DSA/CIS 330/593 -- Intro to Data Science II (Fall 2024)
Prof. Apostolos Kalatzis
Final Project
Due: Proposal, Wednesday October 9, 2024 (11:59 PM)

Final Project :Thursday, December 5, 2024 (11:59 PM)

This project will investigate doing exploratory data analysis and model building and evaluation in a dataset of your choice. The goal is to get you fluent in working with the standard tools and techniques of exploratory data analysis and model building and evaluation, by working with a data set where you have some basic sense of familiarity

Python Installation

Instead of installing python and other tools manually, we suggest installing **Anaconda**, which is a Python distribution with a package and environment manager. It simplifies a lot of common problems when installing tools for data science. More introduction can be found here. Installation instructions can be found here. If you are an expert of Python and data science, what you need to do is install some packages relevant to data science. Packages that I believe you will definitely use for this homework include:

- pandas
- scikit-learn
- numpy
- Matplotlib
- seaborn

Tasks (100 pts)

1. Submit a short proposal where you will identify a dataset of your choise from Kaggle or a dataset you are familiar with or have used in the past. Define a question or multiple questions you are trying to answer using this dataset. Provide a detailed description of the dataset and explain how you will use the dataset to answer your defined question/ questions. (Deadline October 9) (15 points)
2. Complete your analysis and submit a report on your findings. Include the steps you followed to answer your question/questions using this dataset. Your project should include exploratory data analysis(graphs, correlations, data cleaning), model development (i.e., Logistic Regression, Random Forest) and model validation (i.e., accuracy, F1-Score, AUC). (25 points) This will include a five page write up with figures and results explaining the process you followed to answer your question. (65 points)
3. Create a conference style presentations (15 slides of your methodology and results) and present it to the class. (20 points)

Rules of the Game

This assignment must be done in Groups.

Submission

Submit everything through Blackboard. As mentioned above, you will need to upload:

1. The Jupyter notebook all your work is in (.ipynb file), derived from the provided template
2. Python file (export the notebook as .py)
3. Project Proposal and Report in PDF
4. Powerpoint of your presentation