Department of Computer Science
University of Cyprus

DSC510 Introduction to Data Science and Analytics

Fall Semester 2025

Team Project

Weight: 30% (of the total mark)

Description:
Develop a decision support system by applying the fundamental steps of a data science workflow: data acquisition, exploratory data analysis, data pre-processing, model development for either classification or regression task, and model evaluation and validation. The methods will be demonstrated using a dataset selected from the list of open datasets provided below.

Assessment Evaluation:

The project requirements are the following:

- Follow all the steps specified below for the chosen dataset using Python existing libraries like Pandas, Matplotlib, Seaborn, etc.
- Submit in Moodle, word/pdf/powerpoint documents and source code files requested for each deliverable.
- You can work in a team of up to 3 persons.

On the next page, you will find a non-exhaustive list of suggested datasets for your project. If you wish to propose your own dataset, please ensure that it includes multiple features (at least 10 columns) and sufficient observations (several hundred to a few thousand rows). The dataset should also contain a mix of data types, such as numerical values, categorical/text fields, dates. Non pre-processed datasets are especially welcome.

List of Open Datasets:

#	Dataset Name	Domain	Problem	Size (rows)	# Features	Data Types	Cleaned?	Link
1	Adult Income (UCI)	Demographics / Income Prediction	Binary classification (predict whether a person makes over 50K a year)	48K	14+	Numerical, Categorical, String	**Mostly**, needs minor cleanup	Link
2	Covertype (Kaggle)	Environmental / Forestry	Multi-class classification (predict what types of trees grow in an area based on the surrounding characteristics)	581K	54	Numerical, Categorical	**Yes**	Link
3	Ames Housing (Kaggle)	Real Estate / Pricing	Regression (predict house sale price)	2.9K	79	Numerical, Categorical, Text	**Partially**, missing values present	Link
4	Lending Club Loan Data (Kaggle)	Finance / Credit Risk	Binary classification (predict whether a loan will be accepted or rejected)	2M+	145+	Numerical, Categorical, Text	**No**, requires heavy cleaning (multi-file dataset!)	Link

6	Heart Disease Extended (Kaggle)	Healthcare	Binary classification (predict whether a patient suffers from heart disease)	1K	13	Numerical	**Yes**	Link
7	Wine Quality Extended (UCI)	Food Science	Multi-class classification (predict wine quality)	~6.5K	11	Numerical	**Yes**	Link
8	FIFA 15-21 Player Dataset (Kaggle)	Sports Analytics	Not determined e.g. Regression (predict player value)	18K	106	Numerical, Categorical, Text	**Mostly**, some text cleanup needed	Link
9	Online Retail (UCI)	E-commerce	Not determined e.g. Regression (unit price)	500K+	8	Numerical, Categorical, Text	**No**, requires deduplication/missing handling	Link
10	Home Credit Default Risk (Kaggle)	Finance / Risk Scoring	Binary classification (predict whether a customer can repay a loan)	300K+ (main table)	120+	Numerical, Categorical, Text-like	**No**, multiple joins and cleanup needed (multi-file dataset!)	Link

DELIVERABLES AND DEADLINES

See below the different milestones of the project along with deliverables, deadlines and mark breakdown.

Deliverable 1: Dataset selection and brief problem description

Deadline: 26/09/2025 @ 23:59 [15%] – end of Week 3

Form teams of 3 students, preferably coming from different backgrounds (1 student form Math, 1 student from CS, 1 student from Business) and submit a word/pdf file document (feel free to use any template) that includes the following:

1. Project Title and Team Members
2. Brief description of the problem – description of the domain and the specific problem, goal/objective of the project (2-3 paragraphs)
3. Dataset description – explain how, where, and from whom the dataset was collected (if this information is available), describe its features and identify the target variable (the quantity you need to predict)
4. 2-3 references should be given (hint: identify similar papers/reports published relevant to the dataset, the domain. You can find some papers cited in the dataset link, this applies especially for UCI datasets).

Deliverable 2: Data exploration and pre-processing

Deadline: 24/10/2025 @ 23:59 [35%] – end of Week 7

Submit a word/pdf document including:

1. Project Title and Team Members
2. **Data understanding**: Provide an overview of the dataset using tables, graphs, and short explanations. Include key details such as: dataset shape (#rows, #columns), Preview and statistical summary of features, number of classes (binary vs. multiclass classification) if classification problem, distribution of the target class if regression problem, identification of imbalanced data and missing values.
3. **Exploratory data analysis**: use various visualization techniques to explore the dataset, such as: histograms (feature distributions), bar plots,

count plots (categorical features), violin plots, box plots (feature variability), any other relevant techniques for deeper insights.

4. **Data pre-processing**: Describe the **data cleaning** steps taken (handling missing values, outliers, inconsistencies), Explain the **encoding methods** used for categorical variables (if applicable), Discuss **feature scaling** techniques (normalization, standardization), Mention any **feature selection** or **dimensionality reduction** methods applied, Outline strategies for handling **imbalanced data** (e.g., over-sampling, under-sampling, SMOTE), Justify the choices made for each preprocessing step.
5. **Pre-processed dataset versions**: Create different **versions** of the pre-processed datasets using different combinations of techniques (imputation, scaling, encoding, transformations) that will be used to train classification machine learning (ML) techniques.
6. **Selected machine learning techniques**: List at least **four regression/classification techniques** to be explored (e.g., K-Nearest Neighbours, Logistic Regression, Decision Trees, Random Forest, Boosting methods, Naïve Bayes, SVM), and provide a **brief explanation** (2-3 paragraphs) of each technique.
7. **References**: Cite relevant sources that support the selected pre-processing methods and machine learning techniques under investigation.

The source code (e.g. Jupyter Notebook file(s)) should be submitted separately and not included in document.

Deliverable 3: Predictive Model Development and Performance Evaluation

Deadline: 21/11/2025 @ 23:59 [35%] – end of Week 11

The goal of this deliverable is to develop and evaluate machine learning regression/classification models using the pre-processed datasets and compare findings with existing literature. Submit a word/pdf document (final technical report) including the following details:

1. Project Title and Team Members
2. Revisions from previous deliverable (if any)
   - Highlight any updates or modifications to the previous deliverable.
   - If no changes were made, proceed directly to the next step
3. **Model evaluation metrics**: Define the evaluation metrics to be used, such as Accuracy, Recall, Precision, (weighted) F1 Score. Justify the

choice of metrics based on dataset characteristics (e.g. imbalanced data) and project objectives.

4. **Initial model experimentation**: Conduct preliminary training and validation experiments using cross validation on the pre-processed dataset versions. Test at least four classification techniques (as identified in the previous deliverable) using their default hyperparameters to understand which dataset-model combinations yield the best results.
5. **Selection of best-performing models and pre-processed datasets**: Based on initial experimentation, identify the top 2-3 models and the most effective pre-processing techniques.
6. **Pipeline Definition**: Construct pipelines that integrate the best pre-processing techniques with the best-performing models for systematic experimentation.
7. **Hyper-parameter tuning**: Perform GridSearchCV to fine-tune hyperparameters for the selected models. Use a well-defined hyperparameter grid.
8. **Final model evaluation**: Present the final evaluation results using tables and visualizations (e.g., performance metrics, confusion matrices). Report the best-performing model with its optimal hyperparameter values and the best pre-processed dataset version.
9. **Discussion**:
    - **Key Findings**: Summarize the most significant observations from the results.
    - **Comparison with Literature**: Prepare a table comparing your results with previous studies using the same dataset. Compare key aspects such as selected features, algorithms used, and achieved performance metrics.
    - **Domain-Specific Validation**: Discuss the relevance and validity of the findings within the context of the dataset.
    - **Study Limitations**: Highlight any challenges or constraints (e.g., dataset size, computational limitations, class imbalance issues).

The source code (e.g. Jupyter Notebook file(s)) should be submitted separately and not included in document.

Deliverable 4: Presentation of the project

Deadline: 01/12/2025 @ 23:59 [15%] – before last week's lecture

Submit a powerpoint document (see guidelines below)

Presentation day/time/location: Within the last week of the semester during the last two lecture sessions. All team members are required to actively participate to the presentation session.

Presentation guidelines (15-minute presentation + 5 minutes Q/A)

Your powerpoint presentation should be clear, concise, and well-structured, covering the following key points:

1. **Project Title and Team Members**
2. **Introduction to the domain**
   - Provide a brief overview of the field relevant to your datasets.
   - Explain the significance of the problem being addressed.
3. **Dataset overview**
   - Describe the dataset, including its features, and target variable.
   - Highlight any key characteristics (e.g., dataset size, class distribution, class imbalance).
4. **Exploratory data analysis (EDA)**
   - Present key findings from the EDA using visuals (e.g., graphs, tables).
   - Mention any patterns, trends, or challenges identified in the data.
5. **Data pre-processing techniques**
   - Outline the pre-processing steps applied (e.g., handling missing values, feature scaling, encoding, dimensionality reduction).
   - Justify the importance of each technique in improving model performance.
6. **Regression/Classification techniques under study**
   - Briefly introduce the classification techniques chosen for experimentation.
7. **Results presentation**
   - Showcase the performance metrics (e.g., accuracy, precision, recall, F1-score) using tables and graphs.
   - Compare results across different models and pre-processing methods.
   - Include a comparison with related studies (if available) to contextualize your findings.
8. **Interpretation of results for experts in the field**
   - Explain the implications of the results in a way that experts and professionals can understand.

Submit your presentation to Moodle before the deadline.

Ensure your slides are well-structured, visually engaging, and free of excessive text.