# CSE 708: Application of Data Analytics and Engineering

# Fall 2023

## Final Exam

*Due Monday December/11/2023 at 11:59PM*

A late penalty will be applied. If you do not submit your exam on time, you will lose 1 point for each hour late.

1. (**30 pts**) Pattern recognition is a data analytics method that uses machine learning algorithms to automatically recognize patterns and regularities in data. The decision tree is one of the machine learning algorithms used in pattern recognition. Please look closely at the following synthetic dataset and see if you can recognize some patterns. It may be difficult for the human brain even if the dataset is small. Your task is to draw a binary decision tree (by hand, no coding please) which will help to recognize those patterns (classification). The class or the target is this question are Y meaning that one will buy a car and N meaning that he/she will not, based on the color, year, mileage, and model.

| Color | Year | Mileage | Model | Class |
|---|---|---|---|---|
| Blue | 2015 | 110000 | Benz | N |
| Green | 2018 | 70000 | Honda | N |
| Blue | 2016 | 120000 | Honda | Y |
| Purple | 2020 | 8000 | Chevy | N |
| Blue | 2020 | 100000 | Nissan | Y |
| Glay | 2016 | 100000 | Honda | N |
| Bleu | 2016 | 150000 | Toyota | N |
| Red | 2010 | 120000 | Benz | Y |
| White | 2016 | 100000 | Honda | N |
| Bleu | 2017 | 120000 | Nissan | Y |
| Red | 2020 | 50000 | Toyota | N |
| Purple | 2006 | 200000 | Honda | N |
| Red | 2014 | 90000 | Benz | Y |
| Blue | 2015 | 100000 | Toyota | Y |

2. The following is a balanced portion of IRIS dataset. Use the KNN algorithm and Manhattan distance to predict the variety of the new example (in red). Do not use any programming language, this is a hand calculation question, and show all your calculations.

| sepal.length | sepal.width | petal.length | petal.width | variety |
|---|---|---|---|---|
| 7.9 | 3.8 | 6.4 | 2 | Virginica |
| 5.1 | 3.5 | 1.4 | 0.2 | Setosa |
| 6.3 | 2.8 | 5.1 | 1.5 | Virginica |
| 6.1 | 2.6 | 5.6 | 1.4 | Virginica |
| 4.9 | 3 | 1.4 | 0.2 | Setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | Setosa |
| 6.4 | 2.8 | 5.6 | 2.2 | Virginica |
| 4.6 | 3.1 | 1.5 | 0.2 | Setosa |
| 5 | 3.6 | 1.4 | 0.2 | Setosa |
| 7.4 | 2.8 | 6.1 | 1.9 | Virginica |
| 6.6 | 2.9 | 4.6 | 1.3 | ? |

   a. **(30 pts)** What is the variety of the new example? Compute your predictions using 3 neighbors.
   b. (**10 pts)** Then use 5 neighbors.

**3.** (**10 pts**) Choosing the optimal value of k in the k-NN (k-Nearest Neighbors) algorithm is an important task, as it can have a significant impact on the performance of the model. Which of the following methods are not used to choose the best k for KNN classifier? **This question is tricky, you must choose all and only the right answers.**

   a. Grid Search
   b. K-means
   c. Elbow method
   d. Rule of thumb
   e. K-fold cross-validation

4. (**5 pts**) Ensemble learning: a ML engineer trained a logistic regression, a KNN, and a Naive Bayesian model on a dataset, and then combined the models using a neural network, by voting. What kind of ensemble learning is this?

   a. Boosting
   b. Stacking
   c. Bagging

5. (**10 pts**) The process of transforming raw data into informative attributes, or the original features into a new set of features that are more informative, and compact is called (**Only one is correct**):

    a. Dimensionality reduction
    b. Feature selection
    c. Feature Extraction
    d. Feature Creation
    e. Feature normalization

6. (**5 pts**) Is this true or false: SMOTE is Synthetic Majority Undersampling Technique used to address imbalanced datasets.