Machine Learning Intro:

Disclaimer:
The code that we are working with is based on a project/workshop developed by Susan Ibach (hockeygeekgirl on Github), and the powerpoint presentation that she used to explain concepts can be found in her MyFirstSciKitLearnModel repository. This project is linked in the follow instructions step, and was extremely useful while developing my project.

My Engineering Process

Setting Up Your Environment

- Install Anaconda
    - Anaconda is one of the tools used for building machine learning in both Python and R. It installs/updates Python on your computer if you do not already have it, and installs Jupyter Notebook, which we will build our program in.
    - https://www.anaconda.com/
- Open Anaconda and Jupyter Notebook
    - Create a new Notebook
- Follow Instructions
    - Go to the github repository and
    - https://github.com/hockeygeekgirl/MyFirstSciKitLearnModel/blob/master/BasicMLExample.ipynb

Engineering Process

- Pick Your Data!
    - I chose data from the UNICEF open database about child mortality by country, you can find data in an area that interests you, some places to start
        - https://www.freecodecamp.org/news/https-medium-freecodecamp-org-best-free-open-data-sources-anyone-can-use-a65b514b0f2d/
    - Make sure that the data you find is in the .csv format and has a significant amount of data available (5000 lines+)
    - Run it through the code found on the GiHub repo
- Cleaning your data
    - This step takes the longest of anything, ensure that only the data you require is present in your file
    - Get rid of any missing/null/NaN values
- Reassess your data selection
    - There appears to be a connection between geographic area and infant mortality, however when selecting by each country the data is incomprehensible

    - Due to this I will sort by a more generalized geographic area
- Cleaning data 2.0
    - This took even longer!
    - I went through each set of country data and used my judgement of geographical area and also culture to determine what area to classify each country as.
        - I decided on 11 different areas to sort countries into
            - Africa, West Europe, East Europe, East Asia, South Asia, North America, Central America, South America, Caribbean, Oceania
            - If I were starting from scratch, I would divide Africa into different sections to better pinpoint the “problem areas”
- Graphical Analysis
    - I then used the .groupby().mean().plot() command to show a bar graph of Geographic Area and infant mortality
        - The graph appears to have distinct answers for each area, leading me to believe that it likely has some correlation
- Development of my model
    - For this I followed the instructions in the github link that I gave above, however I have a few pieces of advice
        - Ensure that you always pick the same random state value (I chose 16), this will insure your sudo-random separation of test and train values always use the same seed, thus always picking the same values
        - Also, though it can be changed, I would keep the value of test size, at 0.30 (or 30%) as a 70:30 ratio is the industry standard
- Analyzing My Project
    - Using the Mean Error and Regression Coefficient the predictions of the program can be analyzed
        - Mean Error will be lower in more accurate project, as the average error made by the program is smaller, therefore the project is more accurate
        - Regression Coefficient will be larger in more accurate programs

Potential Extensions

If I had more time to further develop my project, I would go deeper into the project by combining multiple factors in the prediction rather than just one. Some of the potential factors that I would test are Sex and Time Period. Additionally, I would look further into this topic to find a more comprehensive data set with more viable lines of information (ideally 100,000+)

How to Improve in the Future

I think the biggest challenge in this project was the selection of data. I didn't really know what I was looking for, and once I found an interesting concept, I didn't know how to test if the concept was viable before getting really deep into the project. Due to this challenge, I am going to spend the rest of the week working on resources for anyone else interested in data science to use to determine the viability of a project early.

Assessing Viability

Selecting a Data Set:

- .csv files are the easiest to work with, and should be used when possible
- Choose something that you find personally interesting, the data scrubbing process will be more tolerable if you are excited by the end prospect you are working towards
- This is an amazing article/resource explaining some of the specifics about choosing/analyzing data sets and choosing modeling types
    - https://towardsdatascience.com/do-you-know-how-to-choose-the-right-machine-learning-algorithm-among-7-different-types-295d0b0c7f60

Conclusion

Due to time restraints and a bought of sickness, I was not able to finish the extensions that I outlined for my project. I will post my finished Mark I of the project as well as the extension that I started (grouping values into bins) in hopes of expanding on this project at a later time. Finally, some of the issues that I faced are more personal, such as not having a strong data management background (course conflicts) and a lack of depth in mathematical knowledge at my school at this time. However, working through this project has allowed me to shallowly delve into areas that are not otherwise offered in highschool, and has given me a surface level understanding of data analysis and machine learning with python, pandas, and other tools often used in industry.