**Progress Report 2:**

**Estimating a Country’s Mortality Rates Using Linear Regression**

Heron Ziegel

**Current Progress:**

After spending last week in the preprocessing stage, this week I used the the [Adult Mortality Rate](https://www.kaggle.com/datasets/mikhail1681/adult-mortality-rate-2019-2021) dataset to create three separate linear regression models. The first model attempted to predict female mortality rate. The second model attempted to predict male mortality rate. The third model attempted to predict average mortality rate, or crude death rate.

First, I had to split the data into testing and training sets for all three models. I did this using sklearn’s train\_test\_split function. Then I normalized the training and testing sets using the StandardScaler normalizer from sklearn. Then I created a linear regression model, also using the sklearn library. I fit the training and testing data to the linear regression model. I checked the bias and coefficients, created a y prediction, the printed out the MAE, MSE and RMSE errors.

For the first model, the MAE, MSE and RMSE were significantly large (40, 3111, and 55 respectively). I decided to add a hyper parameter, but it did not make a significant impact on the prediction accuracy. For the second model, adding a hyper parameter actually made the errors larger. I will be removing the hyper parameter from this model. The third model, crude death rate, had a much lower error than the previous two models (roughly 2, 7 and 3 for the MAE, MSE and RMSE). However, when I added a hyper parameter the error became several times greater for all three measurements. I will definitely be removing the third model’s hyper parameter as well.

My findings from this week were disappointing. With such large errors, these models are clearly not very accurate. It was also surprising that the crude death rate model was significantly more accurate than the models for either male or female death rate alone. Perhaps there are other factors in these which is not being expressed in the data itself, such as whether the country is currently at war or what type of healthcare system they have.

The largest shortcoming of this dataset is that it only has 157 rows. Since each row represents a country, this number represents a good amount of the global mortality data. However, machine learning models work best with millions of data points. Therefore I do not expect my regression model to be particularly accurate when working with this small dataset.

**Remaining Work**:

As I discovered this past week, two of the models’ hyper parameters will need to be removed because they generated additional errors. After that, I would like to generate a CSV file with data found on the internet pertaining to some of the missing countries. Then I can use the models to make predictions about this new data, and compare it to the actual numbers I found. This will be another way of testing how accurate the models are.

After that, I will need to start preparing the final report and lightning talk. I would like to create visualizations for the data and models which are compelling and interesting. Since I am only using a one layer linear regression model, the data itself might be more interesting than the model. However, I would also like to look into the model weights to determine which columns are most relevant and least relevant to the ground truth values.

**Upcoming Week Plan**:

One of the most crucial things I need to focus on this upcoming week is creating the CSV file. I will set aside some time either Monday or Tuesday between classes to gather the necessary data from online resources and correctly format it. Depending on how much data I can gather about the countries not in the dataset, this could be time intensive or it could end up not working at all. If that is the case and I can’t find additional data, I may have to restructure my project slightly. The final report will still contain all of the information from the original dataset, but I may not be able to apply the models to additional data.

Another focus this week will be data visualization. I will utilize tools such as the matplotlib Python library to create graphs and charts for my findings. These graphs will be useful for both my presentation, and my final report. It could be interesting to show a scatter plot of the original data, along with the line of best fit generated by each model.

Since the final report is due on April 28th and is five pages long, I will need to get started on that soon as well. So far my progress has been well documented between the two progress reports, so much of it I can copy from those. However, there will be a significant portion of new material this week, such as the data visualization and new CSV file. Each will need to be carefully documented for the final report in order to give a well rounded and complete picture of what I have been working on.

I would also like to find out this week what the plan is for the lightning talks. Do I need to create a PowerPoint presentation? If so, I will need to get started on data visualization as soon as possible. I would like to gather all of the necessary materials for the presentation ahead of time so that I have time to rehearse my presentation. I might practice with a friend and ask them to critique the quality of the presentation as well.

**References**

Dataset: Adult Mortality Rate. <https://www.kaggle.com/datasets/mikhail1681/adult-mortality-rate-2019-2021>