****

**Assessment Report**

on

**“Fake Job Prediction”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AIML)**

By

Name: Amit Kumar Verma

Roll No.: 202401100400032

Section: A

**Under the supervision of**

“Bikki Kumar”

**KIET Group of Institutions, Ghaziabad**

**May, 2025**

**1. Introduction**

With the rise of online job platforms, fake job postings have become a growing concern. These postings can lead to scams and waste users' time. In this project, we aim to classify whether a job post is real or fake using machine learning models and analyze the job market by grouping similar job posts using clustering.

**2. Problem Statement**

To detect fake job postings using machine learning classification techniques and perform segmentation of similar postings using unsupervised clustering.

**3. Objectives**

1. **Classify job postings** as real or fake using machine learning.
2. **Extract key features** like title length and company profile presence.
3. **Train and evaluate models** using accuracy, precision, and recall.
4. **Visualize results** with confusion matrix heatmaps.
5. **Perform clustering** to segment similar job posts using K-Means.
6. **Gain insights** into patterns of fake job listings.
7. **Build a complete ML pipeline** from preprocessing to evaluation.

**4. Methodology**

* Data Preprocessing: Cleaned and encoded data, converted target labels to binary.
* Feature Selection: Selected numeric features such as title length, description length, and company profile flag.
* Classification: Trained a Random Forest model to predict fake job postings.
* Evaluation: Evaluated using accuracy, precision, recall, and confusion matrix.
* Clustering: Used K-Means for unsupervised segmentation to group job posts into clusters for pattern discovery.
* **Model Evaluation**:  
  + Evaluating accuracy, precision, recall, and F1-score.
  + Generating a confusion matrix and visualizing it with a heatmap.

**5. Data Preprocessing**

The dataset is cleaned and prepared as follows:

* Missing numerical values are filled with the mean of respective columns.
* Categorical values are encoded using one-hot encoding.
* Data is scaled using StandardScaler to normalize feature values.
* The dataset is split into 80% training and 20% testing.

**7. Evaluation Metrics**

The following metrics are used to evaluate the model:

* **Accuracy**: Measures overall correctness.
* **Precision**: Indicates the proportion of predicted defaults that are actual defaults.
* **Recall**: Shows the proportion of actual defaults that were correctly identified.
* **F1 Score**: Harmonic mean of precision and recall.
* **Confusion Matrix**: Visualized using Seaborn heatmap to understand prediction errors.

**8. Results and Analysis**

* The model provided reasonable performance on the test set.
* Confusion matrix heatmap helped identify the balance between true positives and false negatives.
* Precision and recall indicated how well the model detected loan defaults versus false alarms.

**9. Conclusion**

This project successfully demonstrated the use of machine learning to detect fake job postings with reasonable accuracy. By extracting simple yet effective features, we built a Random Forest classifier and evaluated it using standard metrics. Additionally, K-Means clustering revealed hidden patterns in job post structures, helping us understand how fake postings differ from real ones. Overall, this project proves that data-driven approaches can play a vital role in improving online job safety and user trust.

**10. References**

* Dataset: Provided by the instructor.
* Libraries: scikit-learn, pandas, seaborn, matplotlib
* Online Resources: Scikit-learn documentation, Medium articles on fake job detection





