**DAT540 Group 10**

**Project draft: Preliminary problem statements and techniques**

**1.Data cleaning and preprocessing.**

-> summary statistics

-> check the null values

-> converting string columns to datetime objects

-> add new features from TimeSunSet and TimeSunRise columns

-> outlier detection/removal with different methods (Isolation Forest, IQR, Z-score)

Number of anomalies detected for Radiation = 654 out of 32686

Number of anomalies detected for WindSpeed = 467 out of 32032

**2.Data visualization and analysis.**

-> 4 month time interval covering both summer and winter season

-> Pairplot

-> Check the distribution of the columns using histograms

-> Oulier detection using boxplots

-> Created line plots to see the fluctuation over different day/time interval

-> We found that 6th and 7th of December dates are missing and perhaps more (?)

Chart, bar chart

Description automatically generated

Chart, diagram, box and whisker chart

Description automatically generated

**3.Select features of interest.**

-> Common sense and feature engineering

-> Correlation heatmap

-> Feature importance property of the model

Chart, waterfall chart

Description automatically generated

**4.Define a problem statement based on your analysis of the dataset and the identified features of interest.**

From this exploration of the data, we see the following patterns in this dataset:

- Higher **temperatures** correlate to more radiation throughput. This is confirmed by a Pearson R-value of 0.73 and the observed behavior of radiation "following" temperature on the daily and weekly time scales.

- **Humidity** has a lesser (but potentially significant) impact on radiation throughput. With a Pearson R-value of magnitude above 0.20, humidity cannot be ignored as a potential driver of the system.

- **Pressure** doesn't correlate much to radiation, but does correlate to temperature and humidity. Since temperature, pressure, and humidity are all characteristics of the atmosphere it is not surprising that they are correlated.

- **Wind direction** has a moderate correlation to temperature (-0.26), pressure (-0.23) and radiation (-0.23)

- **Seasonal changes** are significant. Even though Hawai'i does not see seasons as drastic as the northern continental United States, seasonal changes in temperature and humidity are still severe enough to be taken into account, as shown by the weekly measurement comparisons.

**5.Choose an appropriate ML model based on the dataset and your problem statement.**

**-** Linear Regression,Random Forest, XGBoost, Gradient Boosting, MLP Regressor

**6.Tune the parameters of your model.**

-GridSearchCV and RandomizedSearchCV

**7.Make predictions regarding the future values for some of your chosen variables.**

**Chart, bar chart

Description automatically generated**

**8.Evaluate your model based on different metrics.**

- MSE, RMSE and R^2