Programming Project

Prediction Model for Business Decision-Making

Profesor: Javier Paris

Machine Learning 4-ADE

Contents

1 Task Description

This project involves applying a machine learning algorithm to a dataset **chosen by the team** with the objective of extracting valuable insights for business decision-making. A comprehensive analysis will be conducted, from data exploration to model evaluation, with a focus on presenting the results persuasively, as if in a meeting with a client or potential investor.
Teams will consist of a maximum of 3 people.

2 Learning Objectives

1. **Understand** the workflow of a machine learning project.
2. **Learn** to use Python libraries for machine learning.
3. **Practice** programming in Python and using tools like scikit-learn.
4. **Develop** communication and presentation skills to effectively convey data analysis results to a client or investor.

3 General Instructions

- **Teams**: Maximum of 3 people.
- **Submission**: A self-contained Jupyter notebook that runs in Colab.
- **Defense**: A 10-30 minute presentation per team focusing on the utility and application of the results.

4 Development

4.1 Description

In this practice, each team must **implement a machine learning algorithm** to solve a real-world problem using scikit-learn. The analysis results should focus on their applicability in a business environment, and the presentation should be convincing for a client or investor.

4.2 Phases

The practice must include the following steps:

1. **Dataset Selection**
    - Choose a dataset relevant to the business domain.
    - Ensure online hosting so it is accessible via a URL.
2. **Dataset Study**
    - Perform data visualization and exploratory analysis.
3. **Dataset Preprocessing**
    - Data cleaning.
    - Normalization, categorization, etc.
4. **Algorithm Selection**
    - Choose a machine learning algorithm suitable for the problem.

5. **Model Training**
    - Split the dataset into training and testing sets.
    - Train the model and tune hyperparameters.
6. **Model Evaluation**
    - Evaluate the model with data not used in training.
    - Use appropriate metrics for the problem.
7. **Business-Oriented Presentation of Results**
    - Explain the model's impact on business decision-making.
    - Present findings clearly and persuasively.
    - Include visualizations to support conclusions.

4.3 Outcome

The optimal outcome of this practice will be a **Jupyter notebook (.ipynb)** that includes the dataset analysis, machine learning model application, and result evaluation. Additionally, each team can prepare a compelling presentation that effectively communicates the findings in a business setting.

4.4 Dataset Selection

4.4.1 Characteristics

The dataset for the practice is chosen by the team. It is recommended to use datasets with economic, financial, or business applications. The dataset must meet the following criteria:

- **Minimum of 300 records**.
- **Minimum of 5 attributes** (or 4 plus the target variable).
- At least one **numerical** attribute.
- At least one **categorical** attribute.

It is encouraged to use a **real** dataset with **relevant** applications. *For example: product prices, market values, population or environmental data, etc.* If the dataset is too complex, it can be reduced in terms of attributes and records (with justification).
Confirm dataset validity with the professor before starting the practice. If there are issues with *hosting* (how to download the dataset from the notebook), consult the professor.

4.4.2 Source

The dataset can be obtained from various sources:

- **Public dataset** from an online platform.
- **Private dataset** legally generated or collected.

Several online platforms provide public datasets:

- Kaggle
- USA Government
- UCI Machine Learning Repository
- GitHub repositories

4.4.3 Hosting

Since the notebook must be portable, datasets should be accessible online. Websites like *data.gov* allow direct dataset downloads. If using private datasets or datasets not directly accessible via Python, they can be uploaded to a personal *GitHub* or *Google Drive* account and accessed from the notebook.

5 Follow-Up Sessions

Throughout the course, follow-up sessions will be conducted to review practice progress, team decisions, and set milestones to ensure steady progress.
These sessions will be held outside class hours and do not count for attendance, but they may be relevant to the final grade and are beneficial in assessing the practice's development.

6 Submission and Defense

6.1 Notebook

The notebook must be **self-contained and portable**, executable from start to finish without errors. The allowed libraries are:

- *numpy*
- *pandas*
- *matplotlib*
- *seaborn*
- *scikit-learn*

6.2 Defense

For the practice to be graded, the team must present the project results to the professor in a 10-30 minute meeting. This will simulate a business presentation of results to a client or potential investor.
Therefore, it is important to justify decisions made and, most importantly, present results clearly and persuasively.
A written report or slides showing results and conclusions is optional but may support the presentation.

7 Grading

7.1 Evaluation Criteria

- Quality of analysis and rigor in algorithm implementation.
- Clarity and structure of the presentation.
- Ability to communicate results convincingly.

7.2 Extra Credit

Positive aspects include:

- Real-world applicability of the case study.
- Use of advanced model evaluation techniques.
- Well-designed visualizations and clear explanations.

7.3 Academic Integrity

Using AI tools to generate deliverables is not allowed. However, AI can be used as support for resolving doubts or researching information.

8 Example

A reference notebook will be provided, including necessary sections to guide project development.