Capstone Project

Mohamad Fadil

Machine Learning Engineer Nanodegree

May 18, 2021

Definition

Project Overview

The excessive changes in the lifestyles of inhabitants across the world through the last decades has moved the human societies from farming foods and active lives into fast foods and inactive lifestyles. The grouping of such lifestyle with growing cigarette consuming has improved the risk factors of cardiovascular diseases (CVDs). Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. [1]

Heart failure is a common event caused by CVDs. Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

In this project, I have created a machine learning model that is capable of predicting the heart patient will survive until his/her next follow up session. By utilizing the 13 features from the dataset, I have managed to get best average precision score of 0.82 and recall score of 0.77.

Problem Statement

Healthcare is an inevitable task to be done in human life. Cardiovascular disease is a broad category for a range of diseases that are affecting heart and blood vessels. [2] The health care industry has a lot of data; hence this data should be used with machine learning algorithms in order to enhance the probability of reducing patients' risk by predicting ahead if a patient is more likely to be attacked by a specific illness or not based on their lifestyle and health records data. Hence, the problem to be solved is to predict if a patient is going to decease in the following follow up session.

1) To increase the probability of patient survival by predicting beforehand on either the patient will survive or not.
2) To have predictive capability in predicting any death events for future heart patients

The final model is expected to use the 13 features and machine learning algorithm to predict the death event for heart patients.

Evaluation

For this project, I used recall and precision for evaluation metric as the target output is imbalanced, using accuracy alone will not give us good idea on how well the model performed.

Precision is attempts to find what proportion of positive identifications was actually correct. [3]

$Precision = TP \div TP + FP$

Recall is the attempt to find what proportion of actual positives was identified correctly [3]

$Recall=TP \div TP+FN$

By using both of this metrics, I manage to choose the best algorithm for this problem and also manage to evaluate the effectiveness of the machine learning model in predicting the classes.

Analysis

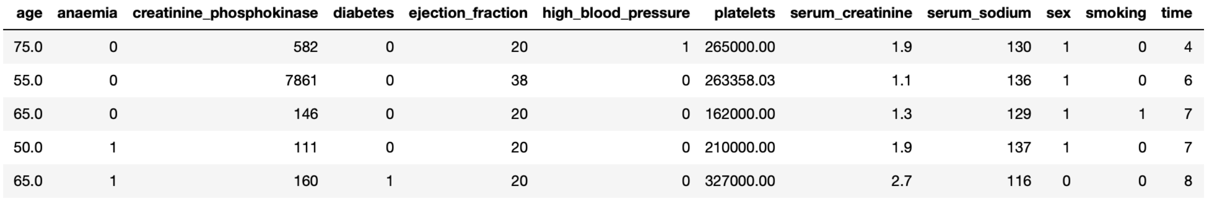
Data Exploration

The dataset is obtained from Kaggle [4]. It contains total of 13 features including the target variable. Below is the list of features and their description.

Table 1. Data Dictionary

Name	Description
age	Age of the patient
anaemia	Decrease rate of red blood cell

creatinine_phosphokinase	Level of the CPK enzyme in the blood
diabetes	Patient has diabetes or not
ejection_fraction	Percentage of blood leaving the heart at each contraction
high_blood_pressure	Patient has high BP or not
platelets	Platelets in the blood (kiloplatelets/ mL)
serum_creatinine	Level of serum creatinine in the blood (mg/dL)
sex	Woman =0 , Man = 1
smoking	If patient smokes or not
time	Follow up period (days)
death_event (target variable)	Did patient decease or not



age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time
75.0	0	582	0	20	1	265000.00	1.9	130	1	0	4
55.0	0	7861	0	38	0	263358.03	1.1	136	1	0	6
65.0	0	146	0	20	0	162000.00	1.3	129	1	1	7
50.0	1	111	0	20	0	210000.00	1.9	137	1	0	7
65.0	1	160	1	20	0	327000.00	2.7	116	0	0	8

Figure 1. Sample data

After the data is loaded, it is time to get some statistical information on each feature, this information consists of getting the count, minimum value, maximum value, standard deviation, mean, 25 percentile, 50 percentile and 75 percentiles of each feature. By doing this we manage to get a higher overview of how the data is distributed and which columns are continuing and categorical.

From the statistics we manage to see that the average age of our patients is 60 years old, almost half of the patient has decrease in red blood cells, almost half of them have diabetes, around 0.35 of patients have high blood pressure and 0.3 patients are smokers.

Explanatory Data Analysis

In this section, I will go thru each of the features and correlate the feature with the target variable in order to find relationship between the feature and target variable. By performing this activity, I manage to find features that are more likely to have more effect towards predicting the target variable. Hence, I will be sharing all the methods used in this section. Starting off, I used the mutual info classification algorithm to find any sort of correlation between the features and target variable. Mutual information (MI) between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency.[5]

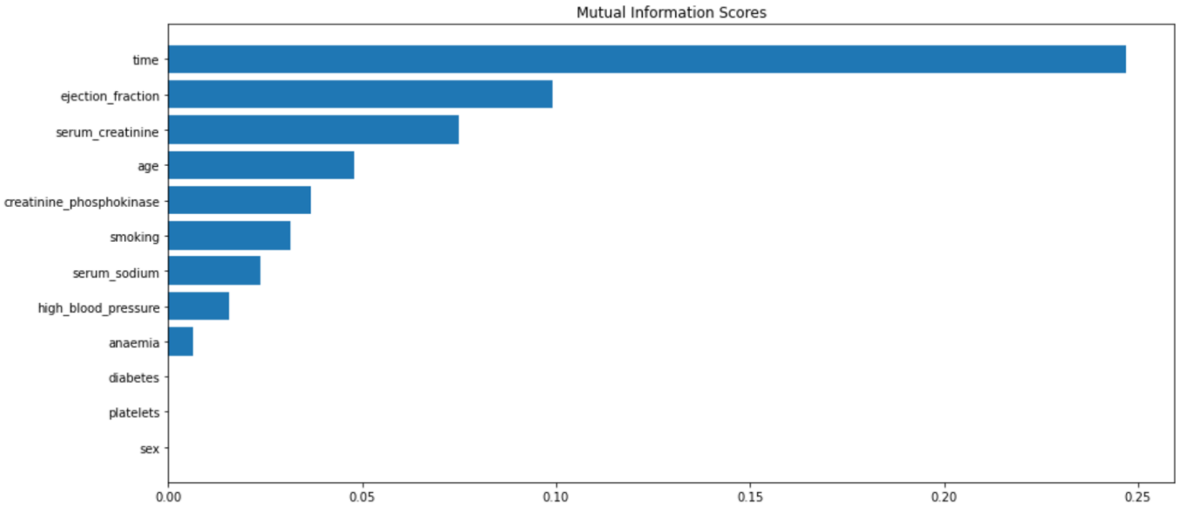


Figure 2. Mutual Info Score Per Feature

From the figure 2, I learned that time of next follow up has the highest score towards the target variable, this could be because patients that are having longer follow up days have higher chances of not surviving due to gap of checkup, or it could be that shorter time frame can have higher probability of causing death to the patient, as mutual info doesn't provide us information on either the feature is positively or negatively correlated, we often need other type of correlations to support its output, hence I performed spearman correlation on the dataset to find supporting information on the mutual information scores.

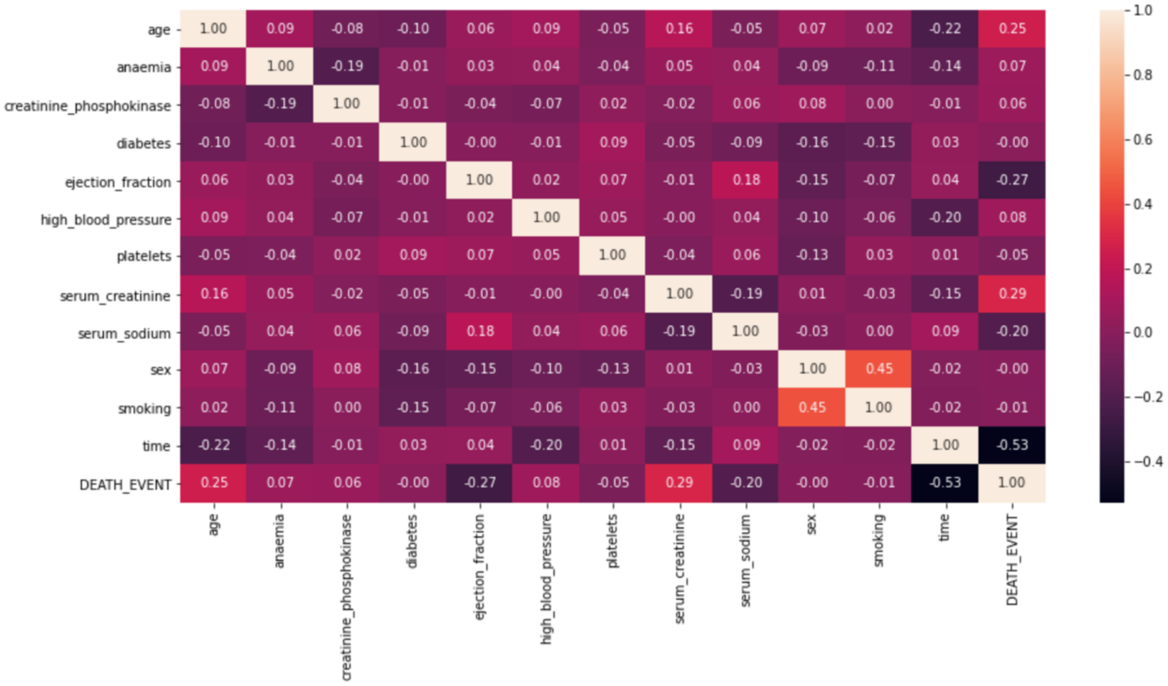


Figure 3. Spearmen Correlation on Dataset

With the help of figure 3, I manage to understand that there are 5 main features that are either positively or negatively correlated towards the target variable. The first one is time which is negatively correlated, and what this means is that as the time to next follow up date is shorten, the probability of survival is less as well, and this could be because patients that are in critical zone are more likely to have shorter gap between follow ups.

The second feature that has some correlation with target variable is ejection_fraction which is also negatively correlated towards target variable, which means that as the percentage of blood leaving out at each pump is reducing, the rate of survival is getting lower and the patient is more likely to decease.

The third feature is serum_creatinine which is positively correlated with target variable, and what this means is that when the level of serum creatinine in the body is increasing it also means that the patient is more likely to not survive until his/her next follow up session.

The fourth feature is age which has positive correlation with target variable, and this is pretty straightforward, where are older patients with heart problem are less likely to survive.

The fifth feature with some correlation with target variable is serum sodium and this feature is negatively correlated, and what this means is as the level of serum sodium in patients body starts decreasing, it will cause the patient to have less probability of surviving.

After getting the high viewpoint of the dataset, it is time to look at each feature correlation with target variable visually. Hence in this part, I will be sharing my findings on visualization of each features.

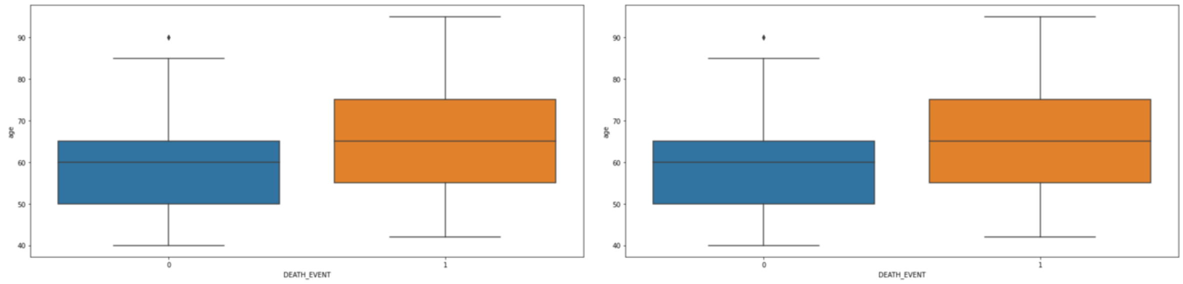


Figure 4. Age and Anaemia chart against target variable

From figure 4 we can learn that patients who survive to the next follow up session tends to have lower median age compared to the ones that do not survive. Besides that, patients that doesn't have decreased in red blood cell are more likely to survive.

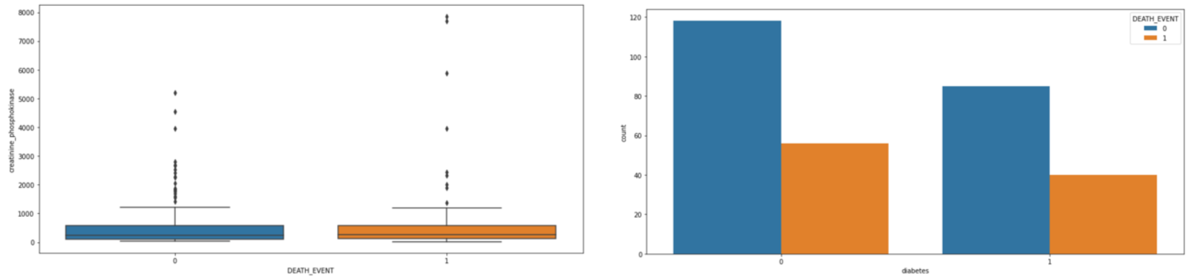


Figure 5. CP and Diabetes chart against target variable

From figure 5 we can learn that having lower creatinine_phosphokinase can boost the probability for the patient to survive and since the number of patients with diabetes is higher than the ones with diabetes, hence the second chart does makes clear sign that not having diabetes will help in increasing the probability of survival.

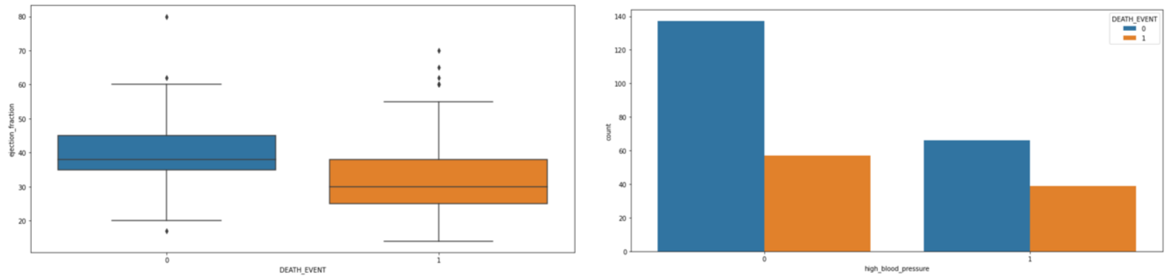


Figure 6. EF and HBP chart against target variable

From figure 6, we can clearly learn that having lower ejection fraction can cause in increase of probability of not surviving and having high blood pressure does increase the probability of not surviving as well.

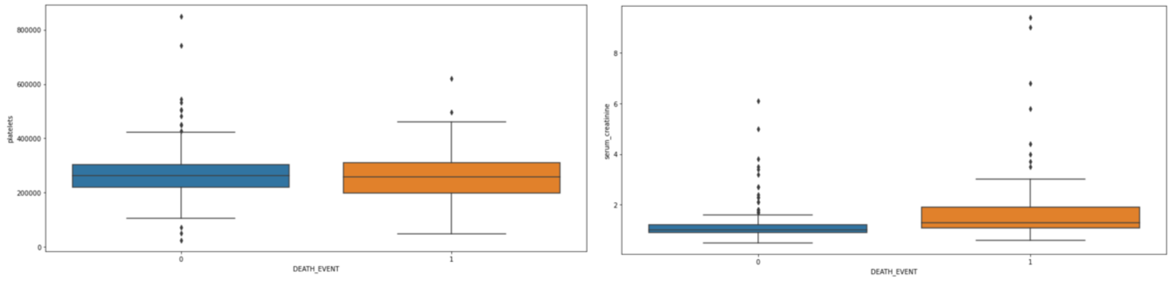


Figure 7. Platelets and SC chart against target variable

From figure 7, we can't differentiate much on the platelets as the median is almost the same for survivors and not survivors, however the lower quartile for none survivors is lower which indicates that being in the 25 percentile zone can be dangerous for the patient. On the other hand, serum crestinine has clear signs that having higher serum level in body can reduce the chances for survival.

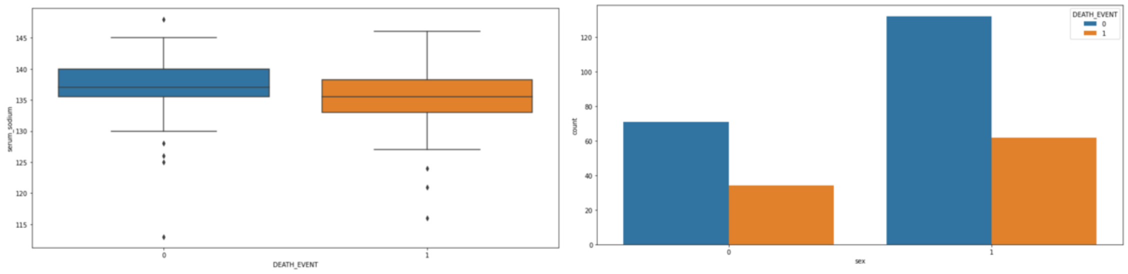


Figure 8. SS and Sex chart against target variable

From figure 8 we can learn that having lower serum sodium level in body does effect the probability of survival and as the data is not balanced between woman and man, hence we can see clearly that man has higher chances of surviving and also not surviving.

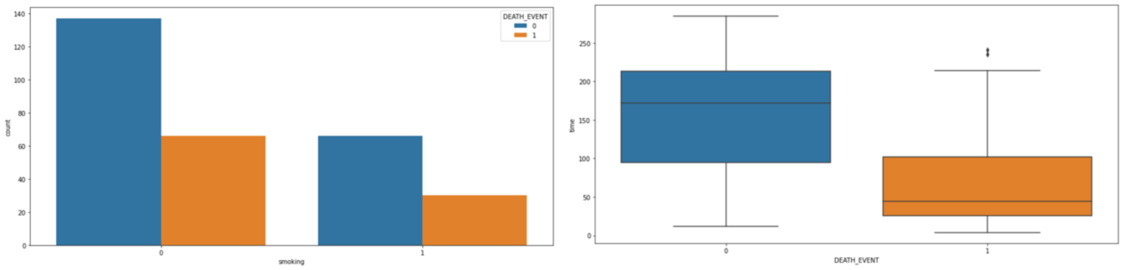


Figure 9. Smoking and Time chart against target variable

From figure 9, we can learn that being a smoker or non smoker still gives high probability for not surviving, this is due to imbalance value counts between the smokers and non smokers. However, the time feature has clear sign that patients with less days to next follow up appointment tend to have higher probability of not surviving. Once I understand each feature relationship with the target variable, I moved to the next phase where I define the algorithm and methods to be used for building the machine learning model.

Benchmark

As for the benchmark model that will be used to compare other models' performance, I decided to use the decision tree model which got third place during the training and model selection phase with precision of 0.7 and recall score of 0.67. The goal will be to beat this models performance using other algorithm.

Algorithms and Techniques

Classification is a common task of machine learning (ML), which involves predicting a target variable taking into consideration the previous data. To reach such classification, it is necessary to create a model with the previous training data, and then use it to predict the value of the test data. This process is called supervised learning, since the data processing phase is guided toward the class variable while building the model.

Predicting either the patient will survive or not is a classification task as the output will be either one of the class 0 (not survived) and 1 (survived).

For this project, I implemented these models and find the best based on validation performance:

Support Vector Machine (SVM) - The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. [6]

Decision Tree Classifier - Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation. [7]

Logitstic Regression - Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. [8]

K Neighbors Classifier - KNN is an algorithm that is considered both non-parametric and an example of lazy learning. Non-parametric means that it makes no assumptions. The model is made up entirely from the data given to it rather than assuming its structure is normal. Lazy learning means that the algorithm makes no generalizations. This means that there is little training involved when using this method. Because of this, all of the training data is also used in testing when using KNN. [9]

Random Forest Classifier - A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. [10]

In the end all models will be compared based on their performance based on validation dataset, and model with the best score will be used for final product

Implementation

In this phase I will describe the steps taken in order to build the machine learning model. During the first step the data was divided into training and testing set. The testing set was 20% of total data size, which means the test set consist of 60 rows and the training set consists of 239 rows which total ups to 299 rows. The train and test split was done by using sklearn train_test_split method, the shuffle parameter was set to true in order to add randomness to the dataset for training and testing.

Once the dataset was splitted, I prepared a list of models that will be used for training the data, the reason for creating this list is to get benchmark on each algorithm and their accuracy on testing set based on the evaluation metrics chosen, which are precision and recall.

Next, each of the algorithms were trained on the same training dataset and tested on the same testing set, the reason of using the same dataset is so that all the models are returning performance metrics based on the same data distribution, that way it is easier to pick the best algorithm. Once all the models are trained and tested, the performance metric is then collected and tabulated for comparison purpose.

After all the models were trained, referring to table 2 in section models evaluation, I learned that SVC was performing bad more than it should, hence I decided to scale the data and retrain the model. The result of this exercise was massive improvement in the performance of the model especially on the precision score.

The data was scaled using sklearn MinMaxScaler method which will scale all the values to be between 0 and 1, and for values that are already between 0 and 1 nothing will change, since we have some features that are continuous and categorical, this sort of scaler method would fit well to the purpose.

Moreover, I decided to retrain all the models on the scaled data, and the result is shared in table 3 in models evaluation section. SVC, Decision Tree and KNN seems to gain the upper hand with scaled data, others were stagnant or even had performance dropped like the Logistic Regression and Random Forest model.

Final experiment done in order to find the best and optimal solution for this problem was to utilize light gradient boosting model algorithm to train the dataset. The idea is to utilize the power of gradient boosting with ensemble capability in order to beat Logistic Regression performance, and as per the result in table 1, we can see that lightgbm is more consistent in both metrics precision and recall compared to Logistic Regression, hence I would say that it would be optimal to go with lightgbm as final choice of algorithm as it will have more confidence in predicting the classes.

Result

Models Evaluation

After the model was developed and all the models had be trained on the dataset, then all the performance metrics were recorded in order to choose the best performing model.

Table 2. Models Performance Comparison

Model	Precision	Recall
Logistic Regression	0.82	0.77
KNN	0.43	0.47
Decision Tree Classifier	0.66	0.63
Random Forest Classifier	0.79	0.71
SVC	0.29 0.75 (scaled data)	0.50 0.65 (scaled data)
Light GBM	0.79	0.79

Table 3. Scaled data models performance

Model	Precision	Recall
Logistic Regression	0.79	0.67
KNN	0.60	0.53
Decision Tree Classifier	0.70	0.67
Random Forest Classifier	0.74	0.70
SVC	0.75	0.65

Based on the table 2, we can see that Logistic Regression algorithm performed the best on the dataset based on precision but not recall, followed by Random Forest and Decision Tree. However, the lightgbm model manage to generalize better and managed to get high score for both of the metrics that is used for evaluation. Hence it is agreed that lightgbm should be used as final model for this problem.

Justification

The reason why I agree to use lightgbm compared to Logistic Regression is because if we look at the scores for the two metrics LR does perform better in terms of precision but lightgbm is generalizing better and performing on both of the metrics as the same capability, hence it shows that lightgbm will have better prediction confidence compard to Logistic Regression. Besides that, lightgbm is also performing better in comparison with our benchmark model which is Decision Tree, in below table 4 we can see how lightgbm has higher score in recall compared to Logistic Regression and higher score in both precision and recall in comparison to Decision Tree.

Table 4. Comparison of LR, DT, and LightGBM Performance on Test Data

Model	Precision	Recall
Decision Tree (Benchmark)	0.68	0.65
Logistic Regression	0.82	0.77
LightGBM	0.79	0.79

Conclusion

In this project I built a classifier model that has the capability to predict either a heart patient is going to survived or not before their next follow up session. By preprocessing the data and visualizing it, we learned that not all features have direct impact towards the target variable. However, some features such as age, level of red blood cell, level of enzyme in blood, level of serum sodium in blood and days to next follow up session does play a huge part in predicting either the patient will survive or not. With this model we have the capability of predicting the death event as macro average precision of 0.79 and macro average recall of 0.79. This might not be the most ideal performance for this model, but it is a good start for the project.

Reflection

The following steps were taken in order to finish the project:

1. Find a problem and dataset
2. Define the problem that the dataset is providing

3. Download the dataset and process it
4. Perform explanatory data analysis on the dataset to understand each feature and how it can help in solving the problem.
5. Train multiple machine learning algorithms on the training and test set
6. Choose the best algorithm as the final algorithm for the solution

The most interesting part of this project is where I have to define a real world problem and find the dataset that can help to solve this problem, and followed by performing visualization on the dataset which gives an detail view on what the dataset is about and how each feature is correlated towards the target variable. Lastly, building the machine learning models is interesting as well, as it shows that certain algorithm behaves in their own way towards the data distribution.

Improvement

To enhance the model in future, I believe having more data will benefit it the most as it will help the model to generalize the data patterns better and enhance the capability of predicting the survival rate. Apart from that, once there is more data we can use Artifical Nueral Network (ANN) to enhance the accuracy of the prediction as ANNs have better capability in finding patterns and generalizing data provided there is enough data for the algorithm to learn from.

References

[1] Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone (2020). Accessed 11 May 2021. Available at https://bmcmedinformdecism ak.biomedcentral.com/article s/10.1186/s12911-020-1023-5

[2] Prediction of Cardiovascular Disease Using Machine Learning Algorithms (2018). Accessed 11 May 2021. Available at https://www.ncbi.nlm.nih.gov /pmc/articles/PMC5304098/

[3] Classification: Precision and Recall. Accessed 14 May 2021. Available at https://developers.google.co m/machine-learning/crash- course/classification/precision-and-recall

[4] Kaggle. Accessed 11 May 2021. Available at https://www.kaggle.com/andrewmvd/heart-failure-clinical- data

[5]https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html

[6] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[7] https://scikit-learn.org/stable/modules/tree.html#tree

[8] https://www.statisticssolutions.com/what-is-logistic-regression

[9]https://medium.com/capital-one-tech/k-nearest-neighbors-knn-algorithm-for-machine-learning-e883219c8f26

[10]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html