**INSY 662: Data Mining and Visualization (Fall 2023)**

**Individual Assignment 2**

**Due by Thursday, September 28, 2023 at 11:59pm on MyCourses**

* This is an individual assignment. The McGill Academic Integrity code applies to this and other assignments!
* You need to submit **TWO FILES to MyCourse**. One, **a Word file with your answers**. You can type or select your answers (highlighted places) directly on this assignment file. Two, you should also submit **your Python code (in .py format)**.
* The goal of this assignment is to understand how to apply **K-NN**.

**Consider the following dataset that has three observations. Each observation has one target variable () and two predictors ( and ):**

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Black | 1 | 1 |
| Blue | 0 | 0 |
| Blue | -1 | -1 |

Task 1. Before using python, let’s approach the K-NN algorithm intuitively. Intuitively, if we develop a K-NN model by using this dataset and specify , the model should predict the target variable () of a new observation where and to be Blue because:

(0.1,0.1) is closest to (0,0) and we are considering only one point since k=1. Hence it should be classified as blue.

Task 2. Now, go to Python and create the dataset above. (Hint: You can first create an array and change it to a dataframe)

Task 3. Develop a K-NN model by using the dataset and specify without any other parameters.

Task 4. Make a prediction with a new observation where and . The algorithm predicts that the target variable () is Black

Task 5. If you use predict\_proba method instead of predict method, the probability that the target variable is Black is 0.5 and the probability that the target variable is Blue is 0.5

Task 6. To make sure that K-NN algorithm uses distance in the classification task, we need to specify a parameter when we build the K-NN model. In such a case, the command to initialize the K-NN model should be:

knn = KNeighborsClassifier(n\_neighbors=2, weights = “distance”).fit(X,y)

\*\*End of Assignment 2\*\*