Prediction of Hepatitis C and Cirrhosis Diagnosis through Japanese Patient Data

Jiraphat Julprapa (612306012)
Kayla Gusti Haruni (062101882)

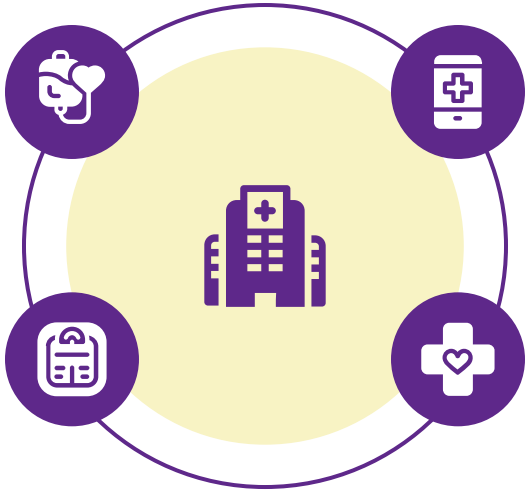
Background

How can machine learning assist in disease diagnosis?

Using machine learning to assist in medical diagnoses - Hepatitis C and Cirrhosis

Hepatitis C is a widespread infectious disease, globally affecting millions

Cirrhosis refers to permanent scarring of the liver, caused by long-term damage



Current methods to diagnose are highly invasive and time-consuming.

Machine learning can predict the diagnosis from electronic health records (EHRs) data of patients

Source

This dataset contains electronic health records of 123 patients diagnosed with hepatitis C, collected at Kanazawa University in Japan.

- https://www.kaggle.com/datasets/davidechicco/hepatitis-c-ehrs-from-japan/data

Dataset (123 rows × 13 columns)

```
pd.set_option('display.max_columns', None)
display(df)
```

	cirrhosis	age	sex	cholesterol	triglyceride	HDL	LDL	PathDiagNum	BMI	ALT	AST	glucose	serogroup01
0	1	63	1	103	147	35	38.6	1	20.0	27	35	117	0.0
1	0	68	2	141	95	38	84	1	23.1	78	74	98	0.0
2	1	79	1	143	71	60	68.8	1	21.3	44	40	95	0.0
3	1	52	1	126	64	39	74.2	1	34.0	18	26	101	0.0
4	1	77	2	126	49	41	75.2	1	25.7	106	97	128	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
118	0	65	2	143	42	53	81.6	1	27.8	74	75	136	NaN
119	0	66	1	223	89	47	158.2	1	24.3	43	39	120	1.0
120	1	64	1	152	82	56	79.6	1	18.0	35	50	91	0.0
121	1	68	2	152	139	76	48.2	1	26.9	84	69	102	0.0
122	1	80	1	168	190	70	60	1	20.8	74	77	134	0.0

123 rows × 13 columns

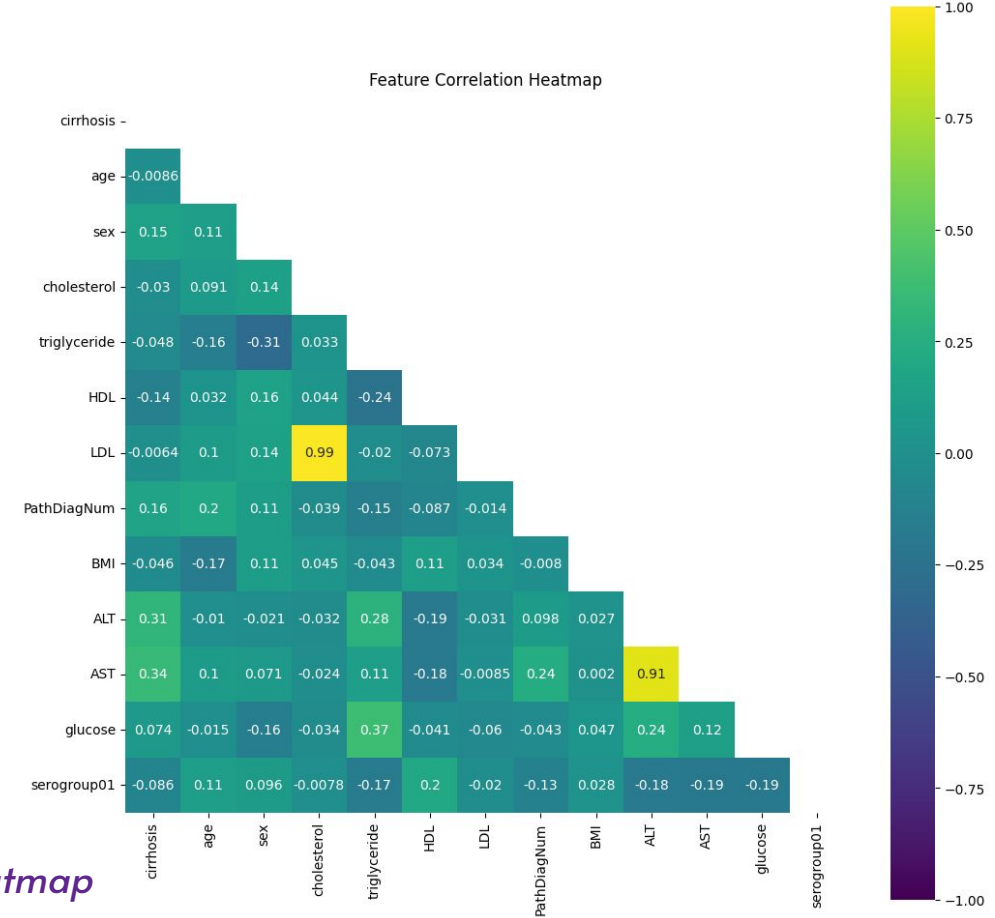
Data Preparation

Data was clean up. All rows with Nan values were removed and data types of all columns were converted to either integers or float.

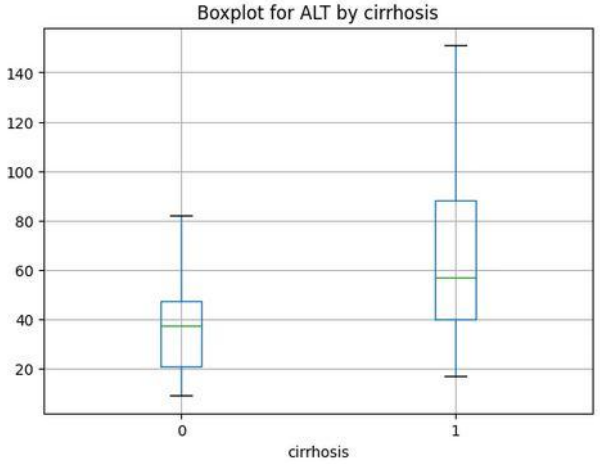
Data Exploration and Visualization

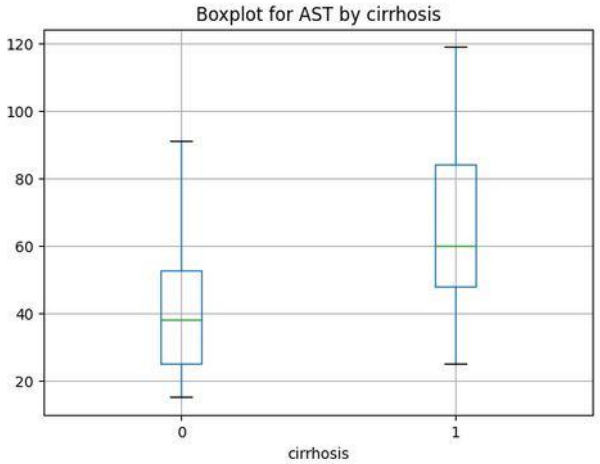
Explore relationship of features to each other as well as the target of having cirrhosis, represented by as follows:

- histogram
- box plot
- correlation heatmap

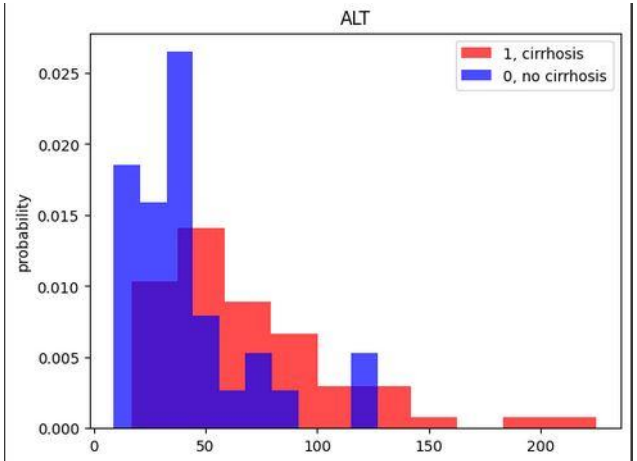


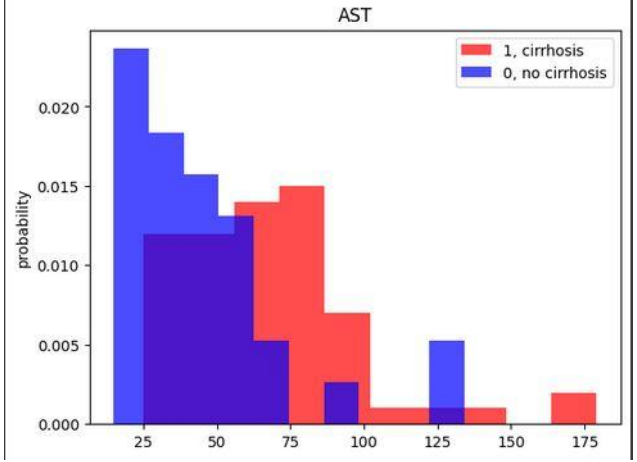
Correlation Heatmap





Boxplot





Histogram

Model Building

Two machine learning algorithms were built based on two different models:

- **Random Forest model**
- **Logistic Regression model**

Model Evaluation

The qualities of the models were tested and measured using various metrics:

- ROC curve
- Accuracy Score
- Precision Score
- Recall Score
- F1 score
- Confusion Matrix

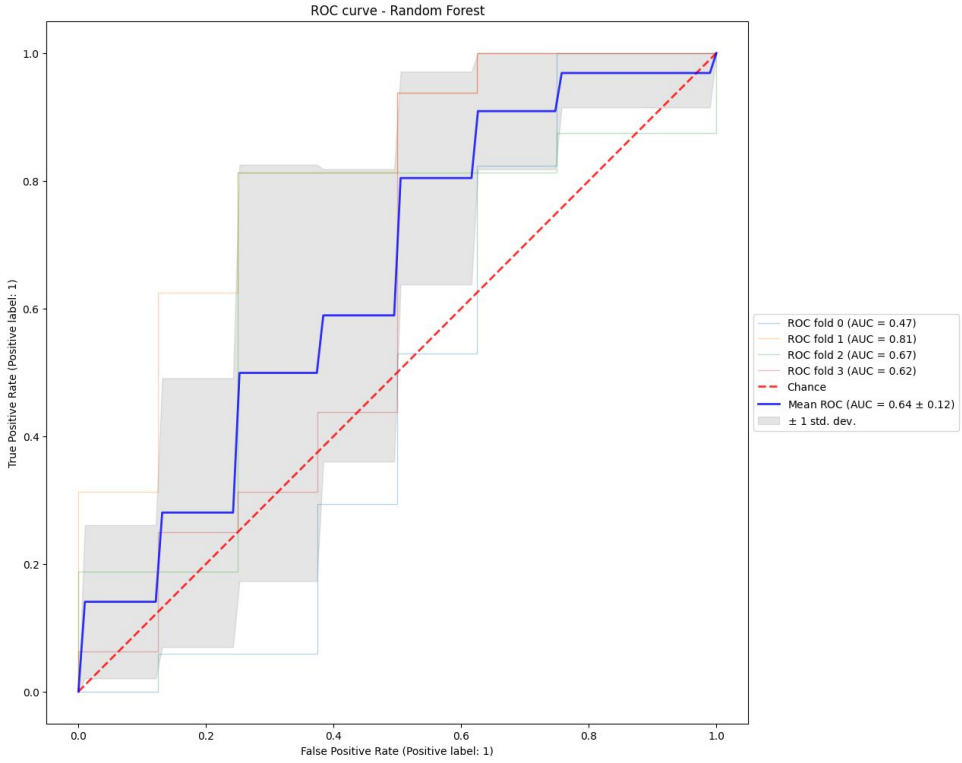
```
Model Evaluation - Random Forest
Accuracy Score : 0.6938775510204082
Precision Score : 0.7045454545454546
Recall Score : 0.9393939393939394
F1 Score : 0.8051948051948052
Confusion Matrix :
[[ 3 13]
 [ 2 31]]
```

Metric Scores

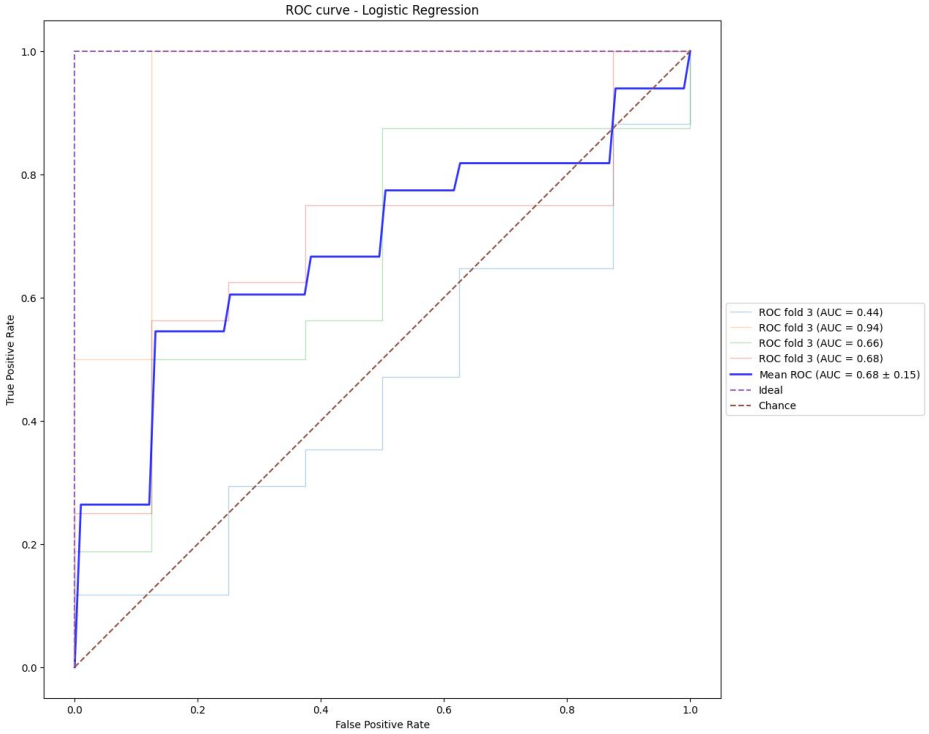
```
Model Evaluation - Logistic Regression
Accuracy Score : 0.7755102040816326
Precision Score : 0.8235294117647058
Recall Score : 0.8484848484848485
F1 Score : 0.8358208955223881
Confusion Matrix :
[[10  6]
 [ 5 28]]
```

Model Building and Evaluation

ROC curve



Random Forest Model



Logistic Regression Model

Conclusion

The features, **AST and ALT levels**, are the most **significant** features, as they are more correlated with having cirrhosis

Logistic regression model seems to be **better**

For example

- its accuracy score of 0.78 is more than Random forest model's score of 0.69.
- its AUC of 0.68 is also more than Random forest model's score of 0.64.

The model obtained can potentially be applied by using as a **preliminary tool for diagnosis of cirrhosis**

Application - *Preliminary Diagnosis*

```
X_diagnosis = pd.DataFrame(data)
display(X_diagnosis)
```

	age	sex	cholesterol	triglyceride	HDL	LDL	PathDiagNum	BMI	ALT	AST	glucose	serogroup01
0	50	2	110	90	50.0	90.4	1	24.3	32	39	99	0

```
#Use Logistic Regression model, as results from evaluation metrics are generally higher
X_diagnosis_norm = norm.fit_transform(X_diagnosis)
y_diagosis=model_LR.predict(X_diagnosis_norm)
print("Prediction:", y_diagosis)
if y_diagosis == 0:
  print("Patient likely does not have cirrhosis")
elif y_diagosis == 1: print("Patient likely has cirrhosis")
```

```
Prediction: [0]
Patient likely does not have cirrhosis
```

THANKS!

RESOURCES

- Random forest codes reference:
  https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc_crossval.html

- Logistic regression codes reference:
  https://github.com/mon2100/Topic-Estimation-of-Prediction-for-getting-heart-disease-using-Logistic-Regression-Model-of-Machine

- Dataset:
  https://www.kaggle.com/datasets/davidechicco/hepatitis-c-ehrs-from-japan/data