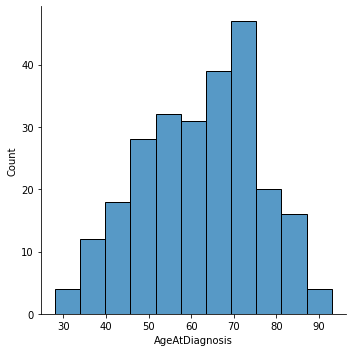
REPORT

The problem statement here is predicting the age of the person at the time of diagnosis based on the genes activity data.

The genes activity data has 24711 features and 251 samples. The target variable here is “AgeAtDiagnosis”. We load the data and read the data in python using the pandas dataframe. The target variable is present in the different file named “Clinical Matrix”. We added the target variable in the genes activity dataset from that file. The original data from the genes activity dataset is not in a proper format. We take the transpose of that data so that we can get the required format which is 24711 features and 251 samples.

We check for null values in the dataset. We see that there are no null values in the dataset. We then check the distribution of our target variable.



After checking the distribution of target variable we see that there are no outliers in it. The data is completely valid. The next step after this is data normalization. There are lot of features which have values with different ranges. We need to normalize the data so that all the features in the data are in a similar range.

Since there are 24711 features in the data, many features in the data will be redundant and would not contribute anything to the data. If we give all the features to a Machine Learning model then it will create a very complex algorithm to understand the data and would not give accurate results. We need to select only those features which are important in predicting the target variable. In order to do this we will be using “ExtraTreeRegressor” from sklearn in python. “ExtraTreeRegressor” has the functionality to calculate feature importance for each feature. After applying the “ExtraTreesRegressor” we eliminate the features which have zero feature importance. Now we have 2928 features available.

Using these features we will create the datasets for training and testing. We split the data into train and test sets. We use 80% data for training and 20% for testing purpose. After the splitting of data we will be using 4 Machine Learning models to predict the target variable. Since this is a regression problem we will first use Linear Regression. Then we will also use Random Forest, XGBoost and SVM regressor. We will evaluate the performance of the models using a metric “R-squared value”.

|  |  |
| --- | --- |
| **Models** | **R-squared Score** |
| Linear Regression | 0.460 |
| Random Forest | 0.230 |
| XGBoost | 0.161 |
| SVM Regressor | 0.022 |

We see that Linear Regression outperformed other models with a 46% R-squared score. The reason the R-squared score of the models is very less because we do not have enough data samples. There are more features than the number of samples in the data. We need more data so that the models can generalize well and produce good results.