**A Comparative Study on Fake Job Post Prediction Using Different Data Mining Techniques**

**Kesireddy Samara Simha Reddy, Jangiti Dwarakamai, Motha Rahul, CH Yogesh Chowdary**

**Mr.M.Srinivasa Reddy(Assistant Professor)**

**[1]** UG Students, Department of Information Technology, Malla Reddy Engineering College, Hyderabad, India.

**[2]** Professor, Department of Information Technology, Malla Reddy Engineering College, Hyderabad, India.

**Abstract**- In recent years, due to advancement in modern technology and social communication, advertising new job posts has become very common issue in the present world. So, fake job posting prediction task is going to be a great concern for all. Like many other classification tasks, fake job posing prediction leaves a lot of challenges to face. This paper proposed to use different datamining techniques and classification algorithm like KNN, decision tree, support vector machine, naive bayes classifier, random forest classifier , multi-layer perceptron and deep neural network to predict a job post if it is real or fraudulent. We have experimented on Employment Scam Aegean Dataset (EMSCAD) containing18000 samples. Deep neural network as a classifier, performs great for this classification task. We have used three dense layers for this deep neural network classifier. The trained classifier shows approximately 98% classification accuracy (DNN) to predict a fraudulent job post. Index Terms--false job prediction, deep learning, data mining.

**I. INTRODUCTION**

In modern time, the development in the field of industry and technology has opened a huge opportunity for new and diverse jobs for the job seekers. With the help of the advertisements of these job offers, job seekers find out their options depending on their time, qualification, experience, suitability etc. Recruitment process is now influenced by the power of internet and social media. Since the successful completion of a recruitment process is dependent on its advertisement, the impact of social media over this is tremendous, social media and advertisements in electronic media have created newer and newer opportunity to share job details. Instead of this, rapid growth of opportunity to share job posts has increased the percentage of fraud job postings which causes harassment to the job seekers. So, people lack in showing interest to new job postings due to preserve security and consistency of their personal, academic and professional information. Thus, the true motive of valid job postings through social and electronic media faces an extremely hard challenge to attain people's relict and reliability. Technologies are around us to make our life easy and developed but not to create unsecured environment for professional life. If jobs posts can be filtered properly predicting false job posts, this will be a great advancement for recruiting new employees. Fake job posts create inconsistency for the job seeker to find their preferable jobs causing a huge waste of their time. An automated system to predict false job post opens a new window to face difficulties in the field of Human Resource Management.

**II. LITERATURE SURVEY**

According to several studies, Review spam detection, Email Spam detection, Fake news detection have drawn special attention in the domain of Online Fraud Detection. A Review Spam Detection-People often post their reviews online forum regarding the products they purchase. It may guide other purchaser while choosing their products. In this context, spammers can manipulate reviews for gaining profit and hence it is required to develop techniques that detects these spam reviews. This can be implemented by extracting features from the reviews by extracting features using Natural Language Processing (NLP). Next, machine learning techniques are applied on these features. Lexicon based approaches may be one alternative to machine learning techniques that uses dictionary or corpus to eliminate spam reviews [11]. B. Email Spam Detection-Unwanted bulk mails, belong to the category of spam emails, often arrive to user mailbox. This may lead to unavoidable storage crisis as well as bandwidth consumption. To eradicate this problem, Gmail, Yahoo mail and Outlook service providers incorporate spam filters using Neural Networks. While addressing the problem of email spam detection, content-based filtering, case-based filtering, heuristic-based filtering, memory or instance-based filtering, adaptive spam filtering approaches are taken into consideration [7]. C. Fake News Detection-Fake news in social media characterizes malicious user accounts, echo chamber effects. The fundamental study of fake news detection relies on three perspectives- how fake news is written, how fake news spreads, how a user is related to fake news. Features related to news content and social context are extracted and a machine learning model are imposed to recognize fake new.

**III. PROPOSED METHODOLOGY**

The target of this study is to detect whether a job post is fraudulent or not. Identifying and eliminating these fake job advertisements will help the jobseekers to concentrate on legitimate job posts only. In this context, a dataset from Kaggle [13] is employed that provides information regarding a job that may or may not be suspicious. The dataset has the schema as shown in Fig. 1.

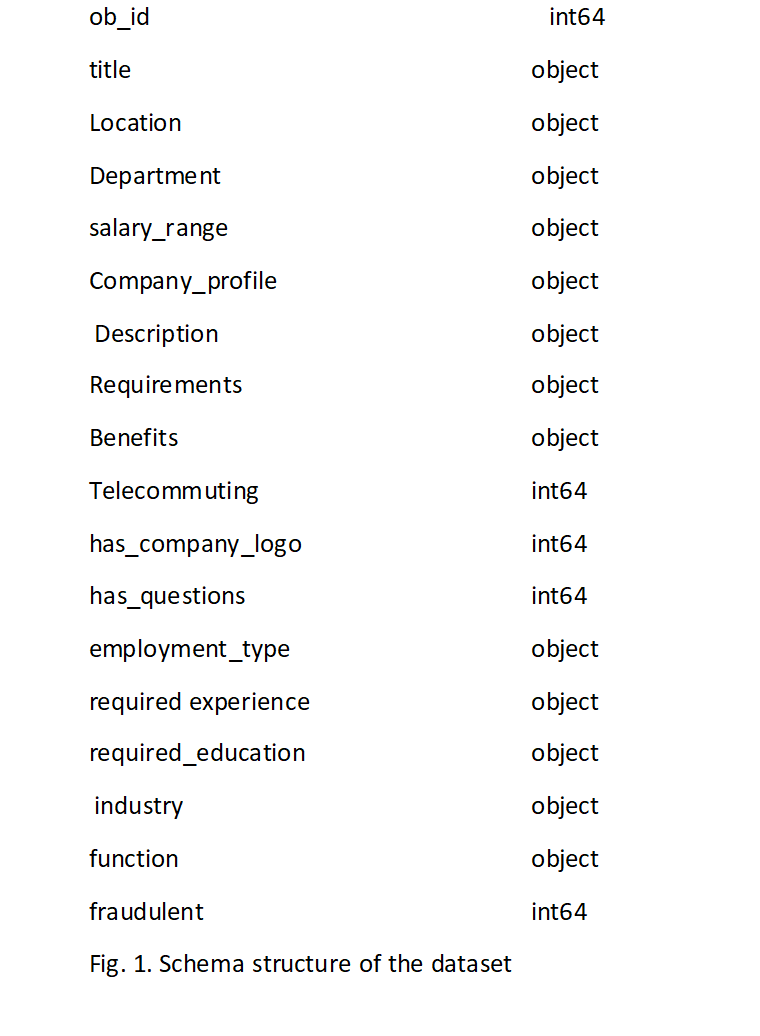


Fig. 1. Schema structure of the dataset

This dataset contains 17,880 number of job posts. This dataset is used in the proposed methods for testing the overall performance of the approach. For better understanding of the target as a baseline, a multistep procedure is followed for obtaining a balanced dataset. Before fitting this data to any classifier, some pre-processing techniques are applied to this dataset. Pre-processing techniques include missing values removal, stop-words elimination, irrelevant attribute elimination and extra space removal. This prepares the dataset to be transformed into categorical encoding in order to obtain a feature vector. This feature vectors are fitted to several classifiers. The following diagram Fig. 2 depicts a description of the working paradigm of a classifier for prediction.

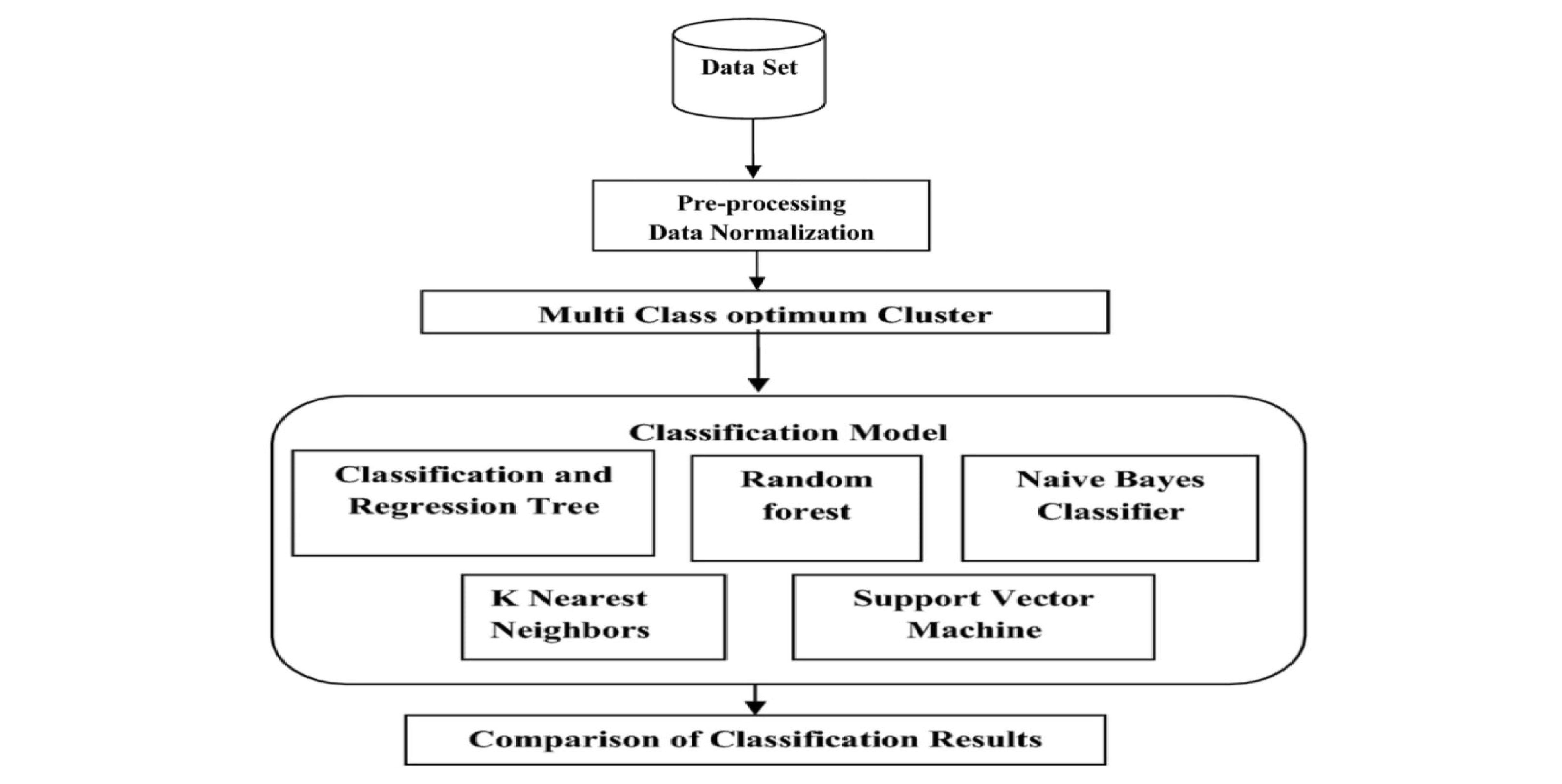


Fig. 2. Detailed description for working of Classifiers

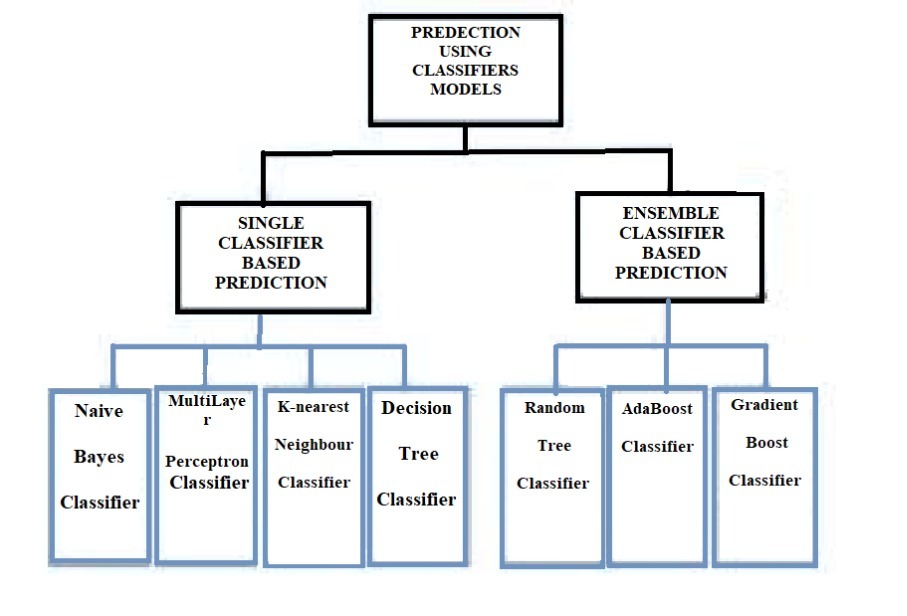


Fig. 3. Classification models used in this framework

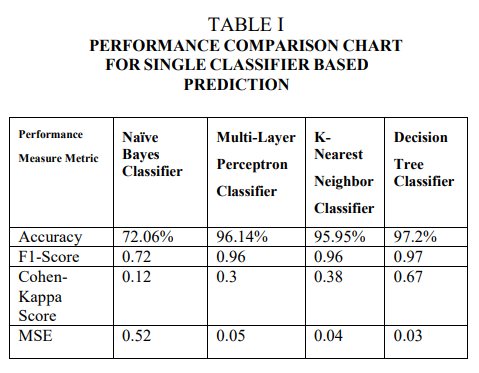
**A. Implementation of Classifiers-** In this framework classifiers are trained using appropriate parameters. For maximizing the performance of these models, default parameters may not be sufficient enough. Adjustment of these parameters enhances the reliability of this model which may be regarded as the optimised one for identifying as well as isolating the fake job posts from the job seekers.

This framework utilised MLP classifier as a collection of 5 hidden layers of size 128, 64, 32, 16 and 8 respectively. The K-NN classifier gives a promising result for the value k=5 considering all the evaluating metric. On the other hand, ensemble classifiers, such as, Random Forest, AdaBoost and Gradient Boost classifiers are built based on 500 numbers of estimators on which the boosting is terminated. After constructing these classification models, training data are fitted into it. Later the testing dataset are used for prediction purpose. After the prediction is done, performance of the classifiers are evaluated based on the predicted value and the actual value.

**B. Performance Evaluation Metrics** While evaluating performance skill of a model, it is necessary to employ some metrics to justify the evaluation. For this purpose, following metrics are taken into consideration in order to identify the best relevant problem-solving approach. Accuracy [14] is a metric that identifies the ratio of true predictions over the total number of instances considered. However, the accuracy may not be enough metric for evaluating model‘s performance since it does not consider wrong predicted cases. If a fake post is treated as a true one, it creates a significant problem. Hence, it is necessary to consider false positive and false negative cases that compensate to misclassification. For measuring this compensation, precision and recall is quite necessary to be considered [7].

# IV. EXPERIMENTAL RESULTS

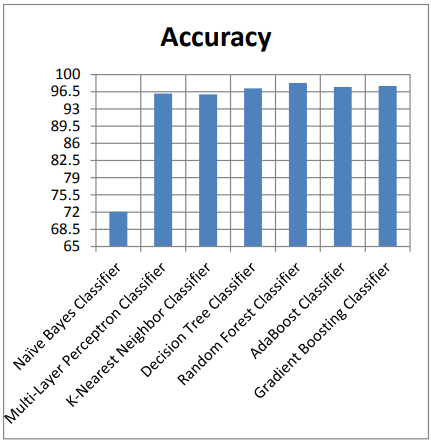
All the above-mentioned classifiers are trained and tested for detecting fake job posts over a given dataset that contains both fake and legitimate posts. The following Table 1 shows the comparative study of the classifiers with respect to evaluating metrics and Table 2 provides results for the classifiers that are based on ensemble techniques. Fig. 4 to Fig. 7 depict overall performance of all the classifiers in terms of accuracy, f1-score, Cohen-kappa score, MSE respectively.



# TABLE II

**PERFORMANCE COMPARISON CHART FOR ENSEMBLE CLASSIFIER BASED PREDICTION**

|  |  |  |  |
| --- | --- | --- | --- |
| **Performance**  **Measure**  **Metric** | **Random**  **Forest**  **Classifier** | **AdaBoost**  **Classifier** | **Gradient Boosting**  **Classifier** |
| Accuracy | 98.27% | 97.46% | 97.65% |
| F1-Score | 0.97 | 0.98 | 0.98 |
| Cohen-Kappa Score | 0.74 | 0.63 | 0.65 |
| MSE | 0.02 | 0.03 | 0.03 |



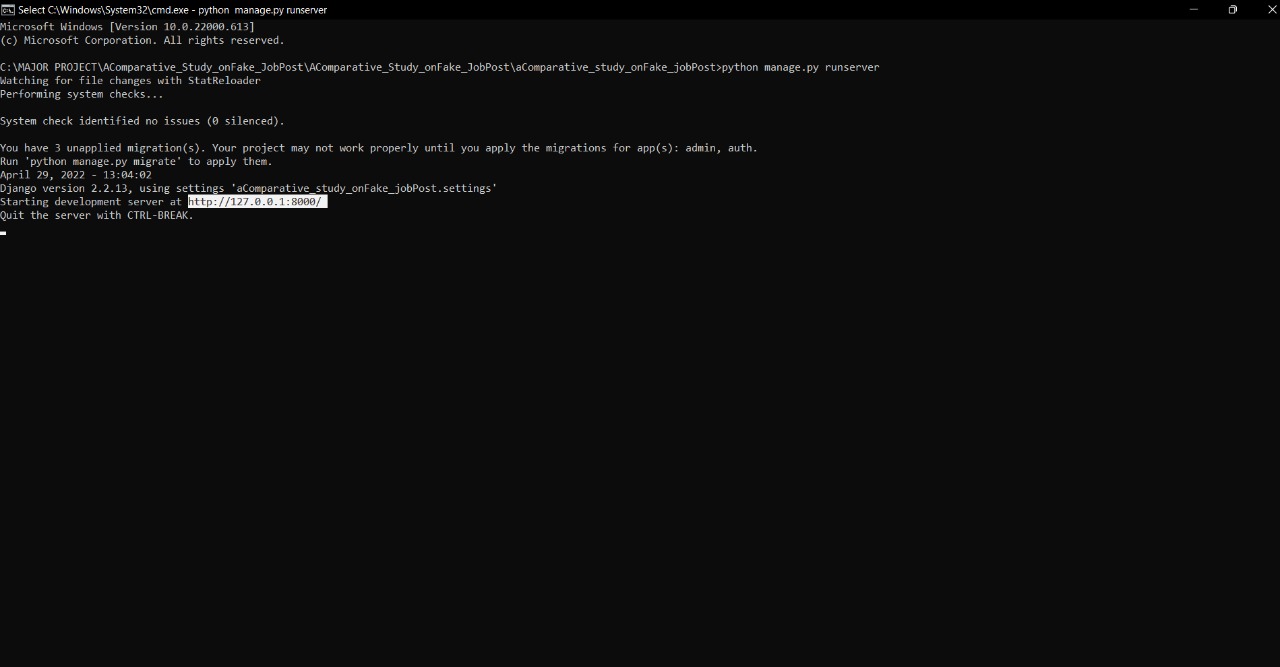


Fig.4. Run command



Fig.5. Admin login

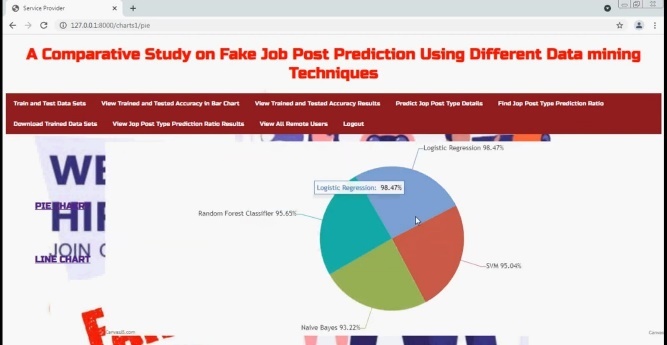


Fig.6. Pie chart of EMDSCAN data set.

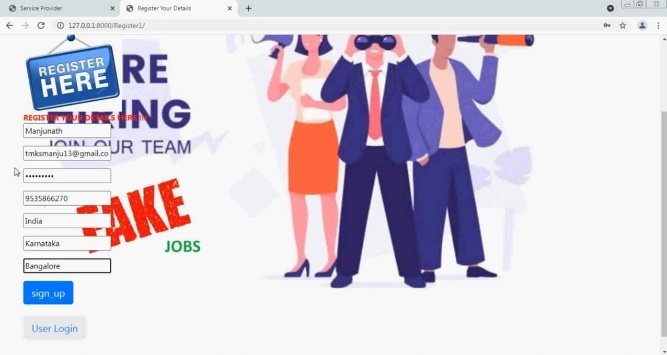


Fig.7. User registration Page

## 

Fig.8. User login

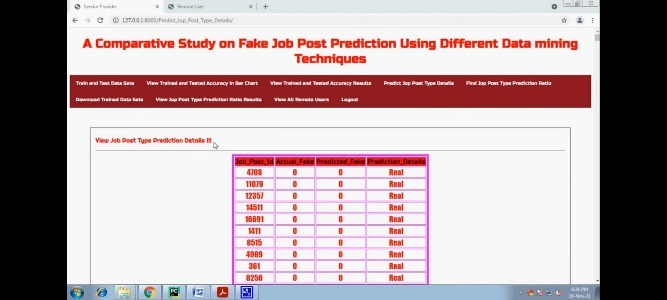


Fig.9. After data cleaning true and fake jobs posts

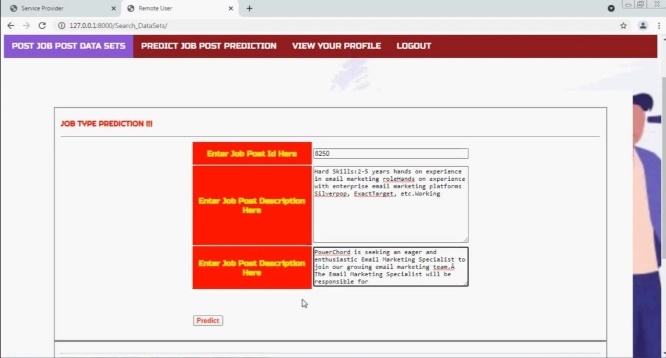


Fig.10. checking the job post

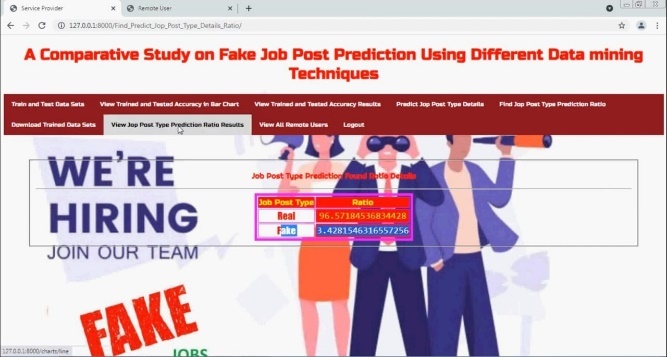


Fig.11.Distinguishing between true and fake job

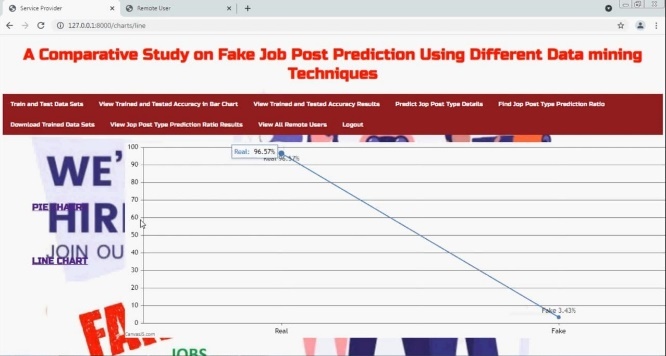


Fig.12. Graphical representation of true and fake jobs

## V. CONCLUSIONS

Employment scam detection will guide job-seekers to get only legitimate offers from companies. For tackling employment scam detection, several machine learning algorithms are proposed as countermeasures in this paper. Supervised mechanism is used to exemplify the use of several classifiers for employment scam detection. Experimental results indicate that Random Forest classifier outperforms over its peer classification tool. The proposed approach achieved accuracy 98.27% which is much higher than the existing methods.

## 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. I. Rish, ―An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier,‖ no. January 2001, pp. 41–46, 2014.
3. D. E. Walters, ―Bayes’s Theorem and the Analysis of Binomial Random Variables,‖ Biometrical J., vol. 30, no. 7, pp. 817–825, 1988, doi: 10.1002/bimj.4710300710.
4. F. Murtagh, ―Multilayer perceptrons for classification and regression,‖ Neurocomputing, vol. 2, no. 5–6, pp. 183–197, 1991, doi: 10.1016/0925-2312(91)90023-5.
5. P. Cunningham and S. J. Delany, ―K -Nearest Neighbour Classifiers,‖ Mult. Classif. Syst., no. May, pp. 1–17, 2007, doi: 10.1016/S0031-3203(00)00099-6.
6. H. Sharma and S. Kumar, ―A Survey on Decision Tree Algorithms of Classification in Data Mining,‖ Int. J. Sci. Res., vol. 5, no. 4, pp. 2094–2097, 2016, doi: 10.21275/v5i4.nov162954.
7. E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, “Machine learning for email spam filtering: review, approaches and open research problems,‖ Heliyon, vol. 5, no. 6, 2019, doi:

10.1016/j.heliyon.2019.e01802.

1. L. Breiman, ―ST4\_Method\_Random\_Forest,‖ Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi:

10.1017/CBO9781107415324.004.

1. B. Biggio, I. Corona, G. Fumera, G. Giacinto, and F. Roli, ―Bagging classifiers for fighting poisoning attacks in adversarial classification tasks,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6713 LNCS, pp. 350–359, 2011, doi:

10.1007/978-3-642-21557-5\_37.

1. A. Natekin and A. Knoll, ―Gradient boosting machines, a tutorial,‖ Front. Neurorobot., vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.
2. N. Hussain, H. T. Mirza, G. Rasool, I. Hussain, and M. Kaleem, ―Spam review detection techniques: A systematic literature review,‖ Appl. Sci., vol. 9, no. 5, pp. 1–26, 2019, doi: 10.3390/app9050987.