DATA SCIENCE CAPSTONE PROJECT

TECHNICAL REQUIREMENTS

Data Selection

Define a business problem worth solving, analyzing, and visualizing.

- Provide a clear description on why you chose the topic and establish your end goal.
- Document data source and include citation, if appropriate.
- Describe limitations and assumptions for data set.

Must be **minimum 1million rows** of data.

- D1 data can be pre-clean.
- D2 data should be raw so you can display data cleaning skills.

Data Storage

Data must be stored on a Database (Postgres, MSSQL or MongoDB)

- Provide an ERD
- Optional: Use Second normal form to store your data.

Database should connect to Jupyter Notebook, Python app or work as data collection from inputs.

EDA – Pre-Processing

Includes: data description, handling of missing values, duplicate values, outliers, Bivariate Analysis with graphs, normalizing and scaling, encoding and any other steps necessary before running a model.

- Use at least one of the following graphing libraries:
  - Altair, Seaborn, Matplotlib or Plotly.
  - Graphs must be clearly titled and labeled
  - Store all graphs in a proper format to use for presentation

Use ML algorithms, in the context of what was learned

Use Regression, Classification, NLP, or Neural Networks

Model Performance Reviews:

- Clear description of Model Selection process
- What confirmed and validated your selection
- Model performance metrics:
  - Regression: R, R^2, MSE, Influence Diagrams, accuracy, etc.
  - Classification: precision, recall, F1 score, accuracy, etc.
- Graph your metrics. (graph trees if using dtrees or rforest)
  - Graphs must be clearly titled and labeled
  - Store all graphs in the appropriate format to use for presentation.
- Model must have a reasonably good performance.

Machine Learning Model deployment

Use Python Flask, HTML and CSS (optional: JavaScript) to deploy model as an app.

HTML page must have user inputs and reasonable styling.

- Optional: Include data validation for your inputs.