Problem Set 4 - (Harsh Gupta)

Steps performed:

- **Data analysis:** Checked the features, min-max values, unique values, etc.
- **Data Preprocessing:**
  - Scaled the values of the columns 'Hour', 'Temperature', 'Humidity’, 'Wind speed (m/s)', 'Visibility (10m)', 'Dew point temperature', 'Solar Radiation (MJ/m2)', 'Rainfall(mm)', 'Snowfall (cm)'
  - Converted **Date** into the day, month, and year numerical fields**.**
  - Transformed **Seasons**, **Holiday**, and **Functioning Day** fields into numerical data. (Assumptions: If a season is not among summer, winter, spring, or autumn -> marked it as summer. If the holiday is not formatted correctly -> marked it as Not a holiday, and, if the value of functioningDay is neither a yes nor a no -> it is marked as a functioning day)
  - Dropped the index column from all three datasets
- **Training and validation**: Trained a Gradient Boosting Regressor on the train data. Performing Kfolds to validate the outcome on the train data. (Tuned the model training parameters, Dropped a few features, performed feature engineering by merging a few features together, trained the model, and validated the model accuracy). Performed the same using other regression models.
- Achieved an RSquared score of 0.89 during the training. With Kfolds and a Rsquare score of 0.90 on the Kaggle private score.