Mehran University Research Journal of Engineering and Technology
Vol. 41, No. 33 - 40, January 2022
p-ISSN: 0254-7821, e-ISSN: 2413-7219
DOI: https://doi.org/10.22581/muet1982.2201.04

Prediction of Insurance Fraud Detection using Machine Learning Algorithms

Laiqa Rukhsar[1a], Waqas Haider Bangyal[1b], Kashif Nisar[2a], Sana Nisar[2b]

RECEIVED ON 23.01.2021, ACCEPTED ON 05.05.2021

ABSTRACT

In current era, people are influenced with various types of insurance such as health insurance, automobile insurance, property insurance and travel insurance, due to the availability of extensive knowledge related to insurance. People are trending to invest in such kinds of insurance, which helps the scam artist to cheat them. Insurance fraud is a prohibited act either by the client or vendor of the insurance contract. Insurance fraud from the client side is encountered in the form of overestimated claims and post-dated policies *etc*. Although, insurance fraud from the vendor side is experienced in the form of policies from non-existent companies and failuew to submit premiums and so on. In this paper, we perform a comparative analysis on various classification algorithms, namely Support Vector Machine (SVM), Random-Forest (RF), Decision-Tree (DT), Adaboost, K-Nearest Neighbor (KNN), Linear Regression (LR), Naïve Bayes (NB), and Multi-Layer Perceptron (MLP) to detect the insurance fraud. The effectiveness of the algorithms are observed on the basis of performance metrics: Precision, Recall and F1-Score. The comparative results of classification algorithms conclude that DT gives the highest accuracy of 79% as compared to the other techniques. In addition to this, Adaboost shows the accuracy of 78% which is closer to the DT.

Keywords: Insurance, Fraud Detection, Supervised Learning, Classification Algorithm, Random Forest, SVM, Decision Tree

1. INTRODUCTION

The major issue faced by insurance companies is a fraud that causes immense loss to insurance companies sometimes beyond repair. The main concern is to avoid fraudulent activities at any cost because combating fraud cases specifically in insurance companies is a challenging task. It is reported that 21% - 36% of cases of auto-insurance claims are suspected to be fraudulent but only 3% of cases are prosecuted [1]. The first step to avoiding fraudulent cases is to detect them which is quite difficult and not "cost-effective" as well because the lengthy and cumbersome investigations may infuriate the authentic customers [2]. Higher investigation costs also cause hindrance in detecting fraud cases. Therefore companies go without carrying out appropriate investigations that lead to several future pitfalls. Manual fraud detection being costly and inefficient is outdated now; we need to investigate the fraud before the claim payment. Different machine learning and data mining techniques have proven to be promising in detecting frauds.

Machine Learning (ML) is a sub-area of Artificial Intelligence with the main aim to mimic human intelligence abilities. ML focuses on constructing models with high prediction capabilities. The most basic feature is "Learning" which is done by looking at the given data. The two basic learning techniques are Supervised and Unsupervised. In supervised

[1] Department of Computer Science, University of Gujrat, Gujrat, Pakistan. Email: [a]Laiqarukhsar123@gmail.com, [b]waqas.haider@uog.edu.pk (Corresponding Author)
[2] Faculty of Computing and Informatics, University Malaysia Sabah, K.K. Sabah, Malaysia. Email: [a]kashif@ums.edu.my, [b]sananisar48@gmail.com

learning, we are provided with fully labeled data that means in the training data against each input we have the desired result as well. It is highly useful for solving problems of classification and regression. In classification, the aim is to predict a discrete value whereas regression deals with continuous data. On contrary, in an unsupervised learning paradigm, we are provided with unlabeled data where results are not known. In a fraud detection scenario in a supervised learning method we can find out fraud and legal cases from training data but in unsupervised learning, we cannot infer which one is a fraud case and which one is legal.

One major task in ML is data classification that is also considered as pattern recognition. A classification problem is encountered when there is an urge to classify an instance in an already defined class based on its similarity to other instances classified into that class [3]. In classification, the aim is to develop such algorithms that dare to create models that can be used to differentiate the exemples/instances based on recognized patterns [4]. Classification is important for several different applications such as voice recognition, image classification, text categorization *etc*. [5]. There are many classification algorithms known that are proved highly beneficial in solving real-world problems. The most famous are K Nearest Neighbor, Support Vector Machine, Decision trees and Neural Networks.

ML can have wide range of applications, the most prominent ones are Social Media Services, Online Customer support, Email Spam Filtering, Fraud detection, Product Recommendation, and the list goes on.

2. LITERATURE REVIEW

Sun *et al* [6] f presented a novel approach for detecting frauds, called Patient Cluster Divergence-based Healthcare Insurance Fraudster Detection (PCDHIFD) in presence of camouflage responses. For the experimental purpose, the health care dataset was chosen and the dataset comprised of around 40M admission records of 10000 patients of the previous five years. The proposed technique worked in 3 steps for three basic records: Life history of patients, diagnosis record, and medical practitioners attended. Steps wee in this sequence: first of all, a patient graph was constructed based on most similar info for the patient level hospital admission. Then a clustering-based graph algorithm was used for finding the peak and real meaning for individual clusters. Lastly, the difference in the patient cluster was found and the probability of fraud for each patient was calculated. The comparison was made with other state of the art algorithms *i.e*. Decision Trees, Support Vector Machines, GridLOF, BP Growth, MLP and LSTM. It was claimed that the proposed approach produced the highest accuracy.

Dhieb *et al.* [7] proposed a method based on an Extreme Gradient Boosting (XGBoosting) algorithm for detecting frauds in insurance agreements. They presented an online learning solution for meeting online time to time requirements. The proposed method combined AI-based methods with blockchain architecture to get better security. The proposed machine learning technique worked in the following way: first preprocessing and cleaning of data was performed then data visualization techniques were used to get insights of data. The third step was to store data privacy by not disclosing personal information and lastly, model building using XGBoosting provided the probability of frauds in the future based on information available. A very fast Decision tree algorithm was designed for online learning solutions. Comparison of XGBoosting was made with other machine learning classifiers *i.e*. Decision Tree, Naïve Bayes, and Nearest Neighbor and the proposed methodology was better than all in terms of accuracy.

Kirlidog and Asuk [8] used the Support Vector Machine (SVM) for the detection of health insurance frauds and anomalies. Different data mining approaches were also discussed in their research. Research was done on the dataset of Turkish insurance companies which contained the total claimed records and other information of clients. The system was implemented in Oracle by using SVM with a linear kernel. The training was done by classifying the records in genuine claims and anomalies. SVM made classification by comparing individual records with genuine and fake claims. The system then calculated the probability for every single record and if the probability was higher than 50% than the record was

considered as an anomaly. Anomalies were considered based on three conditions: how many claims were rejected, how many uncontrolled claims were found in health center types, and how many claims were identified in health centers. Data mining approaches like grouping, classification, and variance detection could be used in insurance fraud detection, and based on the previous data we can predict future fraud claims using these techniques.

Bhowmik [9] applied different machine learning approaches for predicting and assessing fraud in automobile insurance. Machine learning approaches used in this research were Bayesian Networks, Decision Trees, and rule-based algorithms. This work created two Bayesian networks based on the assumptions *i.e.* driver is cheating and the driver is honest. And these two probabilities were calculated separately. And the one with the higher probability was considered as output. Decision trees were based on subtrees or labels known as classes *i.e.* legal and fraud in the research. Further Gini, minority, or entropy measures were used to get the impurity within a class and get the final output. Rule-based system proceeded in with if-then rules and here conditions were driver's age, driver's rating, and auto age. Results and performance were shown in the form of confusion matric and the accuracy was good.

Liu *et al.* [10] proposed a new technique for insurance fraud detection for an imbalanced dataset. The novel technique was based on data partitioning under-sampling with and without replacement on the majority class, and then it merges with the minority class. Tenfold cross-validation is used for testing purposes. The proposed methodology was based on the idea of choosing the best from data partitions under-sampling. The models used for insurance fraud detection were Support Vector Machines, Decision Trees, and artificial neural networks. Experiments were carried on a publicly available dataset containing the records of different automobile insurance claims. Results showed that the Decision Trees were the best among all the classifiers. It was demonstrated that the technique outperformed the previous work done and its accuracy was the best.

In this paper, eight popular ML algorithms are used for detection of Insurance fraud. A comparative analysis is also presented for the application.

3. TYPES OF CLASSIFICATION ALGORITHMS

Many machine learning algorithms are being used in various fields of research to help in solving the real-world problems. Mostly used machine learning classification algorithms are discussed below:

3.1 Support Vector Machine (SVM)

The SVM is a popular machine learning classifier that is used in our research. It is being applied for both linear and nonlinear problems in real-world domain areas [11]. A hyperplane is used to separate instances of classes in SVM. Because of its kernel function which is used to convert low dimension space to high dimension space SVM is best suited for nonlinear classification problems. Summarizing we can say that SVM can be used for classifying instances in complex problems in an efficient way.

3.2 Linear Regression (LR)

Regression helps to find out the relationship between input and target variable. Linear regression is supervised ML algorithm that instead of classifying into different categories predicts a quantitative response within a continuous range of values, output has a constant slope. There are two types of linear regression Simple regression and Multivariable regression. By the term linear we understand that the two variables been on x and y axis are linearly correlated. Linear Regression has been widely used in Price prediction, Trends Prediction and Risk Management *etc*.

3.3 Naïve Bayes (NB)

NB is a very popular classifier based on Bayes Theorem. It works on the probability of instances for each class. Its reasons of its popularity include its simplicity, correctness, and authenticity. Though it has applications in many fields of life but NB has the most implemented work in the Natural Language processing, Hybrid recommender system, text classification, and spam filtering [12]. Its name Naïve

is just because of its simple assumption that each of its attributes has an independent identity and not depends on any other feature. By using past information it computes the probability for each attribute.

3.4 Adaboost

Adaboost which is also called Adaptive Boosting. This algorithm is famous for its quick boosting in machine learning. Boosting algorithms are very suitable for transforming a lazy learner into an eager learner [12]. The basic purpose of adaptive boosting is to enhance the predictive ability of lazy learners with the help of training. To have a strong learner Adaboost merges many weak and slow learners. At the start of the algorithm weights of each attribute are identical and by the further run of the algorithm, weights start to update.

3.5 KNN

K Nearest neighbor algorithm is used to classify the instances to the neighbor with the majority vote. The nearest neighbor *i.e.* the neighbor with the smallest distance is found by using some distance metric. The most common distance measure used is the Euclidean distance [13]. Distance is determined between test and training instances. After determining the distance a feature value is calculated of all the nearest neighbor training examples and the majority of this value is considered as prediction value based on which new test dataset is categorized. KNN is highly recommended in scenarios where accurate prediction is required due to its effectiveness and simplicity.

3.6 Decision Trees

Decision trees enable to present results in the form of a tree. In the decision trees, inner nodes are used to represent the attributes descriptively whereas leaves are labeled with classes. Decision trees are made upside-down top node is called the root. They are widely used in data mining due to their simplicity and robustness. Decision trees work by selecting the best feature that yields maximum information for the classification. The classifier stops when all the leaf nodes have become pure. A leaf node is said to be pure when all instances belong to the same class or decision tree is complete and no further classification is required [14].

3.7 Random Forest (RF)

Leo Breiman and Adèle Cutler proposed the RF classifier in 2001. It works by utilizing the combined effect of two concepts “bagging” and “subspaces” [15]. From the training dataset, a set of decision trees are built and decision *i.e.* label is predicted based on votes collected from these decision tresses [16]. RF provides high accuracy and is mainly used for classifying large datasets due to its ability to handle missing values. The application domains of RF include remote sensing, e-commerce, stock market, fraud detection, network intrusion detection, and so on.

3.8 Multi-Layer Perceptron

MLP belongs to the class of feed-forward artificial neural networks (ANN). An ANN mimics the working behavior of the human brain. The main inspiration behind the ANN is the way the brain receives input, processes it, and produces output. The basic unit of ANN is a Perceptron. Each perceptron has some weight value associated with it and it generates output using the activation function. ANN works by learning representation from training data and further relating it with the desired output variable. ANN has many real-world applications such as Data Compression, Character Recognition, Computer Vision, Pattern Recognition, and Robotics.

4. METHODOLOGY

The methodological approach can be evaluated in three main steps:

4.1 Data Extraction and Preparation

Before the various classification approaches to describe, it is important to introduce the data to be analyzed for predicting the fraud. This study is analyzed with an auto insurance fraud dataset. The raw dataset contains more than a thousand customers with 36 attributes. Fig.1 shows various age groups of policyholders and Fig.2 presents the amount ranges of annual premium. The success of any classifier not only depend upon the type of model to be used. Quality of the training data is also important for Satisfactory

results. To achieve better result, data pre-processing strategy is employed.

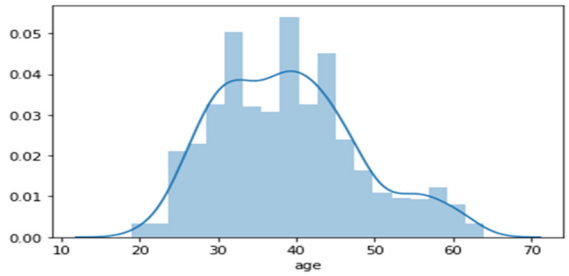

Fig. 1: Age Group of Policyholders

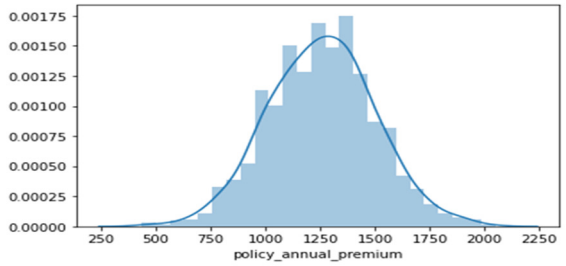

Fig. 2: Range of Policy Annual Premium Amount

4.2 Data Pre-Processing

Data Pre-processing considerably inhibits in Data Mining. Clean data is usually not possible and it may contain impossible combinations, missing values, noise, inconsistencies, *etc*. [17]. The quality of the data is the first and foremost requirement before applying the algorithm [18]. Data pre-processing may affect the way the outcomes of the ultimate is interpreted [19].

Categorical data is the factual information comprising of categorical variables or data. First, the dataset is explored for categorical data. When we consider the dataset, it may have non-categorical data such as insured_sex, PoliceReportFiled, WitnessPresent, insured_hobbies *etc*. which is converted into categorical data by using one-Hot encoder so integer encoded variable is removed and a new binary variable is added for each unique integer value. The model performs progressively when features are on a relative scale close to normally distributed.

Suppose one of the features has an outlier, then the distance will be governed by this feature. Secondly, the gradient descent converges much faster with feature scaling. Here the dataset is exposed to both MinMaxScaler and StandardScaler to make the data featured and close to normal distribution.

4.3 Proposed Work

The fraud detection approach involves number of stages. Fig.3 shows the overview of the fraud detection system.

5. EXPERIMENTAL RESULTS

For performance evaluation, we have computed five metrics: accuracy, Precision, Recall, F1-Score, and confusion matrix. Where Precision is the portion of relative cases among the retrieved occasions, while

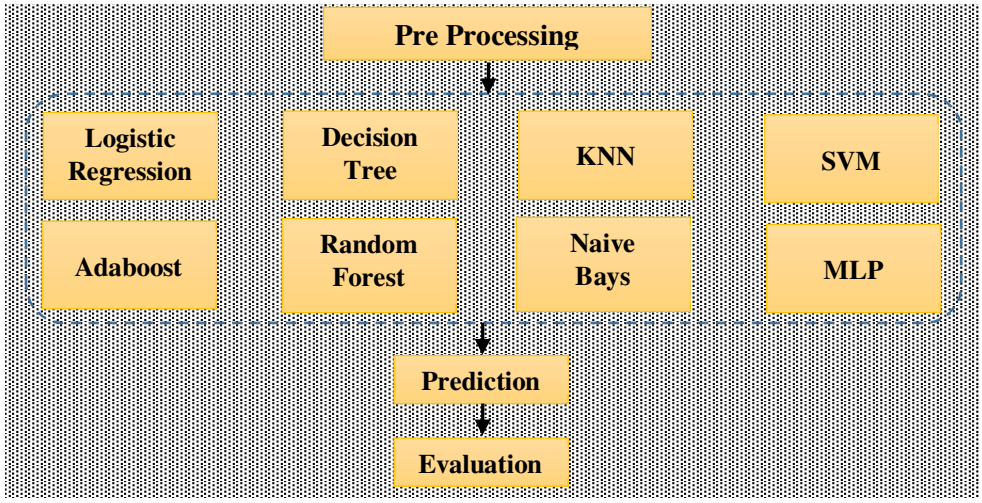

Fig. 3: Fraud Detection System

Recall is the division of the aggregate sum of relative cases that are retrieved. F1-Score is the average of Precision and Recall, while the Confusion Matrix is the measure of performance of an ML algorithm as explained in Table 1 and Table 2.

In this paper, we consider an auto insurance fraud detection dataset and execute a sample that contains 110 customers with corresponding attributes. Table I shows that the eight classification models have been validated using evaluation metrics such as precision, recall, and F1-score with corresponding Macro and weighted average as in Table 2. The results of the experiment have shown that Decision-Tree outperforms in all aspects such as execution time, non-sensitive to outliers, and the reduction of noise. The results obtained using the Classification algorithm outshines using real sample obtained from the reliable repository. For all the experiments in this section, the performance shown is based on the test dataset. Also, the Adaboost almost gave better classification accuracy close to the Decision Tree. The classification accuracy of Adaboost is 78%. The Precision, Recall, F1-Score are computed using the equation. (1), (2) and (3):

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \quad (1)$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} \quad (2)$$

$$\text{F1 Score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \quad (3)$$

Fig. 4 and Fig. 5, demonstrate the Performance Metrics: Precision, Recall, and F1-Score that ranges

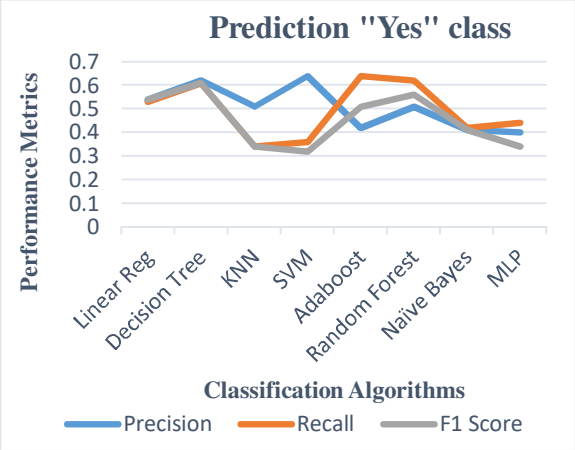

Fig. 4: Performance Metric for Yes class

Table 1: Macro and weighted average of Precision, Recall and F1 Score

Metrics	Average	Classification Algorithms							
		Linear Regression	Decision Tree	KNN	SVM	Adaboost	Random Forest	Naïve Bayes	MLP
Precision	Macro	0.50	0.73	0.50	0.50	0.66	0.70	0.50	0.50
	Weighted	1.00	0.78	1.00	1.00	0.82	0.80	0.80	1.00
Recall	Macro	0.36	0.73	0.36	0.36	0.72	0.72	0.72	0.36
	Weighted	0.72	0.79	0.72	0.72	0.78	0.78	0.77	0.72
F1 Score	Macro	0.42	0.73	0.42	0.42	0.42	0.68	0.71	0.42
	Weighted	0.84	0.78	0.84	0.84	0.84	0.79	0.79	0.84

Table 2: Confusion matrix, Accuracy, Precision, Recall, and F1 score

Metrics	Classification Algorithms															
	Linear Regression		Decision Tree		KNN		SVM		Adaboost		Random Forest		Naïve Bayes		MLP	
Confusion Matrix	[[145 55] [0 0]]		[[123 24] [22 31]]		[[145 55] [0 0]]		[[145 55] [0 0]]		[[132 32] [13 23]]		[[126 29] [19 26]]		[[135 51] [0 0]]		[[145 55] [0 0]]	
Precision	0	1.00	0	0.85	0	1.00	0	1.00	0	0.91	0	0.88	0	0.87	0	1.00
	1	0.54	1	0.62	1	0.51	1	0.64	1	0.42	1	0.51	1	0.41	1	0.40
Recall	0	0.72	0	0.85	0	0.72	0	0.72	0	0.80	0	0.83	0	0.81	0	0.72
	1	0.53	1	0.61	1	0.34	1	0.36	1	0.64	1	0.62	1	0.42	1	0.44
F1 Score	0	0.84	0	0.85	0	0.84	0	0.84	0	0.85	0	0.85	0	0.82	0	0.84
	1	0.54	1	0.61	1	0.34	1	0.32	1	0.51	1	0.56	1	0.41	1	0.34
Accuracy	73%		79%		77%		73%		78%		76%		73%		73%	

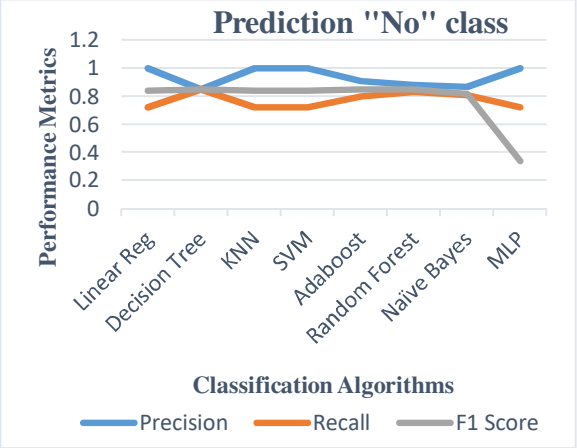

Fig. 5: Performance Matric for No Class

from 0 to 1. The value will be 1 when the system performs well.

6. CONCLUSIONS AND FUTURE WORK

In our research, the classification algorithms namely Random-Forest, Decision –Tree, Support Vector Machine, K-Nearest Neighbor, Adaboost, Linear Regression, Naïve Bayes, and Multi-Linear Perceptron are employed to detect fraud. We audited various techniques and conducted experiments on the auto-insurance dataset from a reliable repository to find or adapt the best classifier for the fraud detection system. Furthermore, the system has been analyzed in the aspects of precision, recall, and F1-score for all the algorithms.

In the future, the fraud detection method can be extended to the Adaptive Neuro-Fuzzy Inference System (ANFIS) which is the combination of both Neuro-Fuzzy and Neural Networks. Hence, the prediction can be made more accurate and Hidden Markov Model (HMM) to predict fraud using internal factors.

REFERENCES

1. Nian K., Zhang H., Tayal A., Coleman T., Li Y. (2016). "Auto insurance fraud detection using unsupervised spectral ranking for anomaly", *The Journal of Finance and Data Science*, Vol. 2, No.1, pp. 58–75, 2016.
2. Kirlidog M., Asuk C., A Fraud Detection Approach with Data Mining in Health Insurance. *Procedia - Social and Behavioral Sciences*, Vol. 62, pp. 989–994, 2012.
3. Sathya R., Abraham A., "The Science and Information Organization Editorial Preface", *International Journal of Advanced Research in Artificial Intelligence,* Vol.2, No.2, pp. 34–38, 2013.
4. Rätsch G., "A brief introduction into machine learning", *Proceedings of the 21st Chaos Communication Congress*, 1–6, Berlin, Germany, 27-29 December 2004.
5. Wang H., Shi Y., Zhou X., Zhou Q., Shao S., Bouguettaya A., "Web service classification using support vector machine", *Proceedings of the International Conference on Tools with Artificial Intelligence,* Vol. 1, pp. 3–6, Arras, France, 27-29 October 2010.
6. Sun C., Li Q., Li H., Shi Y., Zhang S., Guo W., "Patient Cluster Divergence Based Healthcare Insurance Fraudster Detection", *IEEE Access*, Vol. 7, pp. 14162–14170, 2019.
7. Dhieb N., Ghazzai H., Besbes H., Massoud Y., "A Secure AI-Driven Architecture for Automated Insurance Systems: Fraud Detection and Risk Measurement", *IEEE Access*, Vol. 8, pp. 58546–58558, 2020.
8. Kirlidog M., Asuk C., "A Fraud Detection Approach with Data Mining in Health Insurance", *Procedia - Social and Behavioral Sciences*, Vol. 62, pp. 989–994, 2012.
9. Bhowmik R., "Detecting Auto Insurance Fraud by Data Mining Techniques", *Journal of Emerging Trends in Computing and Information Sciences*, Vol. 2, No.4, pp. 156–162, 2011.
10. Liu S., Yang B., Wang L., Abraham A., "Advances in Nature and Biologically Inspired Computing", *Advances in Intelligent Systems and Computing,* Vol. 419, 2016.
11. Bennett K. P., Campbell C., "Support vector machines: hype or hallelujah?", ACM *SIGKDD Explorations Newsletter*, Vol. *2, No.* 2, pp. 1–13, 2000.
12. Noor A., Islam M., "Sentiment Analysis for Women's E-commerce Reviews Using Machine Learning Algorithms", *Proceedings of the 10th International Conference on Computing,*

Communication and Networking Technologies (ICCCNT), pp. 1–6, Kanpur, India, 6-8 July 2019.

13. Priya B. G., "Emoji Based Sentiment Analysis Using Knn", *International Journal of Scientific Research and Reviews,* Vol. 7, No.4, pp. 859-865, 2019.
14. Suresh A., Bharathi C. R., "Sentiment Classification using Decision Tree Based Feature Selection Sentiment Classification using Decision Tree Based Feature Selection", *International Journal of Control Theory and Applications*, Vol. 9, No. 36, pp. 419–425, 2016.
15. Mohamed L., Kamal E. E. K., Yassine A. A., "Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis", *Procedia Computer Science*, Vol. 127, pp. 511–520, 2018.
16. Ankit, Saleena N., "An Ensemble Classification System for Twitter Sentiment Analysis", *Procedia Computer Science*, Vol. 132, pp. 937–946, 2018.
17. Han J., Kamber M., Pei J., Introduction. In Han J., Kamber M., Pei J. (Eds.): *Data Mining* (Third Edition), pp. 1-38, Morgan Kaufmann, 2012.
18. Zhang S., Zhang C., Yang Q. (2003), "Data preparation for data mining", *Applied Artificial Intelligence*, Vol. 17, No. 5–6, 2003.
19. Oliveri P., Malegori C., Simonetti R., Casale M., "The impact of signal pre-processing on the final interpretation of analytical outcomes – A tutorial", *Analytica Chimica Acta*, Vol. 1058, pp. 9–17, 2019.