We have taken our first steps into the world of data science. Now we are at the end of the course, and we want to be able to effectively communicate to others what we learned over the course of the last 9 months. This bulk of this project is up to you. You will not be given exact instructions anymore as you are more than capable of doing this type of work on your own.

Your final project is going to be to incorporate either one or more Machine Learning Algorithms of your choosing to provide analysis on a dataset.

IF YOU DO NOT FEEL COMFORTABLE FINDING YOUR OWN DATASETS CHOOSE FROM THESE TWO.

Classification:
https://www.kaggle.com/datasets/uciml/mushroom-classification

Regression:
https://www.kaggle.com/datasets/shivam2503/diamonds

These are the deliverables expected of you.

A Jupyter Notebook containing your code.

Requirements for the Notebook will be outlined later.

A writeup on your analysis and model.

Analyze your results, as well as analyze the benefits and downsides to your model. Which features are key? What visualizations helped proved that? Which features did you keep in/remove, and why? These are just a **FEW** of the talking points that you should be touching upon. 1 Page **MINIMUM,** with no maximum. Please structure your essay correctly, I.E. Paragraphs and flow, do not just submit run on sentences. This is going to be a large part of your grade, ensure that you are reviewing your essay and not just submitting something last minute.

A written script.

This should be in a style that is easy to read and digest. This is going to be something that you should be able to use to present this project, either to your classmates, or to a prospective employer. **DO NOT** simply copy and paste from your essay. It is important that you learn how to summarize and condense your findings in a way that is easy to explain to others. Keep in mind your target audience. Assume you are explaining to your findings to a **NON-TECHINCAL** individual.

Requirements of Deliverables

Jupyter Notebook Code (10 points)

1 Point: Your code should be bug free, and appropriate blocks of code should be encapsulated, either in fuctions, or in their own seperate cells.

2 Points : Exploratory Data Analysis and Cleansing/Wrangling

Your code should demonstrate your ability to use pandas and SKLearn to work through your dataset to complete the following tasks. Data Cleaning, if needed, such as removing Nulls/NaNs, or replacing them, and One Hot Encoding if needed. You should also be looking for anomalies in your dataset, such as highly skewed columns, columns with a high degree of colinearity, etc. You should a

2 Points : Visualizations that help with the following.

A)Feature Selection
B)Model Evaluation (If applicable)
C)Analysis of the model and the data.

Your visualizations should be clearly labeled, and appropriately scaled. The visualizations should **CLEARLY** demonstrate the point that you are trying to make, without the need of clarification.

2 Points : Application of the SKLearn Library to create your model

Your application of the algorithm of your choosing, as long as using SKLearn effectively and appropriately. If you are doing Linear Regression, explain which algorithm you are going to use, SGD or OLS, and why (feel free to use other ones as well). You should also be using the SKLearn library to gather metrics on your model. Which metrics you choose to use and why, are up to your discression, but these choices should be **JUSTIFIED**.

2 Points : Application of the SKLearn Library to cross validate your model

You should explain which method of cross validation you are using, as well as why you are using it. You should be able to clearly and cleanly implement it into your codebase, and make it in a way that is resuable.

1 Points : **COMMENTS AND EXPLANATIONS ARE MANDATORY**. **STYLE GUIDE INCLUDED AT THE END.**

A Single line that does not clearly and thoroughly explain the cell is not useful. Ensure that you are constantly reviewing and looking back at your code base and commenting appropriately. Your target should be to get your codebase into a state where a **NON TECHNICAL PERSON** can work through your notebook and make sense of it.

Project Write Up (10 points)

2 Points : Appropriate grammar and spelling is used.

4 Points : You clearly work through your findings, and it is clear to the reader that you have gone through the process of both evaluating your model, as well as the data. Point to specific parts of your codebase, and explain how they specifically helped you make your analysis.

4 Points : Your essay has a flow and structure to it. It is important to inform the reader at the start what your idea was when you choose the dataset. It's important to explain your intuition and logic, and for it to logically flow from one idea to another. Do not submit run on sentences and run off paragraphs one after another. Essay writing is an interative process.

Script (5 points)

1 Points : Appropriate grammar and spelling is used.

4 Points : There is a sense of flow and structure to your script. You are hitting the key parts of your projects, as well as the insight gained by the project.

Jupyter Notebook Styling Guide

Your code should be clearly segmented, and clearly commented. Get rid of any code blocks that you ran one ofs to perform EDA.

The expectations of how your code should look like is outlined below:

Ensure that each line of logic is commented. Not every line of code needs to be commented on, but if there is some step of logic being applied that should be explained, ensure that the reader can easily understand what your code is doing.

Ensure that code is being segmented. Code that runs together should be unified under one code block, not spread across multiple cells. It is one thing to test and experiment, but the final product should be clean and easy to read.

Ensure that you are using appropriate variable names and placeholder variable names. No more xs,ys,zs. If you need to create a simple loop iterating over a range, that is ok, but if you need to store the resultant call of a function, ensure that the variable that it is being stored in has a name that represents the data stored inside.

Ensure that you are cluing in the reader(s) as to the overall reasoning behind certain methodologies and functions. It is one thing to explain what a function is doing, but explaining the roadmap as to why that function needs to be built, and what that function plans to do further down the line is different.