e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:06/Issue:07/July-2024 Impact Factor- 7.868 www.irjmets.com

# FAKE JOB POSTING USING MACHINE LEARNING APPROACH

S Karthikeyan*[1]

*[1]Department Of MCA, Brindavan College Of Engineering, Bangalore, Karnataka, India.

## ABSTRACT

Due to the development of modern technologies and social media, the problem of advertising new job openings has become increasingly common in today's world. As a result, there will be several reasons for concern regarding fake job advertising for everyone. Predicting fake jobs comes with a number of challenges, similar to numerous further classification issues. This paper suggested using a variety of data mining techniques and classification algorithms, such as KNN, decision tree, support vector machine, naive bayes classifier, random forest classifier, multilayer perceptron, and deep neural network, to determine whether a job posting is legitimate or fraudulent. We conducted our investigations using 18000 examples from the Employment Scam Aegean Dataset (EMSCAD). A deep neural network classifier performs exceptionally well for this classification task. For this profound.

## I. INTRODUCTION

Job seekers in the modern era have access to a multitude of novel and exciting employment prospects. The adverts for these job offers let job seekers learn about their options based on factors such as availability, credentials, experience, appropriateness, etc.

The power of social media and the internet now affects the hiring process. Social media plays a major role in this since effective advertising is key to the success of any recruitment process. Thanks to social media and electronic media marketing, there are more and more options to provide employment facts. Rather, the ease with which job posts can be shared has led to a rise in the quantity of fake job postings, which in turn have harassed job searchers.

Because they wish to maintain the security and consistency of their personal, academic, and professional information, people don't reply to fresh job postings. Therefore, gaining people's trust and credibility through authentic job adverts on social and electronic media is a highly challenging task. The technologies we use every day are there to make our lives easier and better, not to put us in dangerous jobs. If job advertisements can be accurately filtered to identify false job postings, recruiting new employees will improve significantly. It is difficult for job seekers to find the employment they want, which is a major waste of their time when false job postings are present.

### Existing System

K-Nearest Neighbour 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.

### Disadvantages

- Accuracy depends on the quality of the data.
- With large data, the prediction stage might be slow.
- Sensitive to the scale of the data and irrelevant features.
- Require high memory – need to store all of the training data.

### Proposed System

The target of this study is to detect whether a job posting is fraudulent or not. Identifying and eliminating these fake job advertisements will help the jobseekers to concentrate on legitimate job posts only. In this project we are comparing three machine learning algorithms. Such as Random forest Classifier, Decision Tree and Adaboost to select the best predictive model for detecting the Fake Recruitments.

**Advantages**

- By using machine learning algorithm, the whole process of data interpretation and analysis is done by computer.
- No men intervention is required for the prediction or interpretation of data. The whole process of machine learning is machine starts learning and predicting the algorithm or program to give the best result.
- It can handle varieties of data, Even in an uncertain and dynamic environment, it can handle a variety of data. It is multidimensional as well as a multitasker.

**Module description**

- Data collection
- Data Pre-Processing
- Training data and Test data
- Model Creation
- Model Prediction

### ➢ Data collection

The dataset should be collected form the Kaggle website. If the given dataset of the job is true and original, it will recommend the the job. If the given dataset of the job is fake, then the model will find the original job related to the fake job.

### ➢ Data Pre-Processing

Pre-processing refers to the transformations applied to our data before providing the data to the algorithm. Data Preprocessing technique is used to convert the raw data into an understandable data set. In other words, whenever the information is gathered from various sources it is collected in raw format that isn't possible for the analysis.

### ➢ Training data and Test data

- For choosing a model we split our dataset into train and test
- Here data's are split into 3:1 ratio that means
- Training data having 70 percent and testing data having 30 percent
- In this split process preforming based on train_test_split model
- After splitting we get xtrain xtest and ytrain ytest

### ➢ Model Creation

- Contextualise deep learning in your organisation.
- Explore the data and choose the type of algorithm.
- Prepare and clean the dataset.
- Split the prepared dataset and perform cross validation.
- Deploy the model.

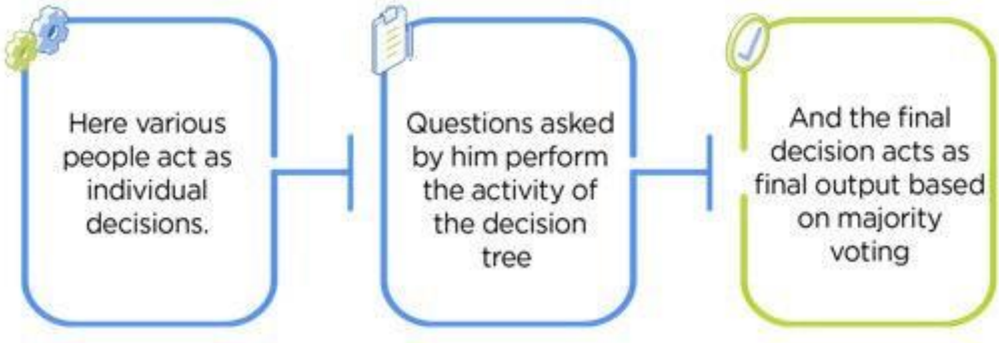
### ➢ Model Prediction

Predictive modeling is a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data. It works by analyzing current and historical data and projecting what it learns on a model generated to forecast likely outcomes. In this Project, our final prediction is model to predict whether a job posting should be Fake or Real and recommend a real job based upon some job types.

**Algorithm Implementation Random Forest Algorithm:-**

Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

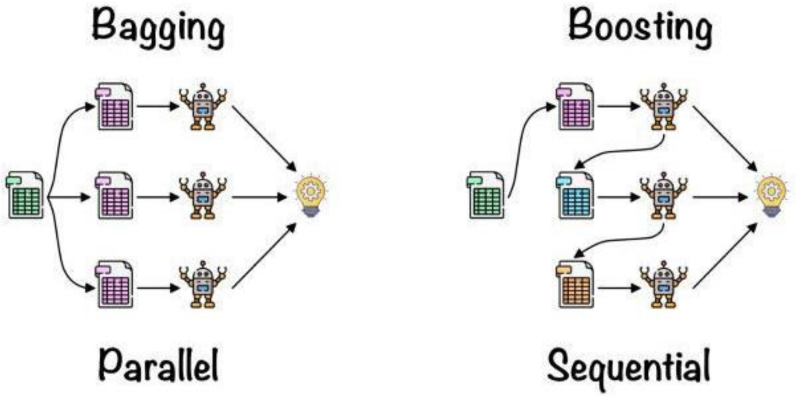
One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification. It performs better for classification and regression tasks. In this tutorial, we will understand the working of random forest and implement random forest on a classification task.


Here various people act as individual decisions.
Questions asked by him perform the activity of the decision tree
And the final decision acts as final output based on majority voting

**Working of Random Forest Algorithm**

Before understanding the working of the random forest algorithm in machine learning, we must look into the ensemble learning technique. Ensemble simply means combining multiple models. Thus a collection of models is used to make predictions rather than an individual model. Ensemble uses two types of methods:

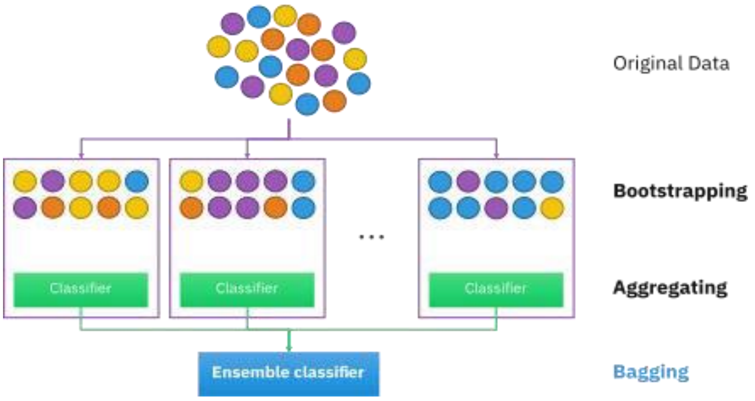
1. Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.
2. Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.



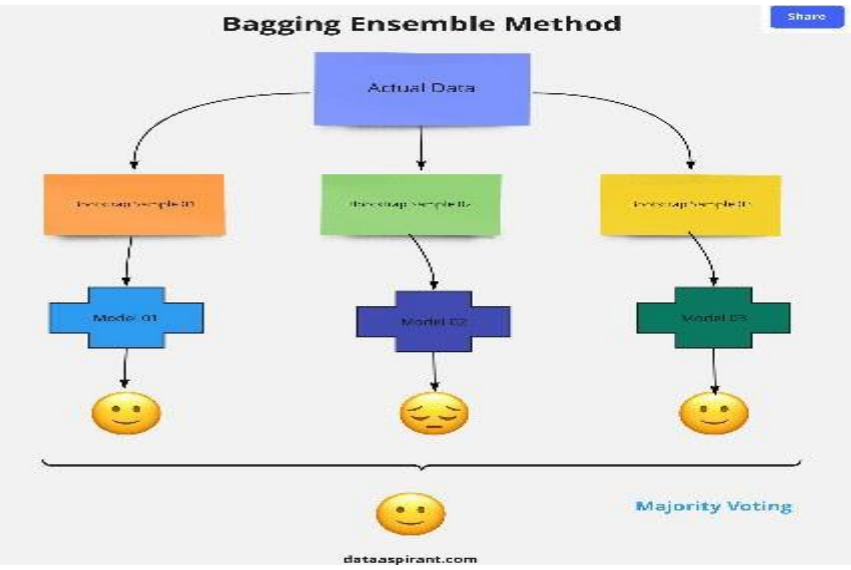
As mentioned earlier, Random forest works on the Bagging principle. Now let's dive in and understand bagging in detail.

**Bagging**

Bagging, also known as Bootstrap Aggregation, is the ensemble technique used by random forest. Bagging chooses a random sample/random subset from the entire data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently, which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting, is known as aggregation.



Now let's look at an example by breaking it down with the help of the following figure. Here the bootstrap sample is taken from actual data (Bootstrap sample 01, Bootstrap sample 02, and Bootstrap sample 03) with a replacement which means there is a high possibility that each sample won't contain unique data. The model (Model 01, Model 02, and Model 03) obtained from this bootstrap sample is trained independently. Each model generates results as shown. Now the Happy emoji has a majority when compared to the Sad emoji. Thus based on majority voting final output is obtained as Happy emoji.



## Boosting

Boosting is one of the techniques that use the concept of ensemble learning. A boosting algorithm combines multiple simple models (also known as weak learners or base estimators) to generate the final output. It is done by building a model by using weak models in series.

There are several boosting algorithms; Boost was the first really successful boosting algorithm that was developed for the purpose of binary classification. AdaBoost is an abbreviation for Adaptive Boosting and is a prevalent boosting technique that combines multiple "weak classifiers" into a single "strong classifier." There are Other Boosting techniques. For more, you can visit

Steps Involved in Random Forest Algorithm

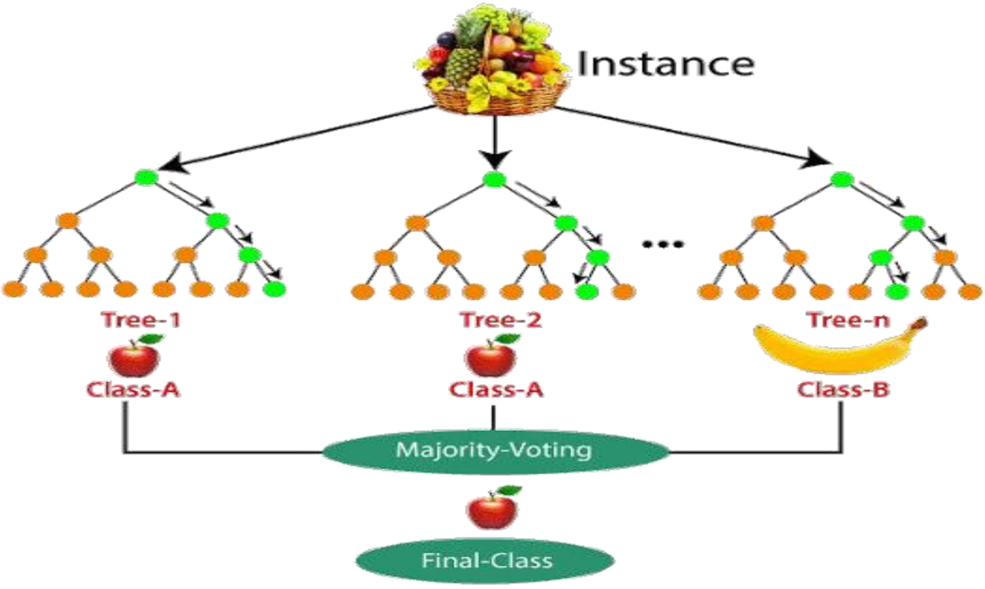
Step 1: In the Random forest model, a subset of data points and a subset of features is selected for constructing each decision tree. Simply put, n random records and m features are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression, respectively.

For example: consider the fruit basket as the data as shown in the figure below. Now n number of samples are taken from the fruit basket, and an individual decision tree is constructed for each sample. Each decision tree will generate an output, as shown in the figure. The final output is considered based on majority voting. In the below figure, you can see that the majority decision tree gives output as an apple when compared to a banana, so the final output is taken as an apple.



## II. CONCLUSION

The identification of job scams has recently become a major problem worldwide. We have examined the effects of employment scams in this paper since they might be a very lucrative topic of study and make it difficult to identify fake job postings. We experimented with the EMSCAD dataset, which contains actual fake job postings.

We experiment with both machine learning (SVM, KNN, Naive Bayes, Random Forest, and MLP) and deep learning (Deep Neural Network) in this study. A comparative analysis of deep learning and traditional machine learning-based classifiers is presented in this article. The most accurate classification is achieved by the Random Forest Classifier when compared to other traditional machine learning techniques. On average, Deep Neural Network and DNN (fold 9) had the highest classification accuracy.

## III. REFERENCES

[1] S. Vidros, C. Kolias , G. Kambourakis ,and L. Akoglu, “Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset”, Future Internet 2017, 9, 6; doi:10.3390/fi9010006.

[2] B. Alghamdi, F. Alharby, “An Intelligent Model for Online Recruitment Fraud Detection”, Journal of Information Security, 2019, Vol 10, pp. 155 176, https://doi.org/10.4236/iis.2019.103009 .

[3] Tin Van Huynh1, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen1, and Anh Gia-Tuan Nguyen, “Job Prediction: From Deep Neural Network Models to Applications”, RIVF International Conference on Computing and Communication Technologies (RIVF), 2020.

[4] Jiawei Zhang, Bowen Dong, Philip S. Yu, “FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network”, IEEE 36th International Conference on Data Engineering (ICDE), 2020.

[5] Scanlon, J.R. and Gerber, M.S., “Automatic Detection of Cyber Recruitment by Violent Extremists”, Security Informatics, 3, 5, 2014, https://doi.org/10.1186/s13388-014-0005-5

[6] Y. Kim, “Convolutional neural networks for sentence classification,” arXiv Prepr. arXiv1408.5882, 2014.

[7] T. Van Huynh, V. D. Nguyen, K. Van Nguyen, N. L.-T. Nguyen, and A.G.- T. Nguyen, “Hate Speech Detection on Vietnamese Social Media Text using the Bi-GRU-LSTM-CNN Model,” arXiv Prepr. arXiv1911.03644, 2019.

[8] 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.

[9] C. Li, G. Zhan, and Z. Li, "News Text Classification Based on Improved BiLSTM-CNN," in 2018 9th International Conference on Information Technology in Medicine and Education (ITME), 2018, pp. 890-893.

[10] K. R. Remya and J. S. Ramya, "Using weighted majority voting classifier combination for relation classification in biomedical texts," International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, pp. 1205-1209.

[11] 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/imsa.2016.8405Journal of Engineering Sciences Vol 14 Issue 02, 2023 ISSN:0377-9254 jespublication.com Page 73

[12] Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan LuuThuy Nguyen. "Emotion Recognition for Vietnamese Social Media Text", arXiv Prepr. arXiv:1911.09339, 2019.

[13] Thin Van Dang, Vu Duc Nguyen, Kiet Van Nguyen and Ngan LuuThuy Nguyen, "Deep learning for aspect detection on vietnamese reviews" in In Proceeding of the 2018 5th NAFOSTED Conference on Information and Computer Science (NICS), 2018, pp. 104-109.

[14] Li, H.; Chen, Z.; Liu, B.; Wei, X.; Shao, J. Spotting fake reviews via collective positive-unlabeled learning. In Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM), Shenzhen, China, 14-17 December 2014; pp. 899-904.

[15] Ott, M.; Cardie, C.; Hancock, J. Estimating the prevalence of deception in online review communities. In Proceedings of the 21st international conference on World Wide Web, Lyon, France, 16-20 April 2012; ACM: New York, NY, USA, 2012; pp. 201-210.

[16] Nizamani, S., Memon, N., Glasdam, M. and Nguyen, D.D. (2014) Detection of Fraudulent Emails by Employing Advanced Feature Abundance. Egyptian Informatics Journal, Vol.15, pp.169-174.
https://doi.org/10.1016/j.eij.2014.