ML Project Description: Predicting Movie Revenue

In this project, you will be tasked with predicting the revenue of a movie based on various features such as numerical values (e.g., budget, runtime) and categorical values (e.g., genre, director). The goal is to develop a machine learning model that can accurately predict the revenue of a movie given its attributes. The project will be evaluated based on a report and the code produced by the students, this can also be presented through a well-structured, commented, and unique python notebook. The project can be done by group of two or alone. It will be based on the evaluation grid bellow that can be used to structure your work. You might not need to apply every subpoint in the list based on your context, but the main steps must be present. **Through the entire keep a critical approach of what you do and discuss your results, don't just take a descriptive approach.** The deadline of the submission on claco is on the **5th of May**.

Evaluation Grid:

Data Understanding and analysis (20%) :

- Understand the structure of the dataset and explore its features.
- Perform exploratory data analysis (EDA) to gain insights into the distribution and relationships among features.
- Visualize key statistics and trends in the data.
- Identify any patterns or correlations between features and the target.
- Analyze the importance of different features in predicting the target (based on logic or on the dataset correlation/mutual information).
- Determine which features are likely to have the most significant impact on revenue prediction.

Feature Selection and Creation (20%) :

- Select relevant features based on their importance and relevance to the prediction task.
- Utilize techniques such as correlation analysis, feature importance ranking, and domain knowledge to guide feature selection.
- Transform existing features or create new ones based on domain expertise or insights from data analysis.

Data Pre-processing (20%) :

- Handle missing values and outliers appropriately.

- Encode categorical variables and handle any categorical data imbalances.
- Scale numerical features to ensure they have similar ranges.
- Split the dataset into training, validation, and test sets.

Model Comparison and Hyperparameter selection (20%) :

- Train and evaluate different machine learning models such as linear regression, decision trees, random forests, ...
- Compare the performance of each model using appropriate evaluation metrics (e.g., mean absolute error, root mean squared error, R-squared).
- Utilize cross-validation to find the optimal combination of hyperparameters that maximize model performance.
- Analyze the possible underfitting or overfitting of your models.

Model Evaluation and Interpretation (10%) :

- Evaluate the final model on the test set to assess its generalization performance.
- Interpret the model's predictions and understand the factors driving its decisions.
- Discuss the limitations and potential biases of the model.

Presentation of Results (10%):

- The report is comprehensive, structured and the graphs are well presented.
- The scientific methodology is followed and results are critically discussed.

Bonus Evaluation Points (up to 15%) :

- **Innovation:** You explore novel approaches or techniques to improve model performance or address specific challenges of this project.
- **External data source:** Import data from external sources to obtain new meaningful features and merge it in the dataset.
- **Model:** Use unusual models or models that were not seen in the machine learning part of the course, e.g. deep learning models and neural networks.