As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 8(05):46-52, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0805008
Available online at: http://www.ijmtst.com/vol8issue05.html

A Machine Learning Approach to Detecting False Job Advertisements

J K V V Durga Bhavani[1] | N Durga Devi[2] | L V Kiran[2]

[1]PG Scholar, Dept of CA, Godavari Institute of Engineering and Technology (A),Rajahmundry,A.P
[2]Assistant Professor, Dept of CA, Godavari Institute of Engineering and Technology (A),Rajahmundry,A.P
Corresponding Author Email ID: bhavanijagarapu1998@gmail.com[1], durgadevi.ansh@giet.ac.in[2], lvkiran@giet.ac.in[3]

To Cite this Article
J K V V Durga Bhavani, N Durga Devi and L V Kiran. A Machine Learning Approach to Detecting False Job Advertisements. International Journal for Modern Trends in Science and Technology 2022, 8(05), pp. 46-52. https://doi.org/10.46501/IJMTST0805008

Article Info
Received: 26 March 2022; Accepted: 25 April 2022; Published: 29 April 2022.

ABSTRACT

Job postings have been more popular in recent years as a result of technological progress and social media communication. The creation of fraudulent job postings and the subsequent solicitation of personal information will provide an opportunity for fraudsters. It will be easier to tackle the problem if bogus job advertising can be identified and avoided altogether. This method makes use of Ensemble Classifiers. This system takes use of a dataset from Kaggle that was previously made accessible to the user. In order to identify fake job postings, ensemble classifiers are utilized. For detecting frauds, these are the most effective classificatory available. To begin, there are 17,880 job postings in the dataset overall. Some pre-processing procedures are used to this dataset before it is used to train a classifier of any kind. Classifiers are created by fitting the data into them after it has been pre-processed. Various Natural Language Processing (NLP) approaches are employed for pre-processing, and ensemble classifiers such as the Random Forest algorithm, Gradient Boosting Algorithm, and Support vector classifier are used to classify the data. To forecast the genuine and actual job postings, the system compares and contrasts machine learning techniques.

KEYWORDS: Fake Job, Online Recruitment, Machine Learning, Ensemble Approach

1. INTRODUCTION

The number of employment frauds is increasing. Over the course of 2018, CNBC reported that the number of job scams has increased. There has been a rise in unemployment due to the present market conditions. As a result of the economic downturn and the corona virus's effect, many people have lost their employment. A scenario like this gives a perfect opportunity for fraudsters. Many individuals are falling victim to these fraudsters leveraging the despair that is produced by an exceptional tragedy. Scammers often use this tactic in order to get the victim's personal information. Personal information might comprise address, bank account data, social security number etc. The fraudsters offer people a rich employment opportunity and then demand money in exchange. Or they need money from the job seeker with the promise of a job. Fortunately, machine learning and natural language processing can help us deal with this critical issue (NLP). This system leverages data supplied by Kaggle. A job posting's characteristics may be found here. There are 17880 job postings in this database. Real and false job advertisements may be found in this section of the site. Fake job advertisements represent a relatively minor part of this dataset. That's what I expected to hear. The number of job advertisements isn't likely to be that high. Ensemble

classifiers are used to identify bogus job postings in machine learning. For fraud detection, ensemble classifiers are the best. Two classifiers are used in this ensemble: Random Forest and Gradient Boosting. This article also makes use of Support Vector Algorithms to sniff out job postings. Evaluation Metrics including Precision, Recall, F1-score and Accuracy were employed. Accuracy is a measure of how well a model really predicts the final result will be. Recall and F1-score are helpful since the Dataset has varying Precision.

2. LITERATURE REVIEW:

Many studies have shown that the fake job prescription and other forms of online fraud, such as spam reviews, spam emails, and false news, have gotten a lot of attention in the field of online fraud detection. Online Recruitment Frauds (ORF)[1] has recently tackled one of the most important concerns in the field of employment scams. Many organizations now choose to list their job openings online so that they may be viewed quickly and efficiently by job searchers. In addition to the current epidemic of the Corona Virus, employers have begun advertising their open positions on the internet. It is almost always possible for machine learning classification programmes to be taught by using supervised learning, which entails providing the model with an existing classification so that it can be compared to the actual class to see how far it has progressed, and then adjusting its mathematical equations until it achieves an acceptable margin between the actual class and the predicted class SVM, NB, and KNN were found to be the most suitable classifiers for text categorization by Khan, Aurangzeb, and colleagues[3]. When a job seeker decided in 2012 to publish a genuine Craigslist job ad in order to discover their competition, they received more than 600 applications in one day. In the same year, the Australian Bureau of Statistics produced a study on personal fraud indicating that 6 million Australians were subject to a variety of scams, including job scams [5].Online job postings may include bogus positions, resulting in theft of sensitive data including social security numbers and credit card numbers. For example, employing sophisticated deep learning as well as machine learning classification algorithms, it is possible to identify and classify both fraudulent and authentic job postings from a job pool.

3. PROPOSED SYSTEM:

The proposed system would use the same dataset that is now being used by the present system. A total of 17,880 jobs are included in the dataset. This data collection is used in the proposed techniques to assess the overall performance of the strategy. This dataset undergoes certain pre-processing operations before it is used to train any classifiers. Procedures used before processing data include the removal of invalid data and the removal of unneeded characteristics and extraneous words. The algorithms used in this system are:

1. Random Forest Algorithm
2. Gradient Boosting Algorithm
3. Support Vector Algorithm

These three algorithms are compared on accuracy and a final model is chosen. Random Forest is classifier that predicts output with high accuracy followed by Support Vector and Gradient Boosting. Random Forest predicts output based on the majority of voting. Gradient Boosting predicts output based on the errors of previous decision tree. Support Vector predicts output based on the extreme cases. Among all the three classifiers, Random Forest Algorithm has higher accuracy of 97.09% followed by Support Vector 96.76% and Gradient Boosting 95.6%

Following are the steps done while implementing the system:

STEPS:

1) Dataset contains fake and real jobs.
2) Pre-process the dataset by removing unwanted columns.
3) Train and test the dataset
4) Apply all three classification algorithms.
5) Predict the Accuracy and evaluation metrics

OUTPUT: Comparing the contrasting machine learning algorithms and choose the algorithms which is having more accuracy.

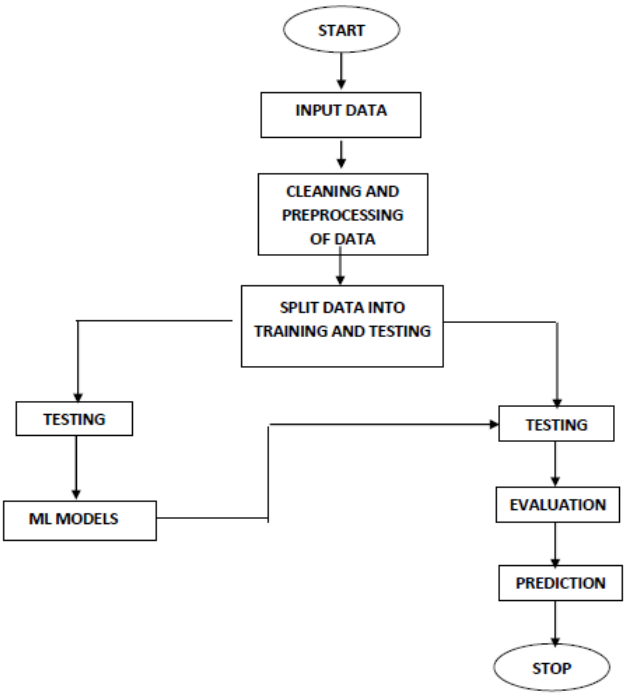


Fig1: Architecture of Proposed Work

4. METHODOLOGY

Random Forest

Random Forest is a well-known supervised machine learning method. In ML, it may be used for both classification and regression. In order to tackle a difficult issue and increase the model's performance, it uses ensemble learning, which involves integrating many classifiers. In order to increase the predicting accuracy of a dataset, Random Forest uses a series of decision trees on different subsets of the data. It forecasts the ultimate output based on the majority of predictions from each tree, rather than one decision tree.

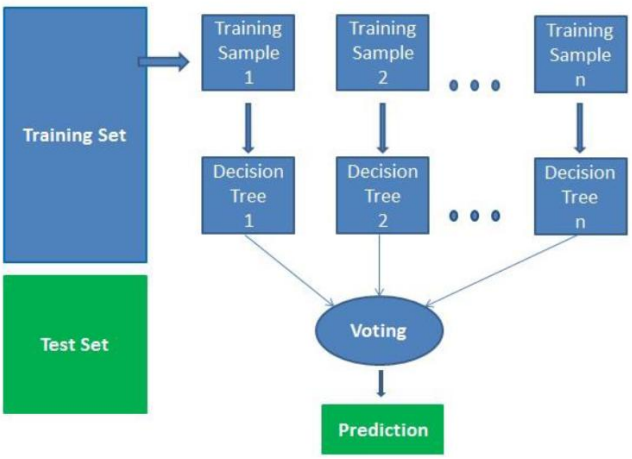


Fig4.1 Architecture Random Forest

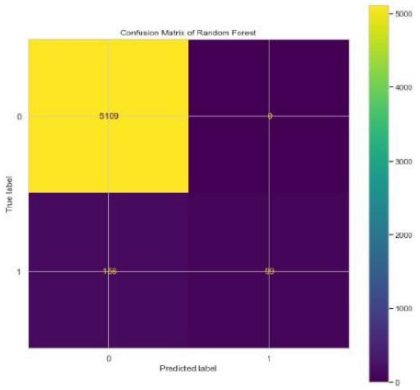


Fig4.2 Confusion Matrix for Random Forest

Random Forest has given the accuracy score of 97.09% and for 5364 test observations, it has correctly predicted the class labels for 5208 job postings. X-axis represents predicted values and Y-axis represents actual values.

Gradient Boosting

One decision tree predicts another, and so on. Remember that decision trees are the weak learners in a gradient boosting machine. But if they use the same technique, how is utilizing 100 decision trees better than one? How can various decision trees collect different signals/information from data? The idea is that each decision tree node uses a different group of information to choose the optimum split. Because the trees aren't all the same, they might catch diverse signals from the data. Also, each new tree considers prior trees' errors. So, each decision tree builds on the preceding trees' mistakes. Sequentially created trees in a gradient boosting machine technique. Gradient boosting methods need a weak learner and an additive component. Decision trees are used as weak learners in gradient boosting systems. Because trees are added to the model over time, their values aren't changed, giving it an additive component. The new tree's output is then added to the prior trees' output. This cycle is continued until a particular number of trees is attained or the loss is decreased.

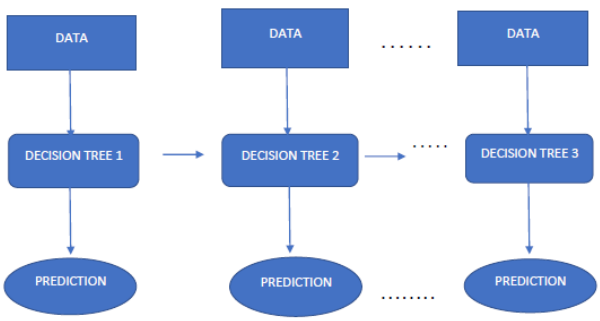


Fig4.3 Architecture Gradient Boosting

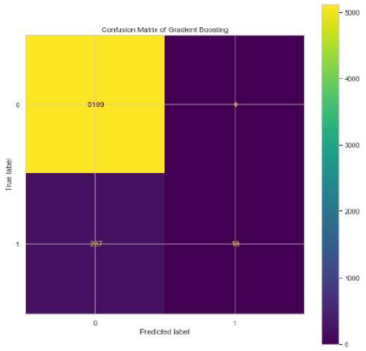


Fig4.4 Confusion Matrix for Gradient Boosting

Gradient Boosting Algorithm has given the accuracy score of 95.5% and for 5364 test observations, it has correctly predicted the class labels for 5127 job postings. X-axis represents Predicted Value and Y-axis represents Actual Value.

Support Vector Machine

SVM is a prominent Supervised Learning technique used for Classification and Regression issues. It is used in Machine Learning for Classification issues. The SVM algorithm's purpose is to find the optimal line or decision boundary that divides n-dimensional space into classes so that subsequent data points may be conveniently placed in the relevant category. A hyperplane defines the optimal decision boundary. SVM picks the hyperplane's extreme points/vectors. These extreme situations are called support vectors, and the method is called SVM.

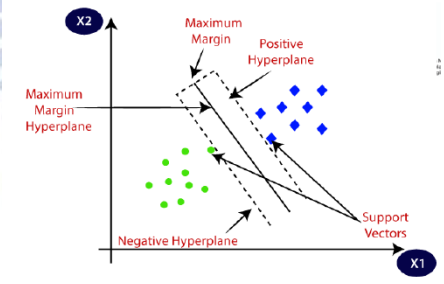


Fig4.5 Architecture Support Vector Machine

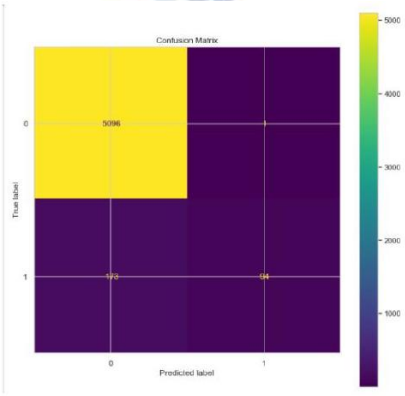


Fig4.6 Confusion Matrix for Support Vector Machine

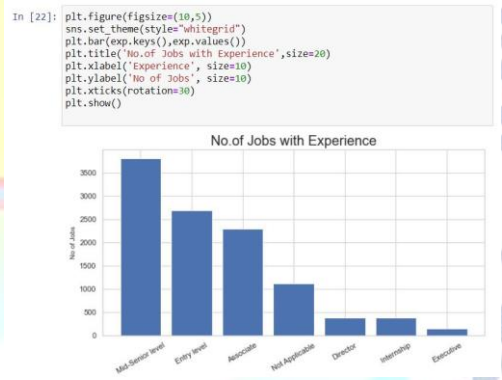
Support Vector Classifier has given the accuracy score of 96.76% and for 5364 test observations; it has correctly predicted the class labels for 5190 job postings.

5. EXPERIMENTAL RESULTS:

Bar Chart of real and fraudulent Jobs

```
In [21]: exp = dict(df.required_experience.value_counts())
         del exp['']
         exp

Out[21]: {'Mid-Senior level': 3809,
          'Entry level': 2697,
          'Associate': 2297,
          'Not Applicable': 1116,
          'Director': 389,
          'Internship': 381,
          'Executive': 141}
```

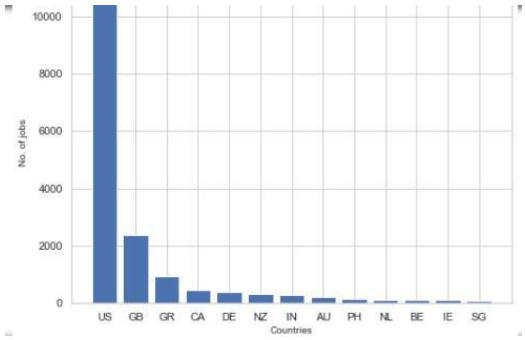
Jobs with Experience



Bar chart of Jobs with Experience

```
In [19]: countr = dict(df.country.value_counts()[:14])
         del countr['']
         countr

Out[19]: {'US': 10656,
          'GB': 2384,
          'GR': 940,
          'CA': 457,
          'DE': 383,
          'NZ': 333,
          'IN': 276,
          'AU': 214,
          'PH': 132,
          'NL': 127,
          'BE': 117,
          'IE': 114,
          'SG': 80}
```

Country Wise Job Postings



Country Wise Job Postings Graph

```
In [60]: pred = rfc.predict(X_test)
         score = accuracy_score(y_test, pred)
         score

Out[60]: 0.970917225950783
```

Accuracy of Random Forest Algorithm

```
In [61]: print("Classification Report\n")
         print(classification_report(y_test, pred))
         print("Confusion Matrix\n")
         print(confusion_matrix(y_test,pred))

Classification Report

              precision    recall  f1-score   support

           0       0.97      1.00      0.98      5109
           1       1.00      0.39      0.56       255

    accuracy                           0.97      5364
   macro avg       0.99      0.69      0.77      5364
weighted avg       0.97      0.97      0.96      5364

Confusion Matrix

[[5109    0]
 [ 156   99]]
```

Confusion Matrix of Random Forest Algorithm

```
In [65]: pred = gb_clf.predict(X_test)
         acc = accuracy_score(y_test, pred)
         acc

Out[65]: 0.9558165548098434
```

Accuracy of Gradient Boosting Algorithm

```
In [66]: print("Classification Report\n")
         print(classification_report(y_test, pred))
         print("Confusion Matrix\n")
         print(confusion_matrix(y_test,pred))

Classification Report

              precision    recall  f1-score   support

           0       0.96      1.00      0.98      5109
           1       1.00      0.07      0.13       255

    accuracy                           0.96      5364
   macro avg       0.98      0.54      0.55      5364
weighted avg       0.96      0.96      0.94      5364

Confusion Matrix

[[5109    0]
 [ 237   18]]
```

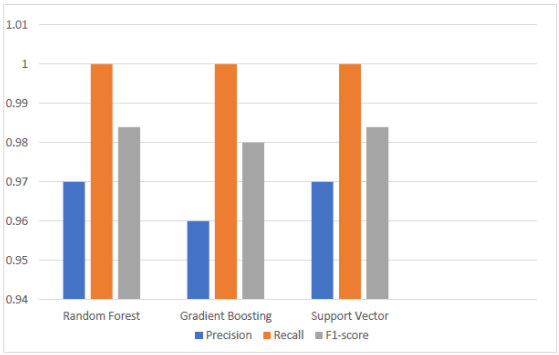
Confusion Matrix of GB Algorithm

```
In [89]: pred2=clf.predict(X_test)
         clf_score=accuracy_score(y_test,pred2)
         clf_score

Out[89]: 0.9675615212527964
```

Accuracy of Support Vector Algorithm

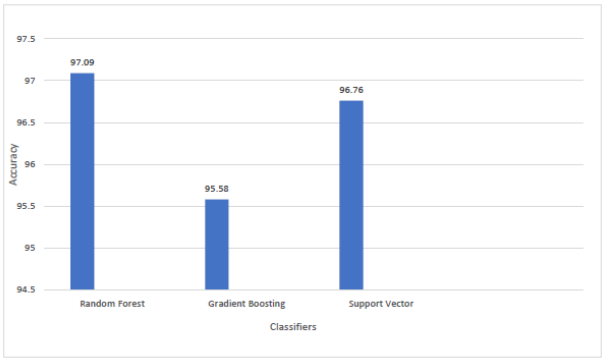
```
In [92]: print("Classification Report\n")
         print(classification_report(y_test, pred2))
         print("Confusion Matrix\n")
         print(confusion_matrix(y_test,pred2))

Classification Report

              precision    recall  f1-score   support

           0       0.97      1.00      0.98      5097
           1       0.99      0.35      0.52       267

    accuracy                           0.97      5364
   macro avg       0.98      0.68      0.75      5364
weighted avg       0.97      0.97      0.96      5364

Confusion Matrix

[[5096    1]
 [ 173   94]]
```

Confusion Matrix of Support Vector



Graphical Representation of Performance Measures

X-axis represents three classifiers Random Forest, Gradient Boosting and Support Vector. Each classifier has precision, recall and F1-score. Blue bar represents precision, Orange represents Recall and Grey represents F1-score. Y-axis represents the values for each of the precision, recall and F1-scores of all the three classifiers.



Accuracy Comparison

X-axis represents classifiers. In X-axis, Random Forest, Gradient Boosting and Support Vector are the three classifiers. Y-axis represents the Accuracy of the three classifiers. Among the three random forest classifier have high accuracy with 97.09%, Gradient Boosting having 95.58% and Support Vector having 96.76%.

6. COMPARATIVE STUDY

Table1: Comparative Study

INPUT	EXPECTED OUTPUT	OBTAINED OUTPUT	REMARKS
Apply Random Forest on Trained dataset.	Model is Trained	Model is Trained	Pass
Apply Random Forest on Testing data.	Able to predict accuracy	Able to predict accuracy	Pass
Apply Gradient Boosting on Trained Data.	Model is Trained	Model is Trained	Pass
Apply Gradient Boosting on Testing Data	Able to predict accuracy	Able to predict accuracy	Pass
Apply Support Vector on Trained Data	Model is Trained	Model is Trained	Pass
Apply Support Vector on Testing Data.	Able to predict accuracy	Able to predict accuracy	Pass

7. CONCLUSION:

Fake job ads are a significant real-world issue that necessitates the development of proactive remedies. The goal of this research is to give a viable answer to this issue. The output of numerous models is merged to get the best possible results. The most intriguing aspect of this investigation was discovering which areas represent the acme of bogus employment opportunities. For example, in Bakersfield, California, the ratio of fake to actual jobs is 15:1. Places like these need more security. Another surprising aspect was that it seems that the majority of entry-level positions are bogus. It seems that fraudsters prefer to target younger individuals who have a bachelor's degree or a high school diploma and are searching for full-time employment. Employment fraud identification can assist job searchers in receiving only authentic employment offers from firms. Several ensemble machine learning methods are employed in the identification of employment scams. A supervised approach is used to demonstrate the use of different classifiers in the identification of job scams. Following a comparison of competing machine learning models, experimental findings show that Random Forest achieved the highest accuracy, followed by Support Vector and Gradient Boosting.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] B. Alghamdi and F. Alharby, —An Intelligent Model for Online Recruitment Fraud Detection," J. Inf. Secur., vol. 10, no. 03, pp. 155–176, 2019, doi: 10.4236/jis.2019.103009.

[2] B. C. Love, —Comparing supervised and unsupervised category learning,‖ Psychon. Bull. Rev., vol. 9, no. 4, pp. 829–835, 2002.

[3] Khan, B. Baharudin, L. H. Lee, and K. Khan, —A review of machine learning algorithms for text-documents classification,‖ J. Adv. Inf. Technol., vol. 1, no. 1, pp. 4–20, 2010.

[4] Auld, E. Man Posts Fake Job on Craigslist, Gets 600+ Resumes. 2012. Available online: http://chemjobber.blogspot.gr/2012/08/man-posts-fake-job-on-craigslist-gets.html (accessed on 19 March 2015).

[5] Gov, K. (2018) Vision of 2030. http://vision2030.gov.sa/en

[6] ACRON (2018) Australian Cybercrime Online Reporting Network (ACORN). http://www.acorn.gov.au/learn-about-cybercrime

[7] Armstrong, A. (2006) Handbook of Human Resource Management Practice. 10th Edition, Kogan Page Limited, London.

[8] Hada, B. and Gairola, S. (2015) Opportunities and Challenges of E-Recruitment. Journal of Management Engineering and Information Technology, 2, 1-4.

[9] Kaur, P. (2015) E-Recruitment: A Conceptual Study. International Journal of Ap- plied Research, 1, 78-82

[10] Prasad, L. and Kapoor, P. (2016) Topic: E-Recruitment Strategies. International Journal of Business Quantitative Economic and Applied Management Research, 2, 80-95.

[11] Panov, P., Soldatova, L. and Džeroski, S. (2013) OntoDM-KDD: Ontology for Representing the Knowledge Discovery Process. 16th International Conference on Discovery Science, Singapore, 6-9 October 2013, 126-140. https://doi.org/10.1007/978-3-642-40897-7_9

[12] Cios, K.J., Pedrycz, W., Swiniarski, R.W. and Kurgan, L.A. (2007) Data Mining: A Knowledge Discovery Approach. Springer, New York.

[13] Hussain, S. (2017) Survey on Current Trends and Techniques of Data Mining Research. London Journal of Research in Computer Science and Technology, 17, 7-15.

[14] Sinoara, R., Antunes, J. and Rezende, S. (2017) Text Mining and Semantics: A Systematic Mapping Study. Journal of the Brazilian Computer Society, 23, 9. https://doi.org/10.1186/s13173-017-0058-7

[15] Diwathe, D. and Dongare, S. (2017) Classification Model Using Optimization Technique: A Review. International Journal of Computer Science and Network, 6, 42-48.

[16] Singh, G. and Singh, A. (2017) A Review Paper: Using Data Mining Clustering Technique to Predict Criminal Behavior. International Journal of Computer Science and Mobile Computing, 6, 160-167.

[17] Witten, I. and Frank, E. (2005) Data Mining Practical Machine Tools and Techniques. Morgan Kaufmann Elsevier, San Francisco.

[18] Kukasvadiya, M. and Divecha, N. (2017) Analysis of Data Using Data Mining Tool Orange. International Journal of Engineering Development and Research, 5, 836-1840.

[19] Rehman, N. (2017) Data Mining Techniques Methods Algorithms and Tools. International Journal of Computer Science and Mobile Computing, 6, 227-231.

[20] Jyoth, P., Siva Ranjani, R., Mishra, T. and Mishra, S.R. (2017) A Study of Classification Techniques of Data Mining Techniques in Health Related Research. International Journal of Innovative Research in Computer and Communication Engineering, 5, 13779-137876.

[21] Vidros, S., Kolias, C. and Kambourakis, G. (2016) Feature: Online Recruitment Services: Another Playground for Fraudsters. Computer Fraud & Security, 2016, 8-13. https://doi.org/10.1016/S1361-3723(16)30025-2

[22] Yasin, A. and Abuhasan, A. (2016) An Intelligent Classification Model for Phishing Email Detection. International Journal of Network Security & Its Applications, 8, 55-72. https://doi.org/10.5121/ijnsa.2016.8405

[23] Al-garadi, M.A., Varathan, K.D. and Ravana, S.D. (2016) Cybercrime Detection in Online Communications: The Experimental Case of Cyberbullying Detection in the Twitter Network. Computers in Human Behavior, 63, 433-443. https://doi.org/10.1016/j.chb.2016.05.051

[24] Sharaff, A., Nagwani, N.K. and Swami, K. (2015) Impact of Feature Selection Technique on Email Classification. International Journal of Knowledge Engineering, 1, 59-63. https://doi.org/10.7763/IJKE.2015.V1.10

[25] Sornsuwit, P. and Jaiyen, S. (2015) Intrusion Detection Model Based on Ensemble Learning for U2r and R2l Attacks. In: 7th International Conference Information Technology and Electrical Engineering, IEEE, Chiang Mai, 354-359. https://doi.org/10.1109/ICITEED.2015.7408971