DSCC 201/401 Final Project
Due: **Wednesday, May 7, 2025 at 5 p.m. EDT**

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. An agricultural scientist collected over 13,000 observations of physical properties of beans (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/R/beans.csv` contains the measurements for the collection of 13,000+ beans with the classification for the type of bean. **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 (Question1.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. (50 points)

   A. Using R, load the `/public/bmort/R/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?
   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 beans 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. Generate a multi-class confusion matrix for the test data to demonstrate the accuracy of the model. Comment on the accuracy of the model.
   G. Based on your model, classify the beans provided in the unlabeled `/public/bmort/R/beans-unknown.csv` data set. Indicate which

classification of the 7 available types has been assigned to each of the 5 unlabeled beans.

2. As mentioned in class, the Space Weather Prediction Center records data related to the interactions between the sun and the Earth's geomagnetic field. The data sets consist of several fields:

Column	**Description**
Date	Date of report
Radio Flux 10.7 cm	Solar radio flux measured at 10.7 cm
SESC Sunspot Number	Number of sunspots
Sunspot Area	Size of sunspots
New Regions	Number of new regions developing
Stanford Solar Mean Field	Solar mean field measurement
GOES15 X-Ray Bkgd Flux	Background radiation
X-Ray C	X-Ray solar flare of classification C
X-Ray M	X-Ray solar flare of classification M
X-Ray X	X-Ray solar flare of classification X
S	Subflare
Optical 1	Optical solar flare of classification 1
Optical 2	Optical solar flare of classification 2
Optical 3	Optical solar flare of classification 3

Typically, the solar radio flux is measured with instrumentation at a wavelength of 10.7 cm. The flux is always a positive value. Is it possible to predict the value of the solar radio flux using the sunspot number and the number of solar flares? This question will explore the ability to predict the value of the solar radio flux using observations of the number of sunspots on the surface of the sun and the number of solar flares that are observed daily. **Using Python with a Jupyter notebook on BlueHive**, answer the following questions below. **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 2023.07-2)** 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. (50 points)

A. Copy the `solar_data.tar.gz` file from `/public/bmort/python`. Unpack the compressed file. Examine the data structure in the tables in each of the individual files.
B. Load the data from all years into a Pandas data frame in Python. Produce a table of summary statistics on the data set. How do the ranges of the values in the columns with numerical data compare? Does each column of numerical data have similar magnitudes and ranges? Are there any outliers? Perform any data imputation or data modification as part of the data clean up process. Explain the reasoning for all actions taken.
C. For the 10-years of available data, which day(s) had the highest value for the radio flux? Which day(s) had the lowest value for the radio flux?
D. Which day(s) had the greatest number of solar flares? How many days had no solar flares?
E. Create a single boxplot of the radio flux data for all 10 years using the Seaborn library. The y axis should be the solar radio flux and the x axis should be year.
F. Partition the 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.
G. Generate a linear regression model to predict the solar flux from the sunspot number and the individual number of X-ray and optical solar flares.
H. Use the 20% of the data set aside for testing to determine the accuracy of your model. Choose an appropriate accuracy metric. How well does your model predict the solar radio flux from the sunspot number and number and type of solar flares? Comment on why the accuracy is good or poor.
I. What is the predicted solar radio flux for a day when 96 sunspots are observed on the sun and a single solar flare with X-ray radiation that is classified as a C-class flare occurs? How confident are you in the answer? Explain your reasoning.

3. **EXTRA CREDIT (20 points total)**

A. **(10 points):** Use the Keras library as provided by the **Python 3 (miniforge 23.3.1-1 dsc-2024)** kernel to build a simple artificial neural network for the prediction of the type of beans similar to question 1. **Please provide a PDF of your Jupyter notebook showing ALL input and output and upload BOTH the PDF and the Jupyter notebook file (ExtraCreditA.ipynb and ExtraCreditA.pdf) showing all input and output.** Based on your model, classify the beans provided in the unlabeled `/public/bmort/R/beans-unknown.csv` data set. Indicate which classification of the 7 available types has been assigned to each of the 5 unlabeled beans.

B. **(10 points):** Use the Keras library as provided by the **Python 3 (miniforge 23.3.1-1 dsc-2024)** kernel to build a simple artificial neural network for the prediction of the solar radio flux based on the sunspot number and the individual number of X-ray and optical solar flares similar to the response to question 2. **Please provide a PDF of your Jupyter notebook showing ALL input and output and upload BOTH the PDF and the Jupyter notebook file (ExtraCreditB.ipynb and ExtraCreditB.pdf) showing all input and output.** What is the predicted radio flux for a day when 96 sunspots are observed on the sun and a single solar flare with X-ray radiation that is classified as a C-class flare occurs? How confident are you in the answer? Explain your reasoning.