Predicting Patients with Liver Disease

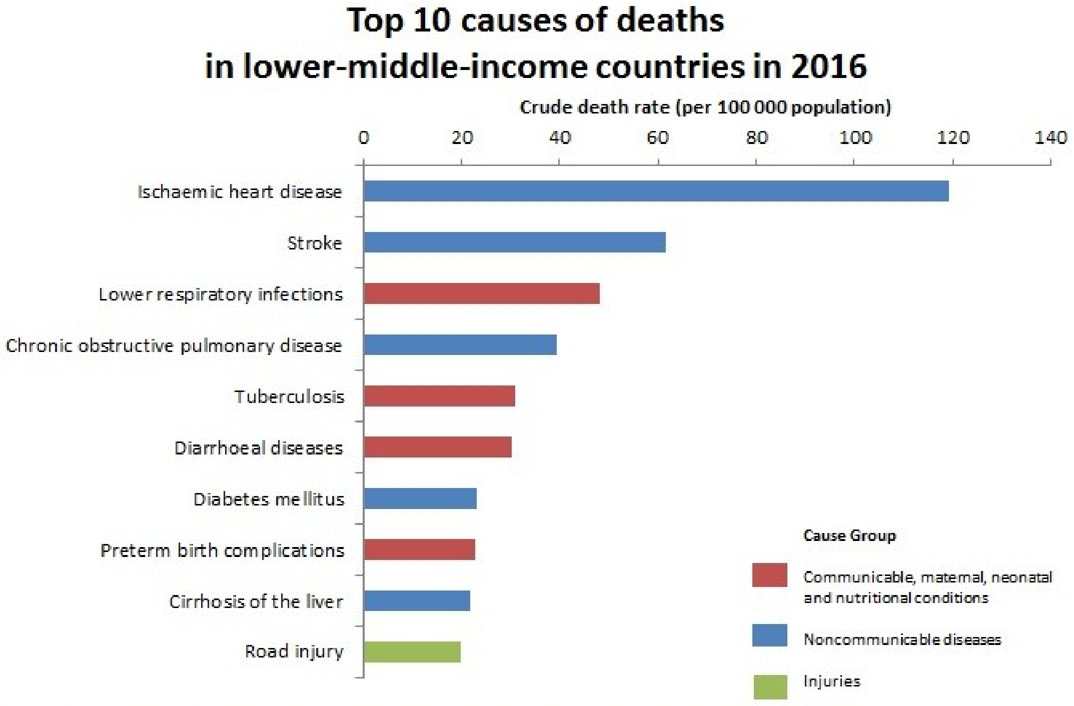
Capstone Project 1 Milestone Report, Springboard

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. In this first capstone project, I am going to use machine learning models and a dataset that contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India.

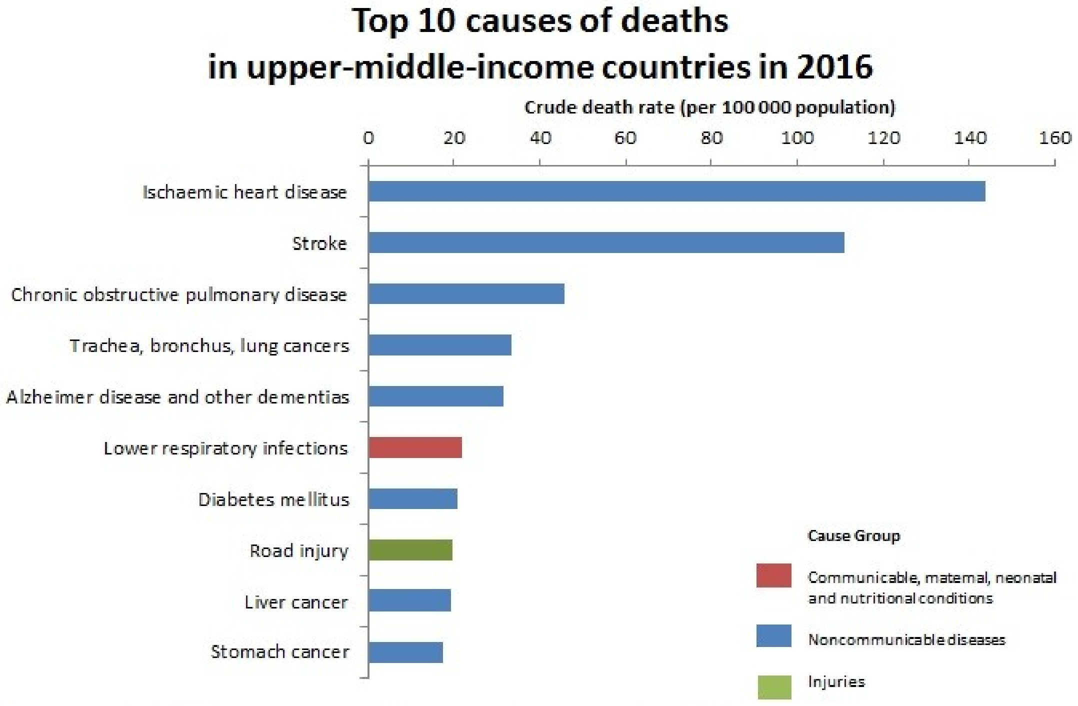
The dataset for patient records will be taken from UC Irvine (UCI) ML repository: (*Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science)*. The current dataset contains 10 columns with patient data and a resulting ‘outcome’ column with the patient liver diagnosis (where 1 means patient has a liver disease, 2 means no disease). Thus, this is a binary classification problem. The techniques used in the project can be applied to other datasets with liver patient records.

It is important to note that the dataset is relatively clean and easy to use. However, the dataset will still require some data cleansing, data wrangling and possibly feature engineering. The exploratory data analyses and statistical learning will be applied to find more insights from data. Based on the above mentioned details regarding the data, several relevant machine learning algorithms (such as logistic regression, random forest and others) will be applied for predicting a liver patient diagnosis.

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. Thus, having an accurate model for predicting patient live disease on early stages based on their records can significantly improve the diagnosis and help in early disease preventive cares. Below are the charts provided by World Health Organization (WHO):



Source: Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva, World Health Organization; 2018.
World Bank list of economies (June 2017). Washington, DC: The World Bank Group; 2017 (https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups).



Source: Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva, World Health Organization; 2018.
World Bank list of economies (June 2017). Washington, DC: The World Bank Group; 2017 (https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups).

(Source: Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Geneva, World Health Organization; 2018.)

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

Data Wrangling

Description and Objectives

The main goal of this project is to build a model that will accurately predict the liver disease for patients. Currently, the liver disease is one of the top 10 causes of deaths in developing countries, which is caused by a number of factors, such as diet, alcohol consumption and smoking. In worst cases, the disease can develop into cirrhosis. An early stage diagnosis of liver associated complications in patients can reduce the mortality rate. The machine learning techniques can help in predictions of liver disease and reduce burden on doctors.

Data Source

The dataset contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India[1]. 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.

Data info

```
data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 583 entries, 0 to 582
Data columns (total 11 columns):
Age                           583 non-null int64
Gender                        583 non-null object
Total_Bilirubin               583 non-null float64
Direct_Bilirubin              583 non-null float64
Alkaline_Phosphotase          583 non-null int64
Alamine_Aminotransferase      583 non-null int64
Aspartate_Aminotransferase    583 non-null int64
Total_Protiens                583 non-null float64
Albumin                       583 non-null float64
Albumin_and_Globulin_Ratio    579 non-null float64
Dataset                       583 non-null int64
dtypes: float64(5), int64(5), object(1)
memory usage: 50.2+ KB
```

There are missing values in Albumin_and_Globulin_Ratio column. They will be filled with median value of the Albumin_and_Globulin_Ratio column.

The Gender column is categorical, meaning that the values from that column are neither continues nor discrete. The hot encoding technique will be applied to split that column into the two new columns (Female, Male) and then the Gender column will be dropped.

Data Statistics

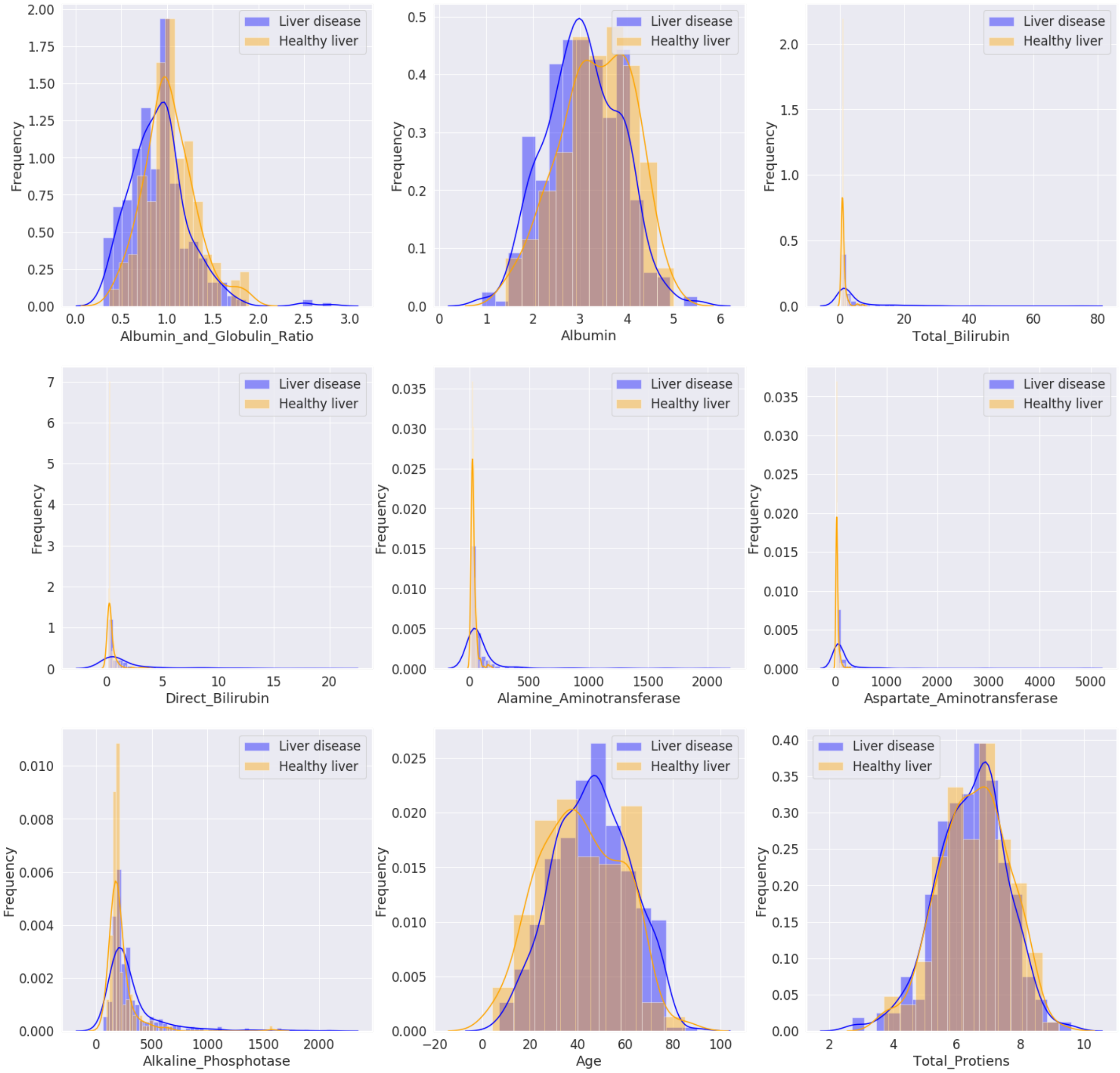
```
data.describe().T
```

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
Age	583.0	44.746141	16.189833	4.0	33.0	45.00	58.0	90.0
Total_Bilirubin	583.0	3.298799	6.209522	0.4	0.8	1.00	2.6	75.0
Direct_Bilirubin	583.0	1.486106	2.808498	0.1	0.2	0.30	1.3	19.7
Alkaline_Phosphotase	583.0	290.576329	242.937989	63.0	175.5	208.00	298.0	2110.0
Alamine_Aminotransferase	583.0	80.713551	182.620356	10.0	23.0	35.00	60.5	2000.0
Aspartate_Aminotransferase	583.0	109.910806	288.918529	10.0	25.0	42.00	87.0	4929.0
Total_Protiens	583.0	6.483190	1.085451	2.7	5.8	6.60	7.2	9.6
Albumin	583.0	3.141852	0.795519	0.9	2.6	3.10	3.8	5.5
Albumin_and_Globulin_Ratio	579.0	0.947064	0.319592	0.3	0.7	0.93	1.1	2.8
Dataset	583.0	1.286449	0.452490	1.0	1.0	1.00	2.0	2.0

We observe from data statistics that features have different range of value. Thus, they need to be normalized before applying machine learning algorithms. This step is particularly important to avoid negative effect from the data on the performance of the logistic regression and KNN algorithms. The min-max normalization will be applied to standardize the range of features.

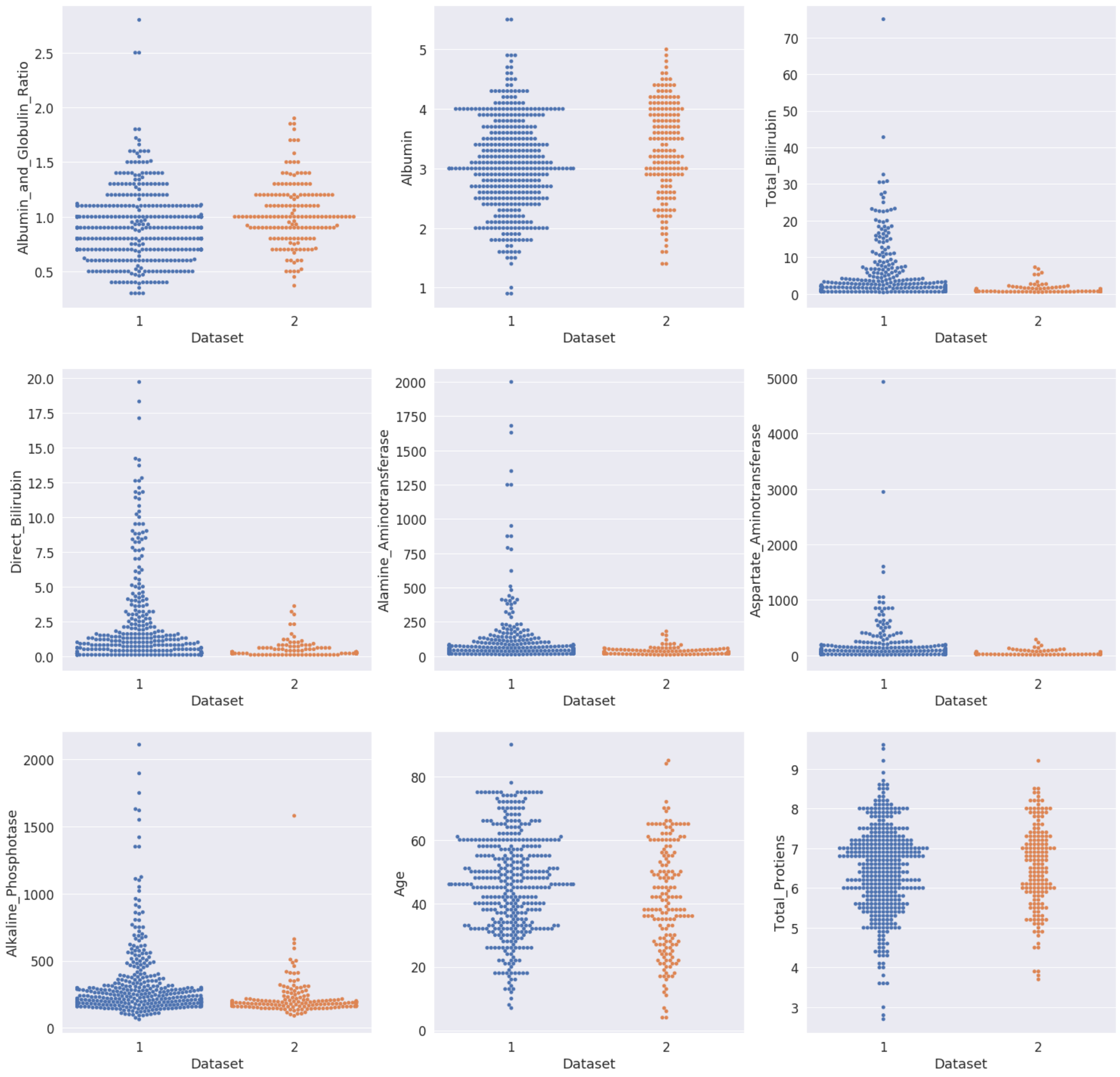
Exploratory Data Analyses (EDA):

Histogram Plots



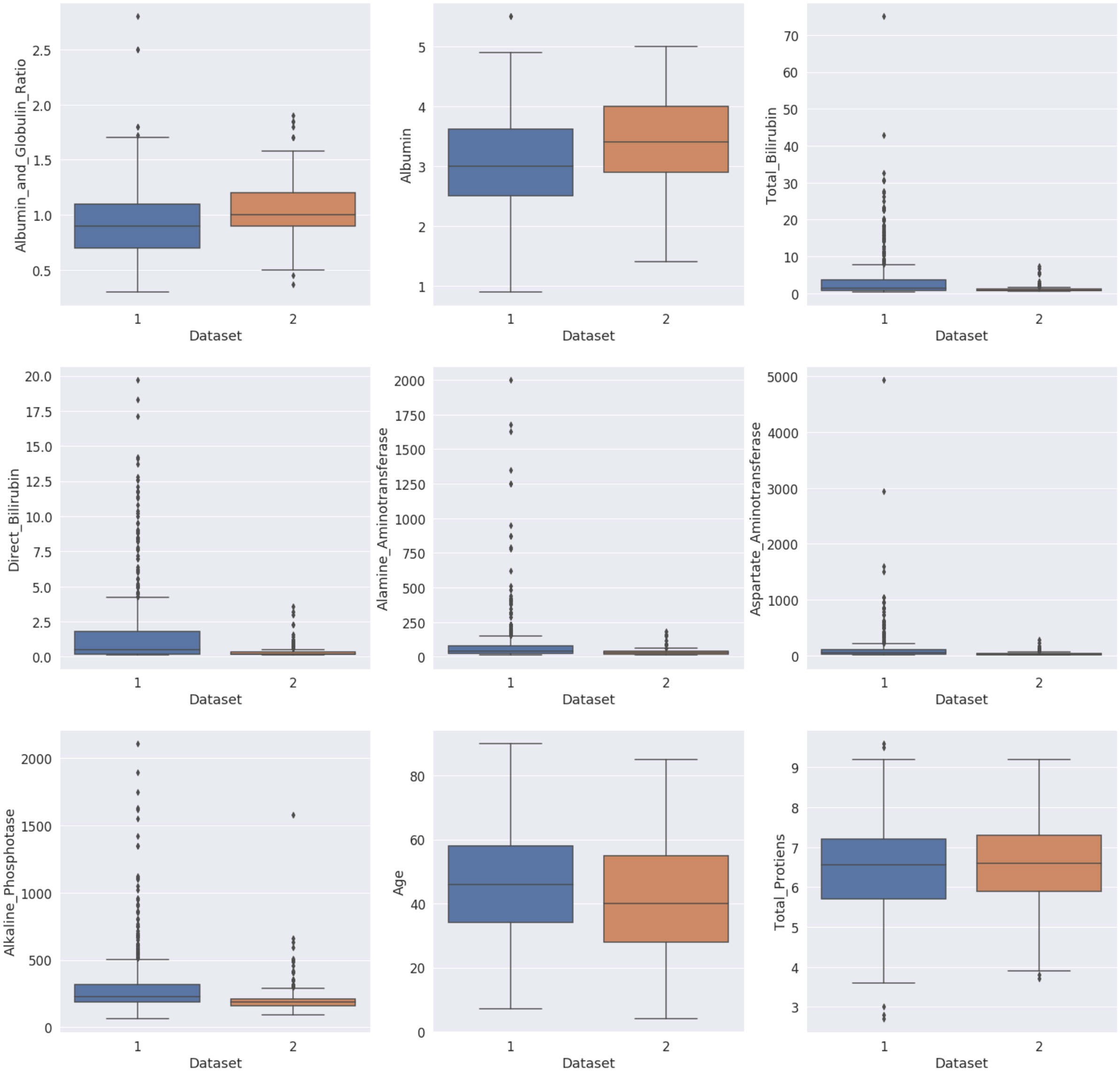
From the histogram plots we observe that the healthy patients have higher frequency of small values in narrow ranges compared to unhealthy patients.

Swarm Plots



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. We will apply feature engineering method to generate new features. Specifically, new hot encoded quantile features will be introduced for alkaline phosphotase, direct bilirubin, alamine aminotransferase, total bilirubin, aspartate aminotransferase, age and albumin and Globulin Ratio features since they revealed some hints on possible outliers in the data.

Bar Plots



From the above 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. Thus, we will use this observation to generate new features for the analysis. Specifically, we will be setting sample value to 1 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. Otherwise, we will set it to 0.

Modeling

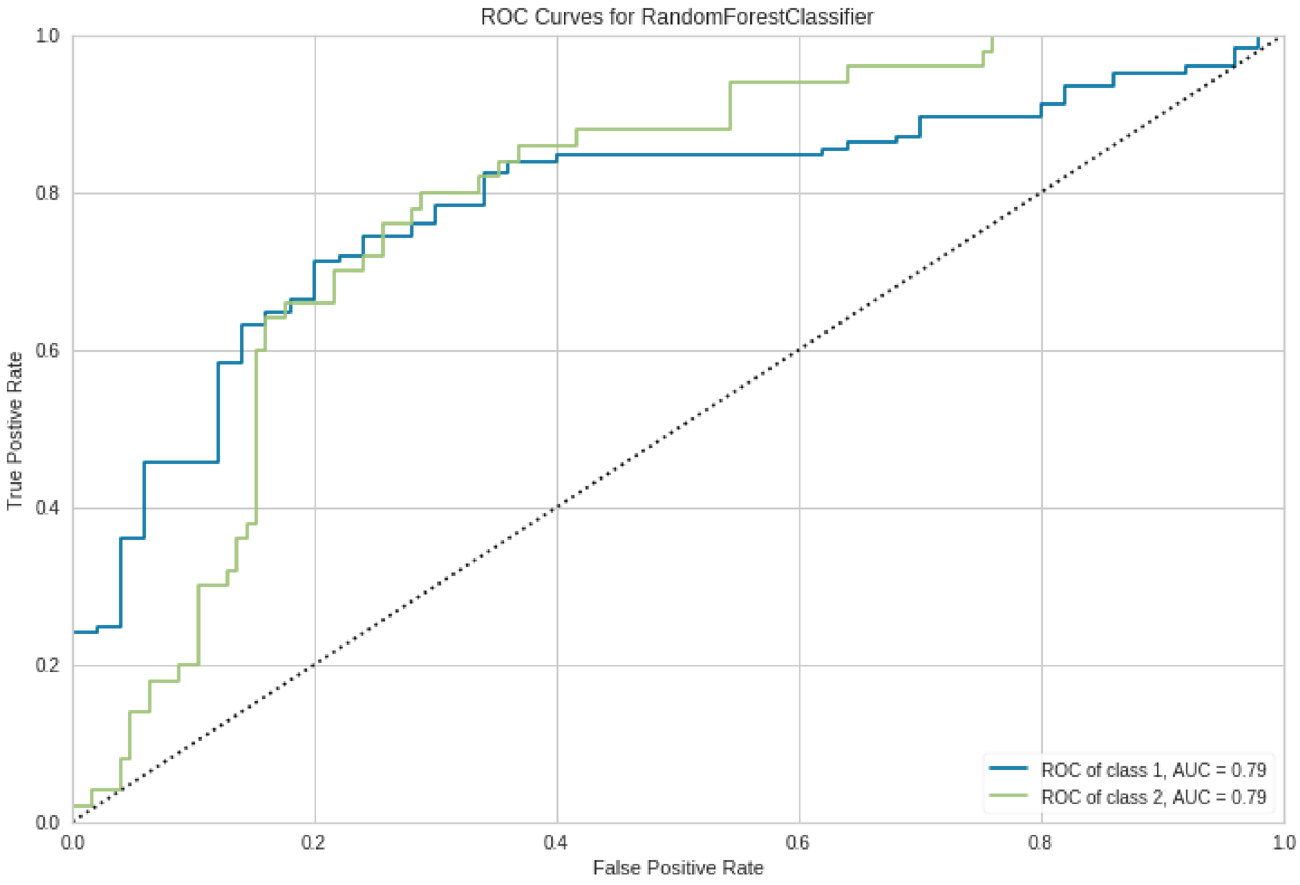
Applied Method	Regular	MinMaxScaled	Quantile	SMOTE	Max
f1 accuracy score for KNN	0.702857	0.714286	0.691429	0.680000	0.697143
f1 accuracy score for Logistic Regression	0.714286	0.714286	0.714286	0.668571	0.714286
f1 accuracy score for Random Forest	0.708571	0.708571	0.720000	0.765714	0.720000
f1 accuracy score for Xgboost	0.662857	0.668571	0.702857	0.731429	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	0.281690	0.281690	0.246154	0.577320	0.140351
f1 score for healthy patients from Xgboost	0.391753	0.395833	0.409091	0.525253	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	0.817204	0.817204	0.828070	0.837945	0.832765
f1 score for unhealthy patients from Xgboost	0.766798	0.771654	0.801527	0.812749	0.838462

The best f1 accuracy score for Random Forest is with SMOTE over-sampled train/test split: 0.7657142857142857

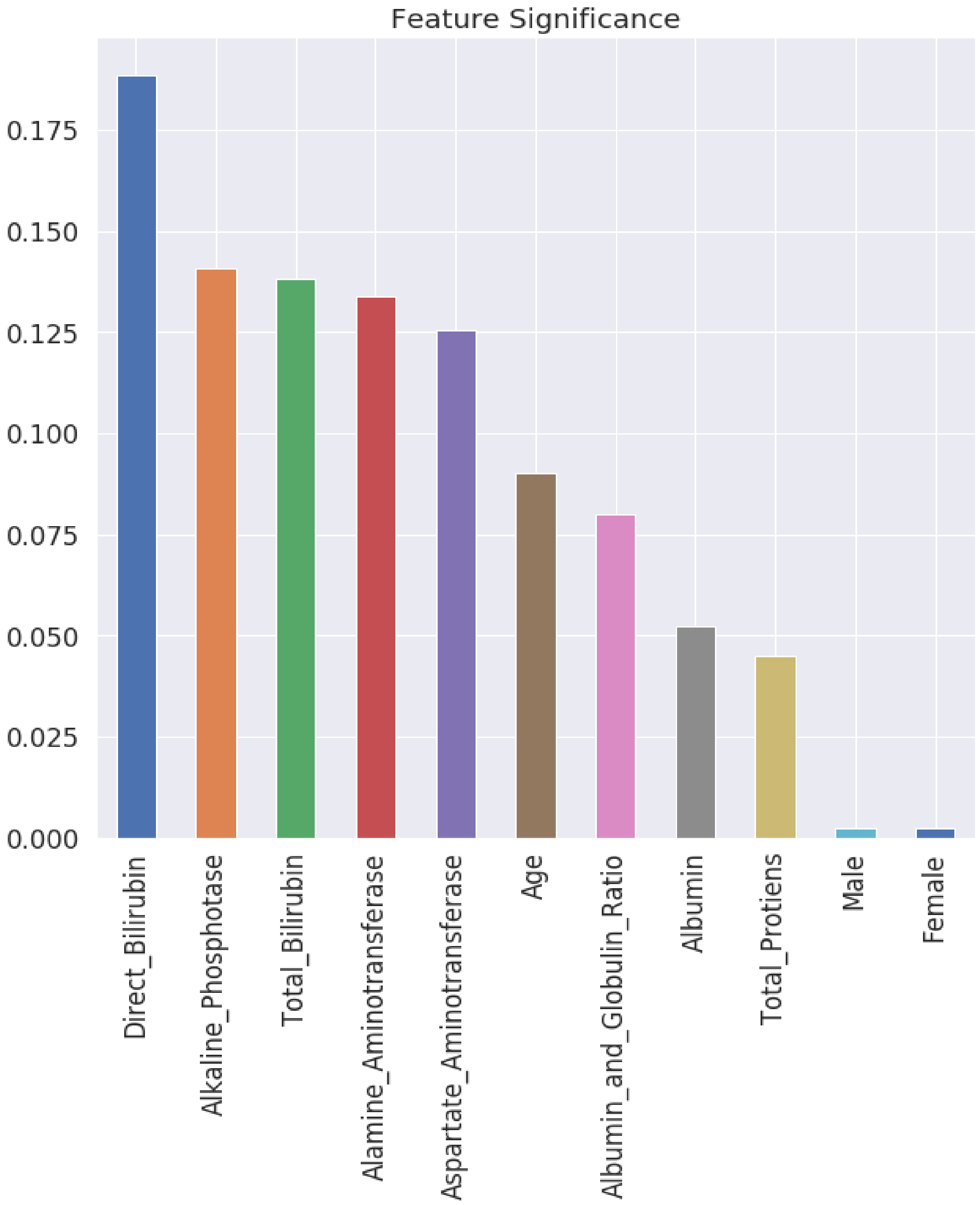
with f1 score for unhealthy patients: 0.8379446640316205 and f1 score for healthy patients: 0.5773195876288660

```
Classification report:
              precision    recall  f1-score   support

           1       0.83      0.85      0.84       125
           2       0.60      0.56      0.58        50

   micro avg       0.77      0.77      0.77       175
   macro avg       0.71      0.70      0.71       175
weighted avg       0.76      0.77      0.76       175

Confusion Matrix:
[[106  19]
 [ 22  28]]
```

ROC Curve for Best Random Forest Classifier with SMOTE



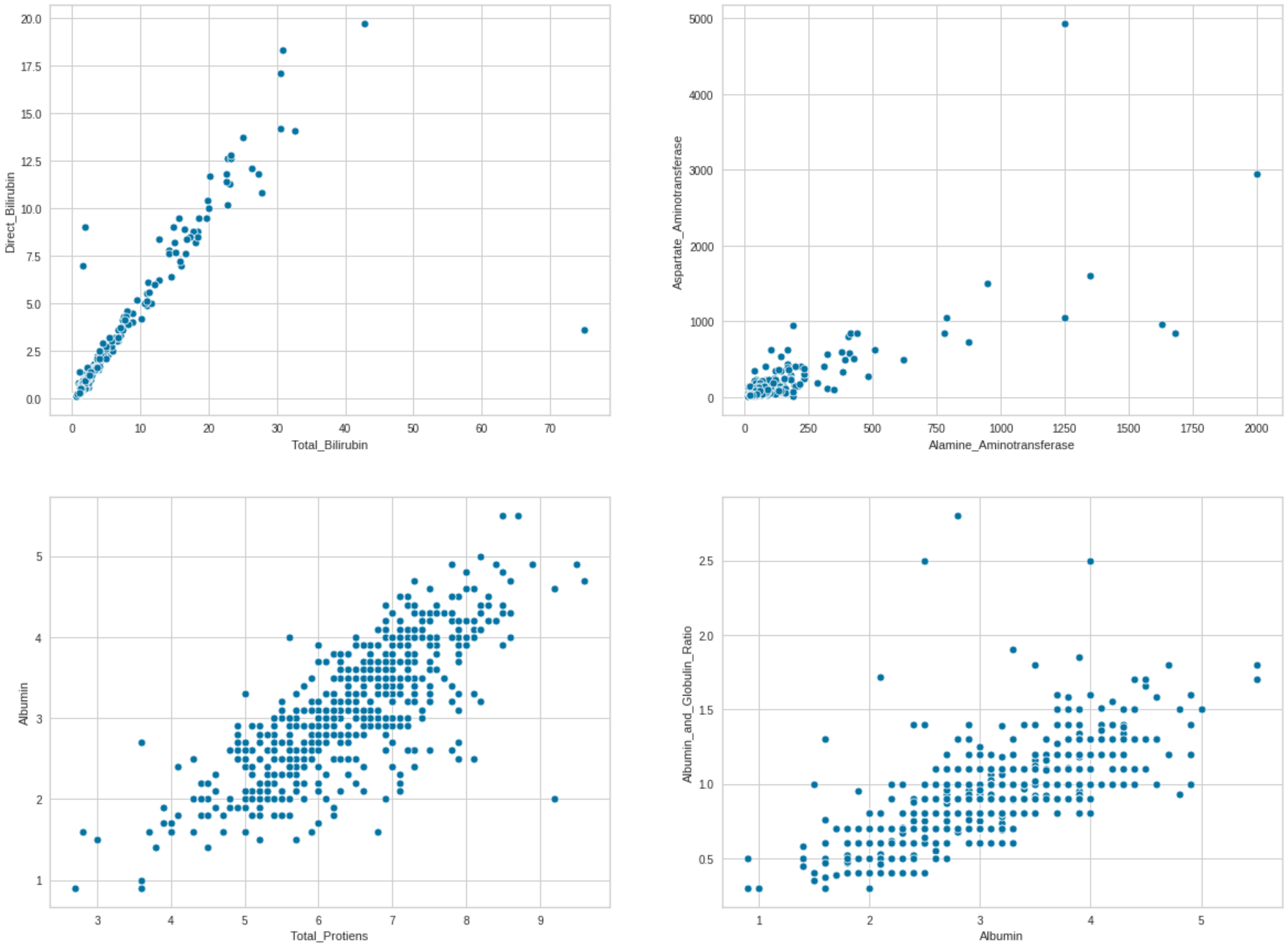
Feature Significance of Random Forest with Regular Method (i.e no extra feature engineering)



Correlations

Strongest pearson correlation was found among the following features:

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



Conclusion

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 on possible outliers in the data.

Another feature engineering is done by setting sample value to 1 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. Otherwise, we will set it to 0.

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.

References

[1] UC Irvine (UCI) ML repository: (Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science).