# Data Mining & Data Warehousing Project

# Real / Fake Job Posting Prediction Final Report

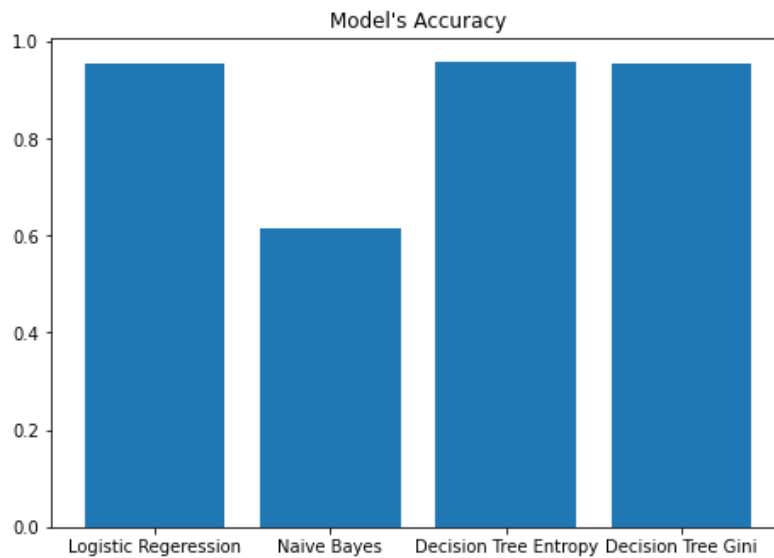Name: **Muhammad Hazique Khatri**                              Seat No: **B18101071**

| Task | Description |
|---|---|
| Source | Link: https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction?resource=download&select=fake_job_postings.csv |
| Columns & Rows Count | 18, 17880 |
| Pre Processing | <ul><li>Deleted unnecessary columns.</li><li>Converted all column text in single text format.</li><li>Preprocessed using nltk library.</li><li>Used word_tokenize for creating tokens.</li><li>Used word_lemmatize for useful words in English.</li><li>Used CountVectorizer to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text.</li></ul> |
| Training Data | 80% |
| Testing Data | 20% |
| Model Evaluation | 1. **Logistic Regression** <ul><li>Accuracy Value = 0.9555369127516778</li><li>AUC Value = 0.8934</li><li>Confusion Matrix<br>`[[3371   24]`<br>`[ 135   46]]`</li></ul> 2. **Naive Bayes** <ul><li>Accuracy Value = 0.6160514541387024</li><li>AUC Value = 0.8269</li><li>Confusion Matrix<br>`[[2050 1345]`<br>`[28  153]]`</li></ul> 3. **Decision Tree Entropy** <ul><li>Accuracy Value = 0.959731543624161</li><li>AUC Value = 0.7853</li><li>Confusion Matrix<br>`[[3325   70]`<br>`[74  107]]`</li></ul> 4. **Decision Tree Gini** <ul><li>Accuracy Value = 0.9555369127516778</li></ul> |

- AUC Value = 0.7778
- Confusion Matrix
  ```
  [[3312   83]
   [76   105]]
  ```

GitHub Link: https://github.com/HaziqueIqbal/Real-Fake_Job_Posting_Prediction

# Accuracy Graph



# ROC Curve