Final Project: Python Project - Data Analytics: Machine Learning

How-To Guide

The Python Project represents 100% of the overall course grade.

Instructions:

Develop a Python project to analyse real-world datasets, generate valuable insights by applying data analytics and visualization techniques, applying machine learning models and evaluate the performance of the models. The project aims to demonstrate your understanding of the following learning outcomes:

1	Demonstrate a strong understanding of machine learning concepts, including supervised and unsupervised learning, ensemble methods, and neural networks.
2	Implement machine learning techniques using Python and the Scikit-Learn library, including data preprocessing, model selection, evaluation, and hyperparameter tuning
3	Create advanced visualisations of data using tools like Bokeh, Matplotlib and Seaborn to effectively communicate insights to stakeholders
4	Apply the CRISP-DM methodology to real-world datasets, including problem formulation, data preparation, model building, evaluation, and deployment
5	Effectively communicate results and insights to stakeholders through presentations, reports, and visualisations

You are required to select a real-world dataset from any open-source dataset available online. Each learning outcome corresponds to a milestone that needs to be achieved. Your project submission should include the following components:

1. Project Report

- Draft a report describing your project process, dataset, data sources, graphs, and insights
- Justify the use of each learning outcome concept, explaining the reasons for your choices
- The report should contain between 1,500 and 2,000 words
- Use the provided template (see Assessments section to download)
- Save as a PDF file

2. Project Files

- Include all the code and data files related to your Python project

- The project should cover all milestones corresponding to each learning outcome to gain full marks
- Compress all the code and data files (but not the PDF) into a ZIP file

3. Submit

- Go to “Submit Final Project”, upload both the PDF file and the ZIP file you prepared earlier

The goal of the project is to demonstrate your ability to apply the course concepts and learning outcomes in practice.

Milestones

The project should address the following milestones:

1. Data:

- Select and utilize a real-world dataset, providing a reference to the source in the report.
- Make sure to check the “How to Source Real-World Datasets for Data Analytics & Machine Learning Projects” file on the LMS for your datasets

2. Importing:

- Explain the method used to import the data into Python.
    - o Import data from a flat file (e.g., CSV, XLS, XLSX, TXT).
    - o Retrieve data using online SQL, APIs, or web scraping.

3. Data Preparation:

- Discuss the steps taken to prepare the data for analysis.
- Explain the creation of pandas DataFrames.
- Describe the sorting, indexing, filtering, and grouping operations performed on the data.
- Explain how duplicate entries and missing values were handled.
- Discuss the definition of custom functions for reusable code.
- Provide details on how multiple DataFrames were merged, if applicable.

4. Data Visualization:

- Generate at least FOUR charts using Matplotlib library.
- Generate at least FOUR charts using Seaborn library.
- Conduct univariate and bivariate analysis using appropriate charts and techniques.

5. Machine Learning:

- Predict a target variable with **Supervised** or **Unsupervised** algorithm.
- You are free to choose any algorithm.
- Perform Model Evaluation using metrics suitable for the choice of ML model/s.
- Perform **hyper parameter tuning** or **boosting**, whichever is relevant to your model. If it is not relevant, justify that in your report and Python comments.

6. Insights:

- Derive at least EIGHT valuable insights from your data analysis.
- Justify your insights with reference to the charts or analysis performed.

Additional Guidance:

Properly reference any quotes or external sources using the Harvard Referencing Style.

https://libguides.ucd.ie/harvardstyle

The word limit for the Project Report allows for a 10% tolerance above or below the stated range. The referencing does not count towards the assessment length limits.

Assessment Criteria:

Your project will be assessed based on the following criteria:

Data Selection and Importing: **Selects and imports a real-world dataset.**	**Data Preparation and Manipulation:** **Prepares and manipulates data effectively using Pandas DataFrames. Handle sorting, filtering, grouping, and missing values appropriately**	**Data Visualization and Analysis:** **Generates interactive charts using Bokeh, Matplotlib and Seaborn libraries. Conducts analysis to identify patterns and relationships.**	**Insights:** **Derives valuable insights from data analysis.** **Justifies insights with reference to analysis.**	**Machine Learning:** **Choice and justification of the algorithm for predicting the target variable, the quality of data preprocessing, successful implementation of the chosen algorithm, effective application of hyperparameter tuning or boosting techniques if relevant, the model's performance on the test dataset**
Distinction Criteria				
Demonstrates a comprehensive understanding of data selection and importing, effectively selecting, and importing real-world datasets.	Displays advanced skills in data preparation and manipulation using Pandas DataFrames, effectively handling sorting, filtering, grouping, and missing values.	Exhibits exceptional proficiency in data visualization and analysis, generating visually appealing and informative charts using Matplotlib and Seaborn libraries.	Derives highly valuable insights from data analysis, demonstrating deep analytical thinking and providing clear justifications for the insights.	Shows a sophisticated understanding of machine learning concepts, algorithm implementation, model evaluation and hyperparameter tuning techniques.
Merit Criteria				
Shows a strong understanding of data selection and importing, successfully selecting, and importing real-world datasets.	Demonstrates effective data preparation and manipulation skills using Pandas DataFrames, appropriately handling sorting, filtering, grouping, and missing values.	Displays good skills in data visualization and analysis, generating visually informative charts using Matplotlib and Seaborn libraries.	Derives valuable insights from data analysis, providing sound justifications for the insights.	Exhibits a good understanding of machine learning concepts, algorithm implementation, model evaluation and hyperparameter tuning techniques.
Pass Criteria				
Demonstrates a satisfactory understanding of data selection and importing, adequately selecting, and importing real-world datasets.	Shows satisfactory skills in data preparation and manipulation using Pandas DataFrames, adequately handling sorting, filtering, grouping, and missing values.	Generates satisfactory charts for data visualization and analysis using Matplotlib and Seaborn libraries.	Derives insights from data analysis, providing reasonable justifications for the insights.	Displays an adequate understanding of machine learning concepts, algorithm implementation, model evaluation and hyperparameter tuning techniques.
Unsatisfactory Criteria				
Displays limited understanding of data selection and importing, with	Demonstrates limited skills in data preparation and manipulation using	Generates insufficient or inaccurate charts for data visualization and	Derives minimal insights from data analysis, with weak or unsupported	Shows limited understanding of machine learning concepts, algorithm implementation,

incomplete or inadequate selection and importing of real-world datasets.	Pandas DataFrames, with deficiencies in handling sorting, filtering, grouping, and missing values.	analysis using Matplotlib and Seaborn libraries.	justifications for the insights.	model evaluation and hyperparameter tuning techniques.
Clear Fail Criteria				
Demonstrates a lack of understanding or knowledge in data selection and importing, with no or inadequate selection and importing of real-world datasets.	Displays limited or incorrect skills in data preparation and manipulation using Pandas DataFrames, with significant deficiencies in handling sorting, filtering, grouping, and missing values.	Fails to generate appropriate charts for data visualization and analysis using Matplotlib and Seaborn libraries.	Derives minimal or irrelevant insights from data analysis, with weak or unsupported justifications for the insights.	Shows a lack of understanding in machine learning concepts, algorithm implementation, model evaluation and hyperparameter tuning techniques accurately describing.
No Attempt Criteria				
No Submission	No Submission	No Submission	No Submission	No Submission