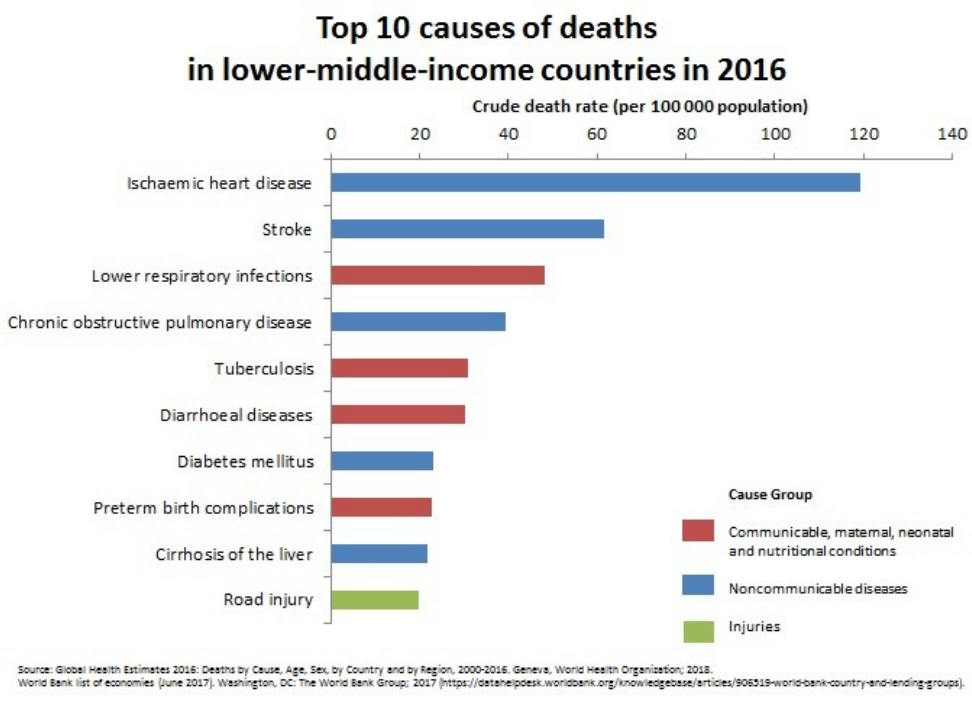
Liver Disease Prediction Using Machine Learning

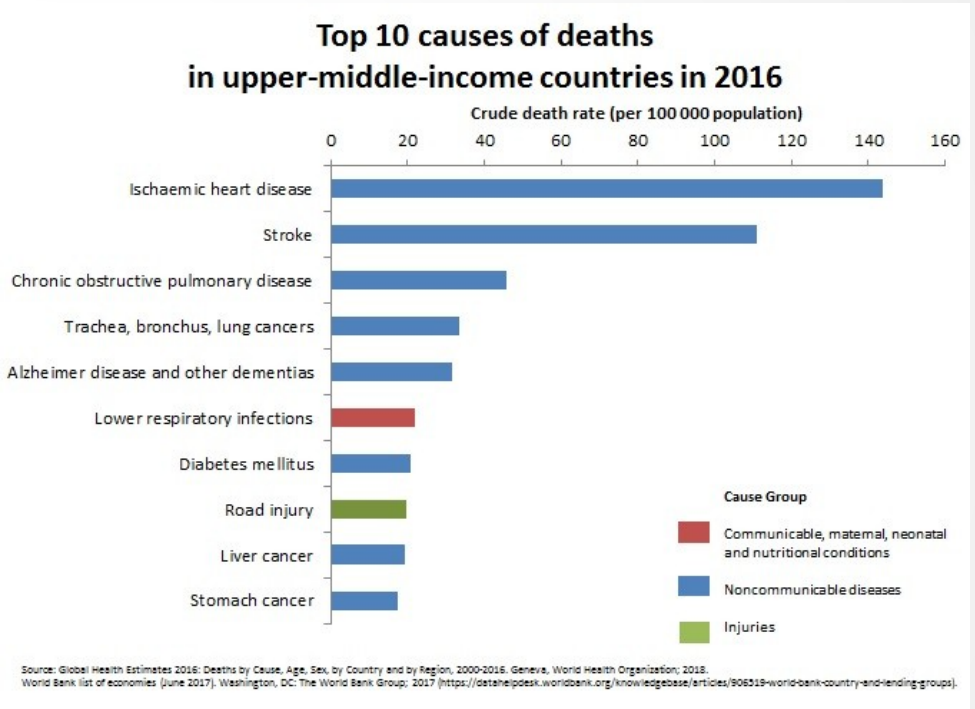
First Springboard Capstone Project

Objectives and Goals

- Predicting whether a patient has liver disease or not based on set of records can significantly reduce burden on doctors in an effort to correctly identify liver disease
- By applying predictive machine learning algorithms against the patients dataset, we can solve the problem of identifying such patients
- Overall, liver disease caused death rate is among top 10 death causes in the lower-middle income and upper-middle income countries in the world, which is caused by a number of factors, such as diet, alcohol consumption and smoking
- Thus, having an accurate model for predicting patient liver disease on early stages based on their records can significantly improve the diagnosis and help in early disease preventive cares

Liver Disease Caused Death Rate



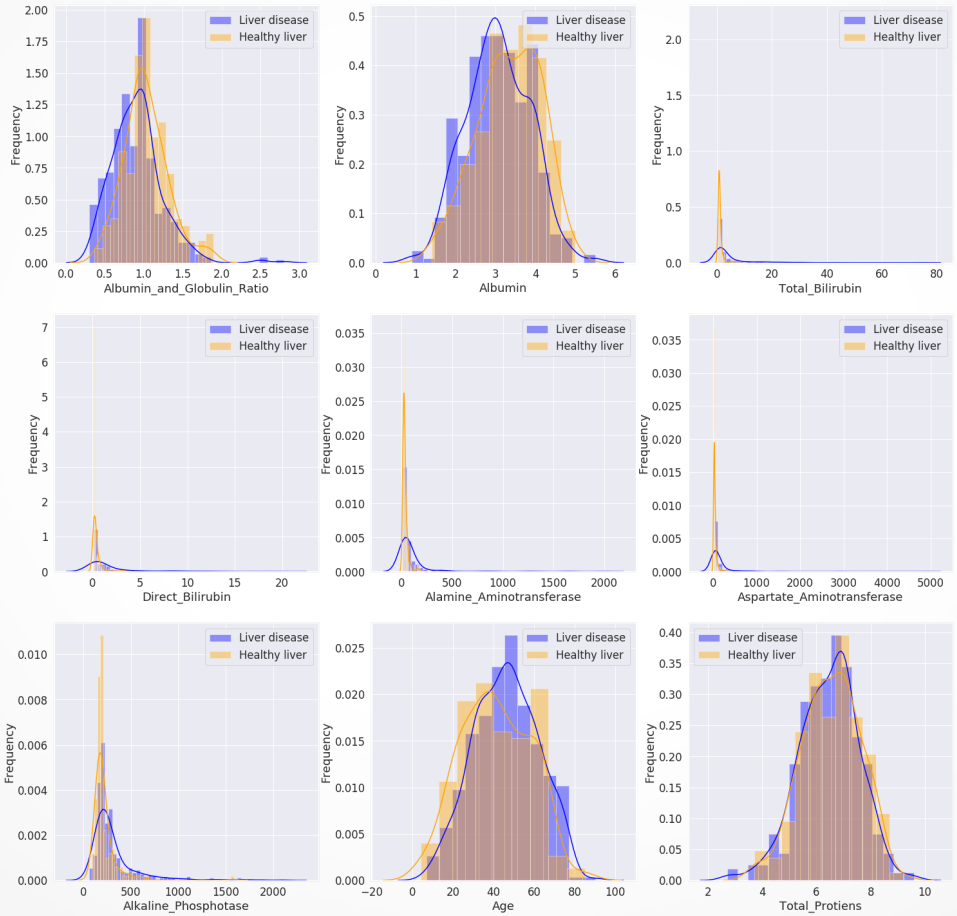


https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

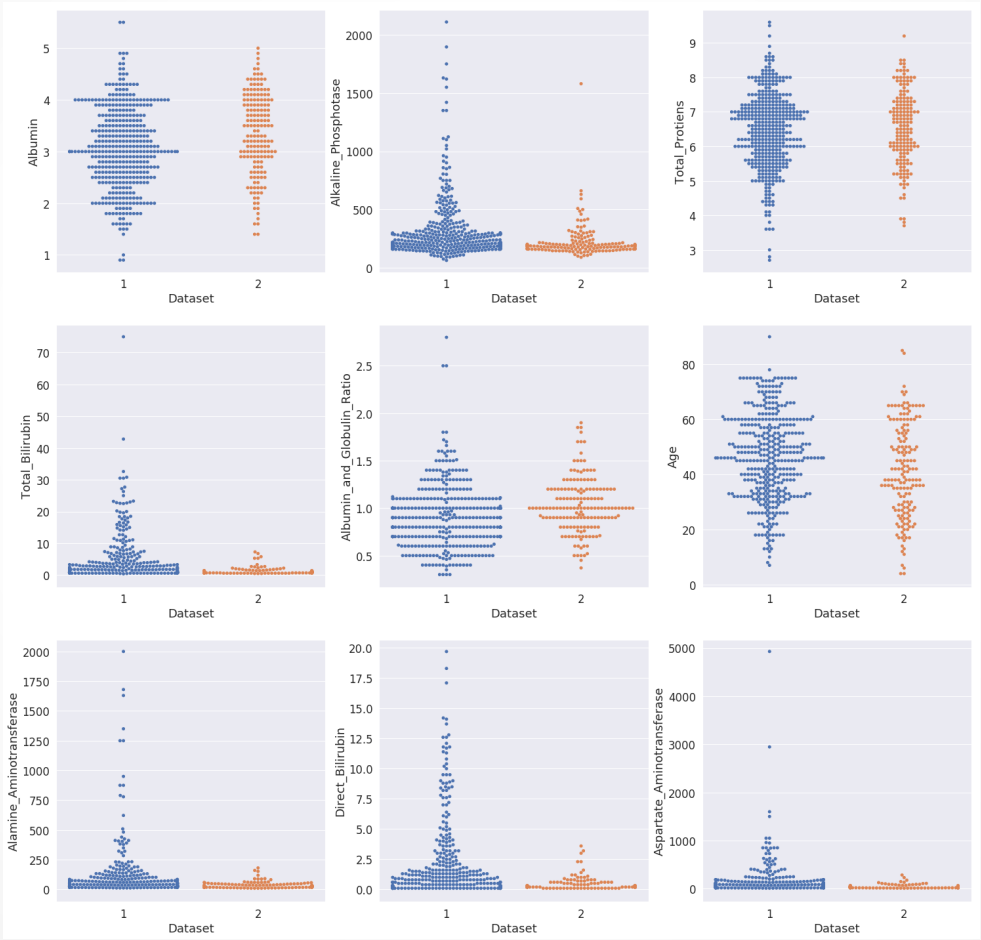
Data Source

- The dataset used for this project contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India
- The dataset consists of 10 columns and a resulting ‘Dataset’ column that contains the patient liver diagnosis (where 1 means patient has a liver disease, 2 means no disease).
- Also, the dataset contains 441 male patient records and 142 female patient records
- There is one categorical variable (gender), which will be split into two: male and female with one hot encoding. Then, the original gender feature column will be dropped
- There are missing values for Albumin_and_Globulin_Ratio feature, which will be recovered with median of the corresponding non-missing values.

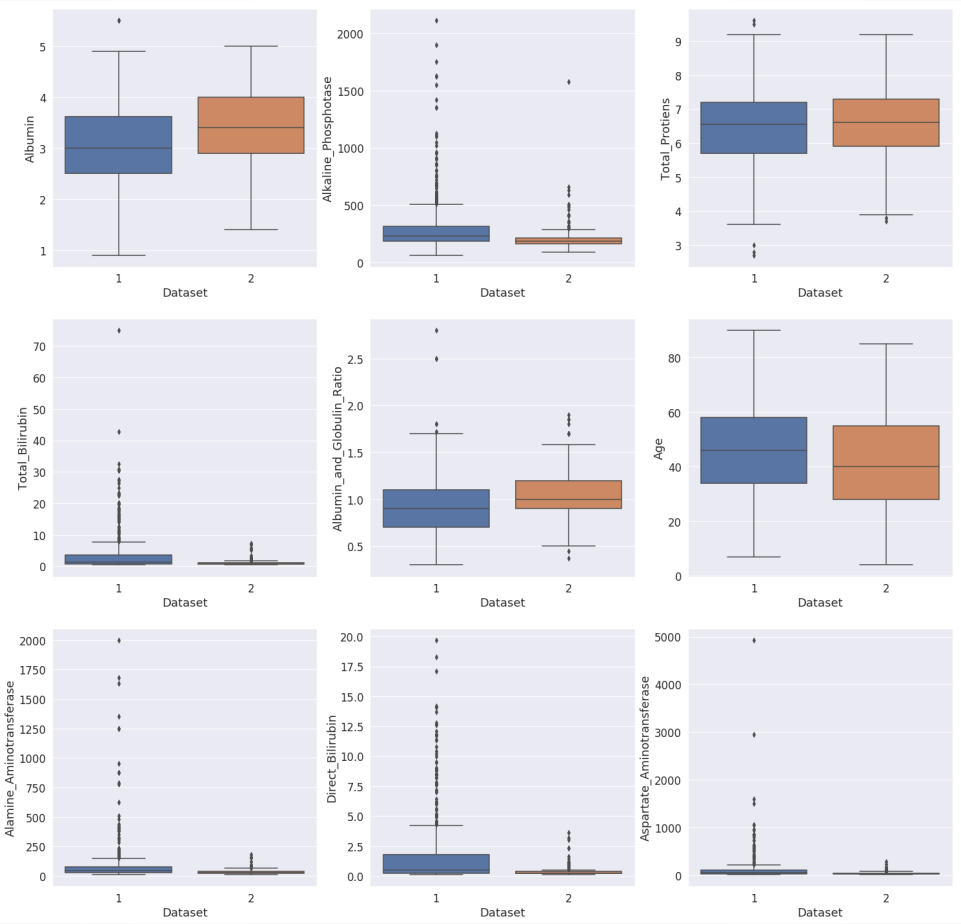
Exploratory Data Analyses (Histogram)



Exploratory Data Analyses (Swarm)



Exploratory Data Analyses (Barplot)



EDA Observations

- From the Histogram plots, we observe that the healthy patients have higher frequency of small values in narrow ranges compared to unhealthy patients
- Swarm plots show outliers for some of the features for both healthy and unhealthy patients. However, we can not claim that these outliers represent erroneous data points
- From the Barplots, we can count outliers of some features for healthy patients to be within the whisker extend of unhealthy patients. For example, Aspartate Aminotransfertase max value for healthy patients is within whisker extent of the corresponding unhealthy patients

Feature Engineering

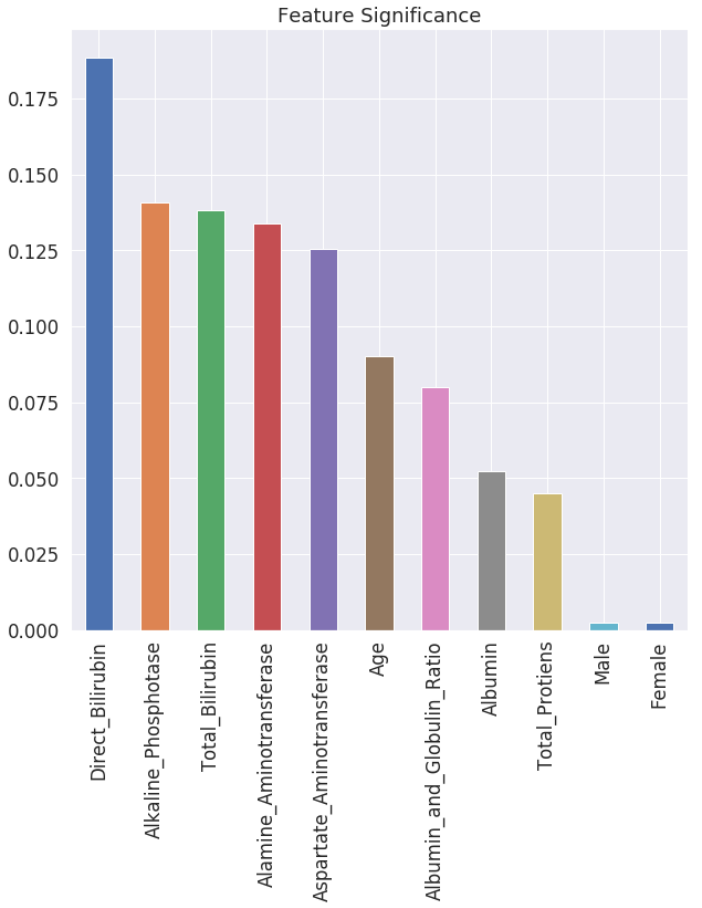
- We apply feature engineering method to generate new features based on EDA observations.
- Specifically, new hot encoded quantile features are introduced for alkaline phosphotase, direct bilirubin, alamine aminotransferase, total bilirubin, aspartate aminotransferase, age and albumin and Globulin Ratio features since they reveal some hints of possible outliers in the data.
- Another feature engineering is done by setting sample value to one for the new feature, if healthy patient's value from the original feature is above the whisker extend of the corresponding un-healthy patient's value, zero otherwise.
- The SMOTE oversampling is applied to improve the imbalance Dataset feature for liver patient disease outcome, since it includes more data for unhealthy patients than healthy ones.

Modeling

- The following ML algorithms with gridsearch parameters are used to predict on test data:
  - logistic regression
  - xgboost
  - random forest
  - knn classifier

- Five study methods are separately applied:
  - Regular with no additional features
  - MinMax feature scaling
  - Quantile feature addition
  - SMOTE oversampling
  - Max value features addition

Feature Importance From Random Forest



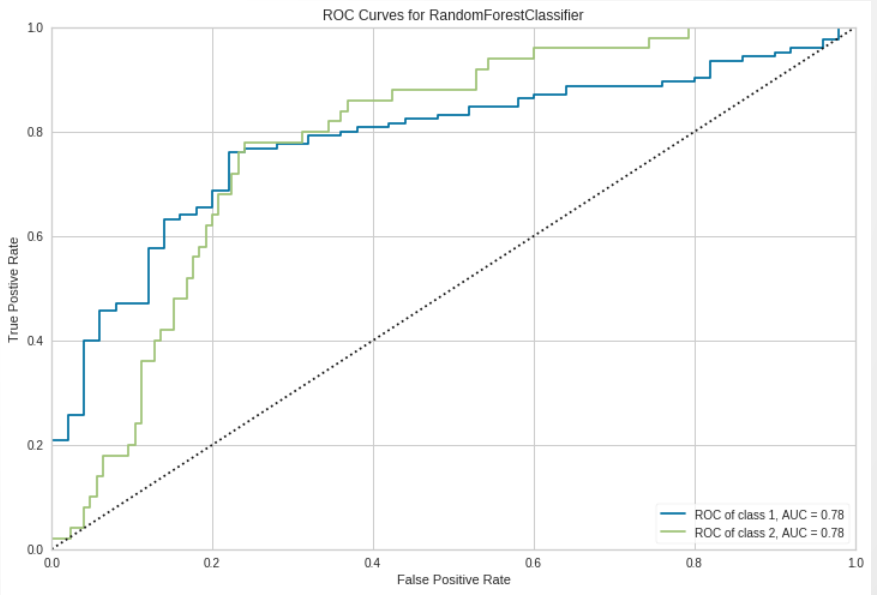
Modeling Results

Applied Method	Regular	MinMaxScaled	Quantile	SMOTE	Max
f1 accuracy score for KNN	0.702857	0.714286	0.691429	0.680000	0.697143
f1 accuracy score for Logistic Regression	0.714286	0.714286	0.714286	0.668571	0.714286
f1 accuracy score for Random Forest	0.720000	0.720000	0.720000	0.754286	0.720000
f1 accuracy score for Xgboost	0.685714	0.685714	0.702857	0.725714	0.760000
f1 score for healthy patients from KNN	0.187500	0.137931	0.156250	0.461538	0.293333
f1 score for healthy patients from Logistic Regression	0.000000	0.000000	0.000000	0.573529	0.000000
f1 score for healthy patients from Random Forest	0.328767	0.328767	0.246154	0.590476	0.140351
f1 score for healthy patients from Xgboost	0.421053	0.432990	0.409091	0.510204	0.533333
f1 score for unhealthy patients from KNN	0.818182	0.828767	0.811189	0.772358	0.807273
f1 score for unhealthy patients from Logistic Regression	0.833333	0.833333	0.833333	0.728972	0.833333
f1 score for unhealthy patients from Random Forest	0.823105	0.823105	0.828070	0.824490	0.832765
f1 score for unhealthy patients from Xgboost	0.784314	0.782609	0.801527	0.809524	0.838462

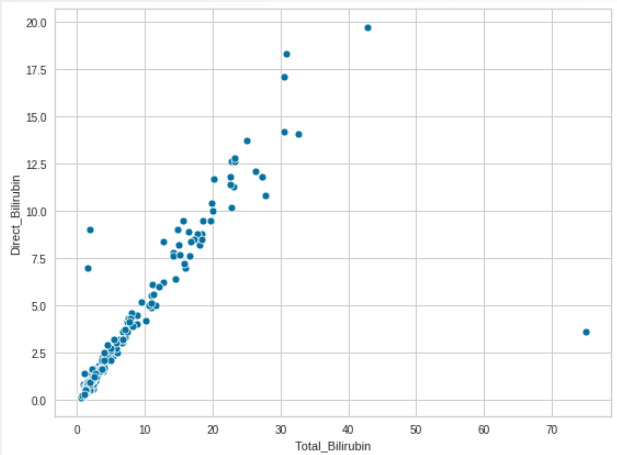
The best f1 accuracy score for Random Forest is with SMOTE over-sampled train/test split: 0.7542857142857143 with f1 score for unhealthy patients: 0.8244897959183675 and f1 score for healthy patients: 0.5904761904761904

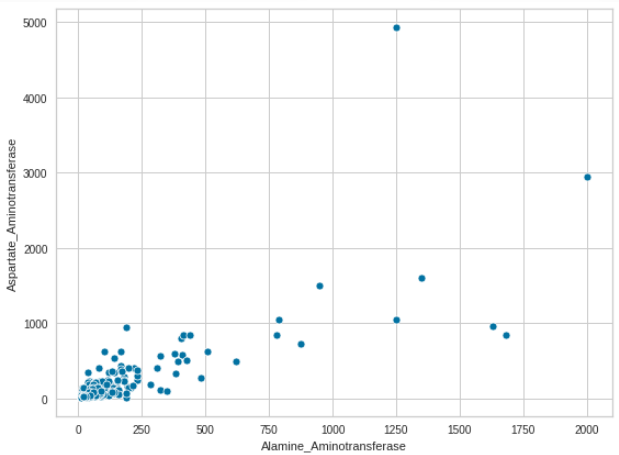
Best Result

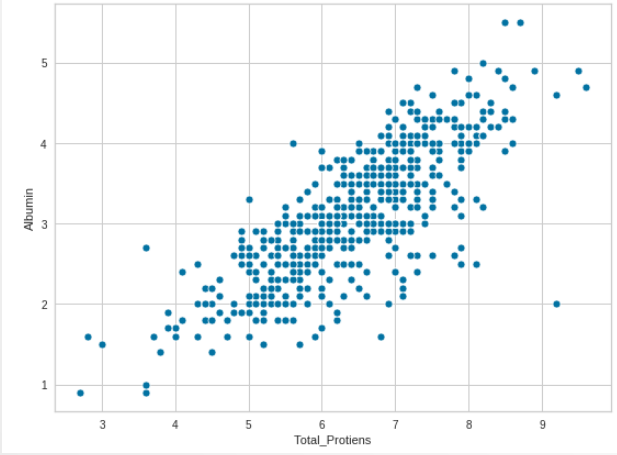
```
Accuracy of random forest classifier on test set: 0.75
Classification report:
              precision    recall  f1-score   support

           1       0.84      0.81      0.82       125
           2       0.56      0.62      0.59        50

   micro avg       0.75      0.75      0.75       175
   macro avg       0.70      0.71      0.71       175
weighted avg       0.76      0.75      0.76       175

Confusion Matrix:
[[101  24]
 [ 19  31]]
```

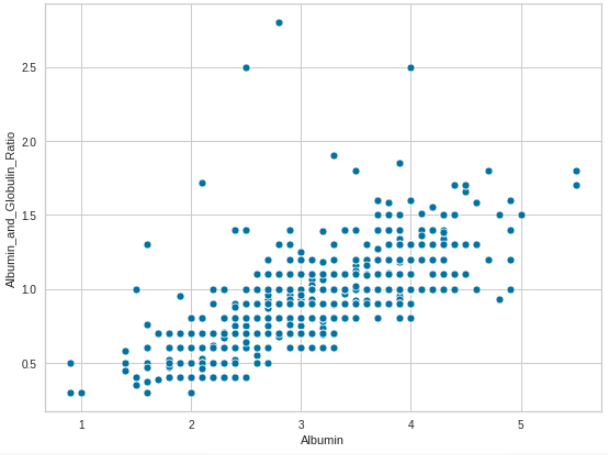


Correlations










```
Total_Bilirubin Direct_Bilirubin 0.8746179301164149
Alamine_Aminotransferase Aspartate_Aminotransferase 0.7919656848536135
Total_Protiens Albumin 0.7840533353871901
Albumin Albumin_and_Globulin_Ratio 0.6860914626301073
```

Summary

- We investigated the Liver patients dataset and applied machine learning algorithms to predict the patient disease. Our observations revealed that the distribution of patients with and without disease significantly differ. Specifically, we observe that the healthy patients have higher frequency of small values in narrow ranges for total bilirubin, direct bilirubin, aspertate aminotransferace and alkaline phosphotase compared to unhealthy patients

- Several ML algorithms were used to predict the outcome on test data. We used logistic regression, xgboost, random forest and knn classifier with gridsearch parameters

- Several feature engineering methods were applied to generate new features. Specifically, new hot encoded quantile features were introduced for alkaline phosphotase, direct bilirubin, alamine aminotransferase, total bilirubin, aspartate aminotransferase, age and albumin and Globulin Ratio features since they revealed some hints of possible outliers in the data

- Another feature engineering is done by setting sample value to one for the new feature, if healthy patient's value from the original feature is above the whisker extend of the corresponding un-healthy patient's value, zero otherwise

- The SMOTE oversampling was applied to improve the imbalance Dataset feature for liver patient disease outcome, since it included more data for unhealthy patients than healthy ones. Our results demonstrated that random forest with SMOTE produced better f1 score on both classes

- We found 4 pairs of strongly correlated features: direct and total bilirubin, aspertate aminotransferace and alamine aminotransferace, albumin and total proteins, albumin and globulin ratio and albumin

- Overall, these analysis and techniques can be applied for liver patient diagnoses and similar medical related problems