Mid Term (Take Home)

Deadline: 16/01/2021 by 6:00 PM [Submit in Google Classroom]

Course: Machine Learning

Fall 2020, North South University

Total Marks: 30 (5 x 6)

(Answer all questions)

1. What do you mean the term 'Curse of Dimensionality'? Explain how regularization can help us to get rid of overfitting problem?

2. Mention how one hot encoding helps to handle the categorical attributes in data. Why do we need to perform scaling of features in a data set before applying machine learning methods? Explain with an example.

3. What do you mean by principal component? What is the difference between the first and the last principal component? Why KNN algorithms is said to work better on small data sets?

4. What's the F1 score? How would you use it? How do you handle missing or corrupted data in a dataset? Discuss your options.

5. Get Iris data set (Iris.csv) from here (https://www.kaggle.com/uciml/iris). Now answer the following question (write only piece of code that is necessary, not the full version)

Write a Python program using Scikit-learn to split the iris dataset into 80% train data and 20% test data. Out of total 150 records, the training set will contain 120 records and the test set contains 30 of those records. Train or fit the data into the *model and calculate the accuracy of the model using the K Nearest Neighbor Algorithm.*