**AIN 380 Homework Assignment #8 Name:**

**Supervised kNN Classification - ML**

Create a supervised ML kNN model for classifying a wheat variety from data characteristics of wheat kernels. There are 3 varieties: 1=Kama, 2=Rosa,   
3=Canadian. The data set to be used is named *wheat.csv*. Make sure that your model is a *sklearn* *KNeighborsClassifier* model. All numeric features in the data set are to be used to predict the “Variety”. Be sure to include the following deliverables.

1. Download the file named *wheat.csv* that contains data collected from over 200 observations of wheat kernels and load the data into a Pandas *DataFrame*. Print the data (all rows and columns) and print the *DataFrame* information using the *info* method. Be sure to appropriately label the output.
2. Drop all categorical columns.
3. Normalize the X feature data (**not** the y data) using the *sklearn* StandardScaler. Be sure to appropriately label the output.
4. Split the dataset training to test at a 70% - 30% split.
5. Print the mean and the standard deviation for the X\_train data.
6. Create a kNN classifier model with k = 3 neighbors.
7. Test the accuracy of the classifier model and print the *sklearn.metrics* accuracy score, classification report, and confusion matrix. Be sure to appropriately label the output.

A screenshot of a computer

Description automatically generated

**This is for k = 3 neighbors**

1. Compute the mean *cross*\_*val*\_*score* for k values from 1 to 50 and plot them in a seaborn lineplot. Print the line plot for credit. Be sure to label the line plot.
2. Analyzing the line plot obtained in #7 above, what is the optimal k value for this model? \_\_8, 10, and 15 all had the same optimal k value\_\_\_\_\_\_

Modify the kNN classification model using the optimal k value identified in #8 above. Test the accuracy of the classifier model and print the *sklearn.metrics* accuracy score, classification report, and confusion matrix. Be sure to appropriately label the output.

**This is for k = 8 neighbors**

A screenshot of a computer

Description automatically generated

**This is for k = 10 neighbors**

A screenshot of a computer

Description automatically generated

**This is for k = 15 neighbors**

A screenshot of a computer

Description automatically generated

1. Explain the confusion matrix obtained in #10 above. Be sure to reference the actual values in the matrix:

**A confusion matrix represents the prediction summary in matrix form, which means that it shows how many prediction are correct and incorrect per class. As we can see in the k=15 output above, we classified 23 Variety 1s correctly and 3 incorrectly, we classified 20 Variety 2s correctly and 1 incorrectly, and finally we classified 15 Variety 3s correctly, and 1 incorrectly.**

1. Using this kNN classification ML model, predict the wheat “Variety” for the following kernel observations, but be sure to ***transform*** scale them first. Print the predictions. Be sure to appropriately label the output.

A table with numbers and letters

Description automatically generated

**Prediction 1 = Variety 2**

**Prediction 2 = Variety 3**

1. Explain in words how confident you are of your predictions in #11 above and explain why: \_\_**I am extremely confident in my predictions because of the accuracy score we got above. For all three k=values the accuracy score for the model was over 90%, so it makes me extremely confident that the predictions are correct.**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Submit all of your output, your answer on this sheet and a listing of your Python code on the due date. Upload your Python code to Canvas with the data set as well. Failure to do both will necessarily result in a loss of points.