**Agenda:**

* Git Hub collaborator access
* PCA data visualisation
* Naïve implementations of machine learning models

**PCA Data Visualisation:**

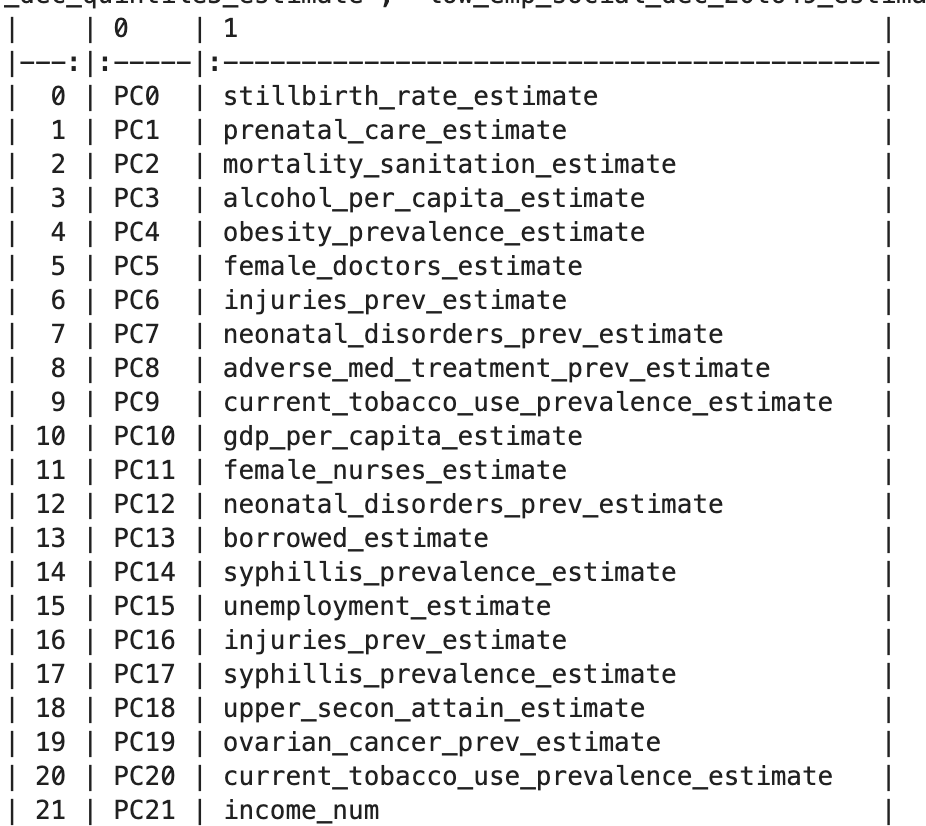
I took in data from the [World Bank](https://genderdata.worldbank.org/en/topics/health#idAllIndicators), rep\_gho, rep\_wb, rep\_ihme\_prev, and rep\_swper about women’s health, economic trends, illness prevalence, and women’s empowerment. Countries reported data at regular intervals (generally every few years), but different countries reported data on different years. Therefore, to get data from each country, I took an average over the values of each indicator for each country.

I fed these averages into the Sklearn’s PCA package. At first, I visualised the proportion of variance in the original dataset that each of the 69 indicators could explain.From the following plot, you can see that the first 5 principal components explain ~58% of variance in the original dataset, with the first principal component explaining ~36%.

A graph of a graph with blue dots

AI-generated content may be incorrect.

I then found the index of the feature that most influenced each principal component.



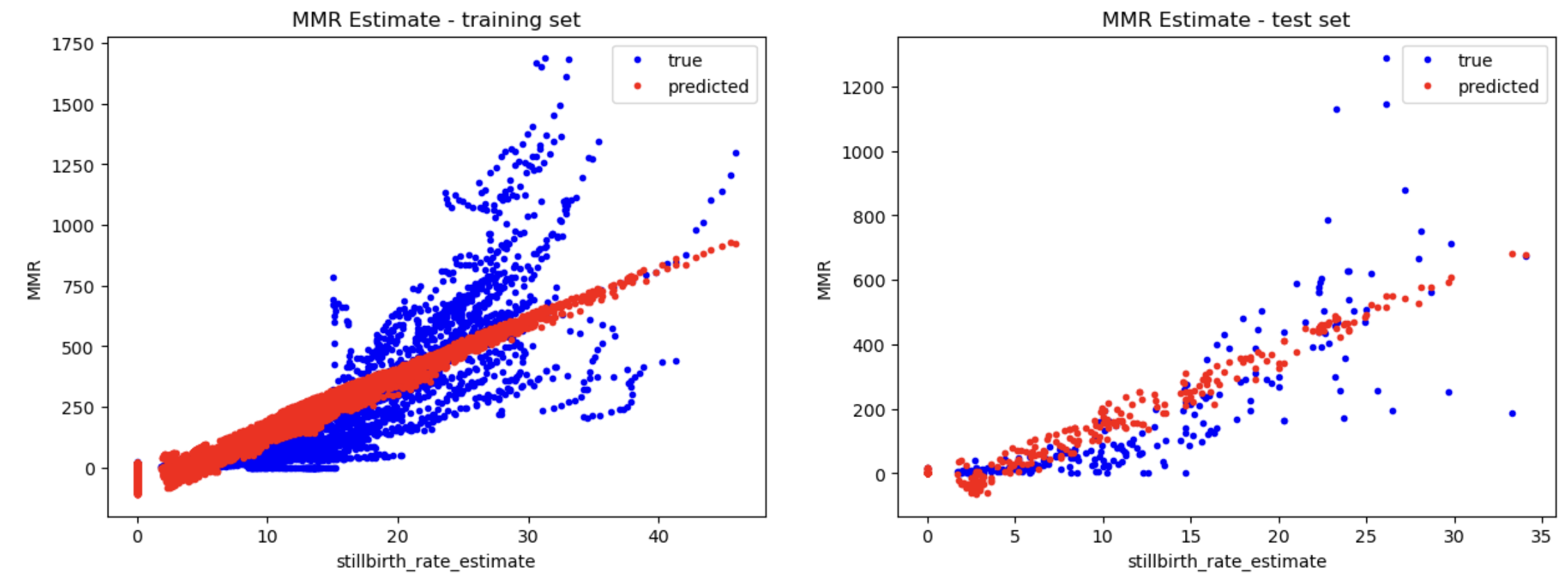
In the following plot, the standardised principal component most influenced by prenatal care (x-axis) is plotted against the MMR estimate, with the colour bar showing income level. As income level increases, MMR estimate and prenatal care decreases.

**A graph of a prenatal care

AI-generated content may be incorrect.**

**Linear Regression:** (using closed form equation to find best parameters)

I used data up to and including 2015 for training the model, with 2016 data used for testing. The first six principal components were used (stillbirth rate, prenatal care proportion, mortality due to sanitation, alcohol consumption per capita (female), obesity prevalence, proportion of doctors who are female. The linear regression model does not fit either the testing or training data, as the data’s trend appears polynomial/exponential.

****

While the model has a regularisation term, changing this term had no impact on RMSE.

**A close up of black text

AI-generated content may be incorrect.**

**Kernel Trick:**

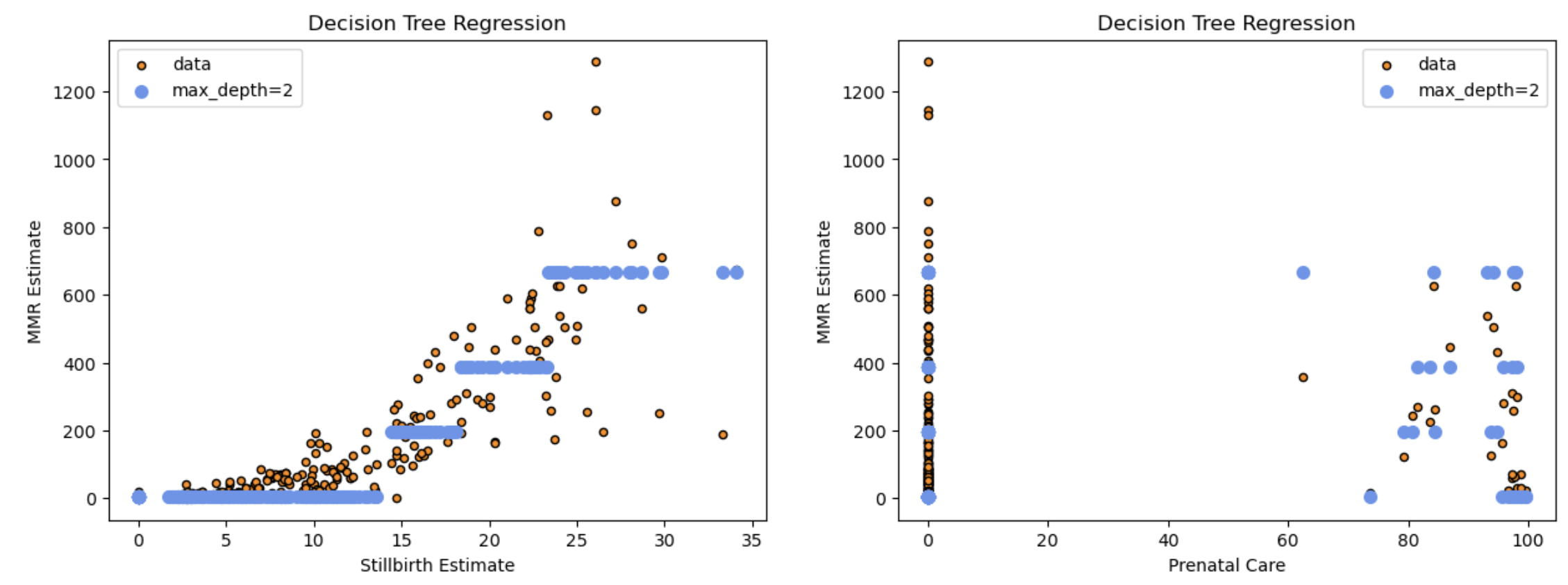
To address the non-linearity of the data, I tried using the KernelRidge method from Sklearn and set the kernel to ‘polynomial’. Unfortunately, my analysis took too long to run and was terminated before it finished. I received the following error:

* UserWarning: Singular matrix in solving dual problem. Using least-squares solution instead.
  + From research, it seems that this error appears when there are more features than instances, which I do not understand in the context of my data.

**Decision Trees:**

I used the Decision Tree Regressor package from Sklearn. I tested three different depths, with the following results:

Depth 2:



Depth 5:

A screenshot of a graph

AI-generated content may be incorrect.

Depth 10:

A close-up of a graph

AI-generated content may be incorrect.

Overall, the mid-depth tree had the lowest error (calculated as: sum((estimate – target)2)), but only by a small amount.

|  |  |  |
| --- | --- | --- |
| **Depth 2 Error** | **Depth 5 Error** | **Depth 10 Error** |
| 7250265782.075951 | 6927763437.535991 | 7304977075.36845 |

**Questions and To-Dos:**

* My data has a large number of NAN values. To be able to run the PCA analysis and machine learning models, I replaced these NAN values with a 0.
  + Is this the best way to solve this problem?
  + This week, I will research imputation methods.
* How to improve the kernel trick method and deal with data that does not appear linear.
* Find an objective metric to compare my error against.
* Have I visualised my methods/models in the standard way?