Job Fair Task Report

For Route Academy

By/ Tarek Muhammed Abdel-Hamid

1. Introduction

This report aims to analyse a dataset containing information about students and predict job offers based on their profiles. The dataset includes attributes such as skills, experience years, course grades, projects completed, extracurricular activities, and whether a job offer was received. This analysis is crucial for understanding factors that influence job opportunities for students.

2. Data Exploration

Summary Statistics

The dataset consists of 20,000 records with 7 columns:

- **Skills**
- **Experience Years**
- **Course Grades**
- **Projects Completed**
- **Extracurriculars**
- **Job Offer**

Key Findings from Data Exploration

- **Skills**: Most common skills include Python, Java, Data Analysis, SQL, and Machine Learning.
- **Experience Years**: Majority of students have between 0 to 5 years of experience, with 4 years being the most frequent.
- **Course Grades**: Grades range widely with a mean of approximately 75.6.
- **Projects Completed**: Students typically complete between 1 to 9 projects, with 5 projects being the most common.
- **Extracurriculars**: Students participate in 0 to 4 extracurricular activities, with 1 and 3 activities being the most frequent.

3. Data Pre-processing

Pre-processing Steps

- **Handling Missing Values**: No missing values were found in the dataset.
- **Removing Irrelevant Columns**: The student ID column was dropped as it does not contribute to predicting job offers.
- **Encoding Categorical Variables**: Skills column was split into separate binary columns to account for each skill possessed by students.

Justification for Pre-processing

- **Scaling**: Not required due to the small range of numerical features.

4. Modelling

Machine Learning Models Used

- **Logistic Regression**
- **Random Forest**
- **Gradient Boosting**
- **Decision Tree**
- **Neural Network (MLP)**

Training and Evaluation

- The dataset was split into training (60%) and testing (40%) sets.
- Models were trained using the training set and evaluated using accuracy score, confusion matrix, and classification report.

5. Results

Model Performance

- **Logistic Regression**: Accuracy = 100%
- **Random Forest**: Accuracy = 100%
- **Gradient Boosting**: Accuracy = 100%
- **Decision Tree**: Accuracy = 100%
- **Neural Network**: Accuracy = 96%

Best Performing Model

All models are over fitted because of the lack of data records and using hyper parameter tuning, but Neural Network model was the one with the least overfitting percentage

- **Model**: **Neural Network**
- **Accuracy**: 96%

6. Discussion

Insights Gained

- Models indicate that skills such as Python, Data Analysis, and Machine Learning are positively correlated with receiving job offers.
- Experience years and projects completed also play significant roles in predicting job offers.
- The dataset suggests a strong relationship between extracurricular activities and projects completed, influencing job offers.
- The plot of (PlotOverfitting) shows identical training and validation scores, which suggests perfect accuracy for both datasets. This can indicate potential overfitting, where models might have learned the training data too well, failing to generalize to new data. However, identical scores could also stem from data leakage or improper cross-validation. Review data processing and cross-validation steps to ensure robust evaluation.

Challenges and Limitations

- Limited dataset size may affect generalizability.
- Assumptions in feature importance and model choice need validation across diverse datasets.

7. Conclusion

This report demonstrates the effectiveness of machine learning models in predicting job offers based on student profiles. Skills, experience, and project involvement are crucial factors influencing job opportunities. Further research with larger datasets and additional features could enhance predictive accuracy.

Thanks for your time!