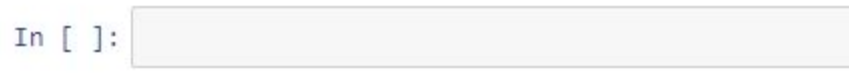
Classification & Regression Coding Project

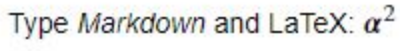
ITCS 3162: Introduction to Data Mining
Due: 5/6/2020 at 11:59PM via Canvas

For this assignment, complete the Jupyter Notebook provided to you via Canvas. Some cells require you to write Python Code and some require written responses using markdown cells. The cells will be preformatted for you and the distinction will be clear. Refer to coding videos on Piazza for help, and feel free to post on Piazza or contact us if you have any further questions.

Code Cells: Appear with square brackets. Run each cell using Shift + Enter



Markdown Cells: Appear telling you to enter markdown text. Double click the cell to start typing in it. Enter your text and "run" the cell using Shift + Enter to display the text.



For this assignment, no formatting or anything is required for the markdown, but if you are interested, here are a couple links for creating text markdown:

- Markdown Basics
- Markdown cheat sheet

Part 1: Classification

In this portion of the project, you will use Decision Tree and Naive Bayes classification techniques to determine if a congress-person is a republican or democrat based on their voting history. The dataset will be available on canvas. Please complete the following tasks.

I. Replace missing values with a ternary value. Meaning if voting no == 0 and voting yes == 1, replace missing values with 2.
II. Create a Decision Tree classifier and a Naive Bayes classifier.
III. Use cross-validation and report the mean F1, precision, and recall scores for each model.

Part 2: Regression

In this portion of the project, you will use features of a bike sharing dataset to predict the number of riders on a given day. This dataset is available on canvas. Please complete the following tasks.

I. Select 3 individual features you think will be good predictors of ridership and create a simple linear regression model for each one.
II. Report the mean-squared-error for each model and create a scatter plot of your results using your best performing single-feature as the x-axis.
III. Create a multiple linear regression model where you utilize all relevant features to predict ridership.