EasyChair Preprint

№ 6587

Prediction of Diabetes Using Various Feature Selection and Machine Learning Paradigms

Simran Gill and Prathmesh Pathwar

September 13, 2021

Prediction of Diabetes using various Feature Selection and Machine Learning Paradigms

Abstract. Many health experts have identified diabetes as one of the most widespread diseases. Not only the underdeveloped but also developed countries have a vast majority of their citizens who suffer from diabetes. In one of the surveys by WHO (World Health Organisation), almost 170 million people are detected with diabetes. It is predicted to increase twofold by the coming decade. Many metabolites for example glucose are considered to be the vital reason for diabetes when present in great amounts. Serious concerns have been raised by health officials around the globe to cure and detect it at an early stage. With the advancement in technology and data mining techniques. This paper aims at developing a classifier and comparing different data mining techniques based on their accuracy for the detection of diabetes based on different symptoms and features. The machine learning techniques were applied to the Diabetes data-set provided by the Biostatistics program at Vanderbilt. The best accuracy (93.95%) was observed with the Genetic algorithm as a feature selection technique along with Random Forest for classification. Thus, Random Forest along with a Genetic Algorithm can be used for efficient diagnosis and prediction of diabetes.

Keywords: Mutual Information, Genetic Algorithm, ANOVA, Naive Bayes, Stochastic Gradient Descent, K Nearest Neighbour (KNN), Random Forest (RF), Support Vector Machines (SVM), Decision Tree, Logistic Regression.

1 Introduction

With the rapid growth and development of the countries, there has been a noted pattern of increase in sedentary lifestyle and the most pervasive disease commensurate with such lifestyle is diabetes [1]. To propel early detection of the disease, researchers have been pointing out human flaws and inefficacy due to human limitations and errors to properly analyze data and give robust results that can be consumed and a stratagem can be developed to solve it for each individual [2].

This paper aims to detect diabetes given the information of a person by analyzing mainly two aspects of every person: biological aspects used for analysis are Cholesterol, HDL, Glucose, Blood Pressure, and their physiological aspects are Height, Weight, BMI, and Waist/Hip Ratio.

The dataset is provided by the Biostatistics program at Vanderbilt. It is cleaned [3] for any missing or incorrect values and irrelevant columns are removed. Normalization is performed on the values to prevent over-fitting or domination of a feature due to high or low values. Then feature selection techniques are employed (ANOVA, Mutual Information, and Genetic Algorithm.) in tandem with multiple classification models (KNN, Decision Tree, Genetic Algorithm, SGD, Random Forest, SVM, and Naive Bayes), to empirically test which approach gives the best output for our use-case.

The parameters and techniques have been shown to provide enough insight to detect diabetes in a person without losing essential features and give an accuracy of about 94%. The impact of this paper can be noted for further research into each model and feature selection techniques to fine-tune hyperparameters for the dataset or employ better deep learning networks and models which have the elasticity to consider multivariate features.

2 Literature Review

With the increase in diabetes rate in the world, diabetes analysis and prediction has become a field of great importance [4].

Saru, S., and S. Subashree [5] applied the machine learning models on The Pima Indian diabetes database. They applied bootstrapping resampling technique and then used Naive Bayes, Decision Trees, and KNN predictive models to predict and compare their accuracy with 10 cross-validation. It was found that the proposed methodology gave an accuracy of 90.36%.

Alkaragole, Mohammed Layth Zubairi, and Sefer Kurnaz [6] analyzed the accuracy of various data-mining techniques, mainly decision tree, Naive Bayes, SVM, and hybrid algorithms. Hybrid algorithms(proposed ensemble SVM + decision tree with an iteration of 100) outperformed all the other algorithms with an accuracy of 94% and sensitivity of 91%.

Sneha, N., and Tarun Gangil [7] studied various classification algorithms to find an optimal classifier for diabetes prediction. The dataset was provided from the UCI machine repository archive and the study was performed on 5 classification algorithms: random forest, KNN, decision tree, Naive Bayes, and SVM. Naive Bayes had the best accuracy of 82.3%.

Aada, A., and Sakshi Tiwari [8] used PIMA Indian diabetes dataset for analysis, KNN, Naive Bayes, and decision tree were applied along with bootstrapping resembling methods. SVM provided the best accuracy of 94.44% after applying bootstrapping.

Srivastava, Suyash [9] applied machine learning algorithms, artificial neural networks on the Pima Indians dataset for the prediction of diabetes. It gave an accuracy of 92% which could further be increased if the size of the training dataset is increased.

Kaur, Harleen, and Vinita Kumari [10] conducted a study on the PIMA Indian diabetes dataset to predict and analyze the trends of diabetes. They used R data manipulation tool and 5 algorithms for prediction: SVM-linear, radial basis function kernel support vector machine, k-nearest neighbor, artificial neural network, and multi-factor dimensions reduction. SVM-linear model provided the best accuracy of 89% for diabetes prediction.

Maniruzzaman, Md [11] used the diabetes dataset from the National Health and Nutrition Examination Survey consisting of 6561 individuals, among which 657 were diabetic. Logistic Regression, Naive Bayes, decision tree, AdaBoost, and random forest were applied for diabetes prediction. The best accuracy of 94.25% was observed with logistic regression applied as feature selection and random forest for classification.

Prasad, K.S., Reddy, N.C.S. & Puneeth, B.N. [12] applied Naive Bayes, decision tree, J48, and random forest with 10 fold cross-validation. Random forest gave the maximum accuracy and Naive Bayes provided the least mean absolute error and root mean squared error.

3 Dataset And Experimental Setup

The analysis and prediction of diabetes are conducted on the dataset originally provided by the Biostatistics program at Vanderbilt and is downloaded from data.world. The dataset was collected by surveying several hundred rural African-American patients by diagnosing different parameters. The dataset contains different levels of the factors that help in predicting diabetes. There are a total of 390 peoples records containing 18 feature vectors including Cholesterol, Glucose, HDL Chol, Chol/HDL ratio, Age, Gender, Height, Weight, BMI, Systolic BP, Diastolic BP, waist, hip, Unnamed: 16, Unnamed: 17, Patient number and Waist/hip ratio. Most of the attributes are numerical in nature. Out of 390 records, 60 patients are found to be diabetic and 330 are found to be non-diabetic.

The dataset is divided in the ratio of 80/20, for training and testing, respectively. 80% of the dataset is used for training purposes, consisting of 312 samples, and 20% for testing, consisting of 78 samples, to establish the accuracy and precision of the algorithms. Table 1 depicts the distribution of the Vanderbilt dataset.

Table 1. Vanderbilt Dataset Distribution

Dataset Distribution	Total number of records	Diabetic	Non-Diabetic
Original Dataset	390	60	330
Training Dataset	312	47	265
Testing Dataset	78	13	65

4 Methodology

4.1 Data Visualization

Data visualization helps in analyzing the trends, relations, and correlations in a dataset. **Correlation Analysis** helps to select features that are strongly correlated to our target and prevents overfitting of our model [13]. Since our dataset is a high-dimensional data set consisting of 18 features, it is of at most importance to find a suitable set of features to avoid overfitting and reduce the time complexity and data redundancy.

4.2 Data Pre-processing

Datasets are present in raw forms and may contain null values, incorrect information, and redundant information, so data needs to be pre-processed before applying any models to increase the productivity and accuracy of the model.

4.2.1 Data Cleaning

Data Cleaning increases the validity and quality of the model by removing incorrect information and refining the data. Following methods were taken into account while cleaning the data [14].

Looking for missing values: In case of any tuple containing missing values or blank columns, the tuple is deleted from the dataset.

Removing redundant columns: To get a high-quality dataset, we try to refine it by reducing its dimensions or removing unnecessary columns.

Renaming the values of a feature: To form data in a consistent and homogeneous format, all data is converted into the numerical format. Table 2 depicts the attributes that were renamed.

Table 2. Modified feature values

Feature	Initial Value	Assigned Value
Gender	Female	0
	Male	1
Target	Non Diabetes	0
	Diabetes	1

4.2.2 Data Normalisation

The range of each feature varies and needs to be scaled so that each feature has equal weight and contribution to the model. Data were normalized to a range of [0,1] by the Min-max normalization method using the following algorithm (equation 1) [15]:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \quad (1)$$

4.3 Feature Selection

High dimensional datasets contain a large number of features which increases the computational cost and disturbs the accuracy of the model [16]. Our dataset consists of numerical inputs and categorical output. All the features after normalisation vary between [0,1] and our target class has values {"Diabetes" : 1 , "No Diabetes" : 0}.

4.3.1 ANOVA

ANOVA is a feature selection technique based on calculating the means between different groups within a dataset. F-scores are calculated for each feature and the features are selected accordingly [17].

4.3.2 Mutual Information

Mutual information (MI) is a measure of calculating the information provided by a variable given another variable. It states the dependency between two variables.

4.3.3 Genetic Algorithm

The Genetic Algorithm is used for reducing the dimensions of a search algorithm and soar the performance of the classifier [18]. The methodologies of "Survival of the Fittest" and biological evolution is used for feature selection in genetic algorithms.

4.4 Data Classification

4.4.1 Logistic Regression (LR)

In linear regression, a threshold is decided for classifying into groups, whereas in binary logistic regression it uses a sigmoid function (equation 2) for defining the thresholds for classification [19]. For Y (output) tending to infinity, it is classified as 1, i.e. "Diabetes", else 0 i.e. "No Diabetes".

$$Y = \frac{1}{1 + e^{-z}} \tag{2}$$

4.4.2 Naive Bayes (NB)

The Naive Bayes Classification uses the Bayes theorem for defining its criteria for classification. It defines probabilities (equation 3 and equation 4) to predict the class and assumes all features to be independent of each other [20].

$$Q(p|y) = \frac{Q(y|p)\, Q(p)}{Q(y)} \tag{3}$$

$$Q(p|Y) = Q(y_1|p) * Q(y_2|p) * Q(y_3|p) * \ldots * Q(p) \tag{4}$$

where prior probability of class : Q(p), posterior probability: given attribute Q(p|y), prior probability of predictor: Q(p), probability of predictor: given class Q(y|p).

4.4.3 Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent is a very popular and widely used classifier. It is an optimization technique and is similar to gradient descent, however, in stochastic gradient descent sparse and irregular samples are selected [21].

4.4.4 K Nearest Neighbours (KNN)

KNN is a simple classification technique that classifies a data point by considering its neighbors [22]. It uses similarity measures (example: distance function) for predicting accuracy. Decreasing the number of neighbors might decrease the accuracy of the algorithm.

4.4.5 Decision Tree (DT)

Decision trees classify data by creating a top-down tree by dividing the dataset into smaller sub-datasets. ID3 along with entropy and information Gain is used recursively for building the decision tree, the root node of the tree signifies the classification [23].

4.4.6 Random Forest (RF)

Random forest classifies a dataset by creating several decision trees. It helps correct the over-fitting problem of decision trees. It selects the class by calculating the mode of the trees. It is a very efficient classifier [24].

4.4.7 Support Vector Machines (SVM)

Support Vector machine implements supervised learning algorithm for classification by creating the best fit hyper-plane with the help of support vectors, to divide the n-dimensional plane into classes for future prediction [25].

5 Experimental Results

The dataset was divided into an 80:20 ratio for training and testing. The target is labeled as "Target" and divided as 0 or 1, where 0 means "No Diabetes" and 1 means "Diabetes". Using this as our target class, analysis was made regarding the present states which can be used to make future predictions based on input parameters. The following attributes were used for prediction: Patient number, Cholesterol, Glucose, HDL Chol, Chol/HDL ratio, Age, Gender, Height, Weight, BMI, Systolic BP, Diastolic BP, Waist, Hip, Waist/Hip ratio, Diabetes, Unnamed 17, Unnamed 18. On data cleaning, 3 attributes (Patient number, Unnamed 17, Unnamed 18) were filtered out, and the following 15 features were used for prediction: Cholesterol, Glucose, HDL Chol, Chol/HDL ratio, Age, Gender, Height, Weight, BMI, Systolic BP, Diastolic BP, Waist, Hip, Waist/Hip ratio, Diabetes.

Before classification and prediction, feature selection techniques were used for reducing the dimensionality of the dataset and increasing the accuracy of classification algorithms.

5.1 Feature Selection Algorithms Analysis

5.1.1 ANOVA

Figure 1 depicts the scores of each feature of the Vanderbilt dataset after using ANOVA for feature selection, with Glucose having the maximum f-score.

A higher f-score means higher weightage during the selection of features. The top 5 ranks of the features based on the f-score are Glucose, Age, Chol/ HDL, Cholesterol, and waist.

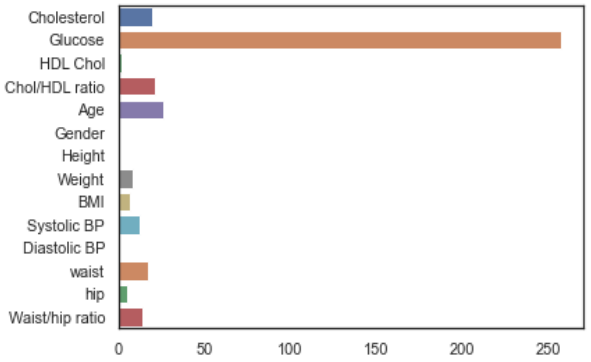

Fig.1 f-score of the features using ANOVA

5.1.2 Mutual Information (MI)

Figure 2 depicts the scores of each feature of the Vanderbilt dataset after using Mutual Information for feature selection, with Glucose having the maximum f-score.

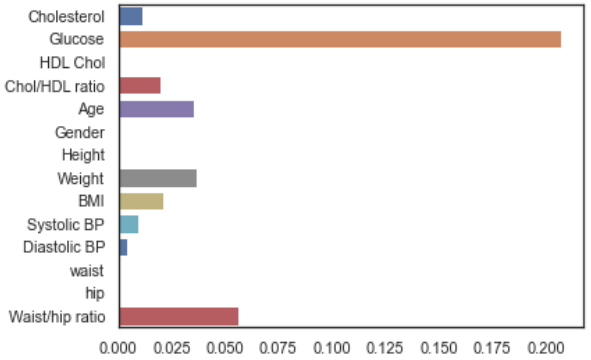

Fig.2 f-score of the features using MI

The top 5 ranks of the features based on the f-score are Glucose, hip, BMI, Chol/ HDL, and waist.

5.1.3 Genetic Algorithm

An individual is formed by a set of features. Scores of each individual were calculated for the Vanderbilt dataset by using the Genetic Algorithm for feature selection. Each algorithm was run on a Genetic Algorithm and individuals formed were scored for accuracy. Genetic Algorithms produced the best results for each algorithm for a specific individual as listed in Table 3.

Table 3. Selected Individual and total number of individuals evaluated by Genetic Algorithm

Algorithm	Selected Individual by Genetic Algorithm	#individuals evaluated
LR	Glucose, Age	491
NB	Cholesterol, Glucose, Diastolic BP , Hip	558
SGD	Glucose, Age, Weight, BMI, and Waist	539
KNN	Cholesterol, Glucose, Height, Weight, Hip	546
DT	Glucose, Age, Gender, Weight, BMI, Waist, Hip	529
RF	Cholesterol, Glucose, Chol/HDL, Systolic BP, Weight, Hip	518
SVM	Glucose, Gender.	562

5.2 Analysis of Classification Algorithms for predicting diabetes

5.2.1 Logistic Regression

The best accuracy for Logistic Regression was with ANOVA feature selection was 89.66%, Mutual Information's accuracy was 89.31%, and accuracy with Genetic Algorithm was 89.42%. Table 4 compares the accuracy and the feature selected.

Table 4. Comparison of the accuracy of logistic regression and features selected

Feature Selection	Features selected	Accuracy
ANOVA	Glucose and Age	89.66%
MI	Glucose, Hip, and BMI	89.31%
GA	Cholesterol, Glucose, HDL Chol, Chol/HDL, Age, Height, BMI, Systolic BP, Diastolic BP, Hip	89.42%

5.2.2 Naive Bayes

On analyzing it was observed that Naive Bayes exhibited the least accuracy with Mutual Information (92.82%), whereas the highest accuracy with Genetic Algorithm as a feature selection (93.59%). Table 5 compares the accuracy and the feature selected.

Table 5. Comparison of the accuracy of Naive Bayes and features selected

Feature Selection	Features selected	Accuracy
ANOVA	Glucose, Age, Chol/HDL ratio, Cholesterol, and Waist	92.91%
MI	Glucose, Hip, and BMI	92.82%
GA	Cholesterol, Glucose, Diastolic BP, and Hip	93.59%

5.2.3 Stochastic Gradient Descent (SGD)

SGD had the best accuracy with Genetic Algorithm feature selection was 93.58%, and 5 features selected were: Glucose, Age, Weight, BMI, and Waist. Mutual Information's accuracy was 91.54%, and accuracy with ANOVA was 90.68%. Table 6 compares the accuracy and the feature selected.

Table 6. Comparison of the accuracy of SGD and features selected

Feature Selection	Features selected	Accuracy
ANOVA	Glucose, Age, Chol/HDL ratio, Cholesterol, Waist, Waist/Hip Ratio, Systolic BP, Weight, and BMI	90.68%
MI	Glucose, Hip, and BMI	91.54%
GA	Glucose, Age, Weight, BMI, and Waist	93.58%

5.2.4 K Nearest Neighbours

On comparing the accuracies, it was found that in the case of KNN, mutual information had the least accuracy of 92.30%, followed by ANOVA, with an accuracy of 92.65%. Eventually, the best accuracy with the Genetic Algorithm feature selection was 93.91%.

Table 7. Comparison of the accuracy of KNN and features selected

Feature Selection	Features selected	Accuracy
ANOVA	Glucose, Age and Chol/HDL	92.65%
MI	Glucose	92.30%
GA	Cholesterol, Glucose, Height, Weight, and Hip	93.91%

5.2.5 Decision Tree

The best accuracy for the Decision Tree was with the Genetic Algorithm feature selection was 90.06%, Mutual Information's accuracy was 88.12%, and accuracy with ANOVA was 88.54%. Table 8 compares the accuracy and the feature selected.

Table 8. Comparison of the accuracy of decision tree and features selected

Feature Selection	Features selected	Accuracy
ANOVA	Glucose, Age, Chol/HDL ratio, Cholesterol, and Waist	88.54%
MI	Glucose	88.12%
GA	Cholesterol, Glucose, Diastolic BP, and Hip	90.06%

5.2.6 Random Forest

Random Forest, when implemented with the Genetic Algorithm as a feature selection technique, provided the best accuracy of 93.95%, On the other hand, Mutual Information's accuracy was 92.22%, and the accuracy with ANOVA was 92.65%. Table 9 compares the accuracy and the feature selected.

Table 9. Comparison of the accuracy of random forest and features selected

Feature Selection	Features selected	Accuracy
ANOVA	Glucose, Age, and Chol/HDL ratio	92.65%
MI	Glucose and Hip	92.22%
GA	Cholesterol, Glucose, Chol/HDL, Systolic BP, Weight, and Hip	93.95%

5.2.7 Support Vector Machine

In all the feature selection techniques, the best results were found with Glucose as a primary feature selected. Moreover, gender along with glucose provided the best accuracy for SVM with the Genetic Algorithm (93.27%), Mutual Information's accuracy was 92.39%, and accuracy with ANOVA was 92.56%.
Table 10 compares the accuracy and the feature selected.

Table 10. Comparison of the accuracy of SVM and features selected

Feature Selection	Features selected	Accuracy
ANOVA	Glucose and Age	92.56%
MI	Glucose	92.39%
GA	Glucose and Gender	93.27%

5.3 Comparative Study

Analyzing the maximum accuracy of each algorithm as shown in Table 11, the best accuracy of **93.95%** for Vanderbilt Dataset for diabetes detection was analyzed for **Random Forest** with **Genetic Algorithm** as a feature selection technique, features

selected as **Cholesterol, Glucose, Chol/HDL, Systolic BP, Weight, and Hip** and **depth of 5**.

Table 11. Comparison of the accuracy of the models

Algorithm	Accuracy
Logistic Regression	89.66%
Naive Bayes	93.59%
SGD (Hinge)	93.58%
KNN	93.91%
Decision Tree	90.06%
Random Forest	**93.95%**
SVM	93.27%

6 Future Work

This work could be further extended as :

- Since our current dataset contains only 312 samples, to analyze the effectiveness and accuracy of the algorithms, this research could be extended to a larger dataset for diabetes prediction.
- This project is worked upon by using records of African-American patients, it could further be extended using the records of more diverse patients, to determine that the predictions are not region-based.
- Advanced feature selection and classification algorithms can be applied.

7 Conclusion

This paper attempts to analyze the diabetes symptoms and gather meaningful insights which can help the health experts in deciding the early symptoms and diagnosis. The data is analyzed using various data mining techniques such as Feature Selection and Classification. All these are used to analyze the trends and predict the symptoms of diabetes. Feature Selection techniques such as ANOVA, Mutual Information, and Genetic Algorithm were used to increase the accuracy and reduce the overhead and training time of the model. Logistic Regression, Naive Bayes, SGD Classifier, KNN, Random Forest, Decision tree, and Support Vector Machine algorithms were used to predict diabetes. A comparative study of all the applied algorithms has been done by computing their accuracy and Random Forest showed the best accuracy of 93.95% with Genetic Algorithm as a feature selection technique, features selected as Cholesterol, Glucose, Chol/HDL, Systolic BP, Weight, and Hip and random forest depth of 5. Hence, it will be very useful to predict diabetes at an early stage, by keeping a proper check on the patients Cholesterol, Glucose, Chol/HDL, Systolic BP, Weight, and Hip. An abnormality in any of the above parameters could suggest the presence of diabetes and an early treatment could be provided to curb it.

References

1. Ramachandran, Ambady, Ronald Ching Wan Ma, and Chamukuttan Snehalatha. "Diabetes in asia." The Lancet 375.9712 (2010): 408-418.
2. Li, Ninghua, et al."Effects of lifestyle intervention on long‐term risk of diabetes in women with prior gestational diabetes: A systematic review and meta‐analysis of randomized controlled trials." Obesity Reviews 22.1 (2021): e13122.
3. Li, Peng, et al."CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks." 36th IEEE International Conference on Data Engineering (ICDE 2020)(virtual). ETH Zurich, Institute for Computing Platforms, 2021.
4. Zhou, Bin, et al."Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4· 4 million participants." The Lancet 387.10027 (2016): 1513-1530.
5. Saru, S., and S. Subashree. "Analysis and prediction of diabetes using machine learning." International Journal of Emerging Technology and Innovative Engineering 5.4 (2019).
6. Alkaragole, Mohammed Layth Zubairi, and Sefer Kurnaz. "COMPARISON OF DATA MINING TECHNIQUES FOR PREDICTING DIABETES OR PREDIABETES BY RISK FACTORS." (2019).
7. Sneha, N., and Tarun Gangil. "Analysis of diabetes mellitus for early prediction using optimal features selection." Journal of Big data 6.1 (2019).
8. Aada, A., and Sakshi Tiwari. "Predicting diabetes in medical datasets using machine learning techniques." Int. J. Sci. Eng. Res 5.2 (2019).
9. Srivastava, Suyash, et al."Prediction of Diabetes Using Artificial Neural Network Approach." Engineering Vibration, Communication and Information Processing. Springer, Singapore, 2019. 679-687.
10. Kaur, Harleen, and Vinita Kumari. "Predictive modelling and analytics for diabetes using a machine learning approach." Applied computing and informatics (2020).
11. Maniruzzaman, Md, et al."Classification and prediction of diabetes disease using machine learning paradigm." Health Information Science and Systems 8.1 (2020).
12. Prasad, K.S., Reddy, N.C.S. & Puneeth, B.N. A Framework for Diagnosing Kidney Disease in Diabetes Patients Using Classification Algorithms. SN COMPUT. SCI. 1, 101 (2020).
13. Kumar, Sunil, and Ilyoung Chong. "Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states." International journal of environmental research and public health 15.12 (2018): 2907.
14. Rahm, Erhard, and Hong Hai Do. "Data cleaning: Problems and current approaches." IEEE Data Eng. Bull. 23.4 (2000): 3-13.
15. Borkin, Dmitrii, et al."Impact of Data Normalization on Classification Model Accuracy." Research Papers Faculty of Materials Science and Technology Slovak University of Technology 27.45 (2019): 79-84.
16. Bennasar, Mohamed, Yulia Hicks, and Rossitza Setchi. "Feature selection using joint mutual information maximisation." Expert Systems with Applications 42.22 (2015): 8520-8532.
17. Elssied, Nadir Omer Fadl, Othman Ibrahim, and Ahmed Hamza Osman. "A novel feature selection based on one-way anova f-test for e-mail spam classification." Research Journal of Applied Sciences, Engineering and Technology 7.3 (2014): 625-638.
18. Babatunde, Oluleye H., et al."A genetic algorithm-based feature selection." (2014).
19. Kurt, Imran, Mevlut Ture, and A. Turhan Kurum. "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease." Expert systems with applications 34.1 (2008): 366-374.

20. Jiang, Liangxiao, et al."Survey of improving naive bayes for classification." International Conference on Advanced Data Mining and Applications. Springer, Berlin, Heidelberg, 2007.
21. Huang, Yuguang, and Lei Li. "Naive Bayes classification algorithm based on small sample set." 2011 IEEE International conference on cloud computing and intelligence systems. IEEE, 2011.
22. Yong, Zhou, Li Youwen, and Xia Shixiong. "An improved KNN text classification algorithm based on clustering." Journal of computers 4.3 (2009): 230-237.
23. Iyer, Aiswarya, S. Jeyalatha, and Ronak Sumbaly. "Diagnosis of diabetes using classification mining techniques." arXiv preprint rXiv:1502.03774 (2015).
24. Babatunde, Oluleye H., et al."A genetic algorithm-based feature selection." (2014).
25. Noble, William S. "What is a support vector machine?." Nature biotechnology 24.12 (2006): 1565-1567.