Step 1: Data Exploration & Processing

1. **Dataset Loading and Initial Exploration** (1 hour)

    - Load the "data_wage.RData" dataset.
    - Perform an initial exploration to understand the dataset dimensions, variable types (numeric, categorical), and a brief overview of the data.

2. **Variable Identification** (30 minutes)

    - Identify the dependent variable (Y) as "wage" from the descriptions provided.
    - Identify potential independent variables (Xs) that might impact "wage", considering demographics, education, experience, and other relevant factors.

3. **Numeric and Categorical Variables Investigation** (1 hour)
    - Investigate the distribution of numeric and categorical variables.
    - Use descriptive statistics and visualizations to summarize the data.
4. **Data Cleaning and Pre-processing** (2 hours)

    - Check and handle missing values (NAs).
    - Identify and treat outliers in the dataset.
    - Convert categorical variables into a format suitable for modeling (e.g., one-hot encoding).
    - Address class imbalances if applicable, especially if the target variable has imbalanced classes.
    - Scale numeric data to ensure that features have a similar range of values.

5. **Feature Selection** (1.5 hours)

    - Determine which variables (features) to include in the model based on their potential relationship with the target variable "wage".
    - Use techniques like correlation analysis and domain knowledge to select relevant features.

Step 2: Model Selection & Training

1. **Model Consideration** (1 hour)

    - This is a regression task (given the continuous nature of the "wage" variable). (not a classification task!)
    - Review potential models (Logistic Regression, Decision Trees, **Random Forest, Gradient Boosting, XGBoost**) and select based on the nature of the data and the expected relationship between Y and X.

2. **Model Training** (2 hours)

    - Split the data into training and testing sets.

- Train selected models on the training set.
- Perform hyperparameter tuning to improve model performance.

Step 3: Testing Performance

1. **Model Evaluation** (1.5 hours)

   - Evaluate model performance using appropriate metrics (RMSE for regression models).
   - Utilize confusion matrix, ROC curve, and AUC for an in-depth performance understanding if any classification models are involved.

2. **Model Comparison and Finalization** (1 hour)

   - Compare the performance of different models.
   - Select the best-performing model based on evaluation metrics.

3. **Model Explainability** (1 hour)

   - Apply model explainability techniques (e.g., feature importance) to interpret the results and understand the main drivers behind "wage" predictions.

4. **Prediction and Interpretation** (1 hour)

   - Use the final model to predict wages for team members as specified.
   - Interpret these predictions in the context of the model's feature importance.

Collaborative Aspects and Finalization

1. **Team Collaboration** (ongoing)

   - Utilize Git and R for version control and collaboration.
   - Ensure all team members are involved in the project stages, as recommended.

2. **Preparation for Presentation** (2 hours)

   - Prepare slides and a presentation to summarize the findings, methodology, model choice, and interpretations.
   - Practice the presentation focusing on argumentation consistency, data visualizations, and the ability to respond to questions.

Total Estimated Time: 18.5 hours

Important:

- Coaching: 23th of April & 14th of May!
- Use your model to predict the wage that each one from the team will earn in the future.