ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 12(22), DOI: 10.17485/ijst/2019/v12i22/144929, June 2019*

A New Method of Data Preparation for Classifying Diabetes Dataset

M. S. Padmavathi* and C. P. Sumathi

SDNB Vaishnav College for Women, Chennai – 600044, Tamil Nadu, India;
padmanivas_2002@yahoo.co.in, drcpsumathi@gmail.com

Abstract

Objective: Millions of people including children and pregnant women are affected by Diabetes mellitus. Undiagnosed diabetes can affect entire body system including cardiac attacks, chronic kidney disease, foot ulcers and damage to the eyes Therefore an intelligent model should be developed for early detection of diabetes. **Method:** Data preprocessing is an important step in building classification models. Pima Indian Diabetes dataset from University of California Irvine (UCI) repository is a challenging dataset with more number (48%) of missing values. Different steps of data preprocessing is performed on Pima Diabetes to improve the accuracy of the classification model. The proposed model includes outlier removal and imputation at stage 1, normalization at stage 2 and balancing the dataset at stage 3. After each stage of preprocessing, the model is evaluated using three classifiers: Support Vector Machine (SVM), Random Forest (RF) and K-nearest neighbor (Knn). **Findings:** It is clearly proved that after each stage of preprocessing, the classification accuracy increases. On completing all 3 stages of preprocessing, the diabetes dataset achieves a highest accuracy (82.14%) and balanced accuracy (81.94%) with Random Forest classifier when compared to SVM and Knn. **Novelty/Improvements:** The preprocessing steps, replacing the outliers using 5 and 95 percentile values with median imputation followed by Z-score normalization and balancing the dataset using smote improves the quality of Pima Diabetes dataset, thereby classification accuracy of the model increases. The same data preprocessing methods can also be applied to different datasets or different classifier models.

Keywords: Balanced Dataset, Imputation, Normalization, Outlier Removal, Random Forest

1. Introduction

Enormous amount of data is available in the area of medical science. The data obtained may not be in a proper format for data analysis, hence raw data need to be preprocessed carefully for proper diagnosis of disease[1]. Data preprocessing is an important step in data mining which involves data transformation, imputation, outlier removal, normalization, feature selection and dimensionality reduction[2]. It is not necessary to involve all the steps of data preprocessing, but according to the nature of the data available, the required steps can be included in the model.

Outlier is a data point that is present far outside from rest of the data or population. They will adversely affect the results of statistical analysis. They typically serve to extend error variance, decrease normality and influence estimates which may affect the model[3]. The most common ways of treating outliers are K-means[4], Statistical Outliers[5], Genetic Algorithms (GA)[6], etc. Missing data occurs when no value is observed for the variable in a dataset. Missing data is common almost in every dataset. If missing range is less than 5%, then it is of no trouble, 5–15% is manageable with subtle techniques to handle the problem and if more than 15% can adversely affect the results of the model, therefore it should be handled in an efficient manner[7]. There are several ways to handle missing data, like single imputation (mean, median, mode, predictive score), multiple imputation, classifier as imputation, etc[8], but deciding the best method should be done according to the characteristics of the dataset.

**Author for correspondence*

Data normalization is transforming different range of variables in the data to a same range. Normalization is applied to make the data points come closer to each other. The techniques available for data normalization are Min-max, Z-score and decimal scaling[9,10]. Imbalanced data refers to classification problems where, the class (output) variables are not equally proportionate. One class outnumbers other class by a substantial proportion. Imbalanced classification happens a lot in binary classification than in multi-level classification. It affects the classifier with over fitting problems; hence it is necessary to balance the datasets before using it[11]. Different types of sampling techniques used to handle imbalanced data are: Over sampling, Under Sampling and SMOTE (Synthetic Minority over Sampling Technique). A classifier is required to map a data item to one of the predefined classes[12]. Classification is the supervised approach, where the model is trained and built using the predefined attributes and values and then it is tested for the unseen data. The most commonly used classification approaches are statistical, mathematical, tree based, fuzzy approach, case based learning, ensemble etc.

This paper emphasizes the different steps in preprocessing and after each step classifier performance is evaluated to see the effect of preprocessing. At stage 1: First, the outliers in the dataset are replaced by 5th and 95th percentile values instead of removing it. As it is the replacement algorithm, it is reliable to consider only the extreme outliers. Since the usage of 5 and 95 percentiles are commonly considered for extreme values in different situations, it is used in the experiment for replacement. Second, the dataset is imputed using median it is proved in the existing study that the accuracy of the model is better when imputed with median[5,6]. At stage 2: The dataset is normalized using Z-score. The attributes in the dataset lies between different ranges of values and the spread of data points lies under a normal curve, hence normalization is done using Z-score to transform the attributes range between (-1, 1). At stage 3: The dataset is balanced using SMOTE, one common sampling technique used for imbalanced datasets[11]. The proposed data preprocessing method for Pima Indian Diabetes is classified using SVM, RF and Knn. The three classifiers are chosen such that one from mathematical, tree-based and instance-based approach is done. Classifiers are evaluated at all the stages and it is proved that each stage of preprocessing has a considerable effect on the classification accuracy. The rest of the paper is organized as follows: Section 2 summarizes about the methodologies used, Section 3 describes the experimental set up and section 4 proves the results obtained followed by concluding remarks and future scope.

This section presents the different existing models available for Pima Diabetes dataset. In[13] presented a classification model with PSO_SVM for feature selection followed by fuzzy decision tree for classification on Pima Indian diabetes dataset. The PSO (Particle Swam Optimization) is used to optimize the SVM and extract reduced features which then applied to fuzzy decision tree improves the accuracy of detecting diabetes. The hybrid combinatorial method of feature selection holds good for diabetes dataset. In[14] classified Pima Indian Diabetes dataset with Fuzzy Genetic Algorithm. The proposed model used SMOTE to handle the imbalanced dataset. Fuzzy and Genetic approaches are combined to enhance the classification performance with 5-fold cross-validation approach. In[15] developed an intelligence system which includes clustering, noise removal and classification approaches. Expectation maximization is used for clustering, Principal Component Analysis (PCA) for noise removal and SVM for classification tasks, respectively. The proposed method is also implemented for incremental situation by applying the incremental PCA and SVM. Experimental results on Pima Indian Diabetes dataset proved that incremental approaches improves accuracy and reduces time compares to non-incremental approaches. In[16] proposes an improved Generalized Regression Neural Network (KGRNN) for Pima Indian Diabetes dataset. The novel KGRNN uses an enhanced K-means clustering technique to produce cluster centers which is used as an input to train the network. The technique outperformed the best known GRNN technique with a classification accuracy of 86% with 83% sensitivity, 87% specificity and roc of 0.87. In[17] proposed a new classifier model with information gain to select the features and Deep Neural Network for classification. The new proposed method in this paper attained a classification accuracy of 90.26% for diabetes dataset which is better when compared with the existing models in the literature. In[18] proposed a novel hybrid classifier combining Logistic regression and Adaptive Network-based Fuzzy Inference System called Logistic Adaptive Network-based Fuzzy Inference System (LANFIS). The experimental results for diabetes dataset obtained a classification accuracy

of 88.05%. It is also proved that the proposed intelligent method obtained 3–5% increase in accuracy compared to the existing models.

2. Methodologies

2.1 5 and 95 Percentile

A statistics that reports *relative standing,* (a place where a data point lies and compared to the rest of data) is called a percentile. Let k is a number between 1 to 100, then k^{th} *percentile* is defined as a value in a data set that splits the data into two parts: First part contains *k* percent of the data and the second part contains the rest of the data ([100 – *k*] percent). The 50^{th} percentile is the median, the point at which 50% of the data falls below the point and 50% falls above it. The percentiles are calculated using the following steps:

Step 1: Sort the data set by value from smallest to largest.
Step 2: Calculate index by multiplying *k* percent by the total number of values, *n*.
Step 3: If the index is not a whole number, round it up to the nearest whole number otherwise index is the same as obtained in step 2.
Step 4: Count the index value in the data set from smallest to largest until you reach the specified index obtained in Step 3.
Step 5: The corresponding value in that particular index of the data set is the k^{th} percentile.

The 95^{th} percentile is a widely used mathematical calculation to evaluate the regular and sustained utilization of a dataset[19]. A 95^{th} percentile indicates that 95% of the time data points are below that value and 5% of the time they are above that value. 95 is a magic number used in almost used in all problem situations.

2.2 Z-score

A Z-score or standard score is employed for standardizing scores on constant scale by dividing a score's deviation by the standard deviation of a data set[20]. It measures the number of standard deviations a given datum is from the mean. The value lies between (-1, 1), where Z-score is negative for values less than the mean, Z-score is positive for values greater than the mean and Z-score is zero for the mean value. The Z-score normalization can be calculated with mean (μ) and standard deviation (σ) of the attributes:

$$Z - score = \frac{x - \mu}{\sigma} \quad (1)$$

2.3 Synthetic Minority Over-sampling Technique (SMOTE)

Imbalanced datasets contains instances with unequal proportion of class labels, which may lead to overfitting problems. To avoid such problems, smote selects a subset of data from the minority class and calculates the synthetic k nearest neighbor instances[21]. One or more of the K-nearest neighbors are selected based on the amount of sampling needed. These synthetic instances are then added to the original dataset, for building the classification models. For each x, using $x_{knn}(x)$ synthetic k nearest neighbors is calculated. New synthetic data point's r is chosen by seeking the vector r on each line segment from x to each x_j such that it has the maximum average distance from the majority class C_i as in Equation (2).

$$r = argmax_{r \in \overline{xx_j}} \frac{1}{k} \sum_{x \in C_i} \|r - x\| \quad (2)$$

2.4 Support Vector Machine (SVM)

SVM performs classification tasks by forming hyperplanes in a multidimensional space that distinguishes the cases of different class labels. The group of data instances used to form the hyperplane is called "Support Vectors". The distance between the hyperplane and the nearest support vector is called as margin. The best hyperplane is chosen based on the maximum – margin separation of distance between the two classes[22]. There are two types of SVMs, 1. Linear SVM is used to separate the data points using a linear decision boundary and 2. Non-linear SVM separates the data points using a nonlinear decision boundary.

For the data points $(\vec{x_1}, y_1) \ldots . (\vec{x_n}, y_n)$, x_i where represents a real vector and y_1 takes the value of 0, 1 representing the class to which x_i belongs. A hyperplane is constructed in order to maximize the distance between two classes y = 0, 1 and is defined as:

$$\underbrace{max}_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j K(x_i, x_j) \alpha_i \alpha_j, \quad (3)$$

Subject to

$$0 \leq \alpha_i \leq C, \quad for\ i = 1,2, \ldots n, \sum_{i=1}^{n} y_i \alpha_i = 0 \quad (4)$$

2.5 Random Forest (RF)

RF is considered as an important ensemble method for classification. Group of trees are built using the Classification and Regression Trees methodology (CART). Each tree is fully constructed using a bootstrapped version of the training data, in which sample of predictors is chosen at each node to find the best split. For any given instance, each tree casts a vote about the predicted class. Once the construction of forest is completed, based on the majority of votes the trees are combined together such that new instances are assigned to a class. In bootstrap sampling, during tree construction only two- third of the samples are included and the remaining is omitted. These omitted samples are called an Out-of-the-bag (OOB) sample which is used to evaluate the performance of the classifier. The parameters used in RF algorithm are the number of predictors and the number of trees[23]. The steps involved are:

Step 1: Draw n_{tree} bootstrap samples from the original data.

Step 2: Grow n_{tree} such that there is no node or fewer nodes at each terminal node. At each node of the tree, m variables are selected in random for splitting.

Step 3: Combine n_{tree} trees for new data prediction based on the majority vote for classification.

Step 4: Calculate an Out-Of-Bag (OOB) error rate by using the data not in the bootstrap sample.

2.6 K–nearest neighbor (Knn)

Knn is a lazy, non-parametric and instance based learning algorithm used for both classification and regression problems. In classification Knn is applied to predict the class for the new unlabelled data. Initially k is chosen and the distance measure between k and each data point is calculated using Euclidean's distance, Hamming distance, Manhattan distance or Minkowski distance. After calculating the distance, the most frequent class occurring in the data points with the minimum value is selected as 'k' nearest neighbors[24]. The steps involved in Knn are:

Step 1: Initialize the value of k.

Step 2: Iterate the following steps from 1 to n of training data points.

- Euclidean distance is used to calculate the distance between the test data and each point in the training data.
- Sort the calculated distances in ascending order.
- Top k rows is chosen as K nearest neighbors from the sorted array.
- The most frequent occurring class in k rows is returned as predicted class.

3. Experimental Setup

3.1 Proposed Model

The block diagram of the proposed model is shown in Figure 1. It involves preprocessing as the main step followed by classifiers to classify the formatted dataset. The preprocessing model is done under 3 stages:

Stage 1: The raw data is processed by replacing the outliers with 5th and 95th percentile values, instead of removing outliers. Thereby there is no loss of instances. The resultant dataset is imputed using median.

Stage 2: The processed dataset from stage 1 is normalized using Z-Score, such that the data lies within the range of (-1, 1).

Stage 3: As Pima Diabetes is an imbalanced dataset, the normalized data from stage 2 is balanced using SMOTE.

After each stage of preprocessing, the classifiers SVM, RF and Knn are evaluated. The preprocessed data obtained after attaining all 3 stages of preprocessing achieves better evaluation metrics compared to the data obtained from stage 1 and stage 2. The RF fits best with the data preparation methods of Pima Diabetes compared to other classifiers (SVM and Knn).

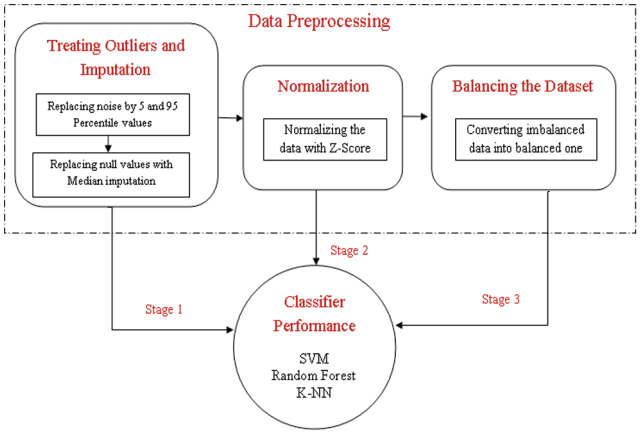

Figure 1. Block diagram of the proposed model

3.2 Data Source

Pima Indian Diabetes[25] is a machine learning database maintained at the UCI (University of California Irvine) repository. The data is based on pregnant women with at least 21 years old to diagnose the presence or absence of diabetes. It is a two-class problem with 8 numerical attributes as input and one output variable. The attribute information present in the dataset is given in Table 1.

3.3 Performance Evaluation

3.3.1 Classification Accuracy

It is the ratio of correct predictions made by the model divided by the total number of instances[25].

$$\text{Classification Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (5)$$

Where True positive (TP) - Positive values correctly predicted, False negative (FN) - positive values wrongly predicted as negative, False positive (FP) – negative values wrongly predicted as positive, True negative (TN) - negative values correctly predicted.

3.3.2 Balanced Accuracy

Imbalanced datasets can provide a high chance for conventional accuracy. To avoid this balanced accuracy can be substituted[26]. It is defined as the average of proportions of correctly predicted patterns (both positive and negative).

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \quad (6)$$

Where, Sensitivity and Specificity is used to measure the fraction of positive/negative patterns that are correctly classified.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (7)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (8)$$

3.3.3 Kappa Statistics

Kappa is a statistical measure considered important for imbalanced datasets[27]. It can be calculated as:

$$\text{Kappa} = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (9)$$

Where, $Pr(a)$ is the percentage of agreement and $Pr(e)$ is the chance of agreement calculated. The value of 1 indicates perfect agreement.

3.3.4 Receiver Operating Characteristic (ROC)

ROC is a widely used performance metric for imbalanced datasets. The area under ROC curve quantifies the overall ability of the model that distinguishes between those individuals with and without the disease[25]. The range of

Table 1. Characteristics of dataset

Pima Diabetes: Total No. of Instances – 768 (Class Class "Absence" – 268, "Presence" - 500)			
S. No	**Attributes**	**Data Description**	**Missing Value (%)**
1.	Number of times pregnant	Numerical	0
2.	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Numerical	5 (0.65%)
3.	Diastolic blood pressure (mm Hg)	Numerical	35 (4.55%)
4.	Triceps skin fold thickness (mm)	Numerical	227 (29.55%)
5.	2-Hour serum insulin (mu U/ml)	Numerical	374 (48.69%)
6.	Body mass index (weight in kg/(height in m)^2)	Numerical	11 (1.43%)
7.	Diabetes pedigree function	Numerical	0
8.	Age (years)	Numerical	0
9.	Class	0 and 1:– Absence/ Presence of diabetes disease	0

values lies between 0 and 1. The value 1 for ROC indicates efficient classifier. If ROC takes value between 0.5 and 1, 50% chance that a classifier can distinguish the classes. The value less than 0.5 indicate the test is made of no use.

3.3.5 F-measure

F-measure combines precision and recall[28]. The effectiveness of classification algorithm increases for higher values of F-Measure. The range of F-Measure is (0, 1).

$$F-Measure = \frac{2*Precision*Recall}{Precision+Recall} \quad (10)$$

4. Results and Discussions

The experiments were set up and conducted on open source tools R version 3.5.2 and WEKA (Waikato Environment for Knowledge Analysis) version 3.7.2. The preprocessing stage is done under R and the classifier evaluation is performed in WEKA. Preprocessing is done by replacing the outliers by 5 and 95th percentile values. Figure 2 shows the presence of outliers and its replacement by the percentile values. After removing noise, the dataset is imputed using median and the descriptive statistics obtained for stage 1 is shown in Table 2. Stage 2 is normalization, where the dataset is normalized using Z-score. Table 3 presents the range of values obtained before and after normalization. All the attributes are scaled such that the value lies between a range of (-1, 1).

In stage 3, the normalized data is balanced using SMOTE. In Figure 3, it is represented that class with "absence" value has the minority instances, therefore subset of data is taken from this class and new synthetic instances are created and then added to the original dataset. Before balancing, the dataset contains 768 instances (268 – "absence", 500 – "presence") and after balancing it were 1036 instances (536 – "absence", 500 – "presence"). The preprocessed data from stage 3 is evaluated using the classifiers in WEKA environment. For classification, 10-fold cross validation method is used where data set is randomly partitioned into 10 equal sized partitions. Each partition is tested against the remaining set of training data. This is repeated for all 10 partitions and a mean accuracy of the results is calculated and reported. The classifiers involved in the experiment are SVM, RF and Knn. Classification is performed for the

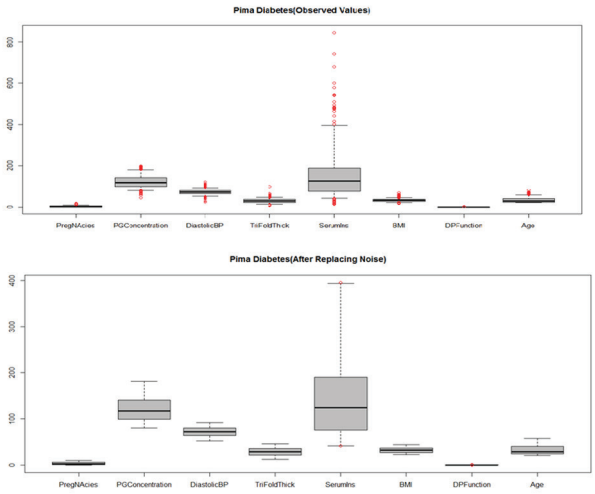

Figure 2. Outlier removal by 5 and 95% percentile.

Table 2. Descriptive statistics of data obtained after replacing noise and imputation (Stage 1)

S. No	Attributes	Observed Values		After Replacing Noise		After Imputation	
		Mean	Std. dev	Mean	Std. dev	Mean	Std. dev
1.	Number of times pregnant	3.84	3.36	3.74	3.12	3.74	3.12
2.	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	121.70	30.53	121.7	28.81	121.64	28.71
3.	Diastolic blood pressure (mm Hg)	72.41	12.38	72.27	10.87	72.26	10.61
4.	Triceps skin fold thickness (mm)	29.15	10.47	28.96	9.44	28.97	7.92
5.	2-Hour serum insulin (mu U/ml)	155.55	118.77	149.13	94.45	137.38	68.68
6.	Body mass index (weight in kg/(height in m)^2)	32.46	6.92	32.33	6.25	32.33	6.20
7.	Diabetes pedigree function	0.47	0.33	0.45	0.27	0.45	0.27
8.	Age (years)	33.24	11.76	32.95	11.00	32.95	11.00

Table 3. Range of values before and after normalization (Stage 2)

S. No	Attributes	Before Normalization		After Normalization	
		Min	Max	Min	Max
1.	Number of times pregnant	0	10	-1	1
2.	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	80	181	-1	1
3.	Diastolic blood pressure (mm Hg)	52	92	-1	1
4.	Triceps skin fold thickness (mm)	13	46	-1	1
5.	2-Hour serum insulin (mu U/ml)	41.65	395.5	-1	1
6.	Body mass index (weight in kg/(height in m)^2)	22.2	44.5	-1	1
7.	Diabetes pedigree function	0.14	1.133	-1	1
8.	Age (years)	21	58	-1	1

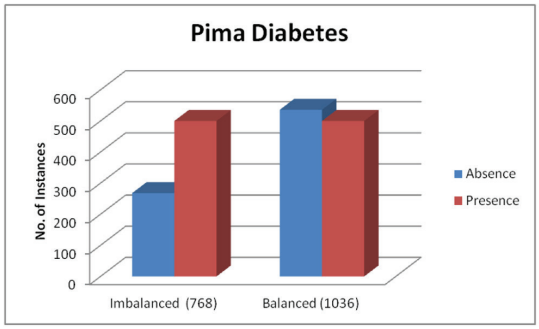

Figure 3. Distribution of class variable for Pima Diabetes dataset (Stage 3).

Table 4. Evaluation metrics obtained by SVM

SVM	Stage 1	Stage 2	Stage 3
Accuracy	76.95	77.08	74.13
Balanced Accuracy	72.08	72.18	74.18
Kappa	0.465	0.468	0.482
ROC	0.721	0.722	0.742
F-measure	0.762	0.763	0.741

Table 5. Evaluation metrics obtained by Random Forest

Random Forest	Stage 1	Stage 2	Stage 3
Accuracy	75	75.65	82.14
Balanced Accuracy	71.79	72.55	81.94
Kappa	0.442	0.457	0.641
ROC	0.823	0.822	0.895
F-measure	0.748	0.755	0.821

Table 6. Evaluation metrics obtained by K-nearest neighbor

Knn	Stage 1	Stage 2	Stage 3
Accuracy	66.92	66.92	80.21
Balanced Accuracy	63.43	63.43	79.84
Kappa	0.269	0.269	0.601
ROC	0.641	0.641	0.796
F-measure	0,669	0,669	0.799

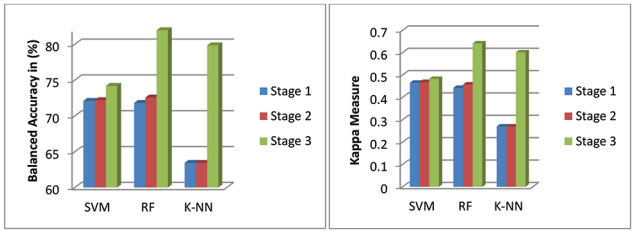

Figure 4. Balanced accuracy and Kappa measure at each stage of preprocessing.

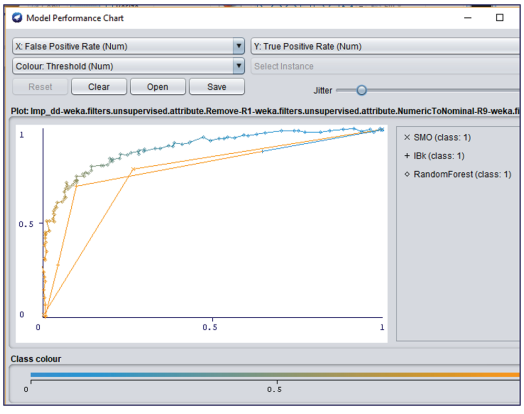

Figure 5. Comparison of area under ROC curve.

dataset obtained at each stage of data preprocessing and the evaluation metrics obtained are reported in Tables 4-6. It is very clearly observed from the table that the evaluation metrics obtained increases for the subsequent steps of preprocessing. Balanced accuracy and Kappa which is considered to be an important measure for imbalanced datasets has a tremendous increase in value after applying smote (stage 3).

Balanced accuracy and kappa is considered to be an essential metric for imbalanced datasets. Figure 4 shows the comparison of balanced accuracy and kappa at all the stages of preprocessing. For all the classifiers tested, it is clearly proved that balanced accuracy and kappa increases at stage 3 after balancing the dataset. Balanced accuracy and kappa has an average of 9.2% and 7% increase in value after balancing the dataset. It is proved that among the tested classifiers, RF provides better accuracy and roc compared to SVM and Knn. The proposed model using RF, obtains a classification accuracy of 82.14% which is a 8.01% high using SVM and 1.93% high using Knn. The reliability of the classifier is obtained using 10 fold cross validation technique. Figure 5 shows the roc obtained by the classifiers at stage 3. It was very clear that RF has a highest roc when compared with other tested classifiers.

5. Conclusion and Future Work

Pima Diabetes dataset contains missing values and noisy data which can adversely affect the performance of classifier. Such defective data need to be pre-processed and then fed into the classifier model. The aim of this proposed system is to implement a data preprocessing model that eliminates the poor quality of data present. Hence a data preparation model is identified and developed for Pima Diabetes dataset, which involves noise removal, imputation, normalization and balancing the dataset. The results from the experiment show that, each step of data preprocessing has a significant effect on the performance of the classifiers. Random forest classifier works well with Pima Diabetes dataset, with accuracy = 82.14, balanced accuracy = 81.94, kappa = 0.641 and roc = 0.895. F-measure takes the highest value of 0.821 for the RF classifier. Balanced accuracy is to be considered important for imbalanced datasets which increases for every stage of outlier+median imputation, Z-score normalization and balancing the dataset by SMOTE.

In future the data preparation methods can be applied to other data sets and the same can be applied for model construction. Different sampling techniques like Modified Synthetic Minority Oversampling Technique (MSMOTE) and bagging can be applied for balancing the dataset. Depending on the nature of spread of data in the dataset, normalization like min-max can also be attempted.

6. References

1. Myatt GJ. Making sense of data a practical guide to exploratory data analysis and data mining. John Wiley and Sons; 2007. https://doi.org/10.1002/0470101024
2. Han J, Kamber M. Data mining concepts and techniques. 2nd Ed. Morgan Kauffmann Publishers; 2006.
3. Anbarasi MS. Outlier detection for multi dimensional data. IJCSIT. 2011; 2(1):512–51.
4. Pachgade SD, Dhande SS. Outlier detection over data set using cluster-based and distance-based approach. International Journal of Advanced Research in Computer Science and Software Engineering. 2012; 2(6):1–5.
5. Santhanam T, Padmavathi MS. Comparison of K-means clustering and statistical outliers in reducing medical datasets. International Conference on Science Engineering and Management Research; 2014. p. 1–6. https://doi.org/10.1109/ICSEMR.2014.7043602
6. Santhanam T, Padmavathi MS. An efficient model by applying genetic algorithms for outlier detection in classifying medical datasets. Australian journal of Basic and Applied Sciences. 2015; 9(27):583–91.
7. Acuna E, Rodriguez C. The treatment of missing values and its effect in the classifier accuracy. D. Banks, L. House, F.R. McMorris, P. Arabie, W. Gaul, editors. Classification, Clustering and Data Mining Applications: Springer-Verlag Berlin-Heidelberg; 2004. p. 639–48. https://doi.org/10.1007/978-3-642-17103-1_60
8. Humphries M. Missing data and how to deal. An overview of missing data. Population Research Center; 2017. p. 1–45.
9. Shalabi LA, Zyad Shaaban. Normalization as a preprocessing engine for data mining and the approach of preference matrix. International Conference on Dependability of Computer System. IEEE; 2006. p. 2007–14. https://doi.org/10.1109/DEPCOS-RELCOMEX.2006.38
10. Gopal Krishna Patro S, Sahu KK. Normalization: A Preprocessing Stage. 2015.
11. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002; 16:1–37. https://doi.org/10.1613/jair.953

12. Michie D, Spiegelhalter DJ, Taylor C. Machine learning, neural and statistical classification. EllisHorwood; 1994. p. 1–298.
13. Choubey DP, Paul S, Bala K, Kumar M, Singh UP. Implementation of a hybrid classification method for diabetes. Intelligent Innovations in Multimedia Data Engineering and Management. IGI Global; 2019. p. 201–40. https://doi.org/10.4018/978-1-5225-7107-0.ch009
14. Thungrut W, Wattanapongsakorn N. Diabetes classification with fuzzy genetic algorithm. Unger H, Sodsee S, Meesad P, editors. Recent Advances in Information and Communication Technology. Advances in Intelligent Systems and Computing. Springer; 2018. p. 107–14. https://doi.org/10.1007/978-3-319-93692-5_11
15. Nilashi M, Ibrahim OB, Mardani A, Ahani A, Jusoh A. A soft computing approach for diabetes disease classification. Health Informatics Journal. 2018; 24(4):379–93. PMid: 30376769. https://doi.org/10.1177/1460458216675500
16. Ndaba M, Pillay AW, Ezugwu AE. An improved generalized regression neural network for Type II diabetes classification. Gervasi O. et al. Editors. Computational Science and its Applications. Lecture Notes in Computer Science. Springer. Cham; 2018. p. 10963. https://doi.org/10.1007/978-3-319-95171-3_52
17. Huang L, Lu C. Intelligent diagnosis of diabetes based on information gain and deep neural network. IEEE. International Conference on Cloud Computing and Intelligence Systems (CCIS); Nanjing. China. 2018. p. 493–6. https://doi.org/10.1109/CCIS.2018.8691378
18. Ramezani R, Maadi M, Khatami SM. A novel hybrid intelligent system with missing value imputation for diabetes diagnosis. Alexandria Engineering Journal. 2018; 57(3):1883–91. https://doi.org/10.1016/j.aej.2017.03.043
19. Johansen MB, Christensen PA. A simple transformation independent method for outlier definition. Clinical Chemistry and Laboratory Medicine. 2018; 56:1524–32. PMid: 29634477. https://doi.org/10.1515/cclm-2018-0025
20. Standardize Function: Return a normalized value (z-score) based on the mean and standard deviation. CFI. 2019. https:// corporatefinanceinstitute.com /resources /excel / functions /z-score-standardize-function/
21. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research. 2002; 16:321–57. https://doi.org/10.1613/jair.953
22. Courant R, Hilbert D. Methods of Mathematical Physics. Wiley: New York, USA; 1953.
23. Liaw A, Wiener M. Classification and Regression by Random Forest. R News. 2002; 2(3):18–22.
24. Dudani SA. The distance-weighted K-nearest-neighbor rule .IEEE Transactions on Systems, Man and Cybernet. 1976; 6(4):325–7. https://doi.org/10.1109/TSMC.1976.5408784
25. Kaur H, Kumari V. Predictive modelling and analytics for diabetes using machine learning approach. Applied Computing and Informatics. 2018. https://doi.org/10.1016/j.aci.2018.12.004
26. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. International Conference on Pattern Recognition. Institute for Empirical Research in Economics, University of Zurich, Switzerland; 2010. https://doi.org/10.1109/ICPR.2010.764
27. Onan A. A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. Expert Systems with Applications. 2015; 42:6844–52. https://doi.org/10.1016/j.eswa.2015.05.006
28. Kumar N, Khatri S. Implementing WEKA for medical data classification and early disease prediction. 3rd International Conference on Computational Intelligence and Communication Technology (CICT); Ghaziabad. 2017. p. 1–6. https://doi.org/10.1109/CIACT.2017.7977277