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
Null and Missing Values	1. Job Id -> Missing = 0, Null = 0 2. Title -> Missing = 0, Null = 0 3. Location -> Missing = 346, Null = 0 4. Department -> Missing = 11500, Null = 65% 5. Salary Range -> Missing = 15000, Null = 84% 6. Company Profile -> Missing = 3308, Null = 19% 7. Description -> Missing = 0, Null = 0 8. Requirements -> Missing = 2694, Null = 15% 9. Benefits -> Missing = 7206, Null = 40% 10. Telecommuting -> Missing = 0, Null = 0 11. Has Company Logo -> Missing = 0, Null = 0 12. Has Questions -> Missing = 0, Null = 0 13. Employment Type -> Missing = 3471, Null = 19% 14. Required Experience -> Missing = 7050, Null = 39% 15. Required Education -> Missing = 8105, Null = 45% 16. Industry -> Missing = 4903, Null = 27% 17. Function -> Missing = 6455, Null = 36% 18. Fraudulent -> Missing = 0, Null = 0 (Important Attribute)
Pre Processing	• Deleted Salary Range and Job Id because of Uniqueness, where salary includes ranges like xxxxx-xxxxx. • Filling NULL values as an empty space " ". • Converted all column text in single text format. • Then Deleted these columns. • Preprocessed using nltk library. • Used Regix to get character only. • Lowercase the text. • Used word_tokenize for creating tokens. • Used word_lemmatize for useful words in English. • 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.

Values of X & Y

- Variable X stores output of CountVectorizer.
- Variable Y store fraudulent values.

Training Data

80%

Testing Data

20%

Model Evaluation

1. **Logistic Regression**
   - Accuracy Value = 0.9555369127516778
   - AUC Value = 0.8934
   - Confusion Matrix

```
[[3371   24]
[ 135   46]]
```

2. **Naive Bayes**
   - Accuracy Value = 0.6160514541387024
   - AUC Value = 0.8269
   - Confusion Matrix

```
[[2050 1345]
 [28  153]]
```

3. **Decision Tree Entropy**
   - Accuracy Value = 0.959731543624161
   - AUC Value = 0.7853
   - Confusion Matrix

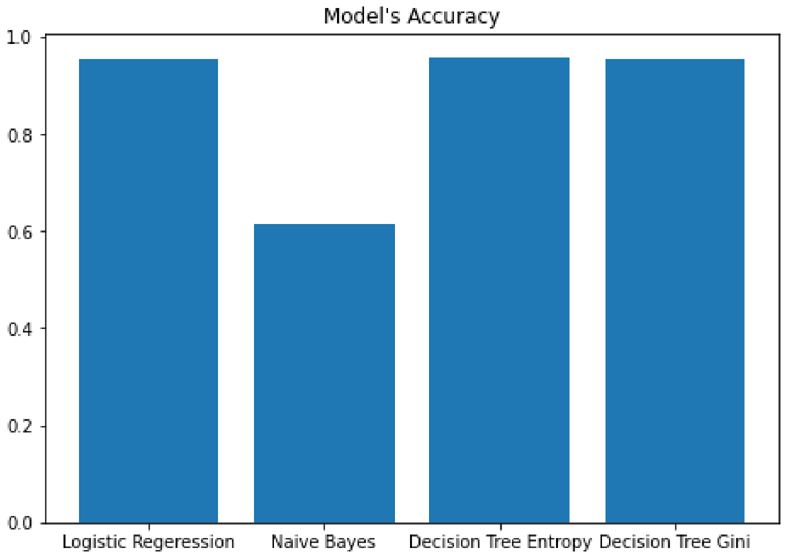
```
[[3325   70]
 [74  107]]
```

4. **Decision Tree Gini**
   - Accuracy Value = 0.9555369127516778
   - AUC Value = 0.7778
   - Confusion Matrix

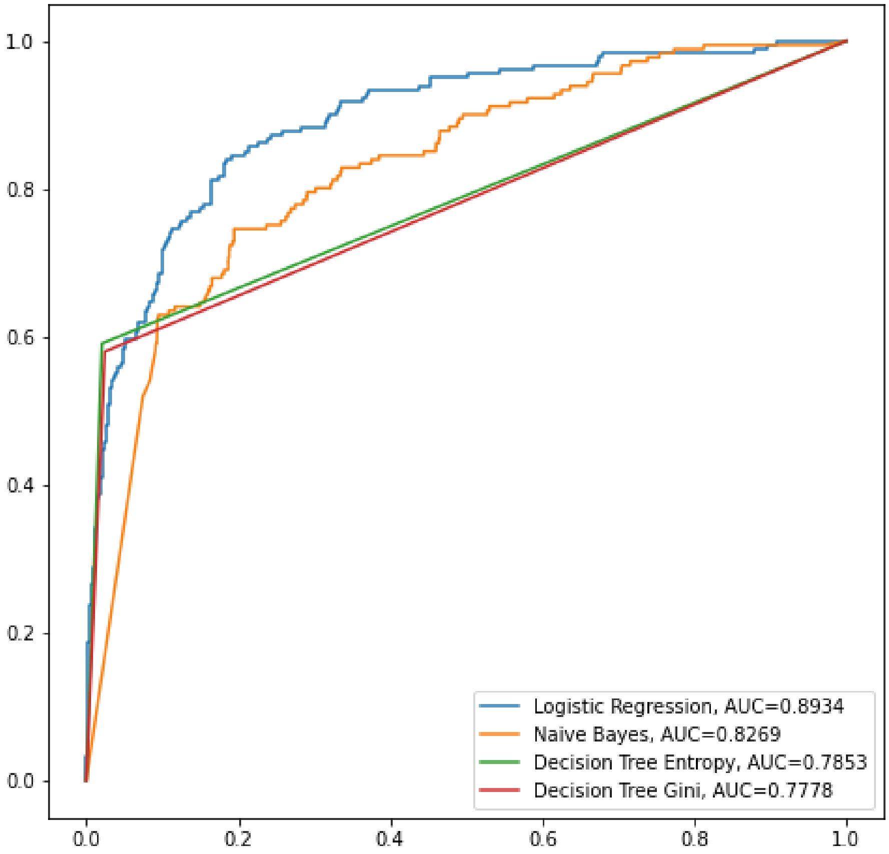
```
[[3312   83]
 [76  105]]
```

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

Accuracy Graph



ROC Curve


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression, AUC=0.8934
Naive Bayes, AUC=0.8269
Decision Tree Entropy, AUC=0.7853
Decision Tree Gini, AUC=0.7778