

# ONLINE FAKE JOB ADVERT DETECTION APPLICATION USING MACHINE LEARNING

**Dr. Y. Venkataramana Reddy<sup>1</sup>, B. Sai Neeraj<sup>2</sup>, K. Puneeth Reddy<sup>3</sup>, P. Bhargav Reddy<sup>4</sup>**

Associate Professor, Department of Information Technology Anurag University, Hyderabad,  
India.

<sup>2,3,4</sup> Students, Department of Information Technology Anurag University, Hyderabad, India.

## ABSTRACT

To avoid fraudulent post for job in the internet, an automated tool using machine learning based classification techniques is proposed in the paper. Different classifiers are used for checking fraudulent post in the web and the results of those classifiers are compared for identifying the best employment scam detection model. It helps in detecting fake job posts from an enormous number of posts. Two major types of classifiers, such as single classifier and ensemble classifiers are considered for fraudulent job posts detection. However, experimental results indicate that ensemble classifiers are the best classification to detect scams over the single classifiers.

**KEYWORDS:** *fraudulent post, machine learning, classifiers, employment scam.*

## INTRODUCTION

Employment scam is one of the serious issues in recent times addressed in the domain of Online Recruitment Frauds (ORF). In recent days, many companies prefer to post their vacancies online so that these can be accessed easily and timely by the job-seekers. However, this intention may be one type of scam by the fraud people because they offer employment to job-seekers in terms of taking money from them. Fraudulent job advertisements can be posted against a reputed company for violating their credibility. These fraudulent job post detection draws a good attention for obtaining an automated tool for identifying fake jobs and reporting them to people for avoiding application for such jobs. For this purpose, machine learning approach is applied which employs several classification algorithms for recognizing fake posts. In this case, a classification tool isolates fake job posts from a larger set of job advertisements and alerts the user. To address the problem of identifying scams on job posting, supervised learning algorithm as classification techniques are considered initially. A classifier maps input variable to target classes by considering training data. Classifiers addressed in the paper for identifying fake job

---

posts from the others are described briefly. These classifiers-based prediction may be broadly categorized into -Single Classifier based Prediction and Ensemble Classifiers based Prediction.

#### **A. Single Classifier based Prediction**

Classifiers are trained for predicting the unknown test cases. The following classifiers are used while detecting fake job posts

##### **a) Naive Bayes Classifier-**

The Naive Bayes classifier is a supervised classification tool that exploits the concept of Bayes Theorem of Conditional Probability. The decision made by this classifier is quite effective in practice even if its probability estimates are inaccurate. This classifier obtains a very promising result in the following scenario- when the features are independent or features are completely functionally dependent. The accuracy of this classifier is not related to feature dependencies rather than it is the amount of information loss of the class due to the independence assumption is needed to predict the accuracy.

##### **b) Multi-Layer Perceptron Classifier -**

Multi-layer perceptron can be used as supervised classification tool by incorporating optimized training parameters. For a given problem, the number of hidden layers in a multilayer perceptron and the number of nodes in each layer can differ. The decision of choosing the parameters depends on the training data and the network architecture.

##### **c) K-nearest Neighbor Classifier-**

K-Nearest Neighbor Classifiers, often known as lazy learners, identifies objects based on closest proximity of training examples in the feature space. The classifier considers k number of objects as the nearest object while determining the class. The main challenge of this classification technique relies on choosing the appropriate value of k.

##### **d) Decision Tree Classifier-**

A Decision Tree (DT) is a classifier that exemplifies the use of tree-like structure. It gains knowledge on classification. Each target class is denoted as a leaf node of DT and non-leaf nodes of DT are used as a decision node that indicates certain test. The outcomes of those tests are identified by either of the branches of that decision node. Starting from the beginning at the root this tree are going through it until a leaf node is reached. It is the way of obtaining classification result from a decision tree. Decision tree learning is an approach that has been applied to spam filtering. This can be useful for forecasting the goal based on some criterion by implementing and training this model.

## **LITERATURE SURVEY**

**TITLE:** “An Intelligent Model for Online Recruitment Fraud Detection,”

This study research attempts to prohibit privacy and loss of money for individuals and organization by creating a reliable model which can detect the fraud exposure in the online recruitment environments. This research presents a major contribution represented in a reliable detection model using ensemble approach based on Random forest classifier to detect Online Recruitment Fraud (ORF). The detection of Online Recruitment Fraud is characterized by other types of electronic fraud detection by its modern and the scarcity of studies on this concept. The researcher proposed the detection model to achieve the objectives of this study. For feature selection, support vector machine method is used and for classification and detection, ensemble classifier using Random Forest is employed. A freely available dataset called Employment Scam

Aegean Dataset (EMSCAD) is used to apply the model. Pre-processing step had been applied before the selection and classification adoptions. The results showed an obtained accuracy of 97.41%. Further, the findings presented the main features and important factors in detection purpose include having a company profile feature, having a company logo feature and an industry feature.

**TITLE:** An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier,

The naive Bayes classifier greatly simplify learn-ing by assuming that features are independent given class. Although independence is generally a poor assumption, in practice naive Bayes often competes well with more sophisticated classifiers. Our broad goal is to understand the data character-istics which affect the performance of naive Bayes. Our approach uses Monte Carlo simulations that al-low a systematic study of classification accuracy for several classes of randomly generated prob-lems. We analyze the impact of the distribution entropy on the classification error, showing that low-entropy feature distributions yield good per-formance of naive Bayes. We also demonstrate that naive Bayes works well for certain nearly-functional feature dependencies, thus reaching its best performance in two opposite cases: completely independent features (as expected) and function-ally dependent features (which is surprising). An-other surprising result is that the accuracy of naive Bayes is not directly correlated with the degree of feature dependencies measured as the class-conditional mutual information between the fea-tures. Instead, a better predictor of naive Bayes ac-curacy is the amount of information about the class that is lost because of the independence assump-tion.

**TITLE:** Bayes's Theorem and the Analysis of Binomial Random Variables,

A very practical application of Bayes's theorem, for the analysis of binomial random variables, is presented. Previous papers (Walters, 1985; Walters, 1986a) have already demonstrated the reliability of the technique for one, or two random variables, and the extension of the approach to several random variables is described. Two biometrical examples are used to illustrate the method.

**TITLE:** Multilayer perceptrons for classification and regression,

We review the theory and practice of the multilayer perceptron. We aim at addressing a range of issues which are important from the point of view of applying this approach to practical problems. A number of examples are given, illustrating how the multilayer perceptron compares to alternative, conventional approaches. The application fields of classification and regression are especially considered. Questions of implementation, i.e. of multilayer perceptron architecture, dynamics, and related aspects, are discussed. Recent studies, which are particularly relevant to the areas of discriminant analysis, and function mapping, are cited.

**TITLE:** K -Nearest Neighbour Classifiers,

We analyze a Relational Neighbor (RN) classifier, a simple relational predictive model that predicts only based on class labels of related neighbors, using no learning and no inherent attributes. We show that it performs surprisingly well by comparing it to more complex models such as Probabilistic Relational Models and Relational Probability Trees on three data sets from published work.

## PROBLEM STATEMENT

Given a dataset of job postings, the task is to predict whether a job posting is real or fake. A fake job posting is defined as a job posting that misrepresents the job or the employer, with the

intention of deceiving or defrauding job seekers. Fake job postings can be created for various reasons, such as collecting personal information, promoting a scam or pyramid scheme, or even as a cover for illegal activities.

The task of identifying fake job postings is important for both job seekers and employers. For job seekers, it helps to protect them from potential fraud and scams, while for employers, it helps to maintain the credibility and trustworthiness of their brand and job postings.

## **PROPOSED SYSTEM**

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 job seekers to concentrate on legitimate job posts only. In this context, a dataset from Kaggle is employed that provides information regarding a job that may or may not be suspicious.

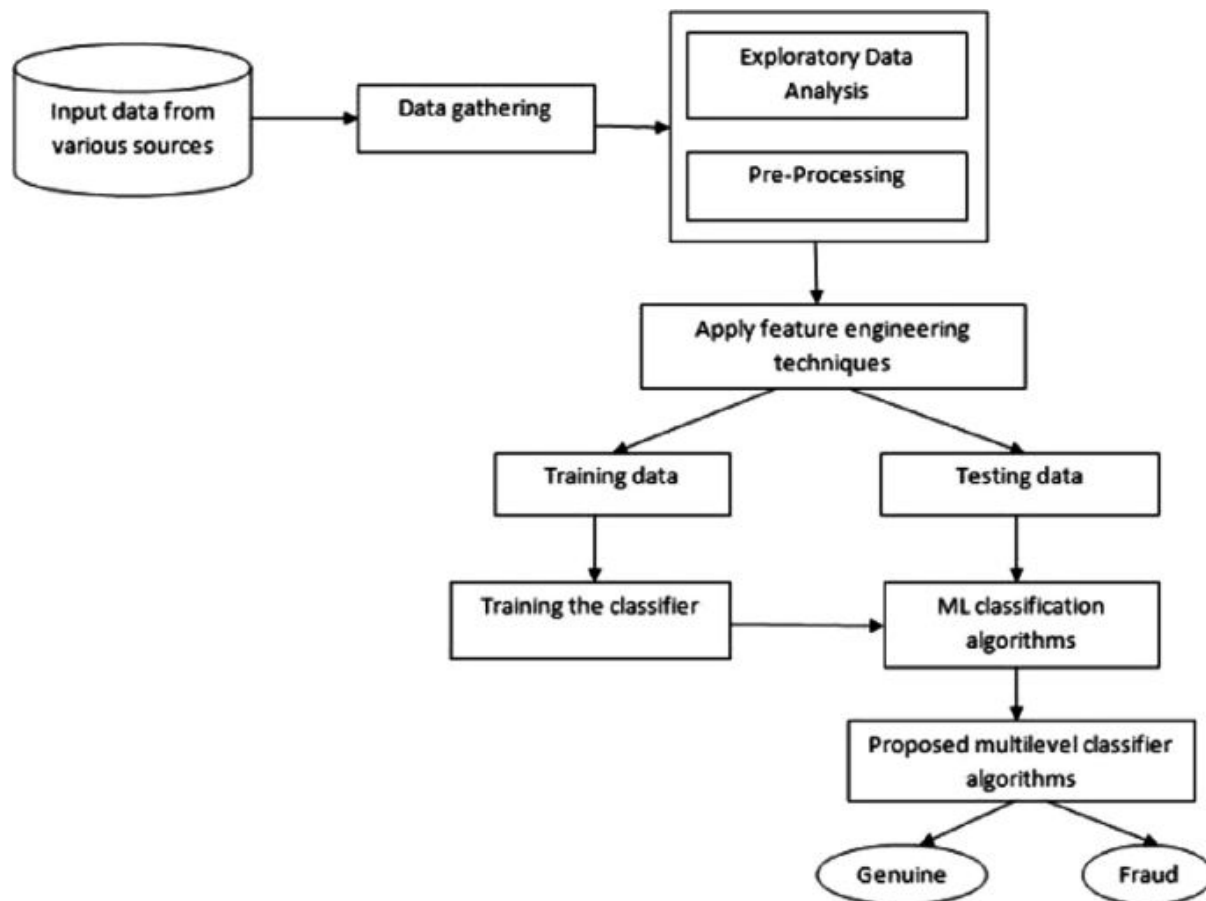
### **A. Implementation of Classifiers**

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.

### **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 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.

## FLOW OF EVENTS



## ADVANTAGES OF PROPOSED SYSTEM

There are several advantages of using machine learning in fake job prediction, including:

**Improved accuracy:** Machine learning algorithms can be trained on large datasets to detect patterns and identify features that are indicative of fake job postings. This can lead to more accurate and reliable predictions, reducing the number of false positives and false negatives.

**Efficiency:** Machine learning models can process large volumes of data much more quickly and efficiently than manual review by human moderators, reducing the time required to review job postings and improving the overall efficiency of the job posting process.

**Scalability:** Machine learning models can be easily scaled up to handle large volumes of job postings, making them ideal for job portals and other platforms that receive a high volume of job postings.

**Consistency:** Machine learning algorithms can provide consistent results across all job postings, eliminating the potential for subjectivity and inconsistencies that can arise with human moderators.

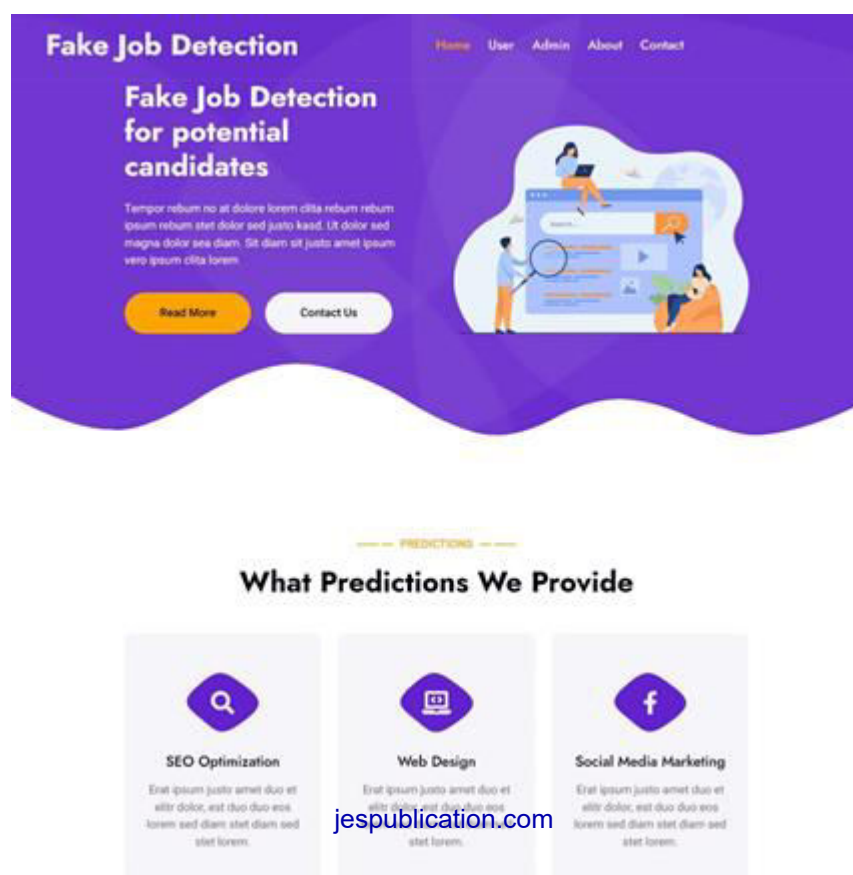
**Adaptability:** Machine learning models can adapt to new fraud techniques and patterns as they emerge, making them more effective at identifying and preventing fake job postings.

**Cost-effective:** While there may be an initial cost to developing and training a machine learning model for fake job prediction, once the model is deployed, it can be much more cost-effective than employing a team of human moderators.

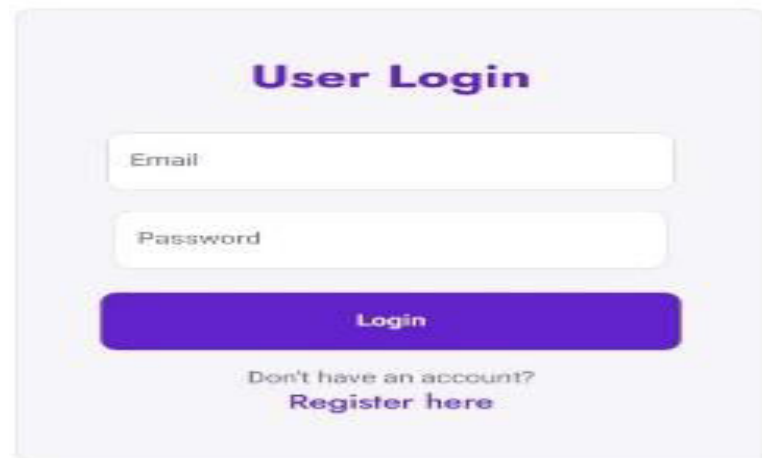
Overall, using machine learning in fake job prediction can provide significant advantages over the existing system, including improved accuracy, efficiency, scalability, consistency, and adaptability, while also reducing costs and improving the overall user experience

## RESULTS AND DISCUSSION

### HOME PAGE



## USER LOGIN



A user login form with a light purple background. At the top, the title "User Login" is displayed in a bold, dark purple font. Below the title are two white input fields with rounded corners. The first field is labeled "Email" in a small, gray font. The second field is labeled "Password" in a small, gray font. Below these fields is a solid purple button with the word "Login" in white. At the bottom of the form, the text "Don't have an account?" is followed by a link "Register here" in a dark purple font.

**User Login**

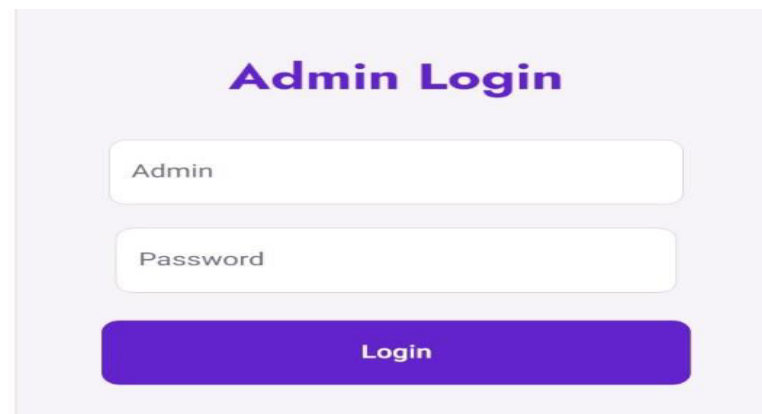
Email

Password

Login

Don't have an account?  
[Register here](#)

## ADMIN LOGIN



An admin login form with a light purple background. At the top, the title "Admin Login" is displayed in a bold, dark purple font. Below the title are two white input fields with rounded corners. The first field is labeled "Admin" in a small, gray font. The second field is labeled "Password" in a small, gray font. Below these fields is a solid purple button with the word "Login" in white.

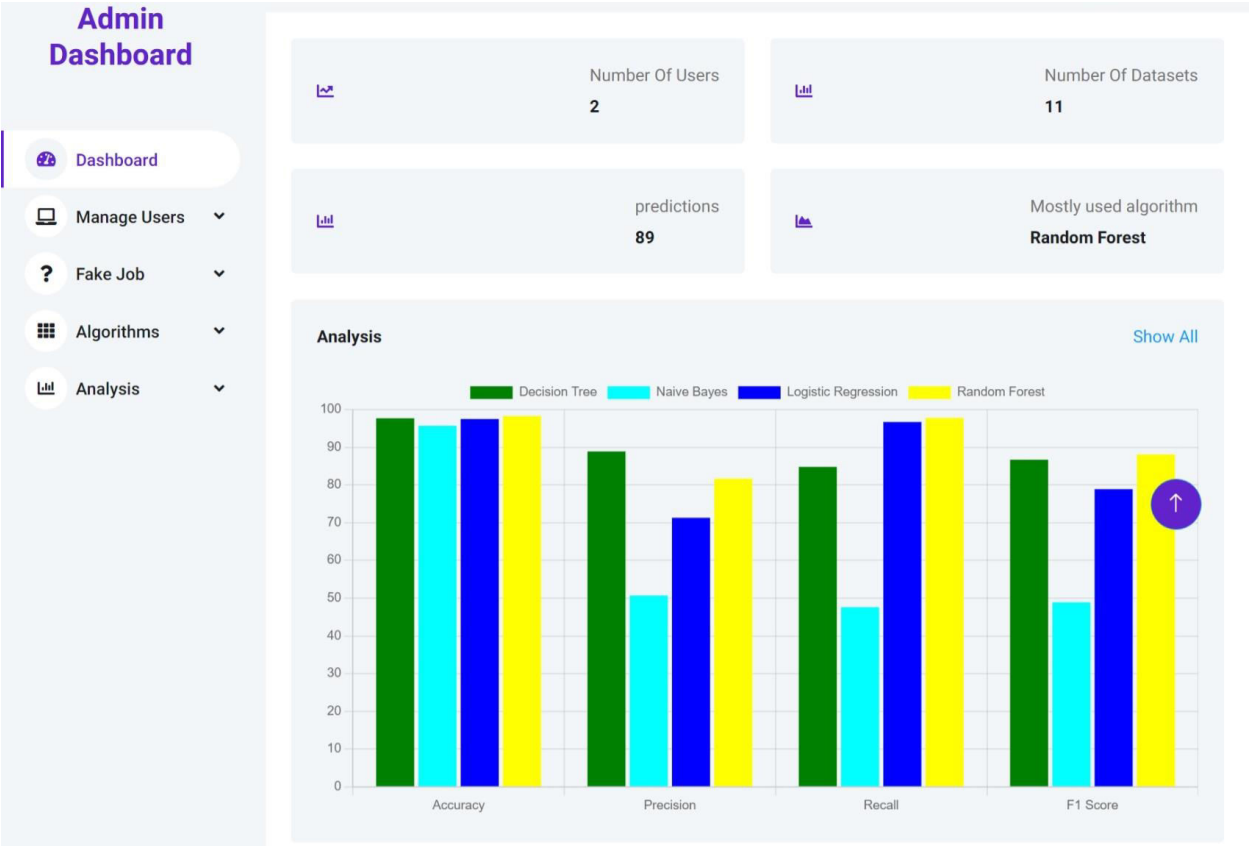
**Admin Login**

Admin

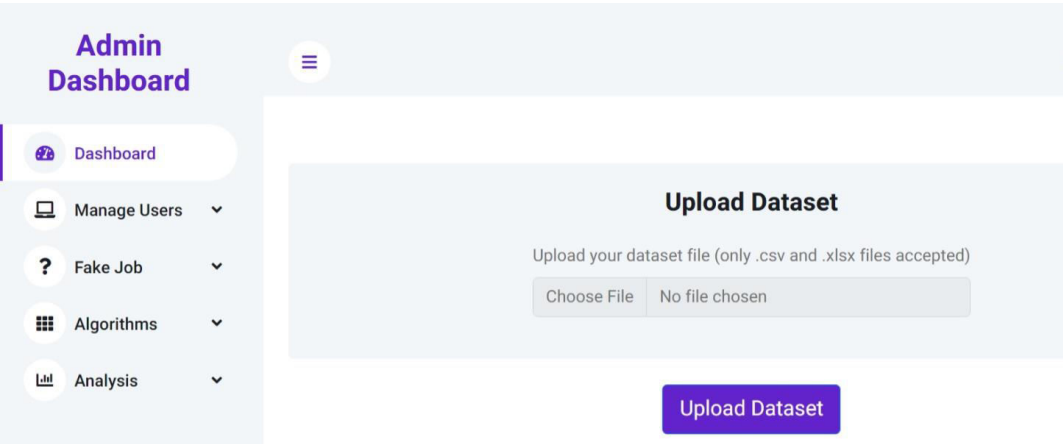
Password

Login

ADMIN DASHBOARD



UPLOAD DATASET





ALGORITHM COMPARISON

Date	Algorithm Used	Accuracy	Precision Score	Recall Score	F1 Score
01 Jan 2023	Logistic Regression	97.50186428038778	71.32670553700844<	96.78476492908649<	78.92893996247655<
01 Jan 2023	Decision Tree	97.81879194630872	89.17677658586449<	85.46277665995976<	87.21626897354007<
01 Jan 2023	Naive Bayes	95.76808351976138	50.0<	47.88404175988069<	48.9191505570898<
01 Jan 2023	Random Forest	98.37807606263982	81.67912844449742<	97.85301981429282<	87.97983193277311<

USER FILLING THE FORM

PREDICT FORM

Fill up the form

Job Title

Data Entry

Job Location

US

Department

Marketing

Salary Range

Salary Range

Description

A-fast-Growing-compan

company profile

We-are-food52

Requirements

Experience-CRM

Required Education

Graduation

Required Experience

Internship

Benefits

Full-Benefits-Offered

Employment Type

Full-time

Industry

Computer-Software

Function

SALES

Predict Job



## PREDICTION OF JOB



## CONCLUSION

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.
- [8] L. Breiman, —ST4 Method Random Forest," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1017/CBO9781107415324.004.
- [9] 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.
- [10] A. Natekin and A. Knoll, —Gradient boosting machines, a tutorial," *Front. Neurorobot.*, vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.
- [11] 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.
- [12] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, —Fake News Detection on Social Media," *ACM SIGKDD Explor. Newsl.*, vol. 19, no. 1, pp. 22–36, 2017, doi: 10.1145/3137597.3137600.
- [13] Shivam Bansal (2020, February). [Real or Fake] Fake JobPosting Prediction, Version 1. Retrieved March 29, 2020 from <https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction>
- [14] H. M and S. M.N, —A Review on Evaluation Metrics for Data Classification Evaluations," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 01–11, 2015, doi: 10.5121/ijdkp.2015.5201.
- [15] S. M. Vieira, U. Kaymak, and J. M. C. Sousa, —Cohen's kappa coefficient as a performance measure for feature selection," 2010 IEEE World Congr. Comput. Intell. WCCI 2010, no. May 2016, 2010, doi: 10.1109/FUZZY.2010.5584447