Project #3 Guidelines
End to End Data Analytics Project

Find below the general guidelines. Please check in with your instructor to confirm their specific requirements (if they have any). Good luck!

Introduction

The goal of this project is for you to apply your data science or data analytics skills by completing an end-to-end analysis on a real-world dataset.

The project will also assist you in demonstrating your proficiency in the skills required for your desired career path and developing your portfolio to showcase your abilities to potential recruiters.

For this project, you have the option to work either individually or in pairs, depending on your interests and goals. If you and a partner share a common interest and have ambitious plans for what you want to achieve, working in pairs is a great way to collaborate and tackle more complex challenges. However, please note that if you choose to work in pairs, we expect a higher level of output and quality.

If you're struggling to come up with a topic, we have provided a list of datasets for you to consider. However, we highly recommend that you explore and select a topic and dataset that personally interests you, as this will make the project more engaging and rewarding.

You may select one of two case studies, each designed to focus on specific skills and cater to your interests and career aspirations. This will enable you to tailor the project to your individual needs and provide a valuable learning opportunity that aligns with your long-term goals.

Case Study 1 - Data Science End to End Project

The goal is to build, on top of a business case, a predictive model in Python.

This project will require you to select a business case of your choice and iterate through the whole data science process, by doing data collection, data cleaning and wrangling, exploratory data analysis, feature engineering, preprocessing, model selection, evaluation and data visualization. The project should be structured as a complete pipeline that includes every step of the process.

Case Study 2 - Data Analytics End to End Project

The goal is to do data analysis and build a dashboard for decision making, using mainly SQL, some Python, and a visualization tool such as Tableau or Power BI.

This project will require you to select a business case of your choice and apply the full data analysis process to it, from data collection, data cleaning and wrangling, feature engineering, statistical analysis, to data visualization and dashboards. The project should be structured as a complete pipeline that includes every step of the process.

Prerequisites

In order to successfully complete the upcoming project, you should possess a strong understanding of several key concepts, including Python programming, data wrangling, and exploratory data analysis (EDA), as required for Project 1 and Project 2. Additional to these, the following are essential prerequisites that you should have before beginning this final project:

Case Study 1 - Data Science End to End Project

- Machine Learning Key Concepts: these concepts include supervised and unsupervised learning, classification and regression, bias-variance tradeoff, train-test-cross-validation, feature engineering (including encoding, scaling, and selection), model evaluation, dealing with unbalanced datasets, and hyperparameter tuning.
- Machine Learning Supervised Models such as Linear Regression, Logistic Regression, Decision Trees, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Naive Bayes, and Ensemble Models.
- Unsupervised Models, such as KMeans,Hierarchical Clustering and PCA.
- Strong understanding of the most appropriate metrics and preprocessing techniques based on the specific problem and context.
- Experience with model enhancement iteration.

Nice to have: same as for Project 2. Additionally:

- Time Series
- NLP
- Deep Learning

Case Study 2 - Data Analytics End to End Project

- SQL
- Visualization tools such as Tableau or Power BI

Nice to have: same as for Project 2.

Suggested ways to get started

To get started on your project, it's important to plan and organize your tasks effectively. One way to do this is by creating a Kanban or Trello Board to track your progress and stay on top of your objectives. You may also choose to use data from previous projects.

Once you have selected a business problem, you will need to locate and gather the necessary data. It's important to explore the data and gain an understanding of what each field represents. Use statistical techniques, including data visualization, to examine the relationships between features in the dataset. Based on your observations, make informed guesses about which features should be investigated further.

Next, perform data cleaning and manipulation. This involves handling outliers, missing values, type casting, feature selection, and converting categorical data to numerical, among other techniques. Use appropriate statistical methods to analyze the data.

If you are using machine learning techniques, perform it on the objective variable data that you want to predict, classify, cluster, etc. Remember that this is an iterative process, so experiment with different models and hyperparameters to improve your model's performance. Select the simplest model that produces the best results, and be sure to clearly define the metric you are using to define "best".

Present your findings in statistical summary and data visualizations. If you are not using machine learning, remember to create a database to store your data, use SQL for analysis, and include a dashboard for decision-making purposes.

Deliverables

You must submit the following deliverables in order for the project to be deemed complete:

- A new repo on your github account.
    - A **working code** that **meets all technical requirements**, **built by you**.
        - Jupyter Notebook, Python, SQL files, Tableau/PowerBI report, or any additional needed files for your work
    - A **README with the completed project documentation**.
    - The URL of the **slides for your project presentation**.
- **Presentation:** when presenting your work, there are many important factors to consider, such as the content of your presentation and the way you deliver it. The following link offers valuable advice on how to make a strong presentation: Presentations.
- Paste your own repo's link in the Student Portal Project Activity

Rubrics

In order to assess your project and ensure all requirements are met, a **rubric** will be used. This rubric is used to **evaluate your project** by your teaching staff but also to **communicate** what constitutes incomplete, acceptable and excellent performance across each of the learning outcomes for the project. Take some time to review the rubric here and ask your lead teacher or TA any questions about it if necessary.

Optional Advanced Features

While completing the basic requirements of your project is a great start, taking advantage of some advanced features can really take your work to the next level. Here are some options to consider if you want to go above and beyond:

- Same as for Project 2
- Use sophisticated methods for data cleaning such as handling missing or noisy data with advanced techniques such as Machine Learning algorithms.
- Use advanced techniques for handling unbalanced data (if needed)
- Considering time series analysis if applicable
- If using Machine Learning:
    - Implement more complex and sophisticated models, such as neural networks, nlp, advanced time series techniques.
    - Use of stacking techniques to combine multiple models to improve the predictions.
    - Use of advanced feature engineering techniques such as dimensionality reduction with principal component analysis (PCA), and advanced feature selection methods.
    - Deployment of the model to a web application or cloud service.