Python for Data Science Assignment

Total Marks: 30

Dataset Description: You are provided with a dataset containing information about individuals, including their job roles, education levels, gender, and English-speaking status.

Task 1: Data Loading and Preprocessing (5 marks)

1. Load the dataset from the given CSV file into a Pandas Data Frame.
2. Perform basic data preprocessing steps, including handling missing values and removing duplicate rows.
3. Display the first few rows of the cleaned dataset.

Task 2: Exploratory Data Analysis (6 marks)

1. Create visualizations to show the distribution of job roles, education levels, gender, and English-speaking status.
2. Calculate the percentage of individuals belonging to different job roles, education levels, genders, and English-speaking groups.

Task 3: Gender and English speaker Analysis (7 marks)

1. Calculate the average education level for each gender group (Male, Female, Others).
2. Compare the distribution of job roles among different gender groups using a stacked bar chart.
3. Create a histogram to show the distribution of education levels among English speaking and non-English speaking individuals.

Task 4: Predictive Modeling (12 Marks)

1. Encode categorical variables (job, education, gender, English speaker) using appropriate techniques (e.g., one-hot encoding).
2. Split the dataset into training and testing sets (80% training, 20% testing).
3. Build a classification model to predict the gender of individuals based on job role, education level, and English-speaking status.
4. Evaluate the model's performance using accuracy, precision, recall, and F1-score metrics.
5. Use feature importance techniques (e.g., feature importance scores, permutation feature importance) to identify the most influential features for gender prediction.
6. Visualize the ROC curve and AUC score for the gender prediction model.
7. Discuss the implications of the model's performance and the significance of the features in a concise summary.

Submission:

Submit your Python scripts (.py files) along with any necessary data files. Ensure that your code is well-commented and organized. Additionally, include a brief explanation of the logic and approach you used to solve each problem.

Important Notes:

· Your code should be well-documented and easily understandable.
· Feel free to use external libraries such as Pandas, Matplotlib, Seaborn, and Scikit-learn as needed.
· Remember to provide explanations for any assumptions or choices you make during data preprocessing and analysis.