Ciências
ULisboa

Cátia Pesquita

Foundations of Data Science

2026/2025

Project

The goal of this project is to design, implement, and critically evaluate a data science project, end-to-end. You should prepare a Jupyter Notebook with all code and evaluations, an 8-page research paper, and a presentation. Some steps can be performed manually (M) and others automatically (A) by developing your own Python scripts and using existing libraries.

1. Define a question that you want to answer (they can be vague for now and can be inspired by the following step)
2. Select two or more datasets to analyse related to the question
    a. You can search online or use one of the following sources:
        i. Portal Nacional de Dados Abertos: www.dados.gov.pt
        ii. Lisboa Aberta: dados.cm-lisboa.pt
        iii. Amazon AWS: http://aws.amazon.com/datasets
        iv. Kaggle: https://www.kaggle.com/datasets
        v. Google:https://datasetsearch.research.google.com/
        vi. EU Open Data Portal: data.europa.eu/euodp
        vii. US Government's Open Data: data.gov
        viii. United Nations Data: data.un.org
        ix. OECD Data: data.oecd.org
        x. Open Data Network: opendatanetwork.com
        xi. World Bank Data Catalog: datacatalog.worldbank.org
    b. Select structured data (do not select text corpora - let's focus on tabular data as much as possible!)
3. Select a clear analytical goal based on the datasets. For example:
    Predictive: Predict a numeric outcome (e.g., movie box office revenue, actor popularity score, etc.)
    Descriptive / Exploratory: Identify key factors affecting an outcome (e.g., which genres correlate with high ratings).
    Classification: Predict a categorical outcome (e.g., success/failure, hit/flop, award/non-award)
4. Discuss ethical and regulatory issues that your study may be subject to (data privacy and consent, potential biases in data or models, transparency and explainability, fairness and accountability implications)
5. Describe the original data set selected (M+A). Use
    a. Describe the dataset: size, number of attributes, type, etc
    b. Identify the most important information in each data source
    c. Identify missing or incomplete data, and identify possible strategies to solve these issues
    d. Identify possible problems regarding data quality

   e. Define for each attribute in the original data source, if any operation is needed (e.g, concatenation, extraction of a portion, etc.)

6. Develop and implement a strategy for data cleaning that addresses missing data and entity duplicates:
   a. Apply a strategy to solve missing data
   b. Implement one or more strategies for entity similarity to detect and merge, e.g., use string similarity to detect potential duplicates, define rules to solve duplicates
7. Define an integrated data model (M)
   a. ER diagram, relational model, graph, etc.
   b. Define for each attribute in original data sources, if any operation is need (e.g, concatenation, extraction of portion, etc)
8. Develop and implement a strategy that, based on the data models defined, is able to: (A)
   a. extract the data from the dataset according to the data models
   b. use the data and the defined model to produce a single integrated dataset.
9. Design and apply a machine learning methodology to analyze your data using both interpretable and black box models.
   a. Justify model selection in each case.
   b. Consider feature selection, train/test split if applying supervised learning,
   c. Select appropriate evaluation metrics
10. Provide explanations for the selected model:
    a. Use SHAP or LIME to extract explanations from the black-box model
    b. Compare interpretable models with black box models
11. Write a report in the style of a research paper
    a. Use the following sections:
       i. Abstract
       ii. Introduction
       iii. Related Work
       iv. Data
       v. Methods
       vi. Results and Discussion
       vii. Conclusions
    b. Use tables/charts to report any results that can fit a table/chart
    c. Include excerpts of the original data, extracted data and integrated data as annexes (not counting towards page limit)
    d. Discuss choices and decisions
    e. Describe the open source tools and libraries employed in your project, describing briefly how they were used.
    f. Include an estimation of hours each student contributed to the project as an annex.
    g. Use the ACM Conference Proceedings Primary Article Template (8 page limit excl. references and annexes): https://www.overleaf.com/latex/templates/acm-conference-proceedings-primary-article-template/wbvnghjbzwpc