# CSE 5243 – Introduction to Data Mining
# Homework 2

## Objective:

In this homework you will build a classifier based on the "altered_seoulbikesrented" data that you used for homework1. You will use 3 classifiers to predict the class labels – "IsitDay".

The learning objectives are:

1. Learn to implement off the shelf classifiers in Python
2. Learn to tune and evaluate a classifier to achieve good performance
3. Learn to compare and contrast different classifiers based on the characteristics of dataset

## Problem Statement:

You are a recent employee at startup called "SmartBike". Researchers at SmartBike have developed a new set of sensors that can measure weather information like temperature, windspeed, humidity, etc cheaply and accurately. Each bike has a small microcontroller that can record the weather data collected from sensors and it also has access to extra information like whether it's a holiday, functioning day, seasons and number of bikes rented at a particular hour. However, there is one issue - the photoelectric sensor is not working. The researchers are working to fix this issue. Meanwhile you are asked to

build a classifier that can accurately predict whether its day/night outside. Based on your classification the automatic flag in bike would determine whether to turn on the head lights or not. Wrong prediction of day/night would cause bike to switch on/off the headlights at wrong time of the day. Through market analysis the team has deduced that having the lights off at night is costly since it can result in higher chance of accidents, user dissatisfaction, etc. While having lights on during day causes the battery to be used unnecessarily and hence needs to be replaced more frequently. In short, the false positive (predicting day while its night) is costly compared to false negative (predicting night while its day). The costs are quantified in following way

| | Predicted class | | |
|---|---|---|---|
| Actual Class | | Class=Day | Class=Night |
| | Class=Day | 0 | 1 |
| | Class=Night | 10 | 0 |

Example: if your test data had 100 instances and your classification give 10% error then you have 10 instances of wrongly labeled data points. Say out of 10- 7 instances were false negatives and 3 instances were false positives then the total cost of your algorithm would be

$$Cost = 7 * 1 + 10 * 3 = 37$$

The lower the cost of your algorithm the better it is.

## Approach:

1. **Evaluation Method**

Define measure to evaluate your classification algorithm. You will use two types of evaluation measure.

a) Measure that does not include cost measure defined in the problem statement above (Accuracy, Precision, Recall, F1, etc)
b) Measures that include Cost information as defined in the problem statement

2. **Preprocessing data**

You will use "altered_seoulbikedata.csv" to train your model. Before training, split the data into two parts -one for building the model (train data) and other to tune the hyperparameters (validation data).

You should start with the features that you built in the homework 1 to build the classification model. If you wish to change some features, you can go back and change the feature appropriately and build classification model on new feature set. **Mention all the steps that you have taken explicitly as well as the thinking process behind it.**

Note: Some algorithms in python implementation do not accept nominal features. One way around this is to use one hot encoding. For example, a nominal feature called day = {Mon, Tue, Wed, Thurs, Fri, Sat, Sun} can be represented by 7 boolean features – isitMon, isitTue, isitWed, etc. Only one of the 7 features will be set true and thus its called "one" hot encoding.

(https://stackoverflow.com/questions/38108832/passing-categorical-data-to-sklearn-decision-tree )

3. **Classification models**

You will use three off the shelf classification model to train and evaluate. Out of the three classification, one of them has to be KNN – K Nearest Neighbor classification. You can use any classification model (Decision Tree, SVM, Logistic regression, etc) for other two. For each classification model you will do the following-

a) Configure the off the shelf classification models from SciKit Learn library (check references below). Explain all the steps involved – setup, parameters, etc.
b) For each classification method tune the parameters and make observation in your evaluation measures that you defined earlier.
c) Explain what parameters/characteristics of classification model works best for the dataset.

4. **Comparison of the Three Classifiers**

a) Compare the three classifiers to each other. Mention advantage and disadvantage of all.
b) Choose one classifier as best and explain **why.** Remember often the which is best classifier is not always clear as all of them will have their advantages and disadvantages. What is important is the reasoning behind why you considered one over the other.

5. **Conclusion**

Write a short summary of your discovery and learning in the homework.

## Grading Criteria:

1. Overall readability and organization of your report (5%)
2. Evaluation Method (10%)
3. Pre-Processing of the Dataset (10%)
4. Evaluation of the KNN Classifier (20%)
5. Evaluation of the Second Classifier (20%)
6. Evaluation of the Third Classifier (20%)
7. Comparison of the Three Classifiers (10%)
8. Conclusion (5%)

## How to turn in your work on Carmen:

Submit to Carmen the Jupyter Notebook and any supporting to files that you used to process and analyze this data. You do not need to include the input data. All submitted files (code and/or report) except for the data should be archived in a *.zip file and submitted via Carmen. Use this naming convention:

• Project1_Surname_DotNumber.zip.

The submitted file should be less than 5MB.

## References:


1. https://scikit-learn.org/stable/modules/preprocessing.html
2. https://scikit-learn.org/stable/modules/preprocessing.html