EasyChair Preprint

№ 2192

A Comparative assessment of Data Mining Algorithms to predict fraudulent firms

Harshit Monish and Avinash Chandra Pandey

December 18, 2019

A Comparative Assessment of Data Mining Algorithms to Predict Fraudulent Firms

Harshit Monish
Jaypee Institute of Information Technology
Noida
harshitmonish20@gmail.com

Avinash Chandra Pandey
Jaypee Institute of Information Technology
Noida
avinash.pandey@jiit.ac.in

***Abstract*—The process of data mining is helpful in discovering meaningful patterns in historical or unstructured data in order to make better business decisions. It helps in creating a better marketing strategy and also helps in risk management, fraud detection, etc. In this study, we put forward a comparative analysis of data mining models for fraud detection. The goal of the analysis is to find the best model which gives high accuracy and is less compute-intensive. We have implemented Decision Trees, Linear Support Vector Machines, RBF Kernel Support Vector Machines, K-Nearest Neighbor, Artificial Neural Network and logistic regression classification models. Further, we have implemented PCA and Ensemble techniques to improve the accuracy of the model and decrease the computational complexity of the models.**

***Index Terms*—Text mining, Supervised Learning, Classification, Ensemble Learning**

I. Introduction

Data mining is the process of exploring and analyzing the large chunks of data to glean meaningful information by trying to predict patterns and trends in the unstructured data [1]. Mathematical algorithms along with statistical rules are used in Data mining to analyze and find patterns, correlation, and links in the information [2]. From discerning the sentiments or opinion of users to spam filtering, credit risk management, and fraud detection [3], [4]. Businesses can learn more about their customers' behavior using data mining, look for patterns in the large datasets and can develop more effective marketing strategies, decrease the cost and increase sales and profit. Risk management has become one of the important research areas with the advent of data mining [5]. It includes analysis and acceptance or mitigation of uncertainty in investment decisions. There are a variety of methods exist to ascertain risk, one of the most common is standard deviation, a statistical measure of dispersion around a central tendency. Other methods include measures of the volatility, or systematic risk of a stock when compared to the entire market [6].

Trying to predict the risk of investing in a company or an emerging start-up is a tedious job for investors [7]. Many risk factors are pertinent in risk management; hence these factors are investigated from several areas like audit-paras, on-going issues report, environmental conditions report, past records of audit office, firm reputation summary, profit-value records, loss-value records, etc. The risk factors that are important and pertinent are evaluated, along with their probability of existence is computed from past data. Various researches are already going in this area to get efficient and comparable results and big companies are heavily investing in this domain. In this paper, we compare the classification models of Data mining that can predict the fraudulent firm based on current and past risk factors. Further, we apply the ensemble techniques to improve the models and finally compare the models based on accuracy and compute complexity. The dataset collected is multivariate having 18 attributes that have been considered as risk factors.

II. Methodology

A. The Overview of the Implementation

The proposed method aims to compare text mining models [8] and techniques based on various factors like accuracy, execution time and memory usage. Initially, the audit data has been collected having 18 risk factors and cleaned and formatted by removing the duplicates and outliers. Exploratory data analysis (EDA) is process that explores data statistically for finding patterns, anomalies, trends, or relationships [9]. These information can be further used in modeling decision. In short, the goal of EDA is to find the relevant features for decision making. In general, EDA starts with high level outline and then narrows to particular parts of dataset.

[9].

The data was then randomly partitioned into two parts, 70% of the data set was considered as training data and 30% of the data was considered as test data. Initially, Data mining classification models were implemented i.e. Logistic Regression, Support Vector Machines with linear kernel, Support Vector Machines with RBF kernel, K Nearest Neighbor, Decision Trees, Neural Network and Multiple Perceptron Neural Network to predict if a firm is fraudulent or not [10]. Ensemble techniques and feature extraction techniques like Principal Component Analysis were further implemented on these classification models to enhance the efficacy of model and the results were compared for each model.

III. Existing methods in the literature

A. Logistic Regression

Logistic regression is a supervised learning algorithm which predicts the outcome of a dependent variable which is categorical based on single or multiple independent variables [11],

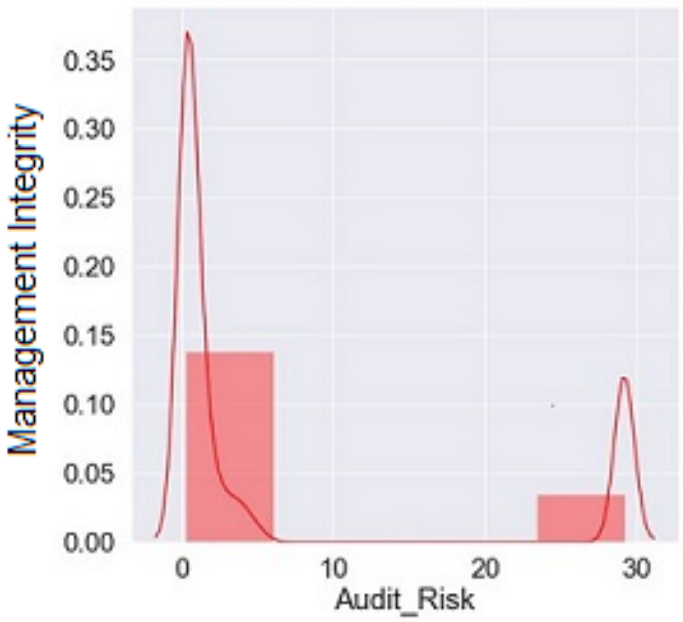


Fig. 1. Audit risk density plot

[12]. Generally, it is used for the problems in which dependent variable is binary. The algorithm uses the maximum likelihood estimation to find the regression coefficient of model and thereby predicts the probability of binary dependent variable or the output variable accurately using Eq. (1). Since the probability of any event lies between 0 and 1, the values above the threshold values are set to 1 and the less than threshold are set to 0. Logistic regression belongs to a larger class of algorithms known as the Generalized linear model.

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta T} x} \tag{1}$$

where, $g(z) = \frac{1}{1+e^{-z}}$

B. K-Nearest Neighbor

The k-nearest neighbor (KNN) is a supervised approach that employs labeled data to learn and produces appropriate result when new unlabeled data is given. In KNN, all similar information are kept in proximity; in other words, similar data are close to each other [13]. The value of K in KNN is empirically decided by its testing its performance for different values of K and the K value that reduces error and predict maximum number of instances correctly, is selected. For each point in the test dataset, we calculate the Euclidean distance for each K values and assign the test data to closest one. KNN Algorithm employs feature similarity to classify the testing data. KNN is also a non-parametric learning algorithm because it does not assume anything about the underlying data.

C. Support Vector Machines

A Support Vector Machine (SVM) is a discriminative classifier that uses hyper-plane to classify test samples. It is a supervised learning method that finds an optimal hyper-plane [14] in two-dimensional space which divides data sample into two parts where each class lay on either side. In SVM, we are looking to maximize the margin between the data point and the hyper-plane [15]. For notational simplicity, we consider the case of linear classifier function which is discussed in Eq. (2).

$$\gamma^i = y^i(W^T x + b) \tag{2}$$

where w and b are unknown and can be determined by optimizing the dual optimization equation as given Eq. (3) using Lagrange duality. We wish to minimize the equation by solving the following dual problem with linear inequality.

$$max_\alpha \quad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{j=1}^{m} y^i y^j \alpha_i \alpha_j (x^i, x^j) \tag{3}$$

here $\alpha_i \geq 0$, where, $i = 1, 2, \cdots, m$ and $\sum_{i=1}^{m} \alpha_i y^i = 0$.

The regularization parameter of SVM is used to find the wrongly classified instances in training data. For the same, partial derivatives of weights computed to find the gradient which updates the previous weights. In SVM, if there is no wrong classification, only regularization parameters are used for updating the gradient whereas when there is a wrong classification, loss along with the regularization parameters are employed to update the gradient.

The above equation can be modified if data belongs to more than two classes. For the same, kernel trick is used in which input parameter is mapped to higher dimensional space via nonlinear mapping. Gaussian kernel and RBF kernel are some of the widely used kernel. In this paper we have used RBF kernel trick to predict the fraudulent company. Kernel function is defined according to Eq. (4).

$$K(x, z) = \phi(x)^T \phi(z) \tag{4}$$

D. Decision trees

Decision trees (DT) creates a training model by learning decision rules from prior (training) data and these rules are used to predict the label or class of target (test) data [16], [17]. DT uses tree representation to solve the problem in which each leaf node corresponds to a class label and features/ attributes are represented by internal nodes. Decision tree for Boolean function is depicted in Fig. 2. The values of the features in DT are preferred to be categorical, if the values are continuous then they are converted to categorical values [18], [19].

The major challenge in DT is to identify the attribute for the root node at each level. For the same, information gain, gini index , and gain ratio are used. To compute the above-mentioned measures entropy for each feature is computed which helps to find the root node for DT. Attribute having higher entropy or information gain is placed to root node of the tree. The same process is repeated until decision tree is built.

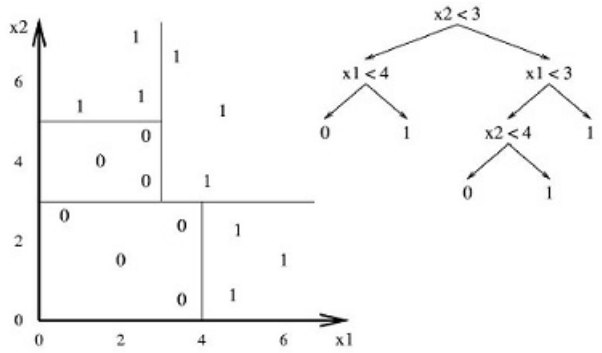


Fig. 2. Decision Tree

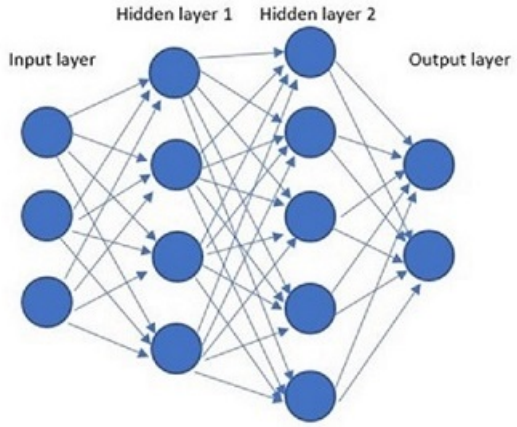
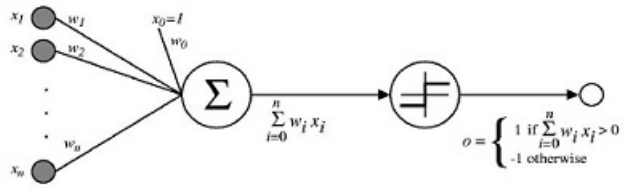


Fig. 3. Artificial Neural Network

E. Artificial Neural networks

Artificial neural network (ANN) is a computational model, based on the structure and functions of biological neural network [8]. It is a non-linear statistical model that establishes a relationship between input and output as shown in Fig. 3. It takes vector of inputs and finds the belongings of data from each classes. For this, a series of hidden layers and the linkage between the nodes created.

At each node the input from the data is combined with a set of weights that either dampen or amplify that input, hence assigning significance to input variables for the underlying hidden function that the algorithm is trying to learn. The product of input and weights are summed and passed further through an activation function in order to determine whether the signal should progress further through the network of layers and to what extent [20]. The training is guaranteed to succeed if the training examples are linearly separable and a small learning rate is used for gradient descent algorithm even when training data contains noise. The technique of backpropagation helps in fine-tuning the weights of a neural net based on the error rate obtained in the previous iteration. In forward propagation, we take the weighted sum of inputs of a unit and plug in the value into the activation function. Using this activation value, we get the input feature for the connected nodes in the next layer. Backpropagation is all about feeding the loss backward to update the weights. The optimization function i.e. Gradient Descent help us find the weights that yield a smaller loss in the next iteration. The same is also depicted in Fig. 4.



$$o(x_1, \ldots, x_n) = \begin{cases} 1 & \text{if } w_0 + w_1 x_1 + \cdots + w_n x_n > 0 \\ -1 & \text{otherwise.} \end{cases}$$

Fig. 4. Weight updating in ANN

F. Ensemble model and Techniques

Ensemble models and techniques are used to combine the decisions from multiple models in order enhance the overall performance of the system. Ensemble techniques have been proven to be very effective and impactful in solving the industrial challenges. In an ensemble-based data mining system there are four major parts: preprocessing the data, generating single classifier, ensemble processing, predicting and evaluating results [21]. For instance, if we consider the case of decision trees there are numerous factors, we should consider like what features we make our decision on, what is the threshold for classification. Taking these factors into account Ensemble methods help us to take a singular decision tree into consideration, calculate what all feature to use and finally make a predictor based on the consolidated and aggregated results of all decision trees that were sampled. Types of Ensemble methods include-

- Max Voting: In this Ensemble method we take multiple classification models to make predictions for every data point and these predictions by every model is considered as a vote. The final prediction is considered from the predictions that get maximum number of votes.
- Bagging: This technique combines the results of multiple models to get a generalized result. Bagging uses Bootstrapping sampling technique which divides the data into subsets of the original data with replacements and the size of each subset is same as the size of the original dataset [22]. Bagging technique make use of these subsets to get a generalized and fair overview of the distribution by running the models concurrently and independent of each other. The final prediction is evaluated by consolidating the predictions from all the models. One of the advantages of the bagging is that diversity comes into account as different bootstrap samples of the training data is used. Along with that the diversity among the member of ensemble is achieved and it has its origin in the statistical fluctuations of the random bootstrap sampling [23]. As

the number of classifiers that are aggregated increases, the error of the bagging technique becomes smaller. Bagging algorithms include bagging meta-estimator, random forest as shown in Fig. 5.

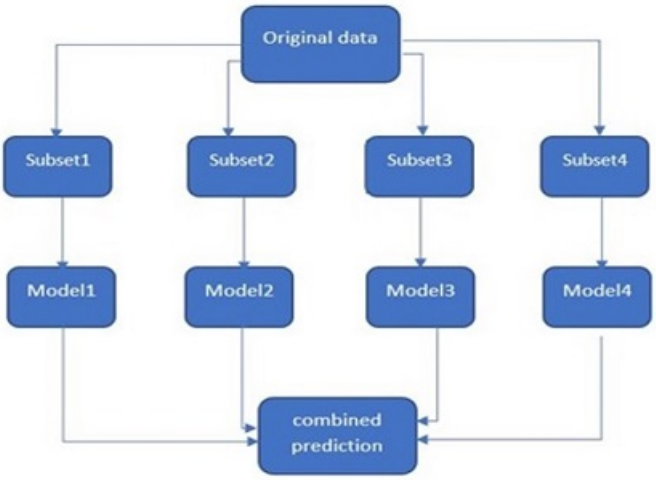


Fig. 5. Bagging Technique

- Boosting: Boosting is a process in which each subsequent model attempts to enhance the error of the previous model. Therefore, it is a sequential process where the models that are succeeding are dependent on the previous model [23].. In boosting we create a subset from original dataset on which a base model is created. This model is used to make predictions on the complete dataset and errors are calculated by comparing the actual values and the predicted values of the model. The incorrectly predicted observation is given higher weights. Similarly, we create multiple models on each subset of the training data which in turns corrects the errors of the previous model. The model that is considered final model is the weighted mean of all the models. One of the advantages of boosting is that since Individual models struggle to perform well on the complete dataset, but they work adequately for some part of the dataset, hence each model boosts the complete performance of the ensemble. Boosting algorithms include AdaBoost, Cat Boost, Light GBM.

G. Principal Component Analysis

Principal component analysis is the process of linear dimensionality reduction by implementing Singular Value Decomposition of the data to project it to a lower dimensional space [24]. PCA tries to identify the subspace in which the data approximately lies. Prior to PCA the data is pre-processed to normalize its mean and variance. The data is known to have zero mean and unit variance. After normalization is done, we try to find the major axis of variation, that is the direction on which the data approximately lies by finding the unit vector which maximizes the variance of the projected data. Hence

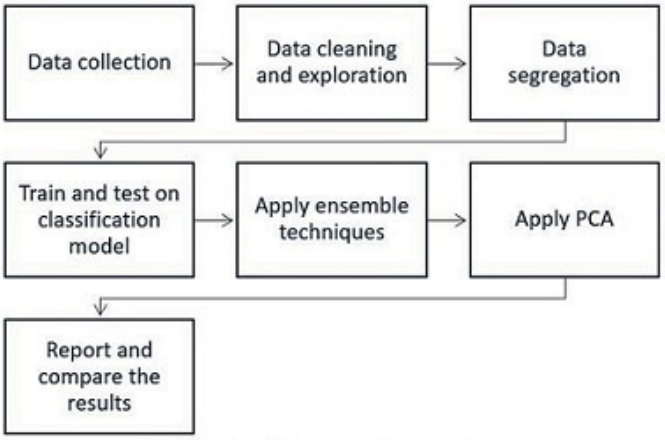


Fig. 6. Flow Chart of the Proposed Implementation

to maximize the variance of the projections, Eq. (5) can be maximized.

$$\frac{1}{m}\sum_{i=1}^{m}(x^{(i)T}u^2) = \frac{1}{m}\sum_{i=1}^{m}u^T x^{(i)}(x^{(i)})^T u = u^T\left(x^{(i)(x^{(i)}T}\right) \tag{5}$$

Maximizing this gives the principal eigenvector which is just the empirical co-variance matrix of the data. PCA have many benefits from compression to reduction in computational complexity, noise reduction.

IV. Proposed Solution and Implementation

The audit data is first collected and randomly segregated into training and test data. The segregation is done in such a manner that training data comprise of 70% of the data and testing data comprise of 30% of the data. After data segregation, data cleaning process is done in which we try to remove the redundant data and remove outliers in the data. Further as part of the data cleaning process we do Exploratory Data Analysis to find anomalies, patterns, trends, or relationships in the data. The data set comprise of various company's risk auditing and the attributes comprise of various risk factors.

First, we implement a basic classification model Logistic regression to try to predict if a company is fraudulent or not. We report the test accuracy and time taken by the algorithm and the memory consumption done by the algorithm. Similarly, we have implemented the K-Nearest neighbor algorithm by taking multiple values of k and the final value taken is 6.

Further we have implemented Support Vector Machine algorithm with Linear and RBF kernel, Decision trees and reported the results. Lastly Artificial Neural network algorithm is implemented with 3 hidden layers having 5 nodes each and single output layer. We have also implemented perceptron learning algorithm with activation function as sigmoid and Adam optimization algorithm. The loss function considered was binary cross-entropy. Second, we have implemented ensemble techniques like max voting – hard voting, soft voting, bagging and AdaBoosting, gradient boosting on the

classification algorithms and reported the results. Third we have implemented Principal Component Analysis with top 2 components, on the data set and then implemented above classification models to reduce the complexity and enhance the accuracy on the test data and reported the results for the same.

V. Experimental results and Conclusion

In this paper, supervised learning models and techniques have been compared for the problem statement of finding the fraudulent company using classification models i.e. Linear Support Vector Machines, RBF Kernel Support Vector Machines, Logistic Regression, Decision Trees, K-Nearest Neighbor and Artificial Neural Networks. Further PCA and Ensemble techniques were implemented to enhance the accuracy and reduce the compute complexity of the models. From this paper it can be shown that though maximum accuracy was given by AdaBoosting and bagging in Decision trees, but bagging is very much computing intensive, whereas AdaBoosting is comparatively less compute intensive and more accurate. PCA with RBF kernel and PCA with Decision Trees is also less compute intensive and comparably accurate but Ensemble AdaBoosting technique has given best accuracy and least computational complexity whereas Artificial Neural Network and Multi-layer perceptron and bagging with grid search indecision trees are highly compute exhaustive models for this kind of dataset. The performance results of all the models and techniques are reported in the following table.

TABLE I
Performance Results of all Models

Classification Model	**Performance Attribute**		
	Time(ms)	***Memory(MB)***	***Accuracy(%)***
Decision Trees	1.994	183.45	98.1
Linear SVM[a]	6.945	183.91	96.0
RBF kernel SVM	13.961	185.84	94.3
K-Nearest Neighbor	6.981	186.45	54.1
Logistic Regression	4.986	187.60	99.3
Ensemble Hard voting	112.698	173.82	98.0
Ensemble Soft voting	84.773	180.15	98.0
Ensemble Bagging RBF SVM	1385.262	190.43	92.6
Ensemble Bagging Linear SVM	783.91	169.58	98.3
Ensemble Bagging logistic Regression	41.89	171.74	98.9
Ensemble Bagging k-nearest neighbor	35.904	172.68	92.7
Ensemble Boosting Decision Tree	1.956	174.53	99.8
Ensemble Boosting logistic Regression	611.361	173.48	99.4
Artificial Neural Network	4709.389	399.663	92.1
NN Multi-layer perceptron	4071.046	422.09	96.3
Gradient Boosting	29185.36	170.73	98.4
PCA K-Nearest Neighbor	5.98	168.80	97.3
PCA Logistic Regression	31.94	168.94	93.7
PCA Linear SVM	7.98	169.85	91.6
PCA RBF SVM	4.986	170.13	99.4
PCA Decision Tree	1.995	169.74	95.8

VI. Future work

The research can be further extended to compare advanced deep learning algorithms for text classification such as rNN. New Advanced Ensemble techniques can be explored to make further comparisons of the models. Furthermore new Data mining techniques that are getting implemented on not just textual data but also on multimedia data like audio, video, images can be explored. These techniques are Multimedia Data mining, Ubiquitous Data mining, Distributed Data mining, Time Series and Sequence Data mining, Spatial and Geographic data mining. The capability of Data mining to integrate the unstructured data like images, text opens the doors to numerous exciting opportunities and possibilities for new research discovery in this domain

References

[1] D. J. Hand, "Data mining," *Encyclopedia of Environmetrics*, vol. 2, 2006.

[2] A. C. Pandey and D. S. Rajpoot, "Spam review detection using spiral cuckoo search clustering method," *Evolutionary Intelligence*, vol. 12, no. 2, pp. 147–164, 2019.

[3] A. Doglioni, A. Galeandro, and V. Simeone, "Data mining and data-driven modelling in engineering geology applications," in *Engineering Geology for Society and Territory-Volume 5*. Springer, 2015, pp. 647–650.

[4] A. C. Pandey, D. S. Rajpoot, and M. Saraswat, "Twitter sentiment analysis using hybrid cuckoo search method," *Information Processing & Management*, vol. 53, no. 4, pp. 764–779, 2017.

[5] V. Deparday, C. Gevaert, G. Molinario, R. Soden, and S. Balog-Way, "Machine learning for disaster risk management," 2019.

[6] M. Haddoud, A. Mokhtari, T. Lecroq, and S. Abdeddaïm, "Combining supervised term-weighting metrics for svm text classification with extended term representation," *Knowledge and Information Systems*, vol. 49, no. 3, pp. 909–931, 2016.

[7] M. F. Grace, J. T. Leverty, R. D. Phillips, and P. Shimpi, "The value of investing in enterprise risk management," *Journal of Risk and Insurance*, vol. 82, no. 2, pp. 289–316, 2015.

[8] A. C. Pandey, M. Garg, and S. Rajput, "Enhancing text mining using deep learning models," in *2019 Twelfth International Conference on Contemporary Computing (IC3)*. IEEE, 2019, pp. 1–5.

[9] A. T. Jebb, S. Parrigon, and S. E. Woo, "Exploratory data analysis as a foundation of inductive research," *Human Resource Management Review*, vol. 27, no. 2, pp. 265–276, 2017.

[10] B. Y. Pratama and R. Sarno, "Personality classification based on twitter text using naive bayes, knn and svm," in *2015 International Conference on Data and Software Engineering (ICoDSE)*. IEEE, 2015, pp. 170–174.

[11] S. Brindha, K. Prabha, and S. Sukumaran, "A survey on classification techniques for text mining," in *2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1. IEEE, 2016, pp. 1–5.

[12] C. Yin, J. Xiang, H. Zhang, J. Wang, Z. Yin, and J.-U. Kim, "A new svm

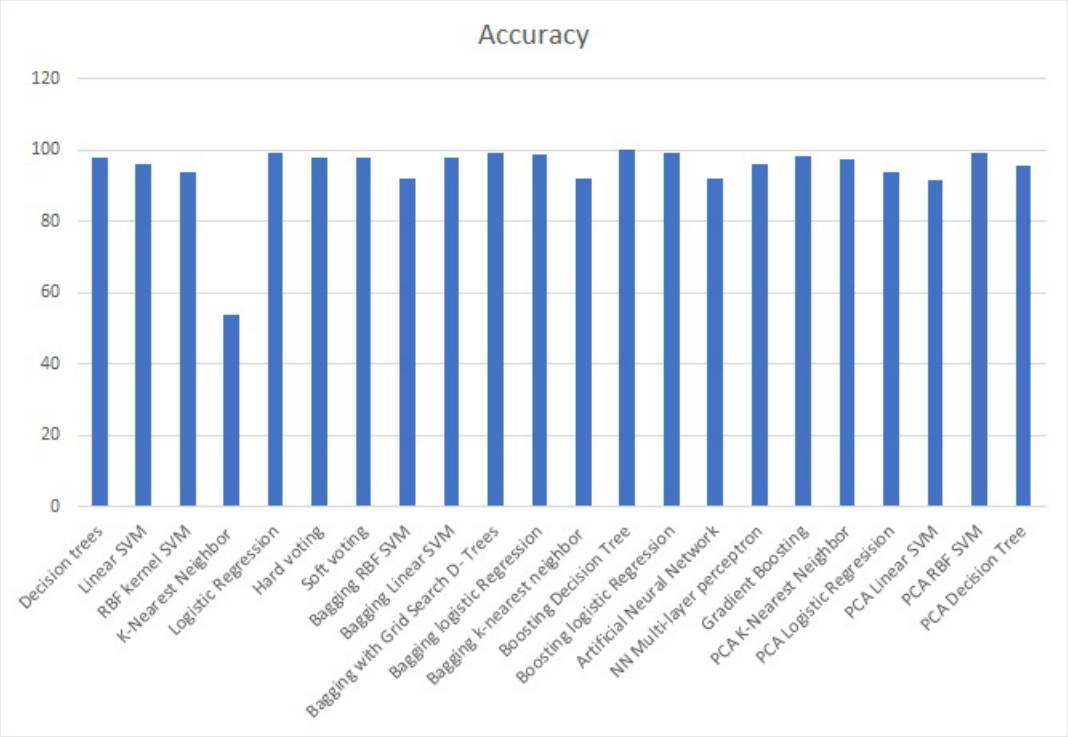


Fig. 7. Accuracy plot

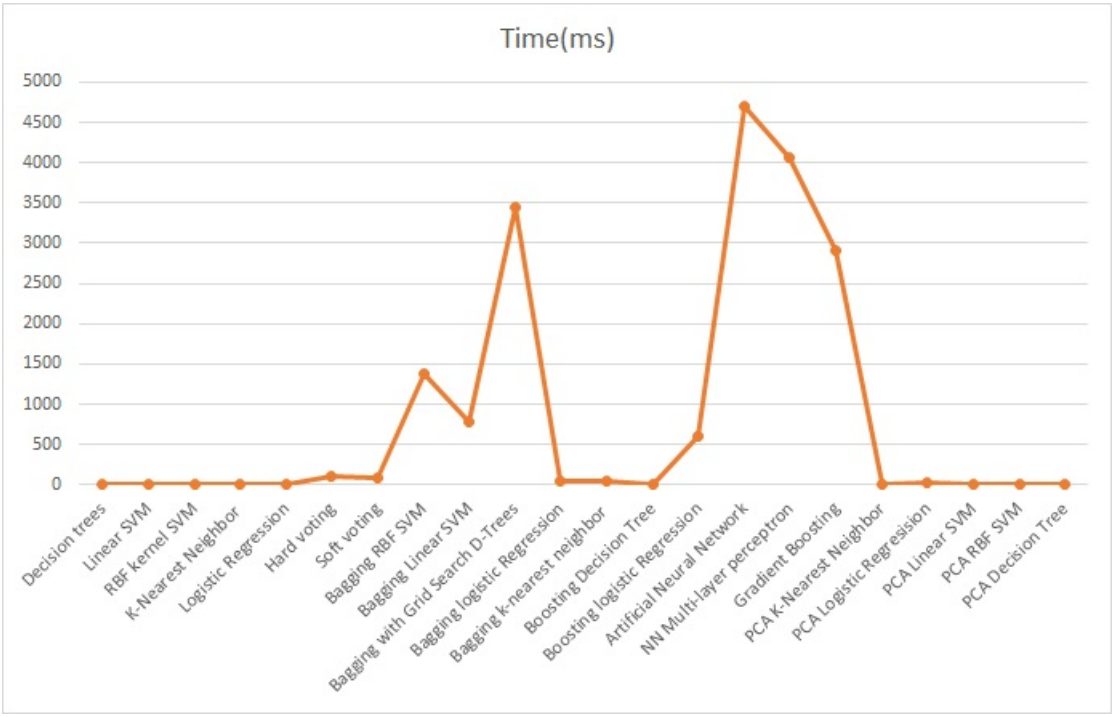


Fig. 8. Time consumption plot

method for short text classification based on semi-supervised learning," in *2015 4th International Conference on Advanced Information Technology and Sensor Application (AITS)*. IEEE, 2015, pp. 100–103.

[13] T. Denoeux, O. Kanjanatarakul, and S. Sriboonchitta, "Ek-nnclus: a clustering procedure based on the evidential k-nearest neighbor rule," *Knowledge-Based Systems*, vol. 88, pp. 57–69, 2015.

[14] B. T. Pham, D. T. Bui, M. Dholakia, I. Prakash, and H. V. Pham, "A comparative study of least square support vector machines and multiclass alternating decision trees for spatial prediction of rainfall-induced landslides in a tropical cyclones area," *Geotechnical and Geological*

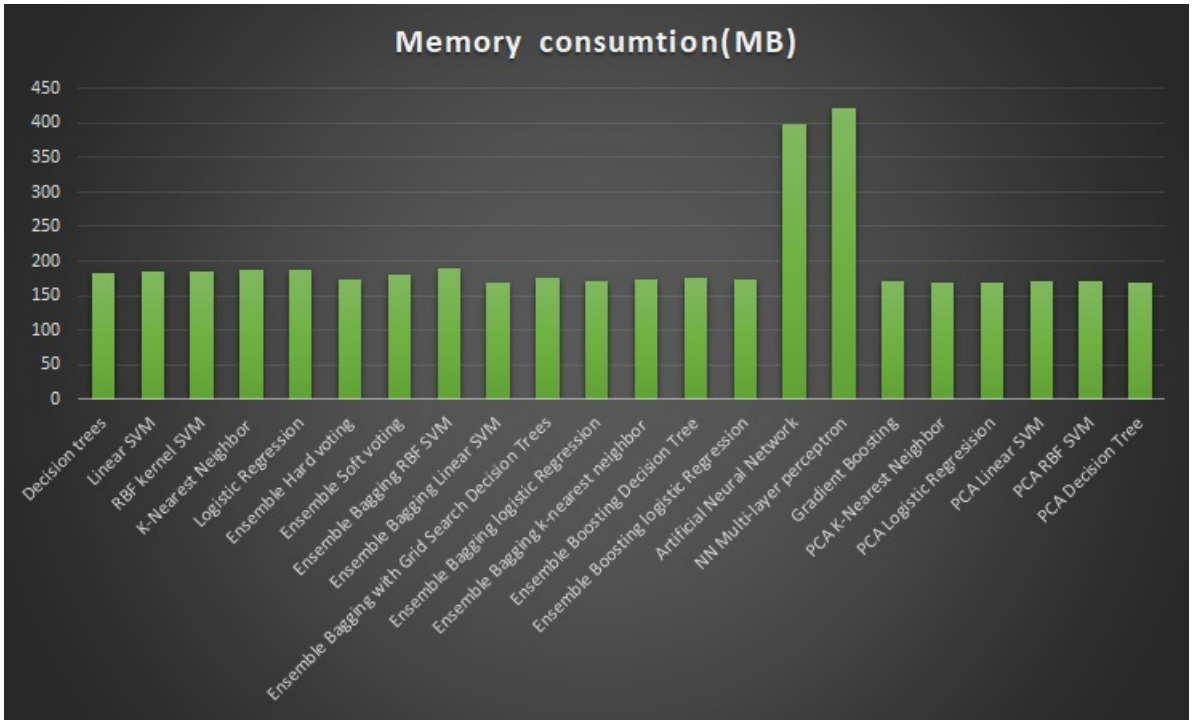


Fig. 9. Memory Consumption plot

Engineering, vol. 34, no. 6, pp. 1807–1824, 2016.

[15] A.-Z. Ala'M, H. Faris, J. Alqatawna, and M. A. Hassonah, "Evolving support vector machines using whale optimization algorithm for spam profiles detection on online social networks in different lingual contexts," *Knowledge-Based Systems*, vol. 153, pp. 91–104, 2018.

[16] A. C. Pandey, S. Misra, and M. Saxena, "Gold and diamond price prediction using enhanced ensemble learning," in *2019 Twelfth International Conference on Contemporary Computing (IC3)*. IEEE, 2019, pp. 1–4.

[17] P. Kaur, M. Singh, and G. S. Josan, "Classification and prediction based data mining algorithms to predict slow learners in education sector," *Procedia Computer Science*, vol. 57, pp. 500–508, 2015.

[18] A. Bifet, J. Zhang, W. Fan, C. He, J. Zhang, J. Qian, G. Holmes, and B. Pfahringer, "Extremely fast decision tree mining for evolving data streams," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 1733–1742.

[19] A. C. Pandey, S. R. Seth, and M. Varshney, "Sarcasm detection of amazon alexa sample set," in *Advances in Signal Processing and Communication*. Springer, 2019, pp. 559–564.

[20] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[21] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," *Neurocomputing*, vol. 174, pp. 806–814, 2016.

[22] S. Hajian, F. Bonchi, and C. Castillo, "Algorithmic bias: From discrimination discovery to fairness-aware data mining," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 2125–2126.

[23] B. Wang and J. Pineau, "Online bagging and boosting for imbalanced data streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 12, pp. 3353–3366, 2016.

[24] A. C. Pandey, D. S. Rajpoot, and M. Saraswat, "Feature selection method based on hybrid data transformation and binary binomial cuckoo search," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–20, 2019.