Capstone Proposal

Appliances Energy Prediction

- **Domain Background:** prediction of energy consumption in a home. Data contains reading from different sensors in different times.
    - Related Academic Research: https://github.com/LuisM78/Appliances-energy-prediction-data
- **Problem Statement:** Predict Energy consumption of appliances in a home based on weather condition inside and outside the home.
- **Datasets and Inputs:**
    - the dataset has 29 attributes:
      -date time year-month-day hour:minute:second
      -Appliances, energy use in Wh
      -lights, energy use of light fixtures in the house in Wh
      -T1, Temperature in kitchen area, in Celsius
      -RH_1, Humidity in kitchen area, in %
      -T2, Temperature in living room area, in Celsius
      -RH_2, Humidity in living room area, in %
      -T3, Temperature in laundry room area
      -RH_3, Humidity in laundry room area, in %
      -T4, Temperature in office room, in Celsius
      -RH_4, Humidity in office room, in %
      -T5, Temperature in bathroom, in Celsius
      -RH_5, Humidity in bathroom, in %
      -T6, Temperature outside the building (north side), in Celsius
      -RH_6, Humidity outside the building (north side), in %
      -T7, Temperature in ironing room , in Celsius
      -RH_7, Humidity in ironing room, in %
      -T8, Temperature in teenager room 2, in Celsius
      -RH_8, Humidity in teenager room 2, in %
      -T9, Temperature in parents room, in Celsius
      -RH_9, Humidity in parents room, in %
      -To, Temperature outside (from Chievres weather station), in Celsius
      -Pressure (from Chievres weather station), in mm Hg
      -RH_out, Humidity outside (from Chievres weather station), in %
      -Wind speed (from Chievres weather station), in m/s

-Visibility (from Chievres weather station), in km
-Tdewpoint (from Chievres weather station), Â°C
-rv1, Random variable 1, nondimensional
-rv2, Random variable 2, nondimensional

- dataset link: https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction#

- **Solution Statement:** it's a supervised problem, specifically I can use linear/multiple regression or SVM.
  - Regression can generally mathematically be expressed as: $Y = mx + b$ where Y is the target, x is the input variable, m is the coefficient, and b is the intercept.
- **Benchmark Model:** in the research paper i've mentioned above it used those models:

  a) Regression with lm

  b) SVM with Radial kernel

  c) Random Forest

  d) Gradient Boosting Machine
  - According to R-squared the GBM achieved the highest score out of all the models.

- **Evaluation Metrics:**
  - R2 Score.

- **Project Design:** .
  - **Data Visualization:** Visualize the data to find the correlations between features and target variable.
  - **Data Pre-processing:** clean the data if necessarily, and splitting data into training, testing, and validation sets.
  - **Feature Selection:** find the relevant features.
  - **Model Selection:** try some algorithms to find out the best one.
  - **Testing:** test the trained model on the testing set.