CMPE 257 Lab 1

Milica Miskovic

March 2019

Introduction

As stated in the EASL[1] Clinical Practice Guidelines, liver cancer is the fifth most common cancer and the second most frequent cause of cancer-related death globally. Hepatocellular carcinoma represents about 90% of primary liver cancers and constitutes a major global health problem. HCC has become increasing concern since the number of occurrences has been growing worldwide.

1 Task1 - Analysis

1.1 Data Structure

HCC dataset was obtained at a University Hospital in Portugal and contains several demographic, risk factors, laboratory and overall survival features of 165 real patients diagnosed with HCC. The dataset contains 49 features selected according to the EASL-EORTC.

This is a heterogeneous dataset, with 23 quantitative variables, and 26 qualitative variables.

1.1.1 Continuous variables

As stated in data description file, this data set has 23 quantitative variables, one of them being ‘Nodule’ (Number of Nodules), which take integer values from 1 to 5 and therefore can be treated as ordinal variable.

[1] European Association for the Study of the Liver

1.1.2 Categorical and ordinal variables

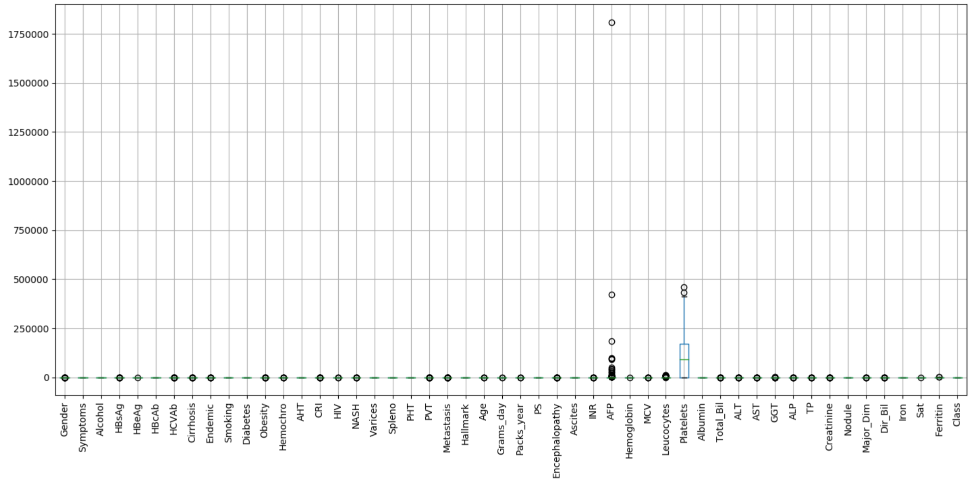
In this dataset we have 26 categorical variables out of which, four are ordinal ('PS', 'Encephalopathy', 'Ascites' and 'Nodule'). To use categorical variables, we have to encode them and make dummy variables and every categorical variable will result in n-1 dummy variables. In order to be useful in logistic regression model, these variables have to be 'Dummy coded', which will have them retain their categorical nature, otherwise the model would consider to be discrete numeric with values 0 and 1.

1.2 Analysis of data using df.info() and df.describe()

Method *data.info* provided us with insights about a DataFrame including the index dtype and column dtypes, non-null values and memory usage.

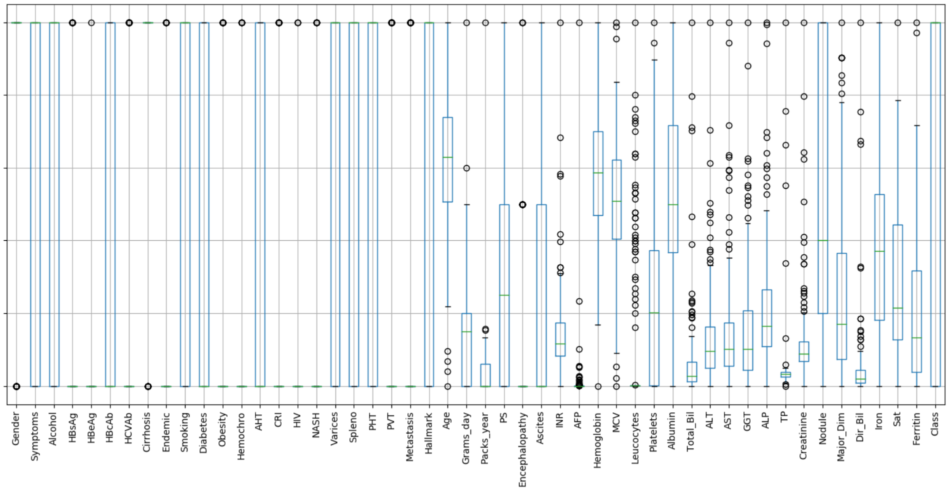
Method *data.describe()* statistical summary of the central tendency measurements, dispersion and distribution of the data, excluding NaN values, as described in method documentation.

Based on the data described, we have following box plot, which is not very informative before min-max standardization of data columns, but as we can see there is an obvious outlier in AFP feature column.



Min-max standardization preprocessing method will homogenize means of features and mitigate influence of outliers on the model.

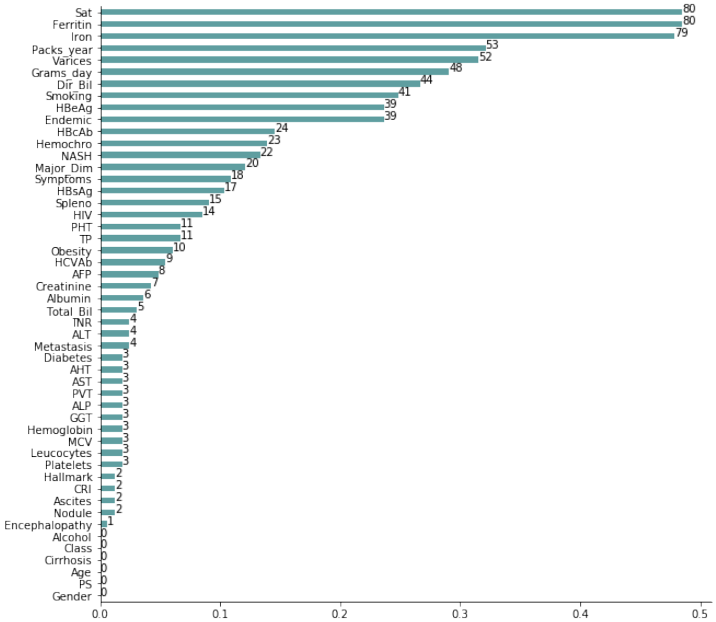
We can also see that normality assumption is violated in some of the continuous variables.



We have 50 columns, 165 entries and our columns have between 165 and 85 observations. Data types are float64 (44 variables), int64 (6 variables). Dataset takes 64.5 KB.

1.3 Null values and missing data analysis

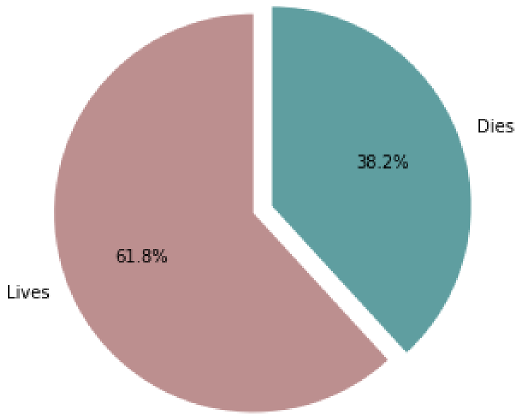
As stated in data description, this data set has 10.22% of the whole dataset and only eight patients have complete information in all fields (4.85%).



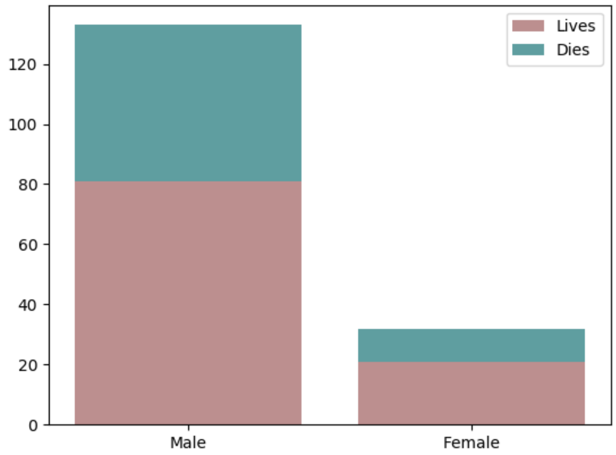
The data set also has 6 complete variables, and total of 27 variables with less than 5% NAs, the amount that is considered to be inconsequential. The potential problem lies in three variables that have over 40% missing data (Iron 79, Ferritin 80, Sat 80 out of 165 in total) although there is no established acceptable cutoff. Imputation (filling in the missing data with predicted values) in cases like this one is not straightforward and can damage data value and quality of statistical inferences. Imputing a fixed value (mean, median, zero or min/max) would introduce considerable bias to the dataset and using imputation methods based on inference would result in data set with heavily correlated features.

1.4 Patients

As already stated, in this data set, we have 165 patients.

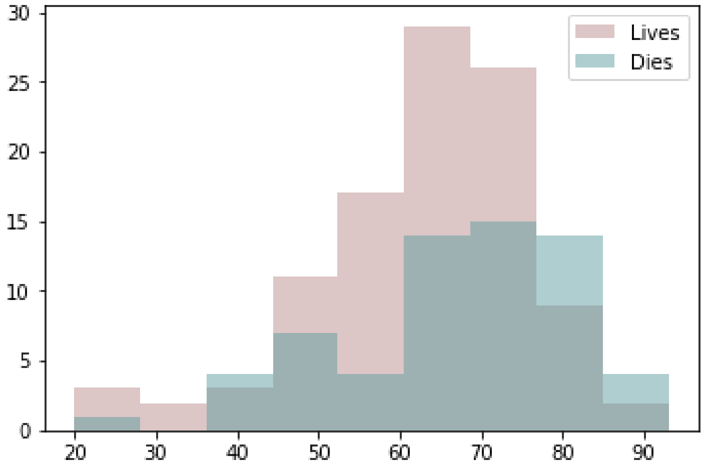


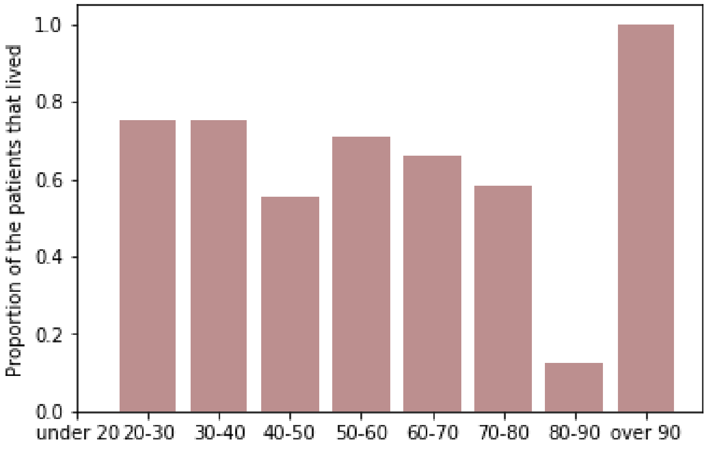
The sample is imbalanced gender wise, it consists out of larger proportion of men (61.81%, or 102 patients) but similar proportion of male and female patients that live or die (slightly more women lived). We have 81 male patients that lived (79.41%) and 21 females (20.59%), similarly, 52 male patients died (82.54%) and 11 (0.17%) of female patients died.



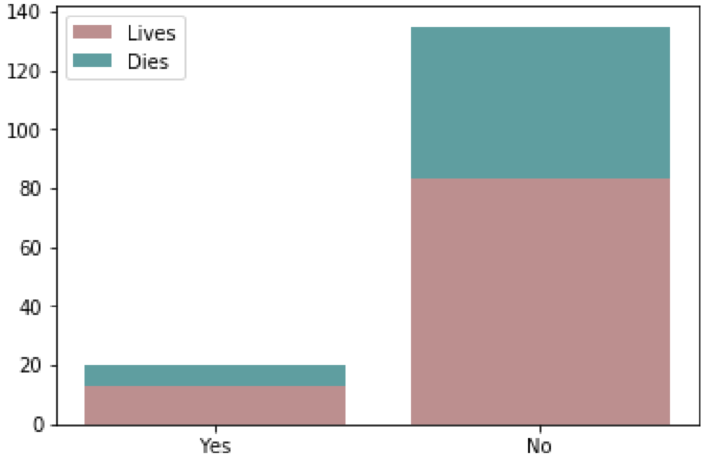
The average in years is 64.69, with youngest patient number 29 being 20 years old and the oldest patient number 100 being 93 years old (both have lived). Mode age (the age that is most frequent) is 71. Median age is 66.

If we split patients into age groups, the most populated one would be 60-70 year-old group with 56 patients, and that is also the group that has the highest number of people that lived (37 patients lived). Interestingly, the group with highest proportion of patients that lived is 90-and-older, since there is only one patient and he lived.

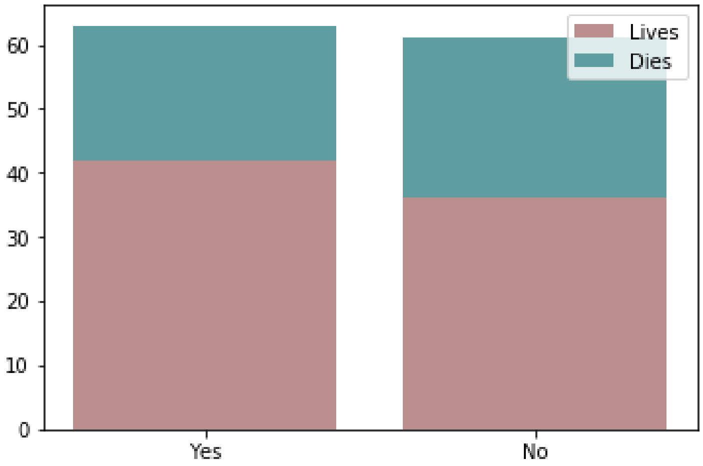




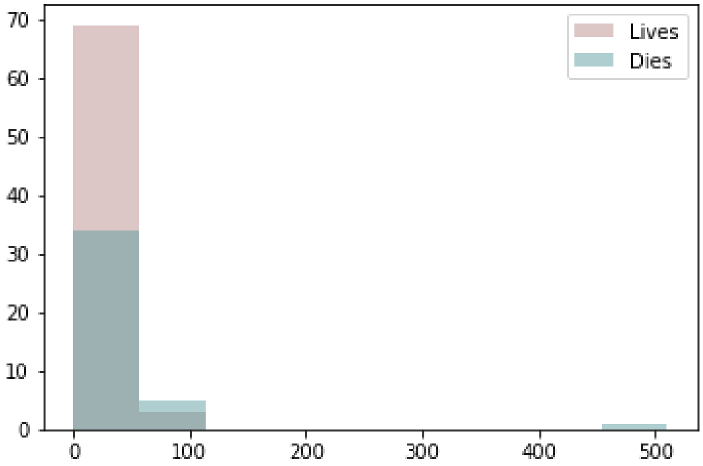
Data is not balanced in terms of obesity either. We have 20 obese patients, out of which 13 live. Out of 135 patients that are not obese, 83 live, which brings us to paradox situation that higher percentage of obese people lived (65% in comparison to the 61.5% of non-obese people).



Much more balanced is the distribution of smokers among our patients. Total number of smokers is 63, non-smokers 61. Out of smokers, 42 lived, which is 53.85%, and 36 or 46.15% out of non-smokers, which is paradoxical but moderately uniform, therefore it can be subject of sampling.

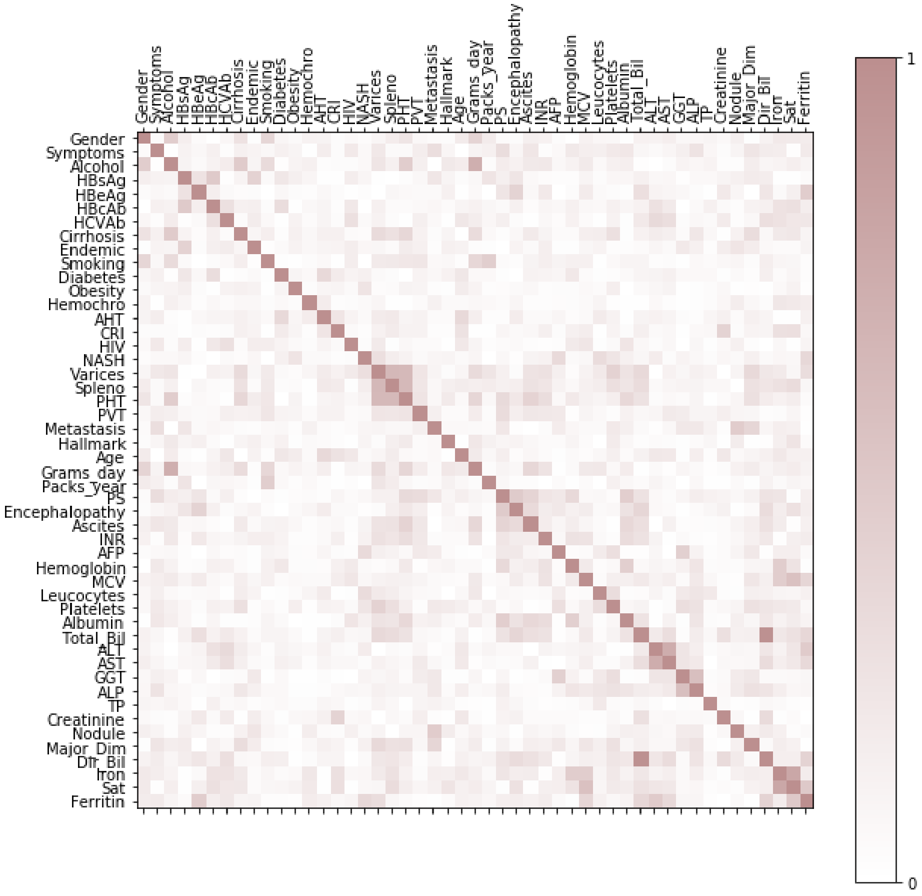


In terms of smokers, our average patient would smoke 20.46 packs of cigarettes a year. Majority of our patients smoke between 0 and 50 packs a day and almost double number lives than dies, in category 50-100 packs a day 30% more people die than live. This variable has an obvious outlier, patient number 128, who smokes 510.0 packs of cigarettes a year and he died during this research.



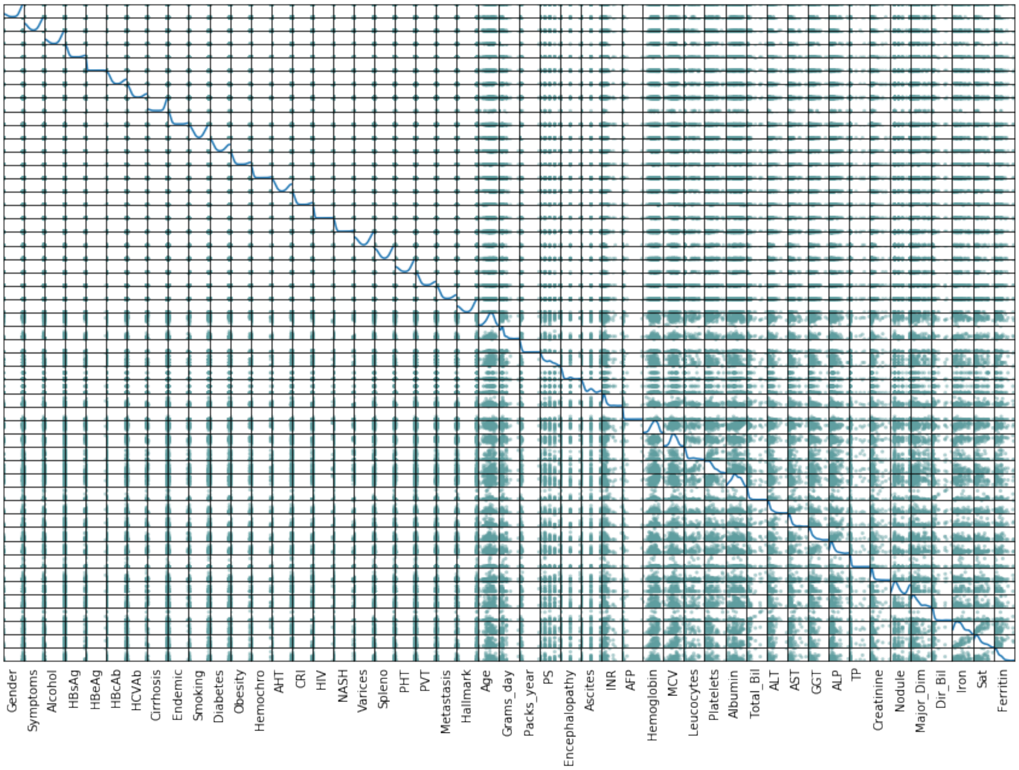
1.4.1 Data distribution

Similarly, to the continuous variables overall data features seem to be reasonably correlated. The most correlated features are direct bilirubin (Dir_Bil) and total bilirubin (Total_Bil) with correlation coefficient of 0.98.



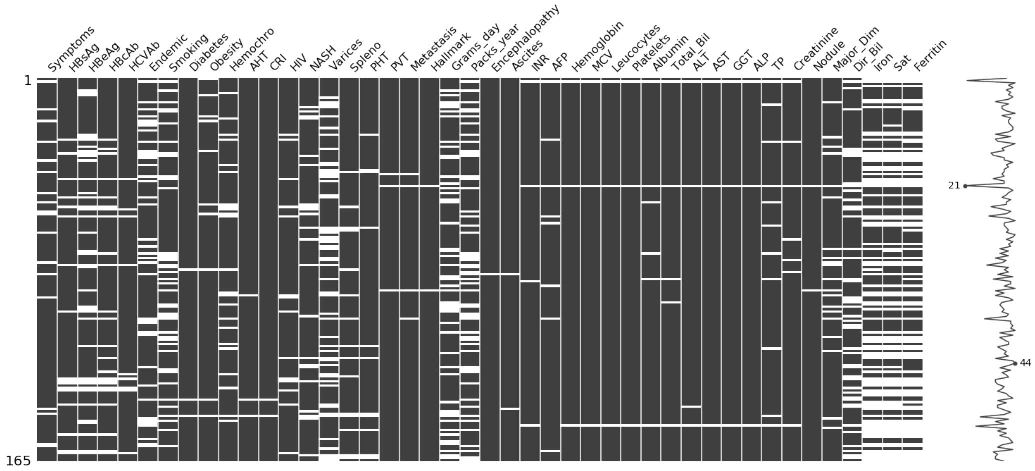
Scatter plot of features seem not very informative, since the number of variables. Exploratory data analysis has to be done on smaller feature sets, or even pairwise.

Data feature selection process has to begin with plotting correlation of our dependent variable ‘Class’. The correlation coefficients will give us some intuition about role that every feature can play in our final model, but we can’t take it as is since in model we will not use only one feature to produce the output, but several and we don’t know how model will perform because of feature interactions and multicollinearity (correlation among the columns). Therefore, some feature selection procedure (as information value method, step wise selection, recursive feature elimination etc.) is advised. More about feature selection for this project will be said in part m) of this report where final regression model will be implemented.

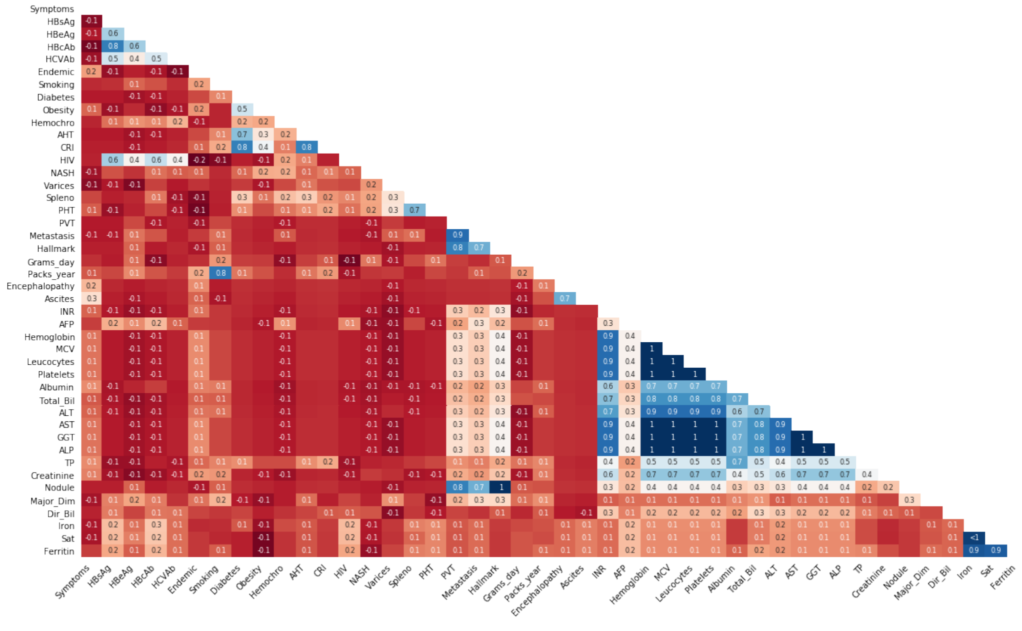


1.4.2 Handling missing values

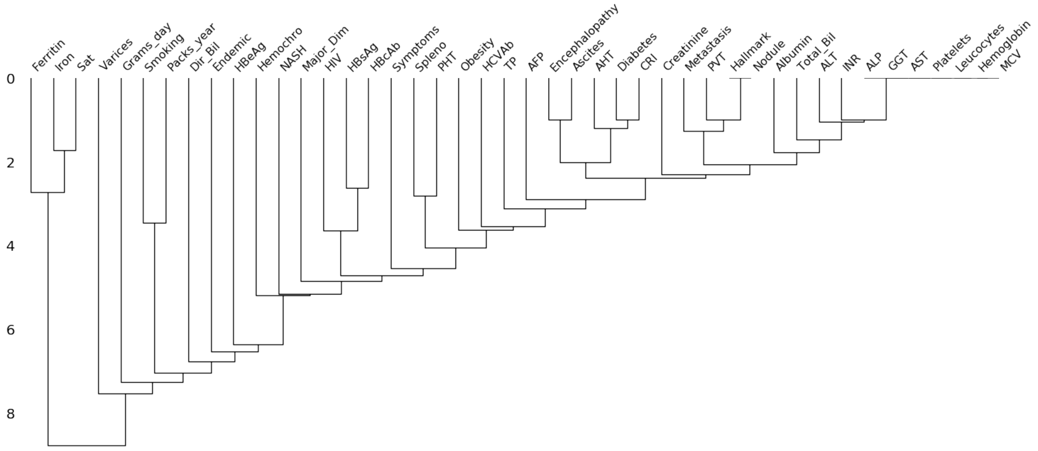
Map of missing value pattern shows us that there are no obvious missing data patterns. Also, the sparkline on the right gives us a general shape of data completeness and indicates patient 46 (counting from 0 to 164) as the most incomplete entry (with 21 missing features), and two more patients with high number of missing data (patients 1 and 149) and so, those observations would be wise to omit from the further inference process.



Correlation coefficient can be presented by a heatmap. Correlation coefficient of 1 would indicate that two features have the same indices of missing values, and -1 the opposite (in row that one variable is missing, the other is non-missing).



Missing value dendrogram depicts the correlation between features in terms of missing data. The lower cluster of variables ALP, GGT, AST, Platelets, Leucocytes, Hemoglobin and MCV indicates that pattern of missing values is closely correlated.



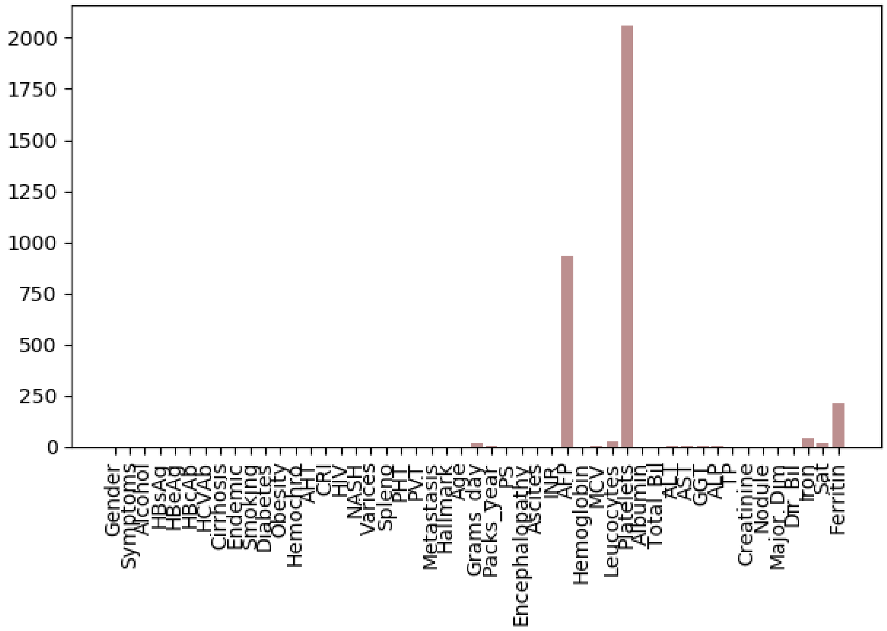
DropNA

This method will omit all data rows (patients), that have at least one measurement missing (at least one NA). Our data set will be reduced to the 8 patients with complete entries. This outcome would result in unnecessary data loss which will probably have underfitted model consequently

Replacement of missing values with zero or max value

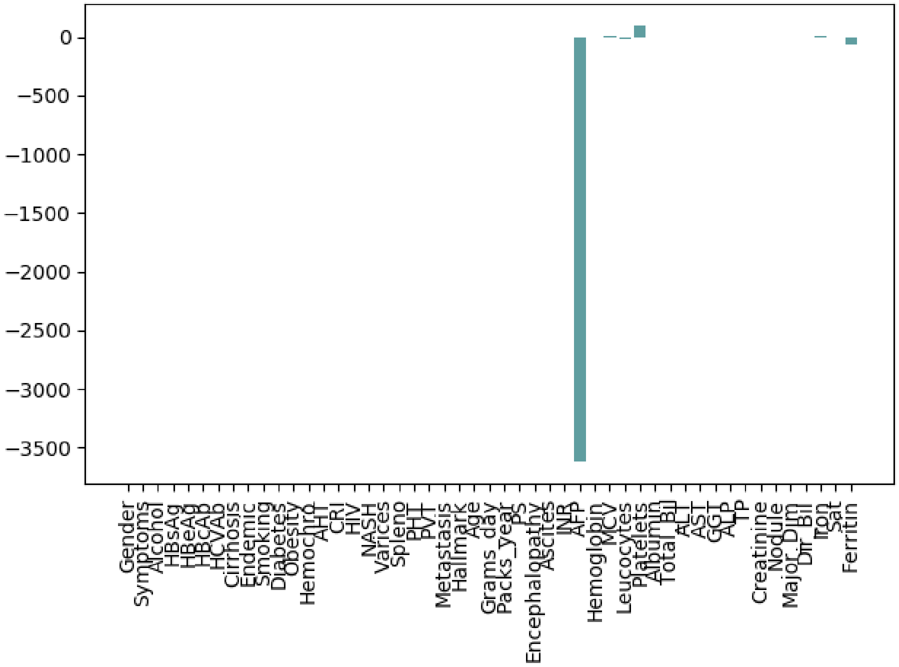
This method will produce highly biased data in comparison to the original one. By imputing zero since we have roughly half of the variables being binary categorical, the distribution to he said variables would be greatly violated and moved towards negative outcome (binary coding is done so 0 indicates negative outcome). For the continuous features, mean will be moved towards zero, proportionally to the amount of entries missing.

Imputation of max value will have similar effect to imputation with zero, but only in opposite direction, since max value of binary categorical variables is 1, and that is code value for positive outcome. Moving of means of the continuous variables, as per expected, will be in the opposite direction.



Replacement of missing values with column mean

This imputation method even though does not move the centroid of the data, in datasets that have considerable proportion missing as does this one, can reduce variability of the data, make our confidence intervals narrower, but in this case, it can change relationships of the variables and therefore change the slopes of our regression planes which can result in small generalizability of our model.



Other Imputation methods

Since the proportion of missing data in our set is not inconsequent, imputation is a non-trivial task. Approach that will preserve inferential power of the data has to assure non-constant imputation and effectiveness of all regression-based methods is questionable since they can introduce even more multicollinearity, which is no desirable if method of choice is a logistic regression.

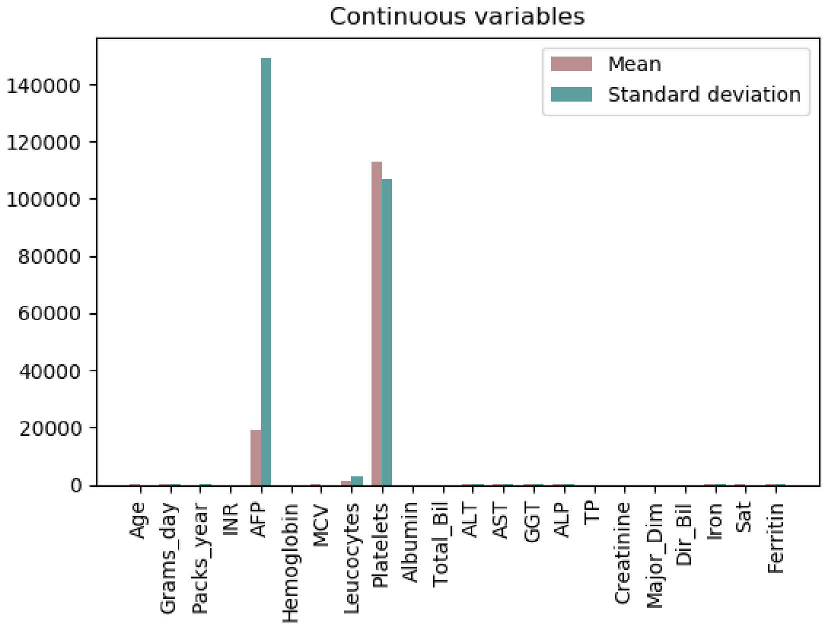
Multivariate imputation by chained equations (MICE) is a method runs multiple regression models for each variable (according to the distribution of the variable) and calculate the missing values by averaging the interpolated regression values for every missing data value. Even though it is considered to be very effective method, it is regression-based and therefore can introduce some amount of correlation between features.

KNN imputation[2] will impute values based on K nearest neighbor clustering method. This method will be used to impute missing values in data set with choice k=10. The rest of the missing values were imputed by mean

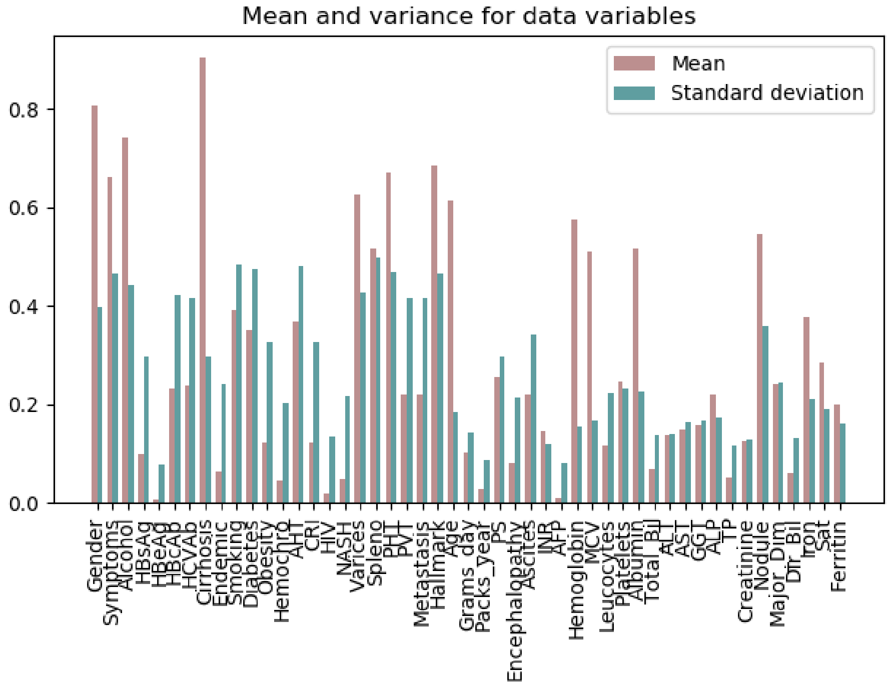
[2] https://gist.github.com/YohanObadia/b310793cd22a4427faaadd9c381a5850

value.

1.5 Standardization of the data

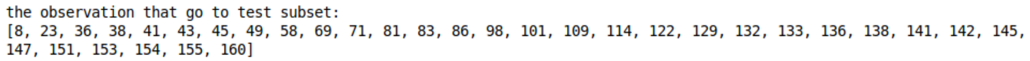


Since we have means and variances of the continuous variables 'APF' and 'Platelets' gravely out of range in comparison with the other variables, min-max scaler function was employed, and means and variances are settled in the interval between zero and one.



1.6 Training and testing data subsets

Data set of features was split into training and testing subsets in proportion 80% to 20%, which gives us 33 observations in test set. Following patients ended up in test group for our model.

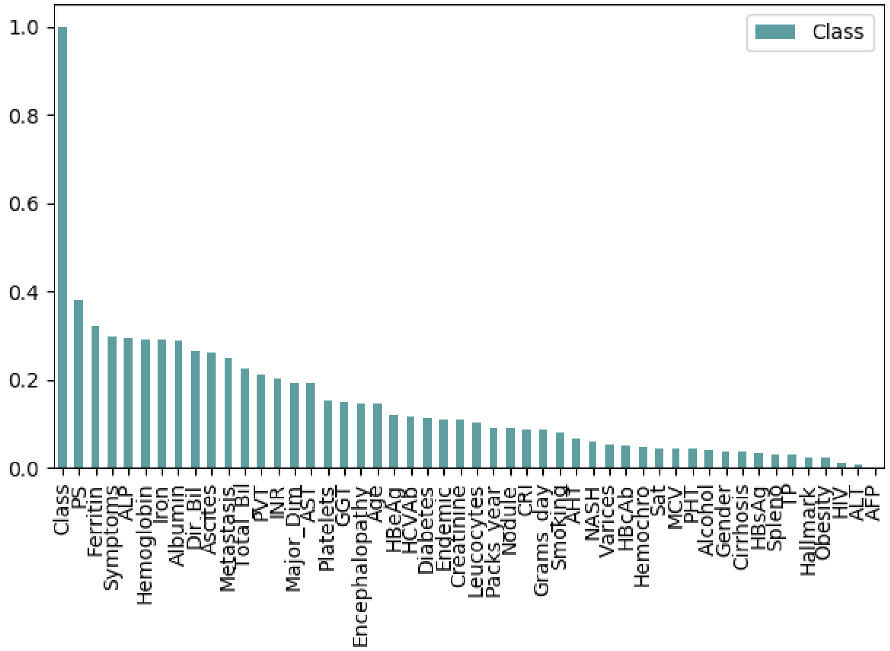


```
the observation that go to test subset:
[8, 23, 36, 38, 41, 43, 45, 49, 58, 69, 71, 81, 83, 86, 98, 101, 109, 114, 122, 129, 132, 133, 136, 138, 141, 142, 145,
147, 151, 153, 154, 155, 160]
```

1.7 Logistic regression model

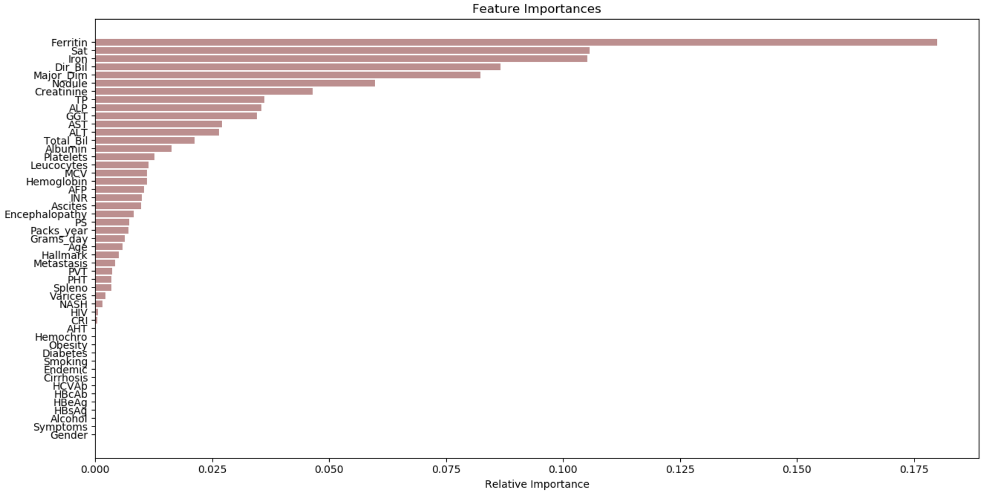
1.7.1 Feature selection

Even though we have 49 features available, only some of them will carry value for the inference. The goal is to make as simple model as possible considering the accuracy. In order to get the ‘rule of thumb’ for variable selection, we have to look into the correlation values of our response variable ‘Class’ and the rest of the predictor variables.



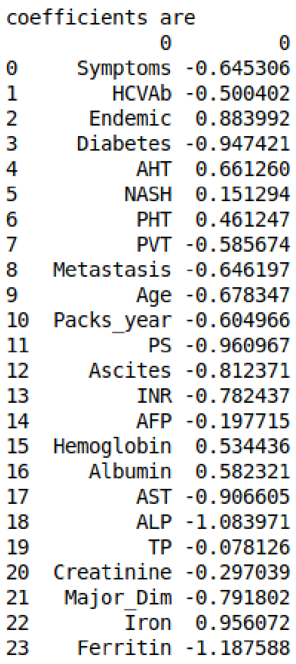
The greatest absolute correlation coefficient variable ‘Class’, as per expected, will have with itself. Than, variables ‘PS’, ‘Ferritin’, ‘Symptoms’ and ‘ALP’ have correlation coefficient around 0.3.

Other solid indicator in feature selection is Random Tree classifier method called ‘Importance’. Although random trees and linear regression are two completely different methods, the graph of feature ‘Importances’ is useful on the informative level.



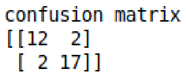
As we can see, similar features have considerable values for both correlation and 'Importances', therefore make good candidates for our model too.

Recursive feature elimination (RFE) is the feature selection method that will start with one randomly picked feature and then randomly add and remove features until it produces the model with maximum accuracy. Since it picks features at random, as with most of the random based methods, local maximum doesn't guarantee global maximum, therefore selected features may be different every time we run the method, even though it is on the same data set. RFE method suggested 24 features that are used to model the data. Logistic regression coefficients of those features are given in the table.



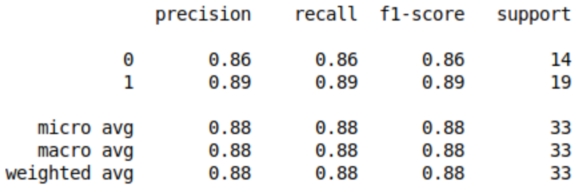
```
coefficients are
              0         0
0     Symptoms -0.645306
1        HCVAb -0.500402
2      Endemic  0.883992
3     Diabetes -0.947421
4          AHT  0.661260
5         NASH  0.151294
6          PHT  0.461247
7          PVT -0.585674
8   Metastasis -0.646197
9          Age -0.678347
10  Packs_year -0.604966
11          PS -0.960967
12     Ascites -0.812371
13         INR -0.782437
14         AFP -0.197715
15  Hemoglobin  0.534436
16     Albumin  0.582321
17         AST -0.906605
18         ALP -1.083971
19          TP -0.078126
20  Creatinine -0.297039
21   Major_Dim -0.791802
22        Iron  0.956072
23    Ferritin -1.187588
```

On given test set out model performed reasonably well.



```
confusion matrix
[[12  2]
 [ 2 17]]
```

We correctly predicted 12 as positive and 17 as negative.

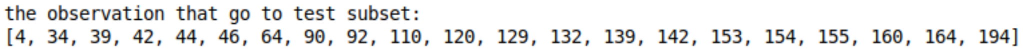


```
              precision    recall  f1-score   support

           0       0.86      0.86      0.86        14
           1       0.89      0.89      0.89        19

   micro avg       0.88      0.88      0.88        33
   macro avg       0.88      0.88      0.88        33
weighted avg       0.88      0.88      0.88        33
```

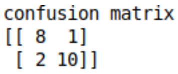
Precision is 86% for patients that died (0) and 89% for patients that lived (1) through the research. Recall was 86% for 0 and 89% for 1. Which gives 0.88 f-score on 33 points test set.

2 Task2 - Model performance on full dataset

Same model from the previous section was used on the full dataset. After loading the data, it was separated into two subsections (training set and test set). 90% of data was used as training set, whereas 10% of data was used as test set. Which gives us 21 following observations in test set.

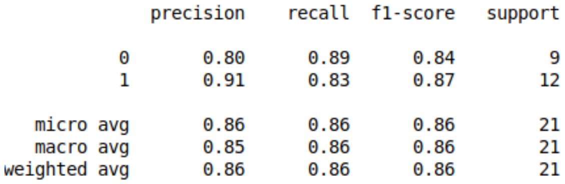


```
the observation that go to test subset:
[4, 34, 39, 42, 44, 46, 64, 90, 92, 110, 120, 129, 132, 139, 142, 153, 154, 155, 160, 164, 194]
```



```
confusion matrix
[[ 8  1]
 [ 2 10]]
```

We correctly predicted 8 as positive and 10 as negative.



```
              precision    recall  f1-score   support

           0       0.80      0.89      0.84         9
           1       0.91      0.83      0.87        12

   micro avg       0.86      0.86      0.86        21
   macro avg       0.85      0.86      0.86        21
weighted avg       0.86      0.86      0.86        21
```

Precision is 80% for patients that died (0) and 91% for patients that lived (1) through the research. Recall was 89% for 0 and 83% for 1. Which gives 0.86 f-score on 21 points test set.

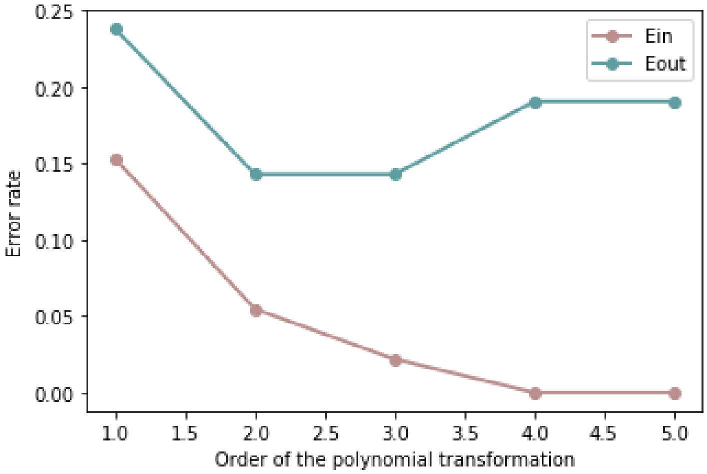
The possible reasons for good performance of model lie in fact that imputation of the missing data was done by KNN method which is random based and values may vary.

Also, picking training and test sets was at random, therefore, fluctuations between consecutive runs are possible.

Loigstic regression should not be method of choice for this dataset due to the fact that normality of the data is questionable, lots of missing data in some feature columns and large number of categorical variables (which would imply some method for categorical data analysis, rather than logistic regression).

3 Task3 - Polynomial transformation

Polynomial transformation from 1^{st} to the 5^{th} order was applied and in sample and out of sample error was calculated and recorded. As per expectations, second and third order polynomial transformations resulted in decrease in both Ein and Eout. Since data is not linear, this kind of behavior of the model was expected.



Two error rates are closest on second degree of polynomial transformation, after which Ein continues to decrease and Eout increases.