Please note that I did this assignment alone :(

Team identification

Name 1:Sisard Fatess Calvis

Number 1:

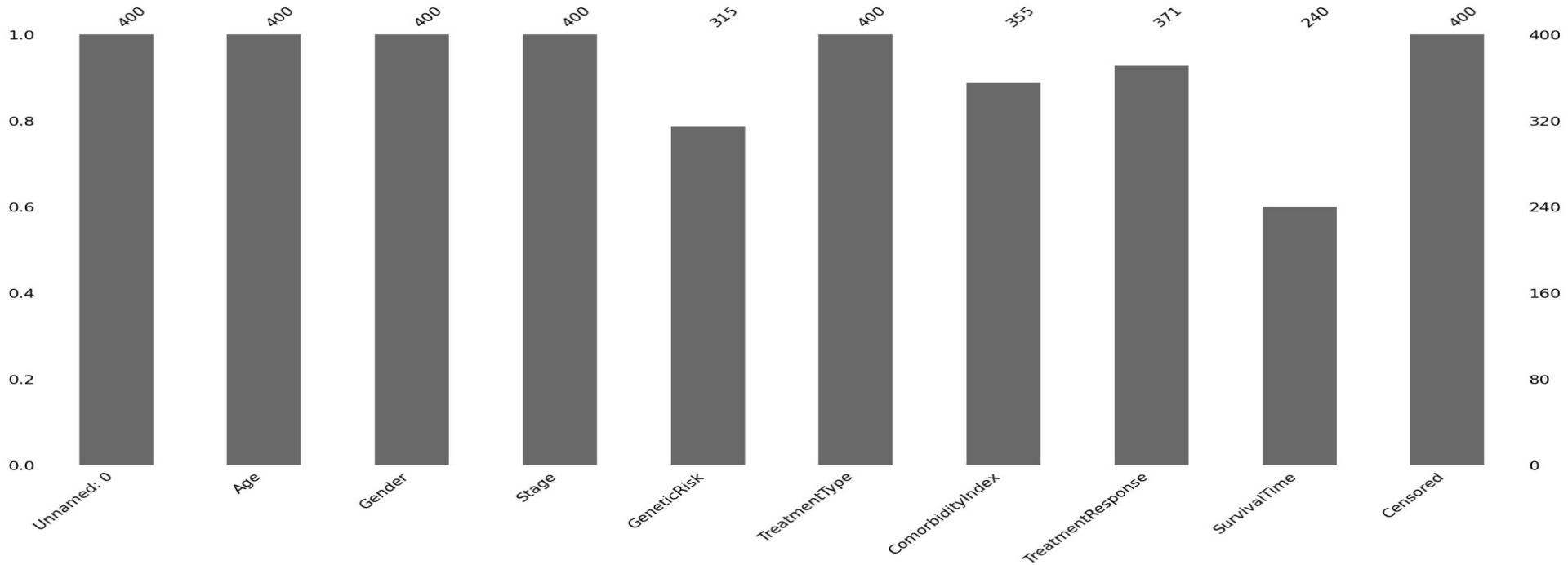
Final score: 2.37

Leaderboard private ranking: 5

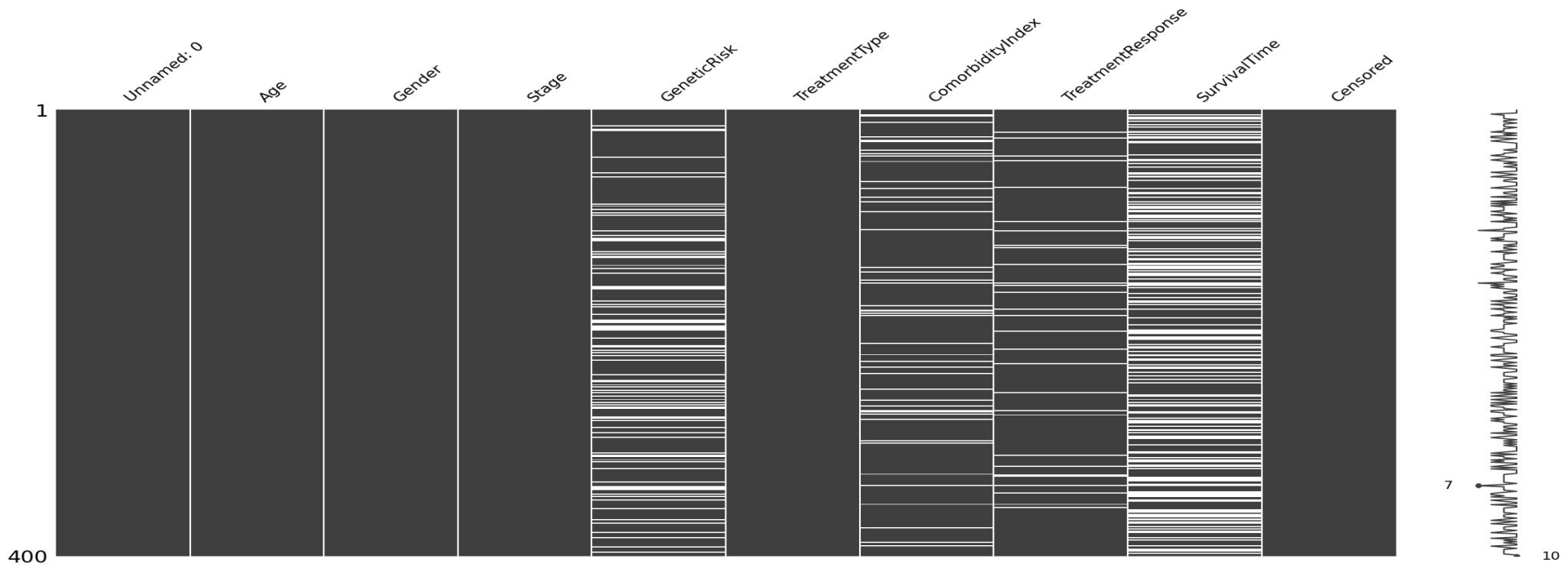
Task 1.1

What was done in task 1.1

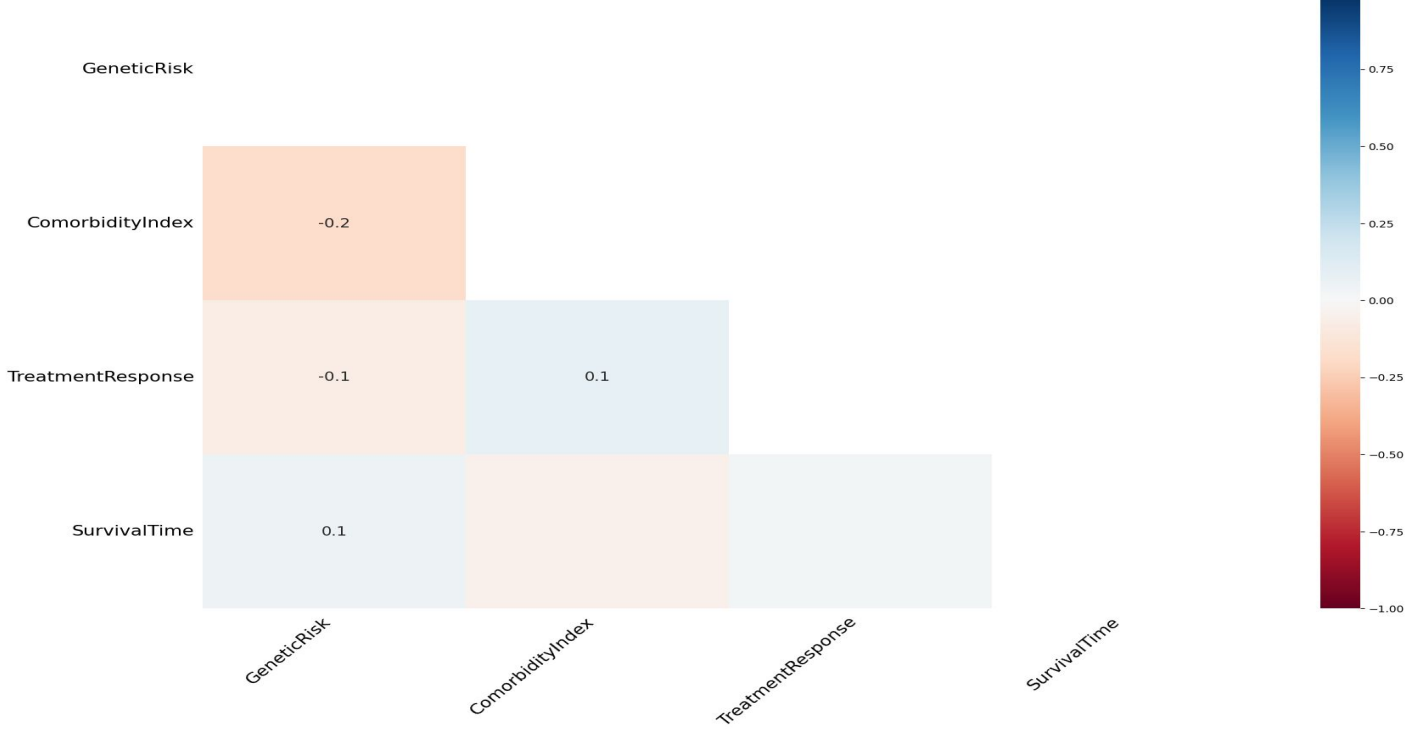
-Some data visualization
-Data preparation
-Chose a split analysis
-Define a metric



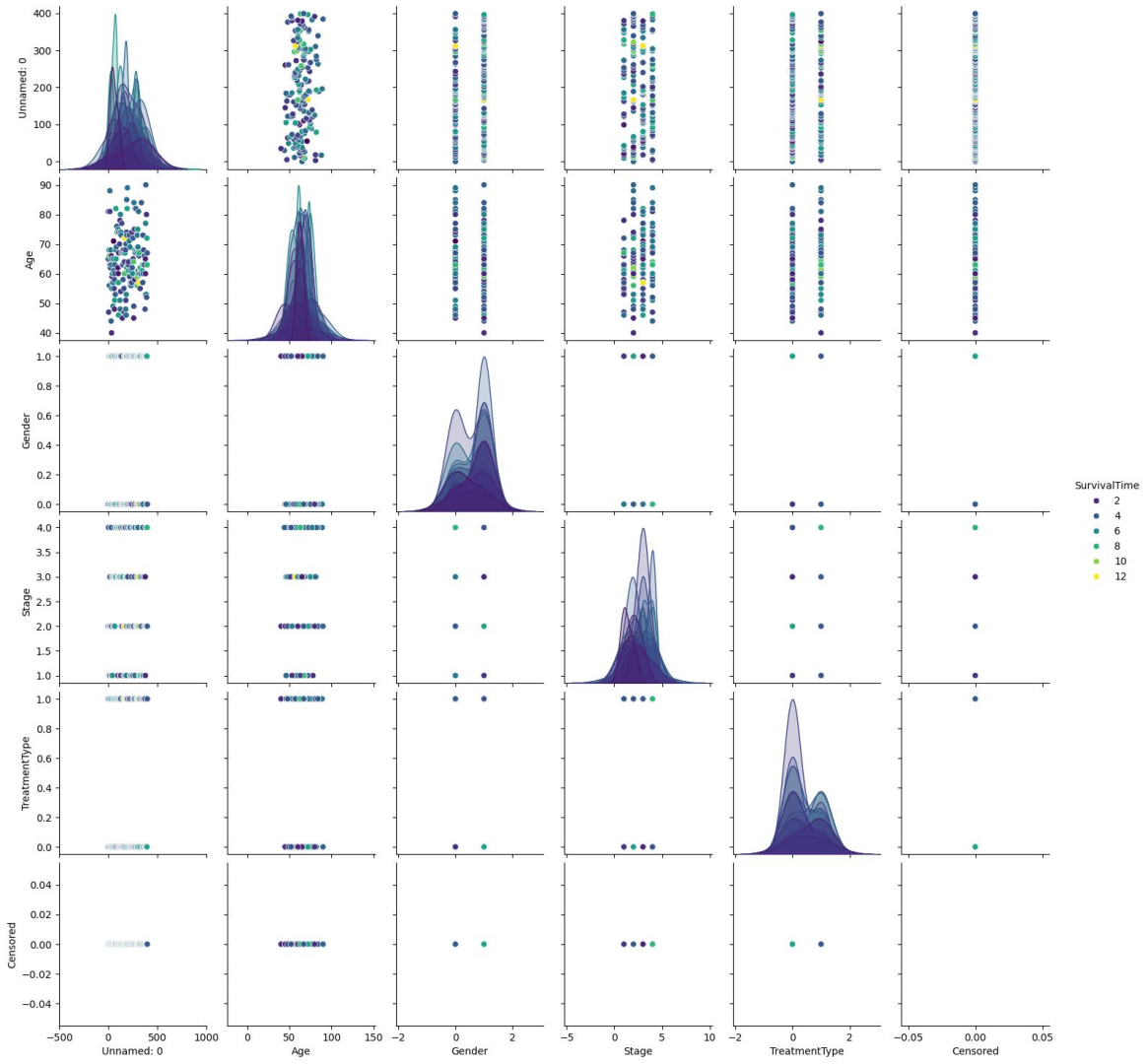
We can clearly see that there is missing data for GeneticRisk, ComorbidityIndex,TreatmentResponse and Survival time for the first three we are missing less than a ¼ of the data the most challenging column is Survival time where we are missing almost half of the data



By looking at the graphic we can’t observe any specific pattern for missing data meaning that at least by a qualitative analysis we can’t conclude that there is any pattern relating missing values between different columns.



Looking at the heatmap we do not see any significant correlation between variables



Data distribution is more or less what we would expect and apart from that we can extract any clear information Surivival time doesen't seem to present any specific pattern as it is evenly distributed.

Cross-validation is more data-efficient because it reuses the training data multiple times for validation, maximizing data usage and reducing variance. A **train-validation-test split** reserves a fixed portion of data for validation, which is less efficient, but works well for larger datasets

Overall assessment

What went wrong

Accidentally loaded test data instead of train data, but appart from that everything was pretty straight forward

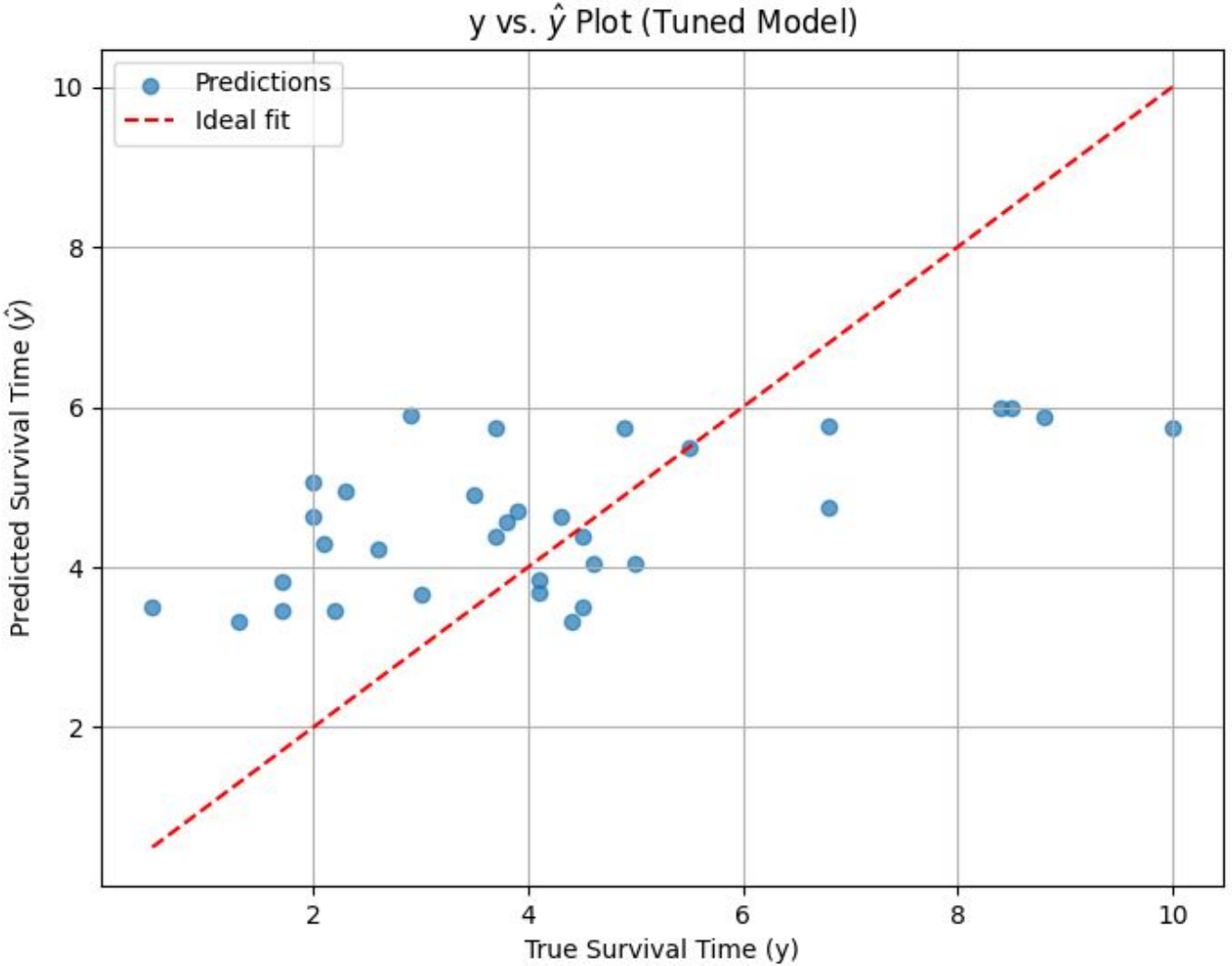
What went great

Libraries where installed without any difficulty and plots where as desired with much effort.

Task [1.2]

What was done in task 1.2

Development of a baseline model Using linear Regression as well as the required assesment.



The model

- Performs as good as it could be spect for a linear Regression model, we can see that it struggles a bit with more extrem values as it only predicts for values between 4 and 6 while the true values span from 1 to 10

What went wrong

I small error was made with the training data I miss understood the assignment indication and input the missing values with the mean of all the column this error was discoverd too late to modify it still I think it is a good baseline model as in fact computing the mean was only a line of code and it provides a good way to compare with other models so I considered that there is no major reason to modify anything all the models evaluated later will also use this imputed data.

What went great

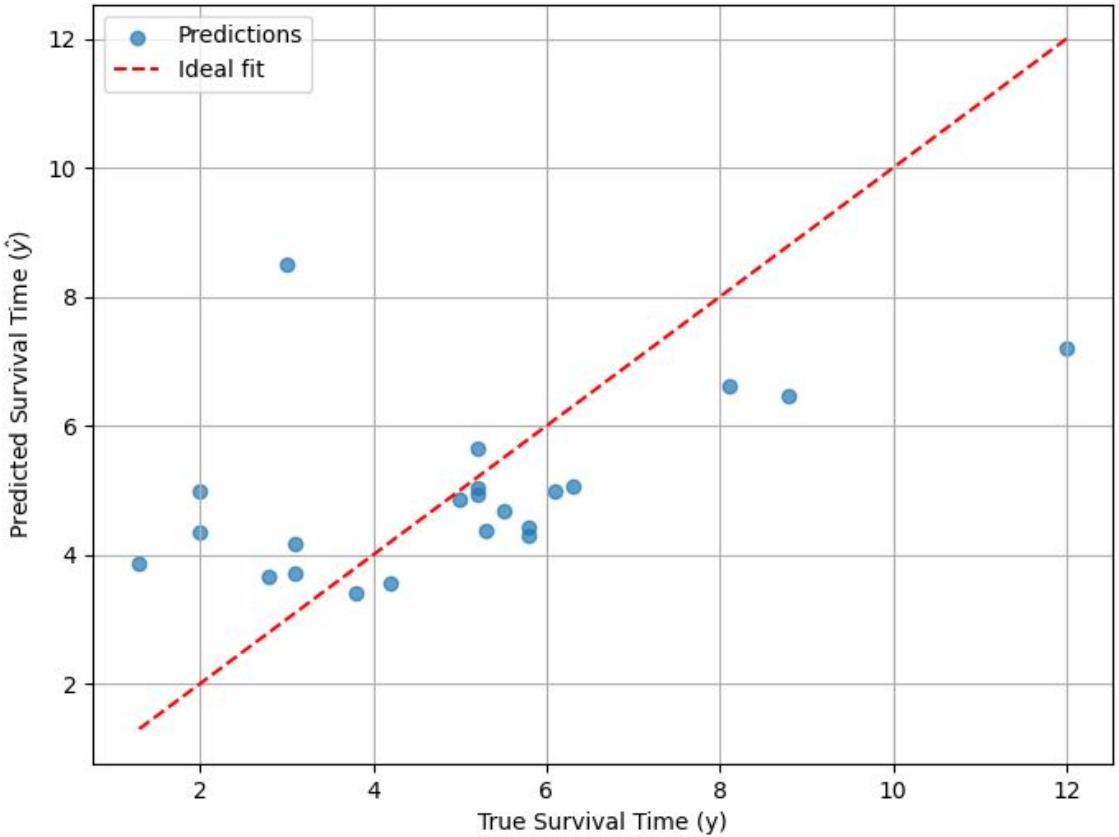
The model performed as expected with any major bug and a very essay implementation(Practice from first assignment is helping here :)

Task 2

What was done in task 2

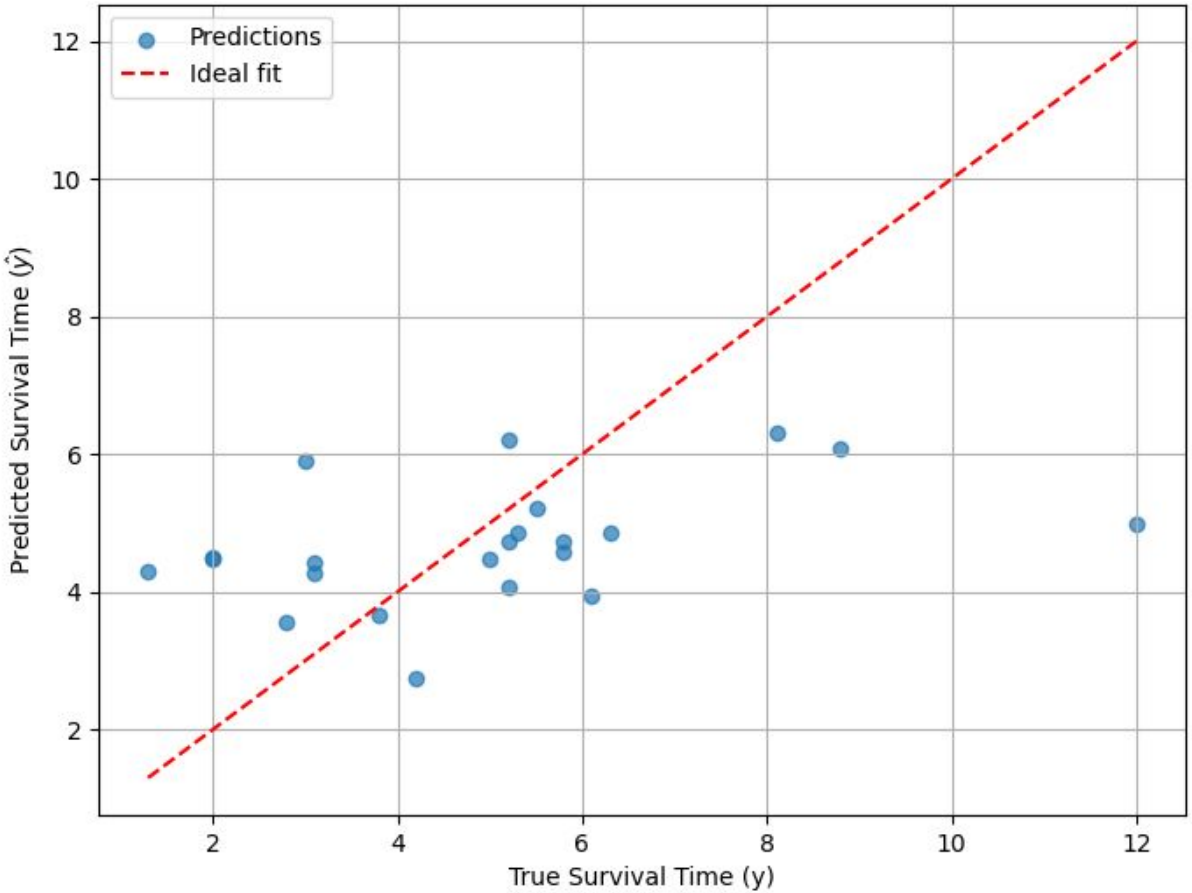
Training of polynomial and k-nearest neighbour models

Polynomial Regression



The Polynomial model performs better than the baseline but not much better with an cMSE of 3.7 and the baseline has an cMse of 3.6

Polynomial Regression



The model does not perform quite well with an cMSE of 4.9

Overall assessment

What went wrong

The Knn model performs quite bad with results much worse than the baseline model making it unuable

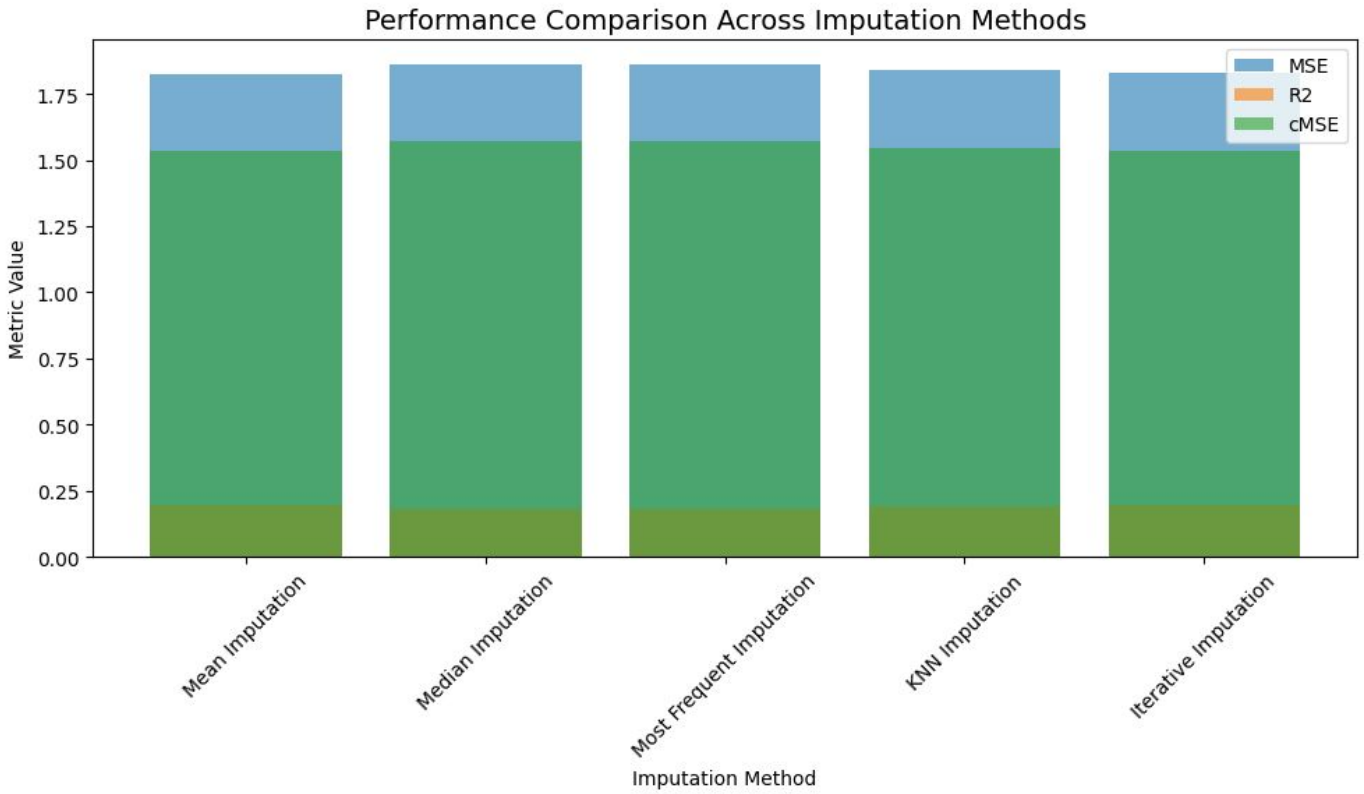
What went great

The polynomial model performs better although the improvement are not huge is a better option than the baseline

Task 3.1

What was done in task 3.1

Different imputations methods where tried and compared between them using the baseline model as reference to decide which is more suitable for our task



Here we have a table with the different methods and how the performed with diferent metrics

Imputer	MSE	R2	cMSE
Mean Imputation	1.8230473497084063	0.19815483914554333	1.5332010450800069
Median Imputation	1.8634115075250246	0.1804011562132828	1.5732714504194931
Most Frequent Imputation	1.8634115075250246	0.1804011562132828	1.5732714504194931
KNN Imputation	1.8386765645983667	0.19128051943603397	1.5454633144496115
Iterative Imputation	1.8278834394818329	0.19602774399198075	1.5340433526349362

For a better analysis we will use the raw data as we although surprising Mean imputations performs better in MSE and cMSE, as cMSE is our metric for this problem we will conclude that MEAN imputation is the best option we could have(this is the one accidentally we have been using for the baseline so ther is no difference at all)

What went wrong

Unexpectedly mean imputation has been the best possible imputation method making. Because we have accidentally been using this method the whole time this means there has been no improvment thanks to the imputation technique.

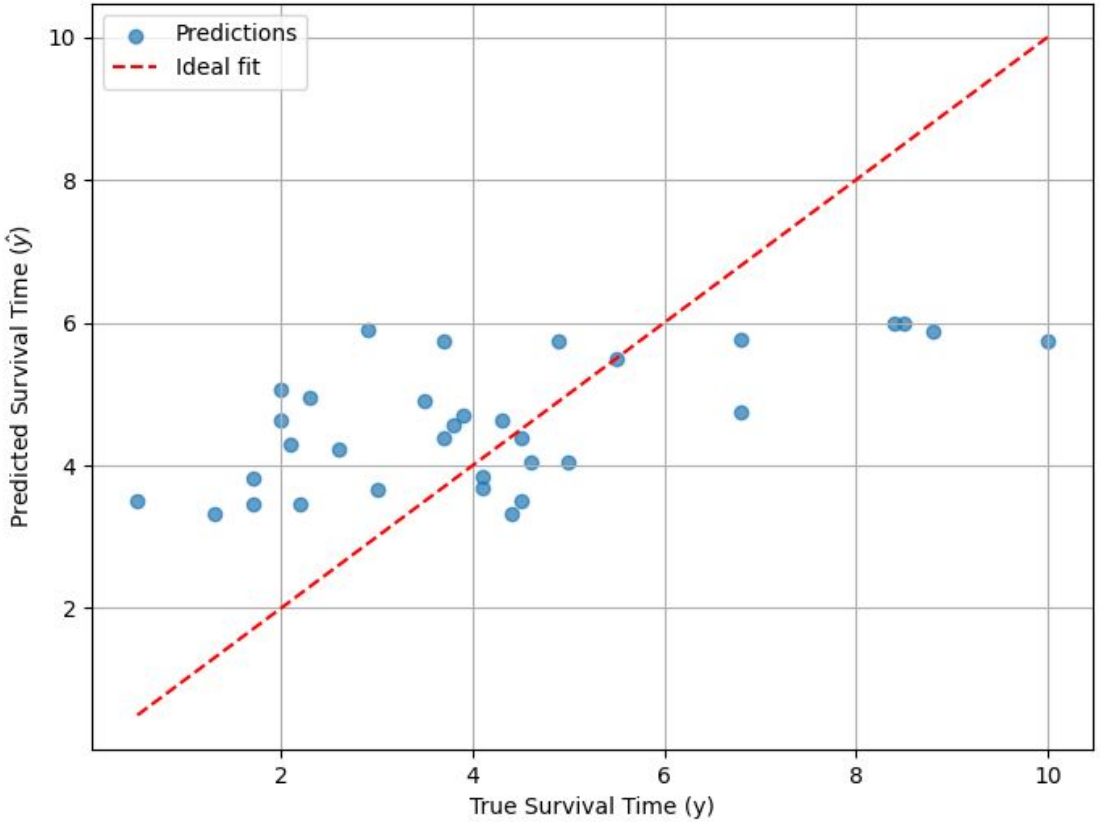
What went great

Although mean imputation being a very simple and essay method is the one that has proven to work the best which almost never happens but we are happy to have a very simple method to understand and implement working so well in our problem

Task 3.2

What was done in task 3.2

- HistGradientBoostingRegressor was implemented on missing data using hyperparameter tunning



cMSE is not really good at arround 3.6 meaning polynomial regression models still is the best performing one

What went wrong

The model has performed worst than the polynomial regression model, also due to problems with dependencies and pip it was not possible to test the catboost model

What went great

The implementation was very straight forward

Task 3.2

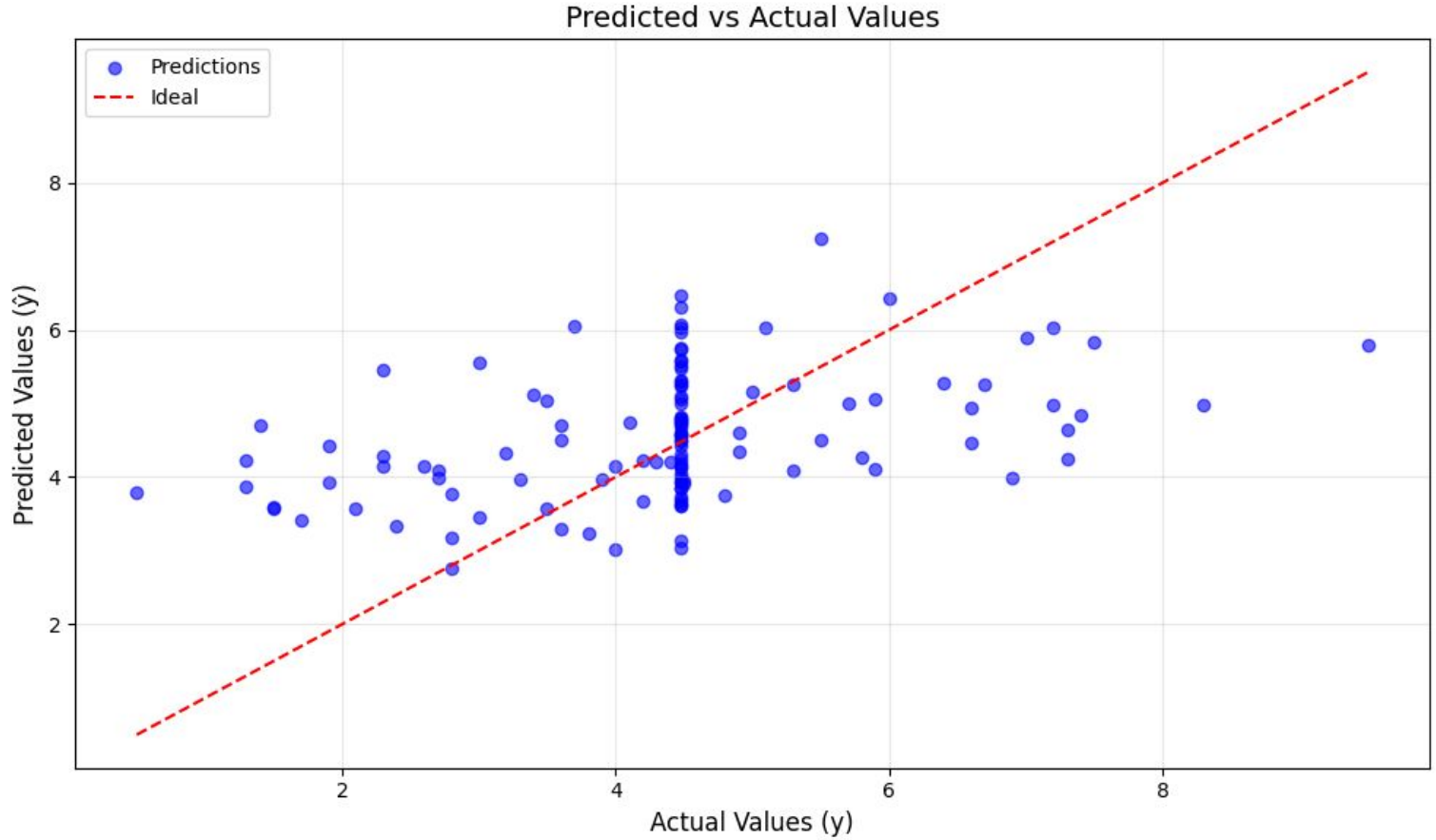
What was done in task 3.3

As we have already tested polynomial regression with mean imputation(the one that worked the best) and all other models performed much much worse I didn’t do a deeper analysis as it would be a complete loss of time

Task 5

What was done in task 5

Training a simple multilayer perceptron



There has been some problem with the training data is at appears to have many values as 0 but I didn't have time to solve it

What went wrong

cMSE is very high but I can’t relly on that information as it is clear that there’s a bug in the code in the y hat plot there a appear a lot of observations centerd at 0 which doesn’t make a lot of sense but i did not have enough time to solve it

What went great

Nothing at all :(