Final Project | with Python

Guidelines

Choose a project that interests and excites you. It is intended as a chance for you to get more practice with Python and to explore its more advanced tools. You may consider working on a problem related to your own research and use your own data. However, you should **focus more on coding rather than on answering a research question**. Through the project you should demonstrate the skills you have learnt by taking this class, but **you are encouraged to implement material not taught in class**.

You are responsible for formulating your own project. However, you should consult with me on the scope of your planned work. Below, I have included a list of topics you may want to consider when planning your project.

The goal of the project is to implement all knowledge acquired in class! You are toke this class because you want to learn about Python, and this is your opportunity to it in any way you like. As a rough estimate, the project should take you around 20 hours of work.

You are welcome to choose a dataset not listed, or data collected as part of a research project, but keep in mind that you may not submit anything twice: any work you do as part of this course may not be submitted for credit in another course and vice versa. If choosing a dataset not listed, make sure it is well-documented, legitimate, and complex enough to support your analysis efforts. Your work should be original; your project should not be a reproduction of published analyses

Project Ideas

- **Oceanographic data** which includes temperature, salinity, oxygen, chlorophyll a, and nutrients measurements are made available through the OCEAN web applications.((http://www.ices.dk/marine-data/data-portals/Pages/default.aspx)
- Continuous Plankton Recorder Dataset (https://www.gbif.org/dataset/67c54f85-7910-4cbf-8de4-6f0b136a0e34)
- Marine predator and prey body sizes (http://www.esapubs.org/archive/ecol/E089/051/default.htm#data)
- Marine/Ocean Data (https://www.ncdc.noaa.gov/data-access/marineocean-data)
- Climate Change: Global Temperature (https://www.climate.gov/news-features/understanding-climate/climate-change-global-temperature)
- Current Observations https://www.ncdc.noaa.gov/crn/current-observations
- Sea Surface Temperature https://www.climate.gov/maps-data/dataset/sea-surface-temperature-map-viewer
- Severe Weather Data https://www.ncdc.noaa.gov/data-access/severe-weather
- Global Mean Sea Level https://www.climate.gov/maps-data/dataset/global-mean-sea-level-graph

Where to Find Datasets

- The best data set is one that you are passionate about. I recommend that you start by finding a question you want to answer and then finding data to answer that question, rather than starting with a data set. That said, here are some helpful websites with large collections of data.

  - Google Data Set Search (https://datasetsearch.research.google.com/)
  - Reddit Datasets (https://www.reddit.com/r/datasets/)
  - U.S. Government's Open Data (https://data.gov/)
  - Kaggle Datasets (https://www.kaggle.com/datasets)
  - List of JSON APIs (https://github.com/toddmotto/public-apis)

Projeto de Análise de Dados e Machine Learning — Requisitos e Tarefas

1. **Seleção e Análise de um Conjunto de Dados Principal**
   a. Escolhe um dataset relevante e de alta qualidade para o problema que desejas resolver.
   b. Justifica a escolha do dataset com base no contexto do problema.
   c. Apresenta o problema que pretendes resolver com clareza.
2. **Exploração e Pré-processamento dos Dados (EDA)**
   a. Realiza uma análise exploratória detalhada (distribuições, valores ausentes, correlações, outliers).
   b. Aplica técnicas de limpeza de dados: tratamento de valores ausentes, codificação de variáveis categóricas, normalização/padronização, etc.
   c. Gera estatísticas descritivas (média, mediana, desvio padrão, etc.) por variável.
   d. Cria gráficos interativos com bibliotecas como Plotly, Seaborn ou Altair.
3. **Visualizações Avançadas**
   a. Cria um conjunto impactante de visualizações que comuniquem claramente as características do dataset.
   b. Utiliza diferentes tipos de gráficos (barras, linha, área, dispersão, boxplot, heatmap, etc.).
   c. Incorpora dashboards interativos (por exemplo, com Dash ou Streamlit).
4. **Análise Estatística Profunda**
   a. Aplica testes estatísticos relevantes (ANOVA, t-test, chi-square, correlação de Pearson/Spearman).
   b. Comenta os resultados dos testes com interpretações práticas para o problema.
5. **Aplicação de Algoritmos de Machine Learning**
   a. Escolhe e aplica pelo menos dois algoritmos de machine learning supervisionados e um não supervisionado.
   b. Avalia os modelos com métricas apropriadas (accuracy, precision, recall, F1, ROC AUC, silhouette score, etc.).
   c. Apresenta os hiperparâmetros escolhidos e o processo de tuning (ex: GridSearchCV).
   d. Aplica feature engineering para melhorar os modelos.
6. **Aplicação em Três Datasets Diferentes**
   a. Repetir a análise e visualização em três conjuntos de dados distintos e contrastantes (ex: saúde, economia, meio ambiente).
   b. Apresentar comparações entre os resultados das abordagens aplicadas a cada dataset.
   c. Explorar diferentes desafios que cada tipo de dado impõe.
7. **Utilização de Todas as Bibliotecas Estudadas**
   Inclui e documenta o uso das bibliotecas: pandas, numpy, matplotlib, seaborn, plotly, scikit-learn, statsmodels, tensorflow/keras ou pytorch (se aplicável), xgboost, streamlit/dash, etc.

1. - The progress report should include the following items.
     a. **Introduction** - Brief overview of your problem.

b. **Literature Review** - Description of other work/papers you've found that are related to your task. Just mentioning a paper is not sufficient; you should at least go into brief detail about what kind of approach they are using/how it relates to your work if it's not immediately clear. When looking at relevant literature examine if there have there been other attempts to build a similar system. Compare and contrast your approach with existing work, citing the relevant papers. The comparison should be more than just high-level descriptions. You should try to fit your work and other work into the same framework. Are the two approaches complementary, orthogonal, or contradictory?
c. **Dataset -** Description of data you are using - size of dataset, distribution of classes, any preprocessing you needed to do.
d. **Baseline** - Description and implementation of your baseline. Please provide a detailed description of your implemented baseline along with evaluation of the baseline using the metrics you define.
e. **Main approach** - Propose a model and an algorithm for tackling your task. You should describe the model and algorithm in detail and use a concrete example to demonstrate how the model and algorithm work. Don't describe methods in general; describe precisely how they apply to your problem (what are the inputs/outputs, variables, factors, states, etc.)?
f. **Evaluation Metric** - Please include what metrics, both qualitative and quantitative, you are using to evaluate the success of your problem. If relevant please include equations to describe your metrics.
g. **Results & Analysis** - At this point, you should have fully implemented your baseline and also have a basic working implementation of your main approach. Please include the performance of your baseline as well as the performance of your main approach so far. Include an analysis of your results, and how this might inform your next steps in fine-tuning your main approach.
h. **Future Work** - This is not mandatory, but it might be helpful for your mentor to get an idea of what next steps you plan to take after the milestone.
i. **References** - Please include a reference section with properly formatted citations. (Not included in 4 page limit).

Project submission

The project report and code should be submitted class room. The report should also be submitted by email. Make sure you submit your work by **20/06/2025**

Best!!!