Homework 10 (Due Nov 14, 11:59 PM, 13 points)

IST 5520 – Fall 2022, Chen

Name: _____

Data

The data file "DC_bike_rental.csv" contains data on bike rental in DC for the years 2011 and 2012. The original data were collected from Capital Bikeshare in DC.
The refined dataset includes the following 10 variables:

- hour: hourly time
- season: 1 = spring, 2 = summer, 3 = fall, 4 = winter
- holiday: whether the day is considered a holiday, 1 = yes, 0 = no
- workingday: whether the day is neither a weekend nor holiday, 1 = yes, 0 = no
- weather: 1 = Clear, Few clouds, Partly cloudy, Partly cloudy; 2 = Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist; 3 = Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds; 4 = Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp: temperature in Celsius
- atemp: "feels like" temperature in Celsius
- humidity: relative humidity
- windspeed: wind speed
- count: number of total rentals (the target we want to predict)

Task A: Data Manipulation (3 points)

1. Read the dataset into a pandas DataFrame. Show the datatype of variables in the dataset.
2. A very important step in predictive analytics is to represent different scales of measurement correctly in the dataset. What variables in the dataset should be represented as categorical variables? List them in the box.

Variables that should be represented as categorical variables:

   Transform data if necessary. Create y and X objects to represent the response variable and predictors. Use one-hot encoding to represent categorical predictors.
3. Use 40-60% data partition strategy, with 40% of data used as test data.

Task B: Predictive Modeling (10 points)

1. Normalize your data to range [0, 1]. Use the training dataset to build the scaler and then use the scaler to normalize both training and test data.
2. Use k-NN algorithm to predict the count of bike rental. Tune the parameter of k based on the RMSE statistic.

   Note: As this is a regression problem, you'll need to use an appropriate method. Refer to the link for all methods offered in the skearn.neighbors module:

   https://scikit-learn.org/stable/modules/classes.html#module-sklearn.neighbors
3. Compare the performance of the k-NN regressors with different values of parameter k. Answer the following questions.

   Q1: What is the value of k that leads to the best performance of k-NN in terms of RMSE?

   Answer:
4. Calculate other performance measures with the best k value and type them in the following table.

Measures	k-NN (k=?)
RMSE	
MAE	
R Squared	

Submission: Submit this document with answers and the Jupyter notebook that shows your analysis.