# Final project

Release: 29 December 2024

## Project Overview

### Team Requirements

- Team Size: 3 students (no less than two different nationalities).
- Condition: The project will be submitted 5 working days before the exam date. Each member can book and take the exam in different dates, but the project cannot be modified after the first member of the team sits at the exam.
- Deadline(s): Five (5) working days before the planned exam date. No exceptions, no supplications.

## Project Activities

### 1. Understanding the Dataset

**Activity**: Dataset Exploration

- **Description**: Students will load the dataset, understand its structure, and perform initial exploratory data analysis (EDA).
- **Deliverables**: D1.1) Section 1 of the final report (titled "Dataset exploration") detailing the dataset's features, their distributions, and any initial observations.

**Guidelines**:

- Describe each feature and its possible values.
- Visualize distributions of numerical features (e.g., histograms, box plots).
- Analyze relationships between features using scatter plots and correlation matrices.

### 2. Data Preprocessing

**Activity**: Data Cleaning and Preparation

- **Description**: Handle missing values, detect and treat outliers, and encode categorical variables.
- **Deliverables**: D2.1) Section 2 of the final report (titled "Data Preprocessing") detailing the preprocessing steps; D2.2) A cleaned dataset.

**Guidelines**:

- Choose appropriate methods to handle missing values (e.g., imputation, removal).
- Identify and treat outliers using techniques such as clipping or transformation.
- Encode categorical variables using one-hot encoding or label encoding.
- Normalize or scale numerical features if necessary.

## 3. Exploratory Data Analysis (EDA)

**Activity**: In-depth EDA

- **Description**: Perform detailed EDA to understand key factors.
- **Deliverables**: D3.1) Section 3 of the final report (titled "Exploratory Data Analysis"), a comprehensive EDA report with visualizations and insights.

**Guidelines**:

- Investigate the relationship between features and the target variable.
- Use visualization techniques like bar charts, box plots, and heatmaps.
- Perform hypothesis testing where applicable (e.g., t-tests, chi-square tests).

## 4. Feature Engineering

**Activity**: Create New Features

- **Description**: Engineer new features that could improve model performance.
- **Deliverables**: D4.1) Section 4 of the final report (title "Feature Engineering"), explaining their creation and potential impact; D4.2) A dataset with new features.

**Guidelines**:

- Combine existing features to create new ones (e.g., interaction terms, ratios).
- Use domain knowledge to add relevant features (e.g., total service usage).
- Justify the inclusion of each new feature with potential benefits.

## 5. Modeling

**Activity**: Build and Evaluate Models

- **Description**: Train different machine learning models. No less than 3 ensemble learning models must be trained.
- **Deliverables**: D5.1) Section 5 of the final report comparing model performances; D5.2) Trained models.

**Guidelines**:

- Split the data into training and testing sets.
- Train several models (e.g., logistic regression, decision trees, random forest, gradient boosting, ensemble learning models).
- Evaluate models using metrics like accuracy, precision, recall, F1-score, and ROC-AUC.
- Perform cross-validation to ensure robust performance estimates.

## 6. Model Tuning

**Activity**: Hyperparameter Tuning

- **Description**: Optimize model hyperparameters to improve performance.
- **Deliverables**: D6.1) Section 6 of the final report (titled "Hyperparameter tuning") on the tuning process and results; D6.2) Best-tuned models.

**Guidelines**:

- Use techniques like grid search or randomized search for hyperparameter tuning.
- Compare the performance of tuned models against default models.
- Document the tuning process and the chosen hyperparameters.

### 7. Model Interpretation

**Activity**: Interpret the Model

- **Description**: Analyze the models to understand which features are most important for predictions.
- **Deliverables**: D7.1) Section 7 of the final report (titled "Model interpretation") on feature importance and model interpretation.

**Guidelines**:

- Use feature importance scores from models like decision trees or random forests.
- Apply SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) for detailed model interpretation.
- Discuss how the insights align with domain knowledge and business logic.

### 8. Presentation

**Activity**: Present Findings

- **Description**: Prepare and deliver a presentation summarizing the project findings and recommendations. Each student has to present a different ensemble learning model in addition to (at least) one algorithm.
- **Deliverables**: D8.1) A 15-minute (20-25 slides max) team presentation in PDF format. Each team member has to discuss a different ensemble learning method

**Guidelines**:

- Summarize the key findings from EDA, modeling, and clustering.
- Highlight important insights and their implications for the business.
- Provide recommendations based on the analysis (e.g., strategies to reduce churn).

## List of deliverables

1. D0.0) The picture of the passports of the team members
2. D0.1) A complete and well documented Jupyter notebook (or Python script) with all the necessary steps to reproduce the results obtained by the team
3. D1.1) Section 1 of the final report (titled "Dataset exploration") detailing the dataset's features, their distributions, and any initial observations.
4. D2.1) Section 2 of the final report (titled "Data Preprocessing") detailing the preprocessing steps.
5. D2.2) A cleaned dataset.
6. D3.1) Section 3 of the final report (titled "Exploratory Data Analysis"), a comprehensive EDA report with visualizations and insights.
7. D4.1) Section 4 of the final report (titled "Feature Engineering"), explaining their creation and potential impact;
8. D4.2) A dataset with new features.
9. D5.1) Section 5 of the final report comparing model performances.
10. D5.2) Trained models.
11. D6.1) Section 6 of the final report (title "Hyperparameter tuning") on the tuning process and results.
12. D6.2) Best-tuned models.

13. D7.1) Section 7 of the final report (titled "Model interpretation") on feature importance and model interpretation.
14. D8.1) A 15-minute (20-25 slides max) team presentation in PDF format. Each team member has to discuss a different ensemble learning method.

# Evaluation Criteria

- **Understanding and Exploration**: Clarity and depth of dataset exploration and initial analysis.
- **Data Preprocessing**: Effectiveness of data cleaning, handling of missing values, and outlier treatment.
- **EDA and Feature Engineering**: Thoroughness of EDA and creativity in feature engineering.
- **Modeling and Evaluation**: Accuracy and robustness of models, choice of evaluation metrics.
- **Interpretation and Insights**: Depth of model interpretation and quality of business insights.
- **Presentation**: Clarity, organization, and professionalism of the presentation.