithms is done through graphical representation of the results.

Keywords: Data Mining, Health Care, Classification Algorithms, Accuracy, Sensitivity, Specificity, Error Rate

Corresponding Author:
Latifa Nass

Received: 22 July 2019
Accepted: 16 September 2019
Published: 19 September 2019

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the PwR Symposium Conference Committee.

1. Introduction

Today the volume of data present in medical datasets is so huge and thanks to the technology that made it possible to store and extract this large voluminous data efficiently and effectively. Medical diagnosis is the process of creating meaningful patterns or evidences from medical data sets [1]. Extracting this useful information from these medical datasets helps the medical practitioner in early diagnosing of diseases which can save a human life. Having adequate tools to handle this big data solves the problem

to a great extent. A large number of research studies have conducted in this area and it is still a topic of great interest. A number of classification algorithms are available in the literature and it is interesting to take a closer look at these existing algorithms and their performance on medical datasets. In this paper we conduct experiments on a number of medical datasets using a number of well-known classification algorithms. The aim is to evaluate whether classifier performance can be improved by applying pre-processing techniques before classification. Medical data is well known to contain missing values, outliers and noise and to the best of the authors knowledge there are few papers that look at the impact of pre-processing.

One of the main factors contributing to high death rate all over the world is heart disease. Heart-Stat log Medical Dataset [2] holds medical data related to heart diseases. It is based on data from the Cleveland Clinic Foundation and it contains 270 instances belonging to two classes: the presence or absence of heart disease. The features used to describe this dataset are listed in the Table 1.

TABLE 1: Heart-Stat log Medical Data Features.

Heart-Stat log Data Set Features
Age
Sex
Chest
Resting blood sugar
Serum cholesterol
Fasting blood sugar
Resting electrocardiographic
Maximum heart rate
Exercise induced angina
Old peak
Slope
Number of major vessels
Thal

The study also considers another medical data set Pima Diabetes Dataset (https://www.kaggle.com/uciml/pima-indians-diabetes-database) [30] which includes 768 instances and 9 attributes. The features used to describe this dataset are listed in Table 2.

This study considers the dataset Heart-Stat log Medical Data Set and Pima Diabetes Data Set using Weka compares the performance of various machine learning algorithms as specified in Table 3. Weka is a collection of data mining algorithms designed in Java for solving real time data mining applications which can be used to perform a wide

TABLE 2: Pima Diabetes Medical Data Features.

Pima Diabetes Data Set Features
Pregnant
Glucose
Pressure
Triceps
Insulin
Mass
Pedigree
Age
Diabetes

variety of tasks like regression, clustering, association, classification and visualization. Also, other existing studies on this dataset are also included in this paper. The main objective of this study is to find out the most suited algorithm for prediction of diseases and also to understand the importance of preprocessing.

TABLE 3: WEKA names of selected classifiers.

Generic Name	WEKA Name
Bayesian Network	Naïve Bayes
Support Vector Machine	SMO
C4.5 Decision Tree	J4.8
K-Nearest Neighbour	1Bk
Random Forest	Random Forest

2. Related Literature

Solving problems in medical domain using different tools, methods and techniques can be defined as Machine Learning. Improving the accuracy of the analyzed data is the ultimate aim. The process of discovering useful patterns from large volume of data is KDD or Knowledge Discovery. The Table 4 lists the important steps of KDD [3].

One of the most important steps in KDD is Data Pre-processing since the datasets are normally not complete due to missing values, noise, non-representable records and in accurate data. This affects the quality of the results. In order to improve the accuracy of the results the preprocessing step is the most important one. The major steps of data preprocessing are

1. Data Cleaning: the process of detecting and correcting erroneous records in a dataset.

TABLE 4: Steps of Data mining.

Steps of KDD
1. Selection
2. Data Preprocessing • Data Cleaning • Data Integration Data • Transformation • Data Reduction • Data discretization
3. Transformation
4. Data Mining
5. Interpretation /Evaluation

2. Data Integration: creating a single dataset from multiple data sources (i.e. heterogeneous relational databases)
3. Data Transformation: the process of converting data from one format or structure to another. E.g. normalisation
4. Data Reduction: transforming data to a simpler and/or more compact form to remove redundancy and to improve algorithm efficiency

Presence of Missing or noisy data can cause inaccurate results. Hence suitable measures should be adopted to deal with these two situations as listed below.

2.1. Missing Data

1. Ignore the tuple: ignore the records with missing values
2. Manually fill in the values: replace them with a global constant
3. Substitute the missing values with a global value
4. Use Mean value: the integral of a continuous function of one or more variables over a given range divided by the measure of the range

2.2. Noisy Data

1. Binning: the process of transforming numerical variables into categorical counterparts.
2. Clustering: the process of making a diagnosis.
3. Regression: A statistical process that allows you to examine the relationship between two or more variables of interest

4. Normalization: the process of organizing data to minimize redundancy.

Data mining refers to the application of algorithms for extracting patterns from data and the two main problem areas under this is Classification and Clustering. A number of algorithms exist in literature that can classify medical data set efficiently.

Some of the algorithms considered this study are explained in detail below: -

C4.5 algorithm

This is a decision tree algorithm that uses divide and conquer strategy. The algorithm eliminates the following problems of unavailable values, continuous attributes value ranges, pruning of decision trees and rule derivation [4].

From the set of training instances select one attribute. Choose the initial subset of training instances and create a decision tree using the instances. Test the accuracy of the constructed tree using remaining instances. If all instances are classified correctly stop else add it to the initial subset and construct a new tree. Repeat the steps.

Advantages and Disadvantages

The advantages of the C4.5 are:

1. Easy to implement and can be interpreted easily
2. Works with noisy data and both categorical and
3. Continuous values

The disadvantages are:

1. Small variation in data can lead to different decision trees
2. Does not work very well on a small training set

2.3. SMO Algorithm

This method usually involves two datasets training data sets and test data set and is generally considered to be a supervised classifier. If the classes are linearly separable a series of lines can be found which divides the classes separately. The best of these is selected as the final separating line which is found by maximizing the distance

to the nearest points of both classes in the training set. Finally, the points on this maximal margin lines are considered to be support vectors. Three important steps of this algorithm Selecting parameters, Optimizing Parameters and calculating the threshold value b. [4]

Advantages & disadvantages

The advantages of the SMO are:

1. Good Prediction accuracy
2. Minimize expected error
3. Works well with few training samples

The disadvantages are:

1. Need to have two data sets: swaps all missing values and converts nominal attributes into binary ones
2. Difficulty in understanding the algorithm

Naïve Bayes

The Naïve Bayes classifier is an estimator algorithm as the algorithm does estimation more than making predictions. First phase is the training phase where the classifier is trained to estimate the parameters needed for classification. Thus, it clearly estimates the probability that a given instance belongs to that particular class. However, the algorithms make an assumption called conditional independence where the effect of an attribute value on a given class is considered independent to the values of the other attributes. It applies Bayes rule in computing the probabilities [5].

Advantages & disadvantages

The advantages of the Naïve Bayes are:

1. Minimum Error Rate
2. Easy to implement

The disadvantages are:

1. Difficult to have learn the interactions between features
2. Works well with big data set but performance can suffer when the dataset is small in size

3. KNN

One of the top 10 ten algorithms for classification, it is easy to implement. In brief, the training portion of nearest-Neighbour does little more than store the data points presented to it. When asked to make a prediction about an unknown point, the nearest-neighbour classifier finds the closest training-point to the unknown point and predicts the category of that training point according to some distance metric. The distance metric used in nearest neighbour methods for numerical attributes can be simple Euclidean distance [6].

Advantages & disadvantages

The advantages of the KNN are:

1. Can be used with very large data sets (scales well)
2. Works comparatively well with noisy data

The disadvantages

1. Lazy Learner as it doesn't learn a discriminative function from the training data but "memorizes" the training dataset instead
2. The success of algorithm depends on the selection of *k* (the number of neighbors)

The table given below shows some existing studies on **Heart-Stat log Data Set**. The performance metrics considered is Classification Accuracy i.e. the percentage of correctly classified instances. The results found in Literature are summarized in Table 5.

TABLE 5: Results from Literature-Heart Stat log Medical Data Set.

Author	Technique	Performace Metrics
Vikas Chaurasia and Saurabh Pa [7]	RBF Network Decision tree	Accuracy.77 Accuracy.75
Amma [8]	Genetic Algorithm	Accuracy.94

Author	Technique	Performace Metrics
Wiharto [9]	SVM	Accuracy.61
Jaganathan P., Kuppuchamy R [10]	Mean selection method	Accuracy.84 Specificity.85 Sensitivity.84
Jaganathan P., Kuppuchamy R [10]	Half selection method	Accuracy.84 Specificity.85 Sensitivity.84
C. V. Subbulakshmi and S. N. Deepa[11]	PSO	Accuracy.86 Specificity.86 Sensitivity.86
Ms. shtake S.H & Prof.Sanap S.A. [12]	Decision Tree Naive Bayes Neural Networks	Accuracy.94 Accuracy.95 Accuracy.94
Chaitrali S. Dangare [13]	Naive Bayes Neural Networks Decision Tree	Accuracy.99 Accuracy.99 Accuracy.90
Jyoti Soni [14]	Decision Tree Naive Bayes Neural Networks	Accuracy.89 Accuracy.86 Accuracy.85
AH Chen, SY Huang, PS Hong, CH Cheng, EJ lin [15]	Neural Networks	Accuracy.80
Vikas Chaurasia, [16]	CART ID3 Decision Table	Accuracy.83 Accuracy.72 Accuracy.82
Andrea D'Souza [17]	Neural Networks K-Means Clustering	Accuracy.79 Accuracy.63
Milan Kumari [18]	Decision Tree Neural Networks SVM	Accuracy.79 Accuracy.80 Accuracy.84
Abhishek Taneja [19]	Naive Bayes Decision tree Neural Networks	Accuracy.86 Accuracy.89 Accuracy.89
Palaniappan Rafiah Awang [20]	Decision Tree Neural Networks	Accuracy.89 Accuracy.85

Table 6 summarizes the results from Literature performed on the Diabetes Dataset.

4. Results and Discussions

4.1. Implementation Platform

The implementation Platform is Weka version 3.9 and the dataset used is Heart-Statlog Medical Data Set and Pima Diabetic Data Set. The Heart-Statlog Medical Data Set contains 270 instances and 13 attributes. The Diabetes Dataset contains 768 instances

Table 6: Results from Literature-Diabetes Medical Dataset.

Author	Technique	Performace Metrics
K. Saravananathan1 and T. Velmurugan [21]	J48 CART SVM KNN	Accuracy.67 Accuracy.62 Accuracy.65 Accuracy.53
Saman Hina, Anita Shaikh and Sohail Abul Sattar [22]	Naïve Bayes MLP J48 Random Forest	Accuracy.76 Accuracy.81 Accuracy.75 Accuracy.79
Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly [23]	J48 Naïve Bayes	Accuracy.74 Accuracy.79
R. Sivanesan, K. Devika Rani Dhivya [24]	J48	Accuracy.73
J. Anitha, Dr.A. Pethalakshmi [25]	J48 Naïve Bayes	Accuracy.79 Accuracy.77
Meraj Nabi, Pradeep Kumar, Abdul Wahid [26]	Naïve Bayes Logistic Regression J48 Random Forest	Accuracy.76 Accuracy.80 Accuracy.76 Accuracy.76

and 9 attributes the implementation algorithms are C4.5 (J48), SMO, Naïve Bayes, KNN and Random Forest.

4.2. Performance Metrics

The actual and predicted classification done by a classification matrix is generated and represented by a confusion matrix. A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.

Once the confusion matrix is generated for each implemented algorithm the following metric values Accuracy, Sensitivity, Specificity and Error rate are calculated from the confusion matrix using the formulas listed below. The table 7 shows the confusion matrix for a two-class classifier [27].

Table 7: Confusion Matrix.

		Predicted	
		No	**Yes**
Actual	**Negative**	A	B
	Positive	C	D

Where: A is the number of True Positives

B is the number of True Negatives

C is the number False Positive

D is the number of False Negatives

1. **Accuracy:** It is the percentage of accurate predictions.

$$\text{Accuracy} = (A + D) / (A+B+C+D)$$

2. **Sensitivity**: It is the proportion of positives that are correctly identified.

$$\text{Sensitivity} = D/ (D + C)$$

3. **Specificity:** It is the proportion of negatives that are correctly identified.

$$\text{Specificity} = A / (A + B)$$

4. **Error Rate**: It is equivalent to 1 minus Accuracy

$$\text{Error Rate} = (B + C) / (A+B+C+D)$$

4.3. Experimental Results Before Preprocessing- Heart Statlog Medical Data Set

The following algorithms C4.5, SMO, Naïve Bayes, KNN and Random Forest were run on the dataset and the generated confusion matrix is listed in Table 8: -

TABLE 8: Confusion Matrix for C4.5 (Heart Statlog).

		Predicted	
		No	**Yes**
Actual	**Negative**	119	31
	Positive	32	88

Accuracy =.77 Sensitivity =.73

Specificity=.79 Error Rate =.23

TABLE 9: Confusion Matrix for Random Forest (Heart Statlog).

		Predicted	
		No	**Yes**
Actual	**Negative**	126	24
	Positive	26	94

Accuracy =.81 Sensitivity =.78

Table 10: Confusion Matrix for Naïve Bayes (Heart Statlog).

		Predicted	
		No	**Yes**
Actual	**Negative**	131	19
	Positive	22	98

Specificity =.84 Error Rate =.19

Accuracy =.85 Sensitivity =.81

Specificity=.87 Error Rate =.15

Table 11: Confusion Matrix for kNN (Heart Statlog).

		Predicted	
		No	**Yes**
Actual	**Negative**	115	35
	Positive	32	88

Accuracy =.75 Sensitivity =.73

Specificity =.77 Error Rate =.25

Table 12: Confusion Matrix for SMO (Heart Statlog).

		Predicted	
		No	**Yes**
Actual	**Negative**	131	19
	Positive	24	96

Accuracy =.84 Sensitivity =.80

Specificity =.87 Error Rate =.16

The results show that Naïve Bayes outperforms the other algorithms when measured in terms of accuracy, sensitivity, specificity and error rate.

4.4. Experimental Results Before Preprocessing- Diabetes Medical Data Set

The following algorithms C4.5, SMO, Naïve Bayes, KNN and Random Forest were run on the dataset and the generated confusion matrix is listed in table 13: -

Accuracy =.74 Sensitivity =.59

Specificity =.81 Error Rate =.26

Accuracy =.76 Sensitivity =.61

Specificity =.83 Error Rate =.24

Table 13: Confusion Matrix for C4.5 (Diabetes).

		Predicted	
		Yes	**No**
Actual	**Negative**	407	93
	Positive	108	160

Table 14: Confusion Matrix for Random Forest (Diabetes).

		Predicted	
		Yes	**No**
Actual	**Negative**	418	82
	Positive	104	164

Table 15: Confusion Matrix for Naïve Bayes (Diabetes).

		Predicted	
		No	**Yes**
Actual	**Negative**	422	78
	Positive	104	164

Accuracy =.76 Sensitivity =.61

Specificity=. 84 Error Rate =.24

Table 16: Confusion Matrix for kNN (Diabetes).

		Predicted	
		No	**Yes**
Actual	**Negative**	397	103
	Positive	126	142

Accuracy =.70 Sensitivity =.52

Specificity=.79 Error Rate =.30

Table 17: Confusion Matrix for SMO (Diabetes).

		Predicted	
		No	**Yes**
Actual	**Negative**	449	51
	Positive	123	145

Accuracy =. 77 Sensitivity =.54

Specificity=.89 Error Rate =.23

The results show that SMO outperforms the other algorithms when measured in terms of accuracy, sensitivity, specificity and error rate.

4.5. Experimental Results After Preprocessing

The algorithms C4.5, SMO, Naïve Bayes, KNN and Random Forest were run on data after applying appropriate preprocessing filters and the results are summarized below. Various modifications were done on data preprocessing and model parameters to achieve the best results. Since the Heart Statlog Medical DataSet did not have any missing values the results after preprocessing stayed the same. However, improvement in performance was exhibited by the Diabetes Data set.

The following algorithms C4.5, SMO, Naïve Bayes, KNN and Random Forest were run on the Diabetes dataset after preprocessing and the results are summarized in the table 18.

TABLE 18: Results after Preprocessing (Diabetes Dataset).

Algorithms	Accuracy	Sensitivity	Specificity	Error rate
C4.5(J48)	.76	.61	.83	.24
SMO	.78	.56	.90	.22
Naïve Bayes	.77	.63	.85	.23
KNN(IBK)	.72	.54	.81	.28
Random Forest	.76	.62	.84	.24

The performance on imputed data (preprocessed data) showed better classification accuracy when measured with respect to sensitivity, specificity and accuracy.

4.6. Comparitive Analysis

The study needs to analyze if the classification accuracy improved after preprocessing (imputation) the data when measured with respect to sensitivity, specificity and accuracy. The Table 19 given below shows the performance measures on Diabetes Dataset before and after preprocessing.

TABLE 19: Performance analysis of Diabetes Dataset before and after Preprocessing.

Algorithms	Accuracy		Sensitivity		Specificity		Error rate	
	BP	AP	BP	AP	BP	AP	BP	AP
C4.5(J48)	.74	.76	.59	.61	.81	.83	.26	.24
SMO	.77	.78	.54	.56	.89	.90	.23	.22
Naïve Bayes	.76	.77	.61	.63	.84	.85	.24	.23
KNN(IBK)	.70	.72	.52	.54	.79	.81	.30	.28
Random Forest	.76	.76	.61	.62	.83	.84	.24	.24

The graph (Figure 1) shows the comparison of performance graphically in terms of accuracy with before preprocessing (BP) and after preprocessing (AP) (imputed and scaled data) for Diabetes dataset

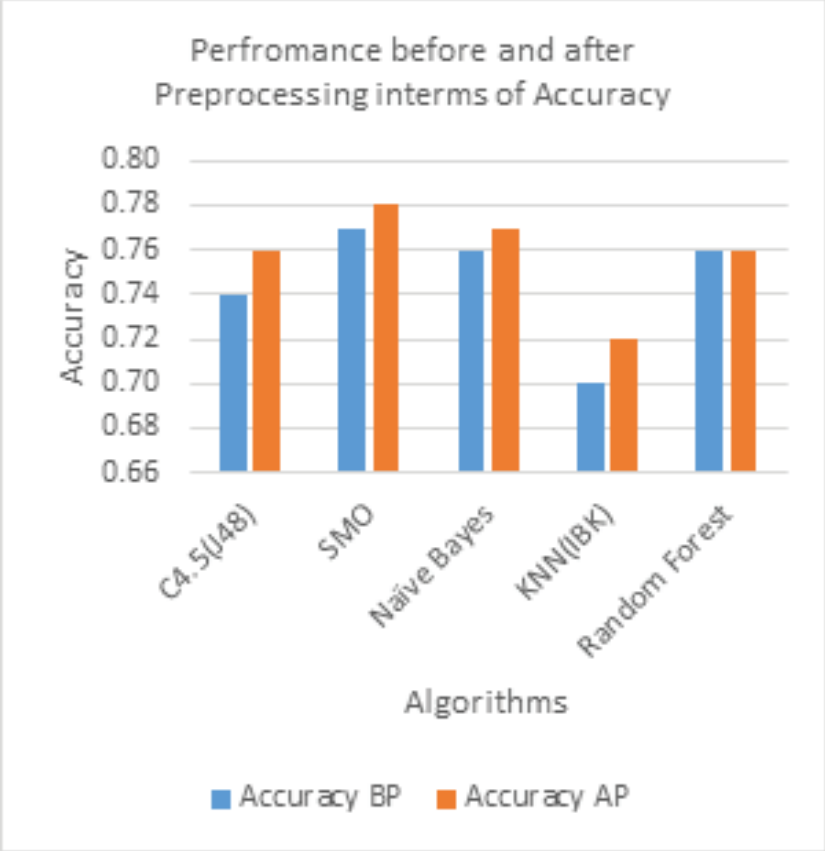

Figure 1: Performance Comparison in terms of Accuracy.

The graph (Figure 2) shows the comparison of performance graphically in terms of sensitivity with before preprocessing and after preprocessing (imputed and scaled data) for the Diabetes dataset.

The graph (Figure 3) shows the comparison of performance graphically in terms of specificity with before preprocessing and after preprocessing (imputed and scaled data) for Diabetes dataset.

The graph (Figure 4) shows the comparison of performance graphically in terms of Error rate with before preprocessing and after preprocessing (imputed and scaled data) for the Diabetes dataset.

The performance is measured in terms of Accuracy, Sensitivity, Specificity and Error rate on Diabetes dataset using the algorithms KNN, Random Forest, SMO and J48. The metrics values are recorded by applying the algorithms on the dataset that has not be preprocessed and again the same algorithms are applied on the dataset after preprocessing. The results clearly show that the performance after imputation has

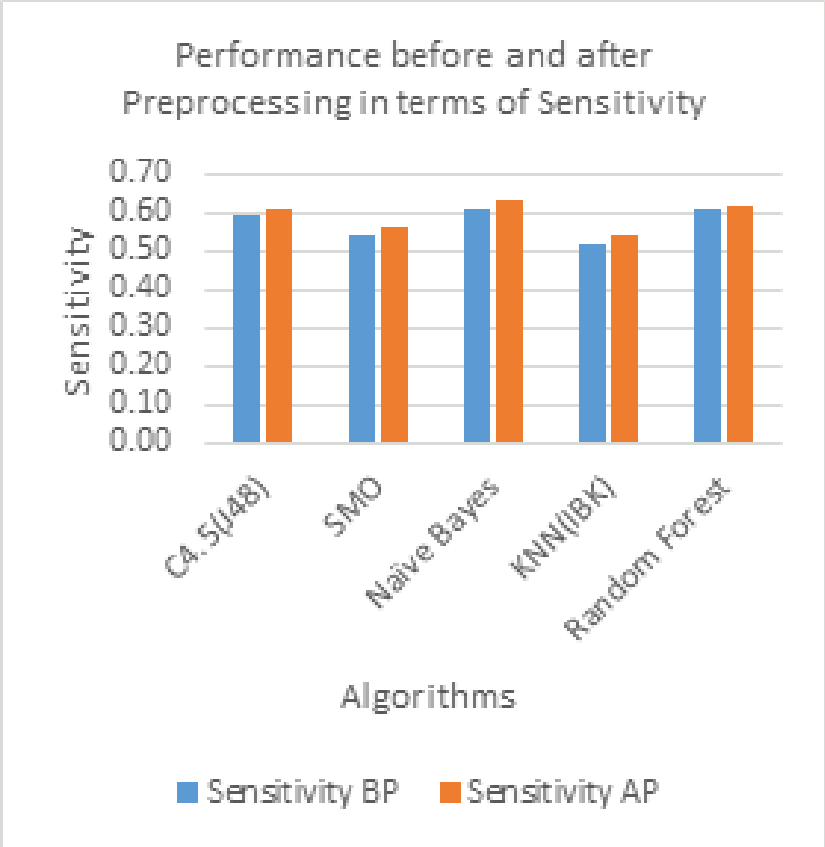

Figure 2: Performance Comparison in terms of sensitivity.

improved significantly. The study showed better classification accuracy when measured with respect to sensitivity, specificity, accuracy and error rate on preprocessed data. This clearly shows the significance of preprocessing step in datamining. The dataset if it has a lot of missing values and noisy data will not give you quality results. A significant improvement is noticed in terms of Accuracy, sensitivity, Specificity and error rate when applied on preprocessed data.

5. Conclusion

Efficient classification of medical dataset is a major datamining problem then and now. Diagnosis, Prediction of diseases and the precision of results can be improved if relationships and patterns from these complex medical datasets are extracted efficiently. This paper analyses some of the major classification algorithms like C4.5 (J48), SMO, Naïve Bayes, KNN Classification algorithms and Random Forest and the performance of these algorithms are compared using WEKA. Performance evaluation of these algorithms is done based on Accuracy, Sensitivity and Specificity and Error rate. The medical data set used in this study are Heart-Statlog Medical Data Set which holds medical data related

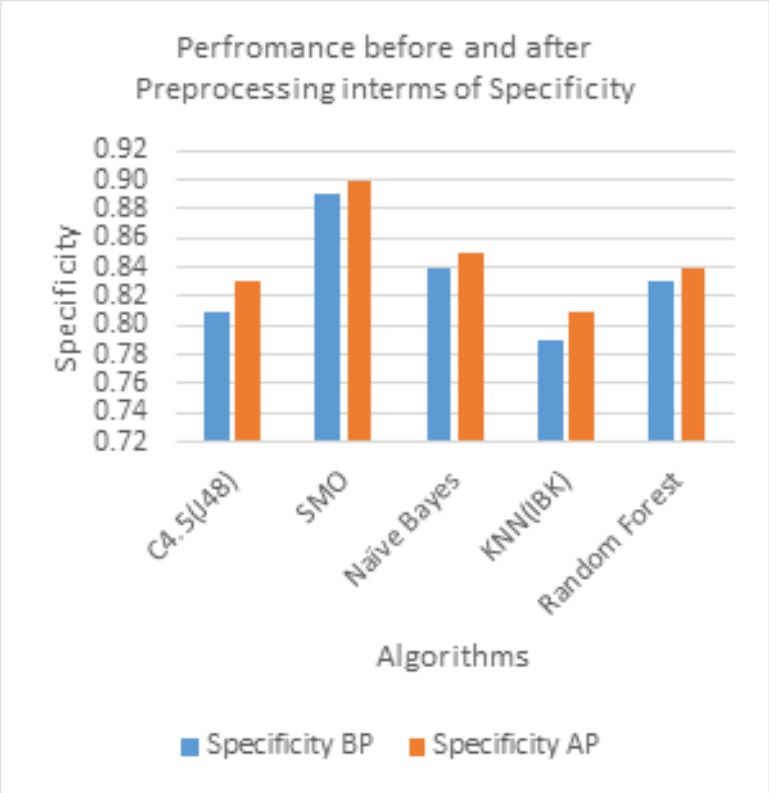

Figure 3: Performance Comparison in terms of Specificity.

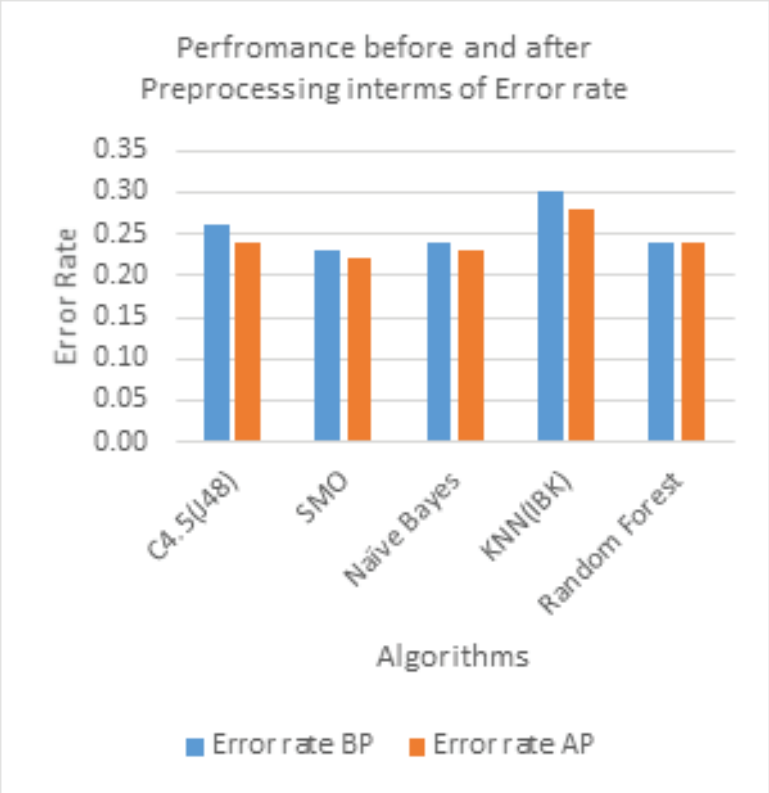

Figure 4: Performance Comparison in terms of Error Rate.

to heart disease and Pima Diabetes Data Set which holds data related to Diabetes. The results showed that SMO outperformed the other algorithms when measured in terms of accuracy, sensitivity, specificity and error rate for Heart-Statlog Medical Data and since the dataset didn't have any missing values the result remained same after preprocessing. For the Diabetes Dataset the results showed that the Naïve Bayes algorithm outperformed the other algorithms when measured in terms of accuracy, sensitivity, specificity and error rate. However, the results improved when appropriate preprocessing namely imputation was done on the dataset. The study showed better classification accuracy when measured with respect to sensitivity, specificity, accuracy and error rate on preprocessed data.

6. Future Work

The main goal of this paper was to explore the different datamining algorithms and to measure their performance on medical dataset using Accuracy, Sensitivity, Specificity and Error rate as the metrics. As a future study the researcher intends to improve kNN algorithm as kNN is considered as one of the top 10 best mining algorithms, and also the researcher intends to take it as a challenge to increase the accuracy percentage of kNN algorithm by improving it and to make it outperform the other algorithms considered in the study.[29]

References

[1] Ms. A. Malarvizhi, Dr. S. Ravichandran," Data Mining's Role in Mining Medical Datasets for Disease Assessments – a Case Study", International Journal of Pure and Applied Mathematics, Volume 119 No. 12 2018.

[2] P. Jaganathan and R. Kuppuchamy, "A threshold fuzzy entropy based feature selection for medical database classification," Computers in Biology and Medicine, vol. 43, no. 12, pp. 2222–2229, 2013

[3] Umair Shafique, Haseeb Qaiser,"A Comparative Study of Data Mining Process Models", - International Journal of Innovation and Scientific Research, Vol. 12 No. 1, Nov. 2014

[4] Dr. T. Karthikeyan, Dr. B. Ragavan, V.A.Kanimozhi, A Study on Data mining Classification Algorithms in Heart Disease Prediction, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5, Issue 4, April 2016

[5] Yong ZENG, H.-m. F.-p.-y. (2016). An Improved ML-kNN Algorithm by Fusing Nearest Neighbor Classification". *International Conference on Artificial Intelligence and Computer Science (AICS 2016)*.

[6] Shaifali Gupta, R. R. (2016). Improvement in KNN Classifier (imp-KNN) for Text Categorization",. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue

[7] Vikas Chaurasia and Saurabh Pa,Performance analysis of Diagnosis and Prediction of Heart and Breast Cancer Disease,Review of reserch, Vol 3,Issue 3 May 2014.

[8] N. Amma, "Cardiovascular disease prediction system using genetic algorithm and neural network," in International Conference on Com- puting, Communication and Applications. Dindigul, Tamilnadu, India:IEEE, Feb 2012, pp. 1–5

[9] W. Wiharto, H. Kusnanto, and H. Herianto, "Performance analysis of multiclass support vector machine classification for diagnosis of coronary heart diseases," International Journal on Computational Science & Applications, vol. 5, no. 5, pp. 27–37, 2015

[10] Jaganathan P., Kuppuchamy R. A threshold fuzzy entropy based feature selection for medical database classification. *Computers in Biology and Medicine*. 2013;43(12)

[11] C. V. Subbulakshmi and S. N. Deepa, Medical Dataset Classification: A Machine Learning Paradigm Integrating Particle Swarm Optimization with Extreme Learning Machine Classifier,Scientic World Journal,September 2015

[12] Ms. Ishtake S.H, Prof. Sanap S.A. "Intelligent Heart Disease Prediction System Using Data Mining Techniques", International J. of Healthcare & Biomedical Research, Volume: 1, Issue: 3, April 2013.

[13] Chaitrali S. Dangare Sulabha, " Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012.

[14] Jyoti Soni, Ujma Ansari, Dipesh Sharma, " Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.

[15] AH Chen, SY Huang, PS Hong, CH Cheng, EJ lin, "HDPS: Heart Disease Prediction System, Computing in cardiology", 2011: 38:557- 560.

[16] Vikas Chaurasia, Saurabh Pal, " Early Prediction of Heart Diseases using Data Mining Techniques", Caribbean Journal of Science & Technology, ISSN 0799-3757.

[17] Andrea D'Souza, "Heart Disease Prediction Using Data Mining Techniques", International Journal of Research in Engineering and Science (IJRES) ISSN (Online): 2320-9364, ISSN (Print): 2320-9356.

[18] Milan Kumari, Sunila Godara, " Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", International Journal of Computer Sci ence and Technology, IJCST Vol. 2, Iss ue 2, June 2011,

[19] Abhishek Taneja, "Heart Disease Prediction System Using Data Mining Techniques", Oriental Journal Of Computer Science & Technology, ISSN: 0974-6471 December 2013, Vol. 6, No. (4).

[20] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques" IJCSNS International

[21] K. Saravananathan1 and T. Velmurugan, Analyzing Diabetic Data using Classification Algorithms in Data Mining, Indian Journal of Science and Technology, Vol 9(43)

[22] Saman Hina, Anita Shaikh and Sohail Abul Sattar, Analyzing Diabetes Datasets using Data Mining, Journal of Basic & Applied Sciences, 2017, 13, 466-471

[23] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, DIAGNOSIS OF DIABETES USING CLASSIFICATION MINING TECHNIQUES, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015

[24] R. Sivanesan, K. Devika Rani Dhivya, A Review on Diabetes Mellitus diagnoses using classification on Pima Indian Diabetes Data Set, International Journal of Advance Research in Computer Science and Management Studies, Volume 5, Issue 1, January 2017

[25] J.Anitha, Dr.A.Pethalakshmi, Comparison of Classification Algorithms in Diabetic Dataset, International Journal of Information Technology (IJIT) – Volume 3 Issue 3, May - Jun 2017

[26] Meraj Nabi, Pradeep Kumar, Abdul Wahid, Performance Analysis of Classification Algorithms in Predicting Diabetes, International Journal of Advanced Research in Computer Science Volume 8, No. 3, March –April 2017.

[27] A. K. Santra, C. Josephine Christy," Genetic Algorithm and Confusion Matrix for Document Clustering", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012

[28] world Health Organization,"Cardiovascular diseases (CVDS)", https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds),May 2017.

[29] Leif E. Peterson," K-nearest neighbor", Scholarpedia,2009.

[30] UCI Machine Learning, (http://archive.ics.uci.edu/ml/index.php).