DSCC 201/401 Final Project
Due: **Friday, December 16, 2022 at 5 p.m. EST**

Answers to these questions should be submitted via Blackboard. Only one submission for this final project will be allowed. Revised submissions will not be allowed. So please make sure you only submit your final answers. All answers must be shown with the corresponding code.

1. The data located at `/public/bmort/python/titanic.csv` contains information from the Titanic data set including information about survival of the passengers. Notice that there are 12 columns in the data set, and they correspond to the following: the ID of the passenger, the survival status (0 = not survived, 1 = survived), the class of the passenger, name, sex, and age as recorded. SibSp represents the total number of siblings and spouses of the passenger on board, and Parch represents the number of children and parents of the passenger who were also on board. Ticket is the ID string for the ticket and Fare is the price of the ticket. Cabin represents the cabin location of the passenger, and Embarked is the location where the passenger embarked the Titanic. **Using Python with a Jupyter notebook on BlueHive**, answer the following questions. **Please provide a PDF of your Jupyter notebook showing ALL input and output and upload BOTH the PDF and the Jupyter notebook file (Question1.ipynb and Question1.pdf) showing all input and output.** It is recommended to use the Python 3 (anaconda3 2021.11) kernel. You may embed your answers to the questions asked below as comments in the code or you may submit a separate text document with the answers to the questions. (20 points)

    A. Load the `titanic.csv` file into a Pandas data frame. Are there any missing values in the data frame? How many missing values occur for each of the columns?
    B. What percent of the passengers survived?
    C. What was the maximum fare that was paid to purchase a ticket by a passenger?
    D. How many unique places did the passengers embark from?
    E. Using Scikit-Learn, normalize the values in the Age, SibSp, Parch, and Fare columns so that the range for each column is [0, 1].
    F. Label encode the values in the Pclass, Sex, and Embarked columns.
    G. Partition the titanic data set so that a random sample of 80% of the data will be used for training and 20% will be used for testing your machine learning model.
    H. Using the Pclass, Sex, Age, SibSp, Parch, Fare, and Embarked features and Scikit-Learn, generate a support vector machine (SVM) machine learning model with a linear basis function kernel to predict if a passenger survives.
    I. Perform k-fold cross validation (with 5 splits) on the model with the training set. What is the average and standard deviation of the accuracy of the model?

J. Use the trained model to predict the survival outcomes of the passengers in the `/public/bmort/python/test.csv` data set. Provide your answer as a Python list of 0s and 1s.

2. An agricultural scientist collected over 13,000 observations of physical properties of beans seeds (e.g. area, perimeter, axis lengths, etc.) from 7 different varieties of beans: Babunya, Bombay, Cali, Dermason, Horoz, Seker, and Sira. The data set located at `/public/bmort/python/beans.csv` contains the measurements for the collection of 13,000+ beans with the classification for the type of bean. Using **Python with a Jupyter notebook on BlueHive**, please answer the following questions. **Please provide a PDF of your Jupyter notebook showing ALL input and output and upload BOTH the PDF and the Jupyter notebook file (Question2.ipynb and Question2.pdf) showing all input and output.** It is recommended to use the Python 3 (anaconda3 2021.11) kernel. You may embed your answers to the questions asked below as comments in the code or you may submit a separate text document with the answers to the questions. (30 points)

A. Load the `/public/bmort/python/beans.csv` data set into a data frame. Are there any missing values? Perform any necessary data imputation on the data set.
B. Produce a table of summary statistics on the data set. How do the ranges of the values in the columns compare? Does each column of data have similar magnitudes and ranges? Are there any outliers?
C. Using the Seaborn library's `heatmap()` function, generate a plot showing the correlations between the numerical data in the data set. Show the commands used to generate the plot and include the plot in your output.
D. Based on the correlation plot, decide which features to include for machine learning. Decide if any of these features need to be standardized or scaled appropriately.
E. Partition the beans data set so that a random sample of 80% of the data will be used for training and 20% will be used for testing your machine learning model.
F. Generate a Random Forest machine learning model for classifying the 7 types of beans based on the chosen features from the data set. Use 50 trees to build the model.
G. Use the test data set (i.e. the 20% of the data that was kept aside earlier) to generate a final validation for your model. Generate a multi-class confusion matrix for the test data to demonstrate the accuracy of the model. Comment on the accuracy of the model.
H. Based on your model, classify the beans provided in the unlabeled `/public/bmort/python/beans-unknown.csv` data set. Indicate which classification of the 7 available types has been assigned to each of the 5 unlabeled beans.
I. **EXTRA CREDIT (5 points):** Use PyTorch (by switching to a different kernel) to build a simple fully-connected artificial neural network for the beans classification based on the chosen features provided in the data. Generate a confusion matrix for the test

data set to demonstrate the accuracy of the model. Based on your model, classify the beans provided in the unlabeled `beans-unknown.csv` data set. Indicate which classification has been assigned to each of the unlabeled beans. How do the results with the artificial neural network compare to the support vector machine model?

3. In a similar agricultural research project, another scientist collected 200 observations of wheat seeds (e.g. area, perimeter, etc.) from 3 different varieties of wheat - type A, B, and C. The data set located at `/public/bmort/R/wheat.csv` contains the measurements for these 200 wheat seeds on the three different varieties of wheat. Seven attributes (i.e. area, perimeter, compactness, length, width, asymmetry, and groove) are provided for determining the classification of the type of wheat seed (either A, B, or C). **Using R version 3.6.1 on BlueHive**, please answer the following questions. **Please provide a PDF of ALL R inputs and outputs and answers to the questions (Question3.pdf).** You may embed your answers to the questions asked below as comments in the code or you may submit a separate text document with the answers to the questions. You may choose to use the R console directly, a Jupyter notebook using an R kernel, or RStudio. (30 points)

A. Using R, load the `/public/bmort/R/wheat.csv` data set into a data frame. Are there any missing values? Perform any necessary data imputation on the data set.
B. Produce a table of summary statistics on the data set. How do the ranges of the values in the columns compare? Does each column of data have similar magnitudes and ranges? Are there any outliers?
C. Using the corrplot library's `corrplot()` function, generate a plot showing the correlations between the numerical data in the data set. Show the command used to generate the plot and include the plot in your output.
D. Partition the beans data set so that 80% will be used for training and 20% will be used for testing your machine learning model. You can do the partition manually at random or use the `createDataPartition()` function in R's caret library.
E. Use the support vector machine (SVM) method with a linear basis function kernel from R's caret library to generate a machine learning model for the 7 types of wheat seeds based on some or all features provided in the data set. Using the caret library's `trainControl()` function, check your model parameter and feature selection by performing repeated cross-validation (with 5-folds) on the training data for your model. Consult the caret library documentation as needed.
F. Use the test data set (i.e. the 20% of the data that was kept aside earlier) to generate a final validation for your model with the `predict()` function in the caret library. Comment on the accuracy of the model.
G. Based on your model, classify the beans provided in the unlabeled `/public/bmort/R/wheat-unknown.csv` data set. Indicate which

classification of the 7 available types has been assigned to each of the unlabeled seeds.

H. **EXTRA CREDIT (5 points):** Using gradient boosting decision trees with R's xgboost library, generate a machine learning model for the wheat seed classification based on the features provided in the data. Generate a confusion matrix for the test data set to demonstrate the accuracy of the model. Based on your model, classify the beans provided in the unlabeled `wheat-unknown.csv` data set. Indicate which classification has been assigned to each of the unlabeled seeds. How do the results with xgboost compare to the support vector machine model?

4. The data set located at `/public/bmort/R/heart.csv` contains a table of data from 300 patients and includes the features: age, sex, chest pain, blood pressure, cholesterol, blood sugar, ECG abnormality, heart rate, angina, ST value, ST slope, major vessel number, thal number, and whether or not the patient was diagnosed with heart disease (1 = yes, 0 = no).). **Using R version 3.6.1 on BlueHive**, please answer the following questions. **Please provide a PDF of ALL R inputs and outputs and answers to the questions (Question4.pdf).** You may embed your answers to the questions asked below as comments in the code or you may submit a separate text document with the answers to the questions. You may choose to use the R console directly, a Jupyter notebook using an R kernel, or RStudio. (20 points)

A. Load the `/public/bmort/R/heart.csv` data set into a data frame. Are there any missing values? Perform any necessary data imputation on the data set.
B. Produce a table of summary statistics on the data set. How do the ranges of the values in the columns compare? Does each column of data have similar magnitudes and ranges? Are there any outliers?
C. Partition the heart data set so that 80% will be used for training and 20% will be used for testing your machine learning model.
D. Using logistic regression as provided by the Caret library in R, develop a model to predict heart disease diagnosis based on the 13 features provided in the data set for each patient.
E. Generate a confusion matrix using the data from your test set to show the accuracy of the model.
F. Write a few sentences providing commentary on the accuracy of the model. What percent are false positives? What percent are false negatives?