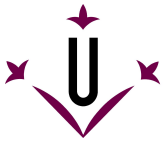
Intelligent Systems
Master's Degree in Informatics Engineering



Universitat de Lleida
Escola Politècnica Superior

Coordinator: Mariano Garralda

Academic Year: 2024-2025

Work Package 3 (WP3)

Open Kaggle Competition

Supervised Machine Learning Model Development and Deployment

Overview

In this work package, your task is to participate in the selected Kaggle competition [1]. You will navigate the entire advanced supervised machine learning pipeline, from problem understanding to model deployment. This process will provide hands-on experience with data science techniques, model development, and deploying solutions for real-world applications. Besides, you will gain valuable experience in presenting your work to a public audience. Finally, **the best performance compared with the other groups, will be recognized with an additional 10% bonus**.

Suggestion:
We strongly encourage you to share your intermediate results—such as metrics, evaluation methods, and validation techniques—on the WP3 forum to keep your colleagues updated on your progress.

Main Points

Your submission should address the following steps:

1. Problem Understanding & Objectives (5%)

- **Explain the Problem**: Summarize the competition problem and its objectives. Provide a brief description of the dataset and the expected outputs.
- **Justification**: Explain how the problem has been addressed, including the motivation for your approach and any specific challenges.

2. Exploratory Data Analysis (EDA) (10%)

- **Data Insights**: Perform EDA to explore the dataset, understand patterns, distributions, missing values, and relationships between variables.
- **Visualization**: Use appropriate visualizations to support your analysis.

3. Feature Engineering (5%)

- **Feature Creation**: Generate new features from the dataset that could improve model performance.
- **Data Processing**: Handle missing values, data normalization/standardization, and feature encoding if necessary.
- **Justify**: Explain your choices for the created or selected features.

[1] Regression with an Insurance Dataset

4. Ensemble Model Selection (15%)

- **Modeling**: Select an ensemble model (e.g., Random Forest, Gradient Boosting, XGBoost, etc.) and justify why you selected this model for the competition.
- **Model Explanation**: Provide a detailed explanation of how the chosen ensemble model works.
- **Tuning**: Highlight any hyperparameter tuning done to improve the model.

5. Model Evaluation (15%)

- **Techniques & Metrics**: Explain the techniques and metrics used to evaluate the model's performance (e.g., accuracy, precision, recall, F1-score).
- **Justify the Metrics**: Discuss why these metrics were appropriate for the given problem.
- **Cross-validation**: Implement cross-validation or other evaluation techniques to ensure model robustness.

6. Model Deployment (20%)

This section involves building and deploying the trained model in a real-world environment. Complete the following tasks:

6.1 Save the Trained Model

Save the trained model in a format that can be loaded for later use (e.g., `.pkl` or `.h5`).

6.2 Create an Inference API

- Use **FastAPI** to create an inference endpoint for the model. This endpoint (API-REST) should take input data and return the model's prediction.

6.3 Build a Web User Interface

- Use a suitable UI framework (e.g., **Streamlit** or **Gradio**) to build a web interface that allows users to upload new data for inference and display the model's output.

6.4 Docker Compose Setup

- **Deployment with Docker Compose**: Deploy the web UI (point 6.3) and API (point 6.2) together using Docker Compose.

Mandatory Oral Defense (25%) + Q&A (5%)

- In-person group defense of the WP3 solution.
- Prepare a clear and concise **20-minute presentation** to justify your solution.
  Note: Minimize reliance on reading directly from the slides and ensure you stay within the allocated time (practice thoroughly to ensure a smooth and confident presentation).
- Allocate approximately **10 minutes for Q&A** with the professor.
- Each team member must participate in the presentation.
- Evaluation will be based on the quality of the presentation, justifications, answers, and the overall technical understanding of your approach.
- The presentation should cover the following aspects:
  - Problem statement and objectives.
  - Methodology and key design decisions.
  - Model selection and evaluation.
  - Short live demonstration of the web interface and API.
  - Discussion of the challenges faced and how they were addressed.
  - Conclusions, future improvements and potential extensions.

Submission Guidelines (ZIP file)

- Include a **Jupyter Notebook** for items 1 to 5, containing all code, visualizations, and explanations.
  Note: Ensure that all cells execute correctly and produce the expected results.
- Provide a **Python script** for model deployment (item 6.1 to 6.3). Include the **Docker Compose setup**, with the `Dockerfile` for both the API and the web interface (item 6.4).
- Attach the oral defense **presentation slides** for the oral defense in PDF format.

Assessment Grading (breakdown summary)

- **Problem Understanding & Objectives**: 5%
- **Exploratory Data Analysis (EDA)**: 10%
- **Feature Engineering**: 5%

- **Model Selection & Explanation**: 15%
- **Model Evaluation**: 15%
- **Model Deployment**: 20%
- **Oral Defense + Q/A**: 30%
- **Extra Plus: Best outcomes compared with other groups:**: 10%

Deadline

- **Submission deadline**: 12th January 2025 at 23:55

- **Oral defense date**: 17th January 2025 at 17:30 (ETS 1.03)
  Note: Please arrive on time and ensure in advance that all necessary equipment (e.g., projector, display connections, etc.) is working properly.