Heart Failure Prediction

Samuel Kim

The problem

Introduction

- Cardiovascular diseases are the number one cause of death.
- eighteen million lives each year (thirty-one percent worldwide).
- Four out of five deaths in cardiovascular diseases are due to heart attacks and strokes.
- one-third of these deaths occur prematurely in people under seventy years of age.

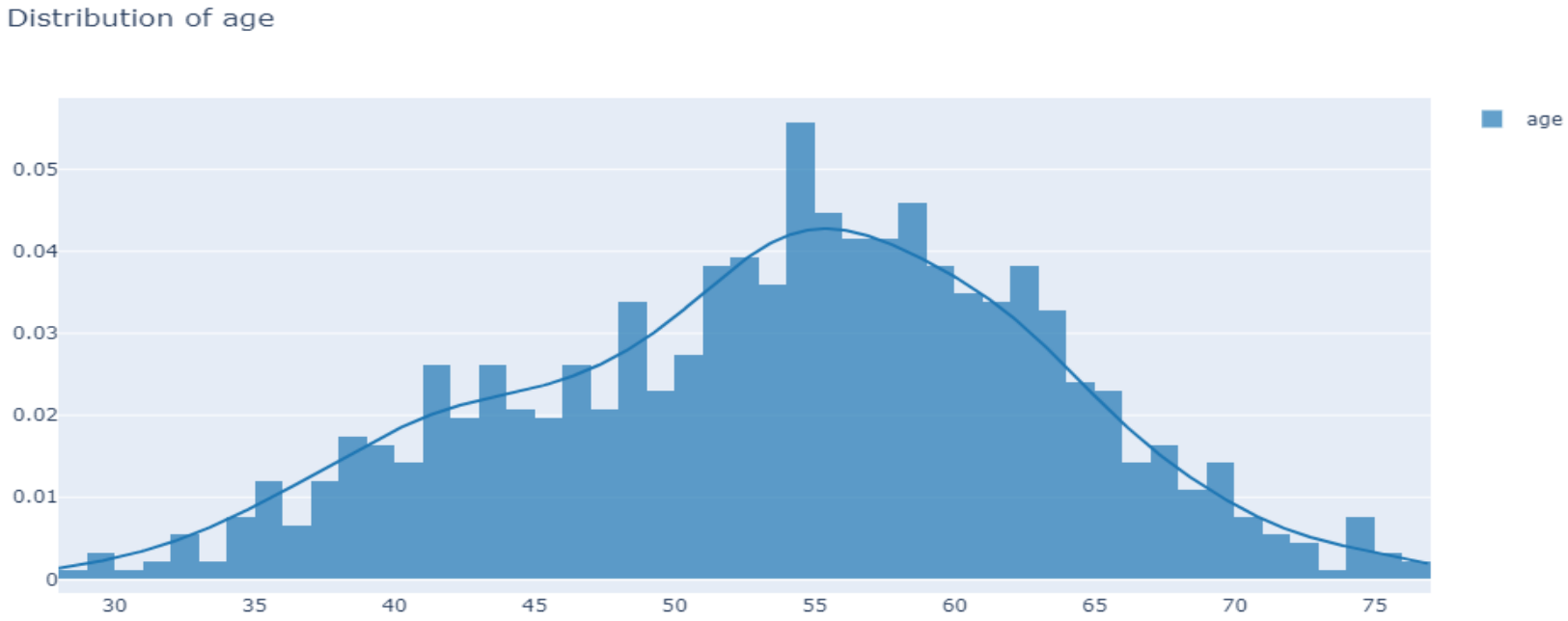
Context

- This dataset contains twelve features that can be used to predict a possible heart disease: Age, Sex, Chest Pain Type, Resting Blood Pressure, Cholesterol, Fasting Blood Sugar, resting electrocardiogram results, Max Heart Rate, exercise-induced angina, Old_Peak, the slope of the peak exercise ST segment, and heart disease.

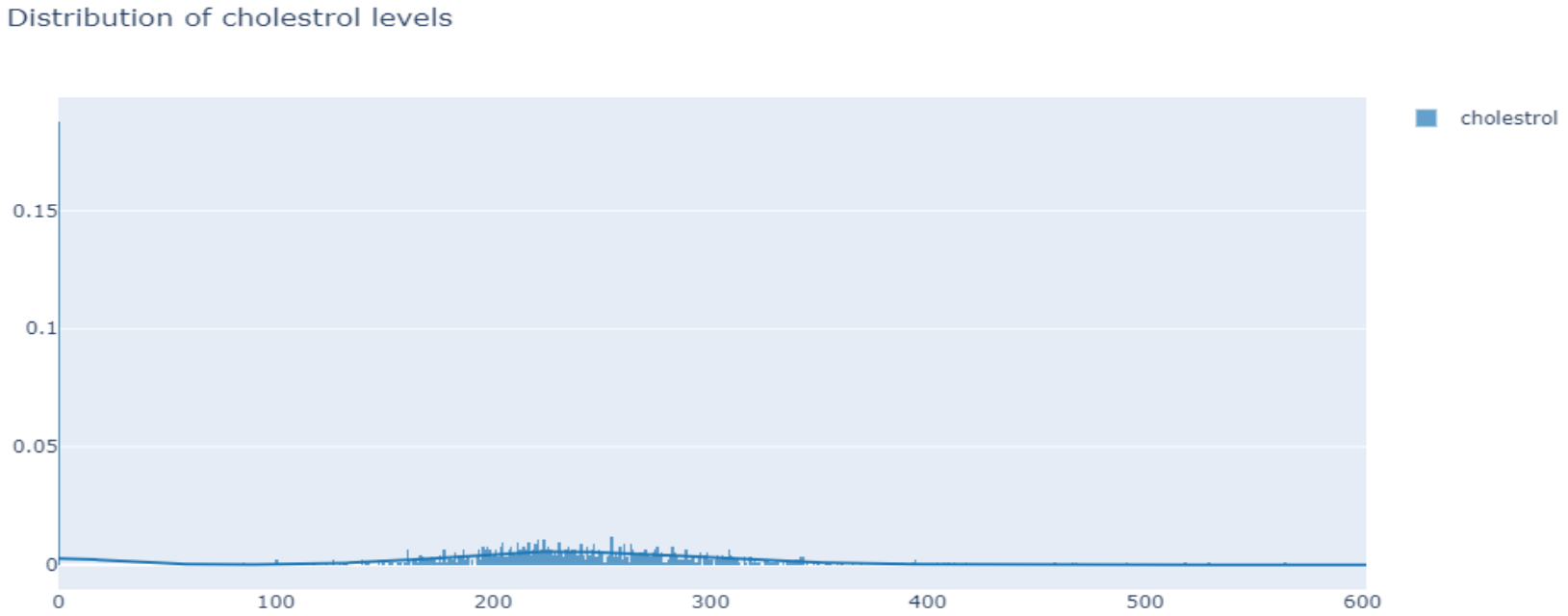
Problem statement

- People with cardiovascular disease or who are at high cardiovascular risk is due to the presence of one or more risk factors. Therefore, machine learning can be a great asset to the early detection and management of people who have cardiovascular disease or who simply have high cardiovascular risk.

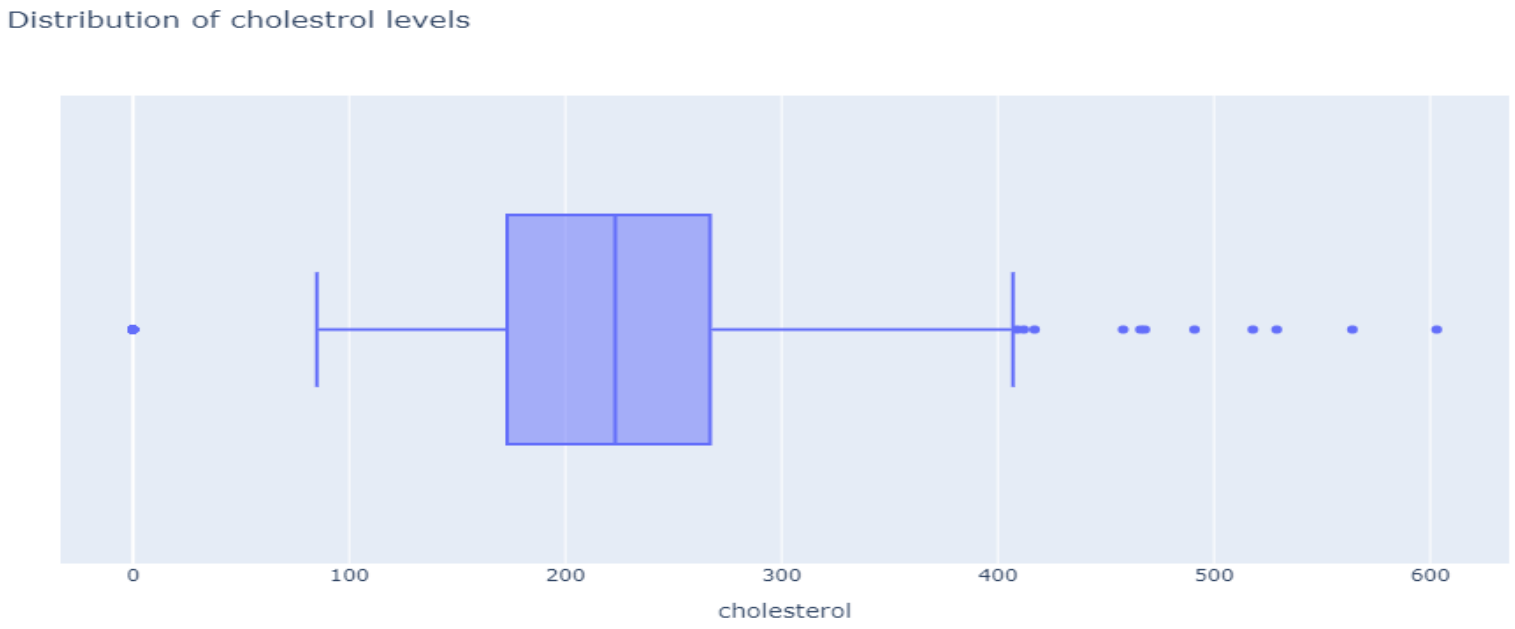
Distribution of Age



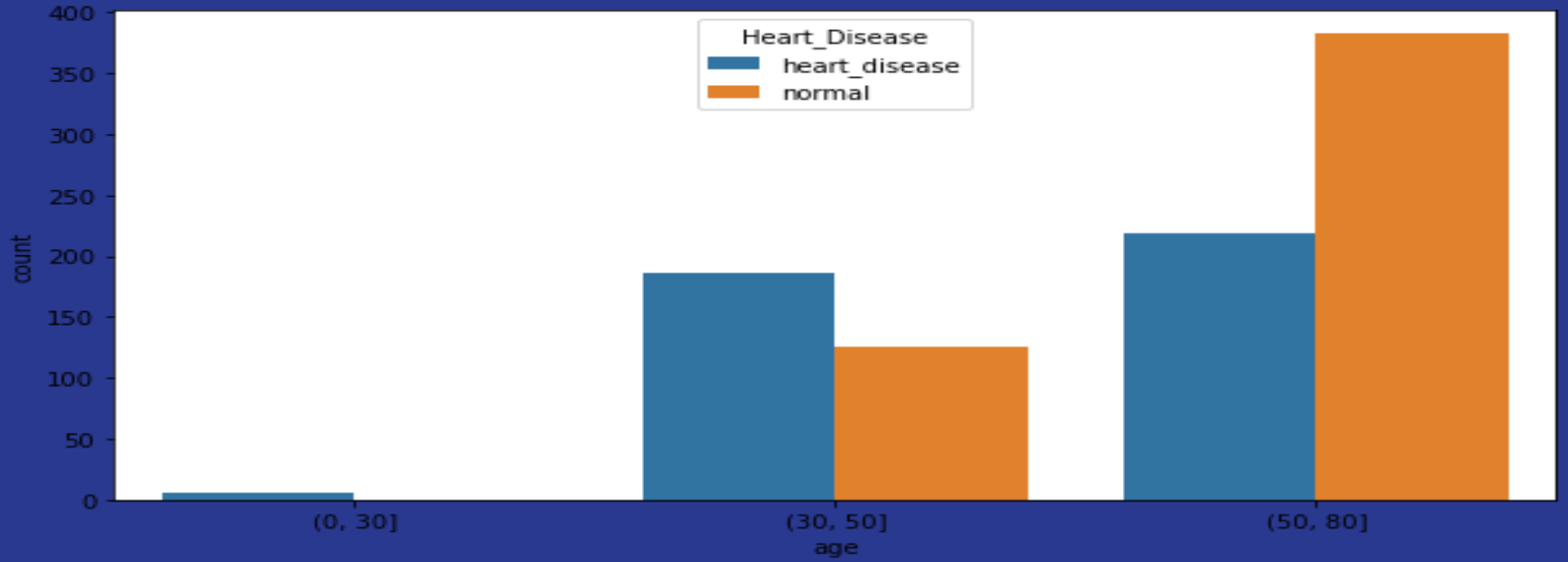
Cholesterol Levels



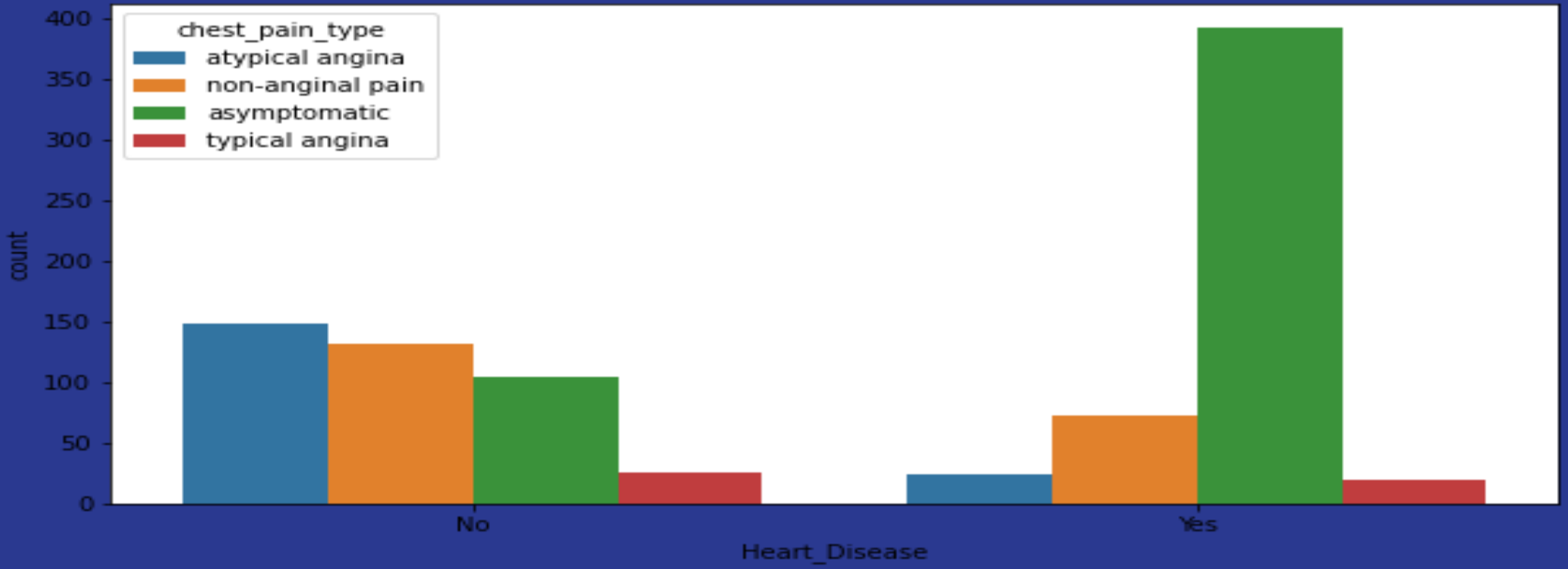
Cholesterol Levels



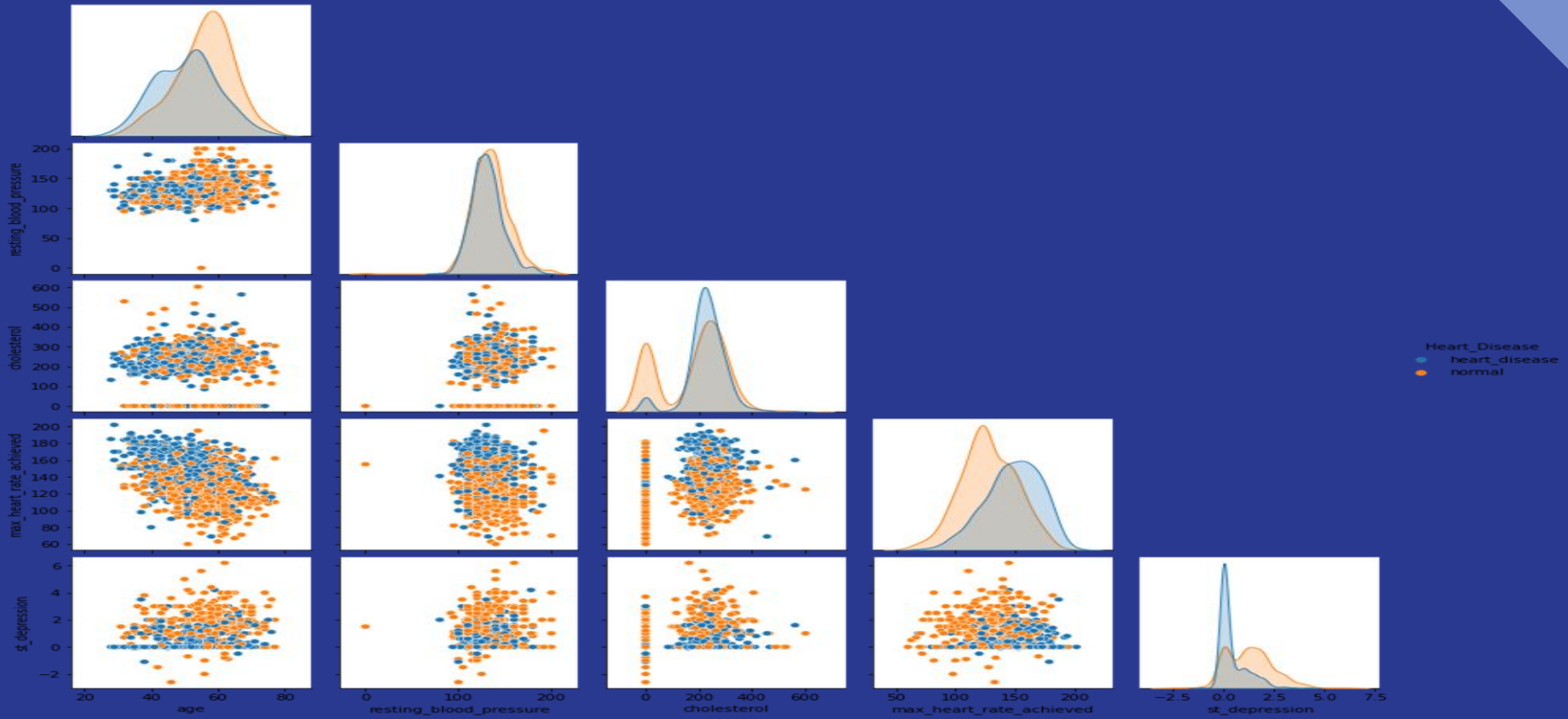
Age 50 older are more likely to have heart disease.



Asymptomatic is the most common within patients.



Pairplot with Heart Disease



Logistic Regression

	precision	recall	f1-score	support
0	0.88	0.83	0.86	119
1	0.88	0.92	0.90	157
Accuracy			0.88	276
macro avg	0.88	0.87	0.88	276
weighted avg	0.88	0.88	0.88	276

Random Forest Classification

	precision	recall	f1-score	support
0	0.88	0.84	0.86	116
1	0.89	0.91	0.90	160
accuracy			0.88	276
macro avg	0.88	0.88	0.88	276
weighted avg	0.88	0.88	0.88	276

SVC

	precision	recall	f1-score	support
0	0.88	0.87	0.87	113
1	0.91	0.91	0.91	163
accuracy			0.89	276
macro avg	0.89	0.89	0.89	276
weighted avg	0.89	0.89	0.89	276

Cat Boost

	precision	recall	f1-score	support
0	0.89	0.88	0.88	114
1	0.91	0.93	0.92	162
accuracy			0.91	276
macro avg	0.90	0.90	0.90	276
weighted avg	0.91	0.91	0.91	276

LGBM Classifier

	precision	recall	f1-score	support
0	0.87	0.83	0.85	117
1	0.88	0.91	0.89	159
accuracy			0.87	276
macro avg	0.87	0.87	0.87	276
weighted avg	0.87	0.87	0.87	276

Best Result

	Model	Validation Score	Cross_Validation Score
0	LogisticRegression	0.880435	0.864496
1	RandomForest	0.884058	0.871900
2	SVC	0.894928	0.866230
3	CatBoost	0.905797	0.877568
4	LGBM	0.873188	0.865816

Conclusion

- As we can see, all of our models obtained decent results within the validation and cross validation score.
- We can see that the Light GBM model performed the worst than the Logistic Regression, SVC, and Random Forest within the validation score.
- It seems CatBoost performed the best within the validation score
- Within the cross validation score we can see that Logistic Regression performed the worst.
- While CatBoost seems to performed the best within the cross validation score as well