Intelligent Systems
Master's Degree in Informatics Engineering

Universitat de Lleida
Escola Politècnica Superior

Coordinator: Mariano Garralda

Academic Year: 2024-2025

Work Package 1 (WP1)

Supervised Machine Learning Pipeline

EDA, Feature Engineering, and Model Development

Overview

In this work package, your task is to carry out a complete supervised machine learning project on a dataset of your choice. You will follow the typical machine learning pipeline, including problem understanding, EDA, preprocessing, feature engineering, model development, evaluation, and selection. This exercise will provide hands-on experience in building predictive models and understanding the intricacies of the machine learning workflow.

Main Points

Your submission should address the following steps:

1. Dataset Selection & Problem Definition (5%)

- **Dataset Selection**: Choose a dataset suitable for a supervised learning problem (classification or regression). Possible sources include Kaggle, UCI Machine Learning Repository, or any reputable open data source.
- **Domain Context**: Provide background information on the domain or industry the data represents.
- **Problem Statement**: Clearly define the problem you aim to solve. Specify the input features and the target variable.
- **Objectives**: Outline the objectives of your analysis and model development.

2. Exploratory Data Analysis (EDA) (15%)

- **Data Overview**: Describe the dataset structure, including the number of samples, features, and data types.
- **Univariate Analysis**:
  - Analyze the distribution of individual features and the target variable using statistical summaries and visualizations.
  - Discuss any observations or patterns found in the distributions.
- **Correlation Analysis**:
  - Compute correlation coefficients between features.
  - Use correlation matrices and heatmaps to visualize relationships.
  - Discuss any strong correlations or multicollinearity issues found.
- **Bivariate Analysis**:

- Examine relationships between features and the target variable using appropriate plots (e.g., scatter plots for regression, box plots for classification).
- Identify significant predictors.

- **Multivariate Analysis**:
  - Explore interactions among multiple features.
  - Use correlation matrices, heatmaps, or pair plots to identify multicollinearity.

- **Initial Findings**:
  - Summarize the key findings from your EDA.
  - Relate the findings back to your problem statement and objectives.

3. Data Preprocessing (10%)

- **Data Cleaning**:
  - Handle missing values based on insights from EDA.
  - Correct data inconsistencies and handle duplicates.

- **Outliers Handling**:
  - Use findings from EDA to identify outliers.
  - Decide on appropriate strategies to handle outliers (e.g., removal, transformation).

- **Data Transformation**:
  - Apply necessary transformations (e.g., scaling, normalization) to prepare data for modeling.
  - Justify why these transformations are appropriate.

4. Feature Engineering (15%)

- **Feature Creation**:
  - Create new features that could enhance model performance.
  - Explain the rationale behind each new feature.

- **Feature Encoding**:
  - Convert categorical variables into numerical formats using techniques like one-hot encoding, label encoding, etc.
  - Discuss any challenges faced during encoding.

- **Feature Selection**:
  - Use methods such as correlation analysis, variance thresholding, or feature importance to select relevant features.
  - Address any issues related to multicollinearity.

5. Model Development (20%)

- **Model Selection**:
  - Choose appropriate machine learning algorithms for your problem (e.g., linear regression, decision trees, SVM, etc.).
  - Justify your choice of algorithms.
- **Training and Validation**:
  - Split the data into training and validation sets using appropriate methods (e.g., train-test split, cross-validation).
  - Explain your strategy for model validation.
- **Model Training**:
  - Train your models using the prepared data.
  - Ensure reproducibility by setting random seeds where applicable.

6. Model Evaluation (20%)

- **Performance Metrics**:
  - Select appropriate metrics for evaluating your models (e.g., accuracy, precision, recall, F1-score for classification; RMSE, MAE for regression).
  - Justify why these metrics are suitable for your problem.
- **Evaluation Results**:
  - Present the performance of your models using the selected metrics.
  - Use visualizations like ROC curves, confusion matrices, or residual plots as appropriate.
- **Model Comparison**:
  - Compare the performance of different models.
  - Discuss which model performs best and why.
- **Cross-Validation**:
  - Implement cross-validation to assess the robustness of your model.
  - Report cross-validation scores and analyze the variance.

7. Conclusions & Recommendations (10%)

- **Summary**:
  - Summarize the overall process, key findings, and the performance of your final model.
  - Reflect on the effectiveness of your feature engineering and model selection.
- **Future Work**:
  - Suggest potential improvements or next steps for further enhancing the model.
  - Discuss any limitations faced during the project.

Assessment (Grading Breakdown)

- **Dataset Selection & Problem Definition**: 5%
- **Exploratory Data Analysis (EDA)**: 15%
- **Data Preprocessing**: 10%
- **Feature Engineering**: 10%
- **Model Development**: 25%
- **Model Evaluation**: 25%
- **Conclusions & Recommendations**: 10%

Submission Guidelines

- Submit a well-documented **Jupyter Notebook** containing your code, analysis, visualizations, and explanations.
  - Ensure that your notebook is organized with clear headings and follows a logical flow.
  - Include all code outputs and make sure the notebook runs from start to finish without errors.
  - Any external data files used should be accessible or instructions provided on how to obtain them.
- Submit a **report** containing your conclusions, recommendations, and any additional insights not covered in the notebook.
- The submission (zip file) should be made via the **Campus Virtual** platform.
- **Note**: You may be requested to provide clarifications or answer questions regarding your submitted work.

Deadline

- **Submission deadline**: 15th December 2024