Final Project Instructions

Objective:

The aim of this project is to apply data science techniques to a biological dataset. You will:

1. Select a dataset.
2. Formulate a data science problem.
3. Perform exploratory data analysis (EDA).
4. Apply and evaluate data science algorithms.
5. Present your findings.

Steps:

1. **Select a Dataset**:
   - Choose a dataset you are familiar with, or find a suitable biological dataset on the UCI Machine Learning Repository or Kaggle or any other site (e.g., datasets on genomics, proteomics, medical imaging, ecological data).
   - Ensure the dataset is sufficiently complex to allow for meaningful analysis.
2. **Formulate a Data Science Problem**:
   - Define a clear problem statement:
     - Classification: Predicting disease presence from gene expression data.
     - Regression: Predicting plant growth based on environmental conditions.
3. **Exploratory Data Analysis (EDA)**:
   - Perform data cleaning: handle missing values, correct data types, and remove duplicates.
   - Summarize the dataset: provide descriptive statistics and visualizations.
   - Identify patterns, correlations, and insights using plots (e.g., histograms, box plots, scatter plots).
   - Highlight any preprocessing steps required (e.g., normalization, encoding categorical variables).
   - Perform statistical tests such as ANOVA, correlation tests, and t-tests using the statsmodels package.
4. **Apply Data Science Algorithms**:
   - Split the data into training and testing sets.
   - Choose at least three different algorithms to apply to your problem. Examples:
     - Regularized linear model, logistic regression, decision trees, random forests.
   - Implement the models using Python libraries such as scikit-learn.
   - Perform hyperparameter tuning to optimize the models (using cross-validation techniques).

5. **Evaluate Model Performance**:
    - Use appropriate metrics to evaluate the performance of your models (e.g., accuracy, precision, recall, F1-score, RMSE, R^2).
    - Compare the results of different models using these metrics.
    - Use visualizations to present the performance of the models (e.g., ROC curves, confusion matrices).
6. **Report and Presentation**:
    - Prepare a comprehensive Jupyter notebook detailing:
        - **Introduction**: Dataset description and problem statement.
        - **Methodology**: EDA, preprocessing, and model implementation.
        - **Results**: Performance metrics and visualizations.
        - **Discussion**: Interpretation of results, challenges faced, and potential improvements.
    - Prepare a presentation summarizing your project. Highlight key findings and insights, as well as challenges faced, such as bugs and performance issues.
    - In addition to your Jupyter notebook report, you are required to create a video presentation summarizing your findings. This video should be based on your notebook and include the key points of your project.
        - The presentation should be **no more than 5 minutes** for those working alone.
        - For pairs **no more than 7 minutes**.
        - For triplets **no more than 9 minutes**.
    - We will stop the video **exactly** after the allowed time.

Submission Guidelines:

- **Report**: Submit a well-documented Jupyter notebook to Moodle in the *final project section*.
- **Presentation**: Record your presentation and submit it through Moodle (or upload to a video service, such as Youtube, and put the link in Moodle) in the *final project section*.
- **Deadline**: 31.9.2024

Resources:

- **Datasets**:
    - Kaggle
    - UCI Machine Learning Repository
    - Other datasets you found
- **Python Libraries**:
    - Pandas, NumPy for data manipulation.
    - Matplotlib, Seaborn for data visualization.
    - Scikit-learn for machine learning models.
    - Statsmodels for statistical analysis.
- **Guides and Tutorials**:
    - Scikit-learn Documentation
    - Statsmodels Documentation