A Machine Learning Approach for Heart Attack Prediction

DongWhan Jun, Jiazhen Cui, Emily Wang, Maopin Yan

Abstract — Big data insights play a fundamental role in predicting future health and providing people with better health outcomes. As we know, heart attacks are one of the worldwide epidemics that have become the leading cause of death in humans over the past few decades. Many studies are using machine learning techniques for predictive analysis to reveal better decision making. Therefore, accurate prediction of myocardial infarction is of great value for early intervention and treatment. More specifically, we consider the common problems of data embedding, data preprocessing, feature selection, and prediction in data sets. All process we did with spark in google collab, which includes mllib and ml libraries. The data collected by Kaggle are preprocessed, and the data are trained and tested by Logistic regression, gradient ascending tree classifier model, SVC, Naive Bayes, and decision tree. Our work achieved these models with different features and found that Random Forest model has 0.87 accuracy score which is the best one compared with others. Also, Random Forest has 0.86 precision and 0.88 recall.

Keywords: Heart Attacks, Machine Learning, Accuracy, Spark, Modeling, RDD

I. INTRODUCTION

Heart attack is one of the more serious circulatory disorders in human society. The heart, the main organ of the body, is used to pump blood throughout the body through the blood vessels of the circulatory system. In the circulatory system, the most crucial role is fulfilled by the heart. If the operation of the heart is damaged due to any condition and does not function properly, then it may lead to serious health problems, including death. It brings a highly dramatic impact not only on the quality of life of patients but also imposes an onerous emergency burden on patients and the country.

As people's living standards improve, their lifestyles and behaviors are inextricably linked to heart attack, with lack of exercise, smoking, poor diet, and obesity are known risk factors for heart disease management. Currently, the diagnosis and treatment process is very challenging due to the shortage of doctors and diagnostic instruments that affect the treatment of heart disease patients. Heart disease can be predicted based on various symptoms such as age, gender, pulse, etc. In healthcare, data analytics helps to predict disease, improve diagnosis, analyze symptoms, provide appropriate medications, improve quality of care, minimize costs, extend life expectancy, and reduce mortality in cardiac patients. Being able to predict the onset of heart disease will help to start early treatment for potential patients. Early treatment will provide a greater chance of survival and recovery. In the meantime, it may reduce the likelihood of permanent disability and the need for extensive rehabilitation.

The purpose of this project is to analyze and predict future heart attacks. Heart disease includes several symptoms, such as weakness, irregular breathing, and swollen feet. Medicine alone cannot reduce the occurrence of heart disease at all, so we must use some technology. In this project, we will develop classification algorithms that can automatically predict the results of a heart attack by using a machine learning approach.

In our study, we will use the dataset collected by the Heart Attack Database. Our approach starts with preprocessing the data to select features that are predictive of a heart attack. The data are trained and validated by using machine learning techniques. After explaining our method and approach, this paper presented and discussed the results of our work and made conclusions to indicate what can be done in the future to improve the method of predicting heart attacks using machine learning.

II. DATASET

This dataset [1] shared by Rahman (2021), is a public source on Kaggle. It includes 14 attributes and 303 instances. The original cleaned dataset has already been transformed into all labeled integer values, which will be convenient for future modeling. However, we transformed the labeled values back into string values for visualization. When we investigated this dataset, we know this dataset is clean with no null values and balanced. 138 (45.5%) out of 303 instances are labeled as ‘less chance of heart attack’. That means the data is quite balanced with respect to the target variable class.

III. METHODOLOGY

The machine learning task is a binary classification that predicts whether a patient will develop heart

disease. However, our task in analyzing these data sets was not only to build a valid predictive machine learning model, but also to discover under what conditions patients had a higher prevalence and the relationship between characteristics and activity success. Therefore, exploratory data analysis will be performed to show the relationships between different attributes. Then, the data is preprocessed to prepare features for the ML model. Finally, we will isolate the data to train and validate the model.

A. Tools

We use Apache Spark (PySpark) as our primary tool. Two machine learning libraries, mllib and ml, are provided in Spark. The operation of mllib is based on RDD (resilient distributed dataset), while ml is based on DataFrame, which is a mainstream machine learning library. The ml package includes three main abstract classes: Transformer, Estimator, and Pipeline.

Transformer classes transform data by appending a new column to the DataFrame. At a high level, when deriving from the Transformer abstract class, each new Transformer class needs to implement the transform() method. This method requires passing a DataFrame to be transformed, which is usually the first and only mandatory parameter.

The pipeline in pyspark ML is used to represent the end-to- end process from transformation to evaluation (with a series of different stages), which can perform necessary data processing (transformation) on some input raw data (in the form of DataFrame), and finally Evaluate the model.

A pipeline can be thought of as consisting of a series of different stages. When the fit method is executed on a Pipeline object, all stages are executed in the order specified in the stage parameter; the stage parameter is a list of transformer and evaluator objects. The fit method of the pipeline object executes the transform method of each transformer and the fit method of all evaluators.

B. Exploratory Data Analysis

We use both Matplotlib and Tableau for data visualization in categorical features and numerical features.

- *Categorical features*

Below are the major categorical features in the dataset, we visualize their relationships with the target variable through bar graphs.

Fig. 1 shows the distribution of the target variable, output. Among all the 303 instances, 45.5% of the value '0', which is less chance of heart attack. Fig.2 shows the chances of a heart attack under different levels of chest pain. Through this distribution, we found that there is a very high possibility of a heart attack under atypical angina, non-anginal pain, and asymptomatic. The chance of heart attack under typical angina is relatively low. Fig. 3 reveals that there is no strong relationship between 'fasting blood sugar' and 'chance of heart attack'. Fig. 4 is the distribution of 'chance of heart attack' under 'Exercise-Induced Angina'. It shows that among those people with less chance of heart attack, nearly 50% of them had exercise-induced angina. However, for those people who have a higher chance of heart attack, only 14% of them had exercise-induced angina. Fig. 5 illustrates the chances of a heart attack under different electrocardiographic resting. It shows people have ST-T wave abnormality have higher chances of heart attack, while people with normal or left ventricular hypertrophy by Estes' criteria have a lower chance of heart attack.

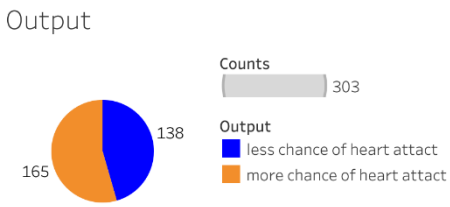

Figure 1 Target Variable Distribution

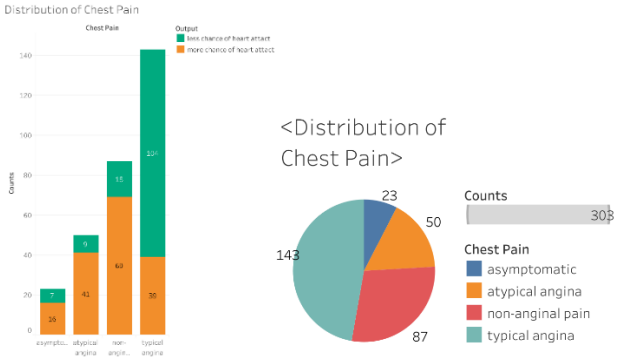

Figure 2 Distribution of Chest Pain

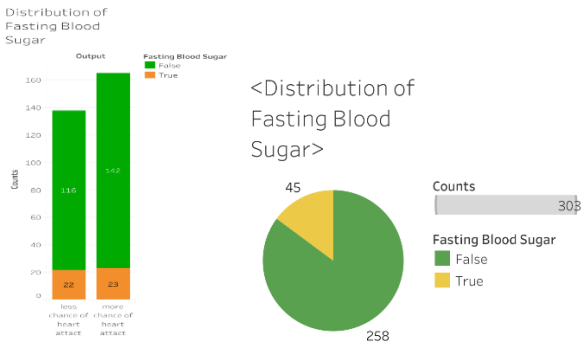

Figure 3 Distribution of Fasting Blood Sugar

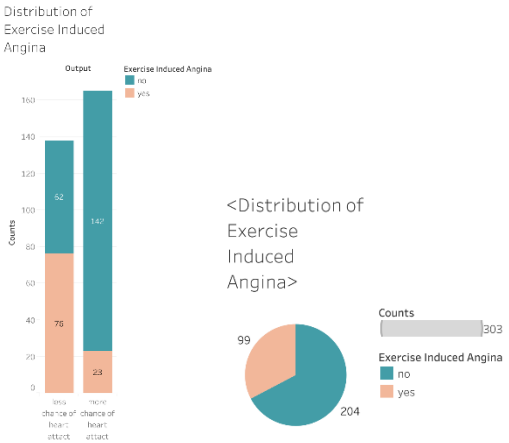

Figure 4 Distribution of Exercise Induced Angina

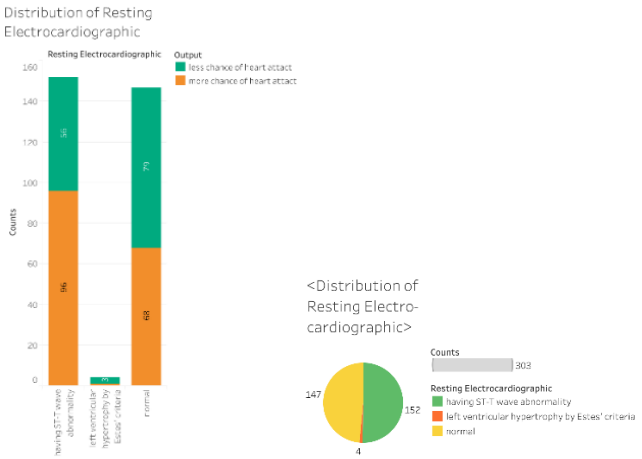

Figure 5 Distribution of Resting Electrocardiographic

- *Continuous Numeric features*

Fig. 6 is the correlation matrix for the continuous features in this dataset. It seems that all these six continuous features have very low correlation coefficients.

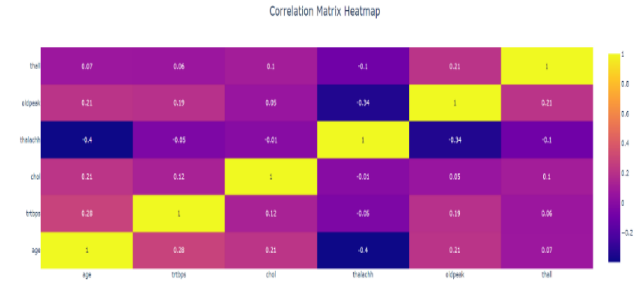

Figure 6 Correlation Matrix for the continuous features

Fig. 7 and Fig. 8 are the scatter plots of some numerical features under 'less chance of heart attack' or 'more chance of heart attack' of the output. Fig. 7 shows that with the age less than 54 or greater than 70, there is a higher possibility of heart attack. For those group of people with age less than 54 and resting blood pressure less than 150, there is a high chance of having a heart attack. Fig. 8 is the scatter plot of the relationship between Previous Peak and Cholesterol. The graph shows while the Previous Peak reaches 2 or higher, there is a less chance of heart attack. There is an outlier that when cholesterol is 550, there will be a high chance of heart attack.

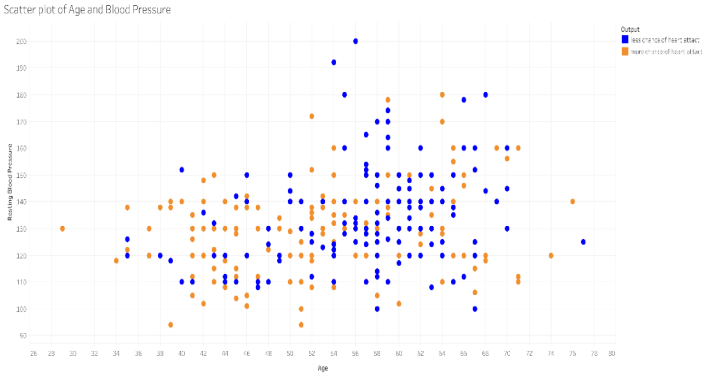

Figure 7 Scatter plot of Age and Blood Pressure

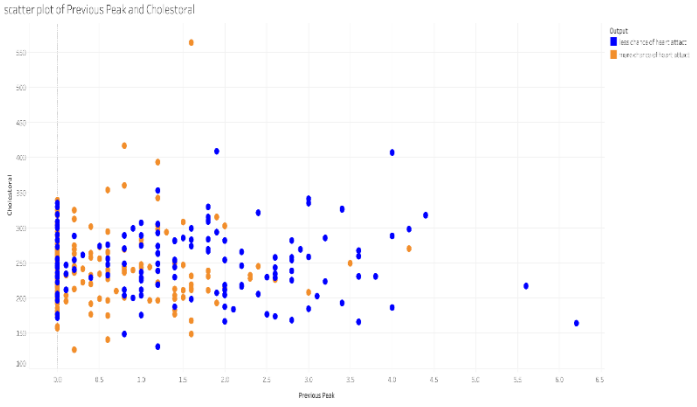

Figure 8 Scatter plot of Previous Peak and Cholesterol

C. Random Forest

Random Forest is a supervised learning technique that, as its name suggests, combines many individual decision trees built during training period to form Ensemble learning (also known as nearest neighbor predictor) and is used to tackle classification and regression problems. The individual tree whose class prediction is upvoted the most or the class with the largest mode value becomes the prediction of model [2] in a swarm of such decision trees. To put it another way, it creates many separate decision trees before combining them to improve accuracy and predictability. By averaging the two extreme problems of high variance and strong bias, it achieves a natural balance [3]. Random Forest has many advantages such as:

- *It can produce data of very high dimensions (many features) without dimensionality reduction and feature selection.*
- *It can determine how important feature is.*
- *It is easy to implement and can avoid overfitting.*
- *For unbalanced dataset, it can balance out errors.*
- *If a significant portion of the feature is missing, accuracy can still be maintained.*

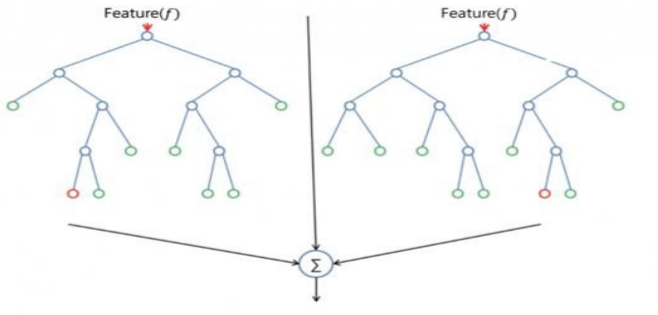

Fig. 9. Basic Random Forest Model

D. Decision Tree

The decision tree classifier with the ID3 algorithm is the second classification we use. A decision tree classifier is a predictive modeling approach that classifies sample data using a tree model. Nodes represent attribute tests, leaves represent class labels, and branches represent attribute values that lead to nodes or class labels. This step is repeated recursively on each extracted subset until the subset at a node has the same value of the target variable or adds no value to the prediction when splitting [4]. Figure 10 depicts an example of a decision tree. The decision tree is constructed by splitting the dataset into subsets based on a value test performed on an attribute.

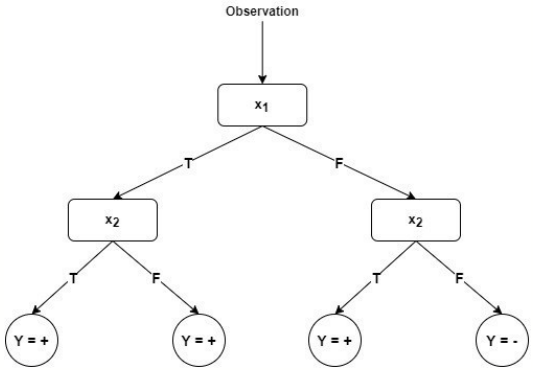

Fig. 10. Decision Tree Example

E. Logistic Regression

Logistic regression is a “supervised machine learning classifier that extracts real-valued features from the input, multiplies each by a weight, sums them, and passes the sum through a sigmoid function to generate probability. The threshold is used to make decision” [5]. It is the baseline classifiers used widely for classification problems which made it suitable for this study.

F. Gradient-boosted Tree classifier Model (XGBoost)

XGBoost stands for “Extreme Gradient Boosting” and it is a supervised learning tree ensemble model that consists of a set of classification and regression trees (CART). XGBoost is an effective machine learning model, even on datasets where the class distribution is skewed [6]. Hence, this classifier is used for this study.

G. SVC

For our study, the next most applicable machine learning algorithm is linear SVC. The goal of linear SVC (support vector classifier) is to adapt to the data provided, returning a "best fit" hyperplane that divides or classifies the data. From there, after obtaining the hyperplane, we can provide some features to our classifier to see what the "predicted" category is. It makes this algorithm quite appropriate for our purposes.

H. Naïve Bayes

The last type of classifier we use for heart attack prediction is Naïve Bayes classifier. The

Naïve Bayes classifier is a probabilistic machine learning model based on the Bayes theorem.

$$P(y|x) = \frac{P(y)P(\mathbf{x}|y)}{P(\mathbf{x})}$$

The Bayes theorem finds the probability of y happening when x occurs. In the equation, y represents the final class, which is stroke or no stroke, and **x** represents all the features or predictors of stroke. $P(y)$ is referred to as the prior and represents the probability of a certain class and $P(x|y)$ represents the probability of class Y generating the observed features x.

For the Naïve Bayes approach, we assume that the features are conditionally independent and rewrite $P(x|y)$ as:

$$P(\mathbf{x}|y) \approx \prod_{j=1}^{D} P(x_j|y)$$

Substituting the $P(x|y)$ into the previous equation we would end up with:

$$P(y|x) = \frac{P(y)\prod_{j=1}^{D} P(x_j|y)}{P(\mathbf{x})}$$

For our method, we are using norm probability density function or the Gaussian Na¨ıve Bayes to calculate $P(x_j|y)$. This method assumes the values of the dataset sample for a Gaussian distribution or normal distribution. As a result, $P(x_j|y)$ can be determined by:

$$P(x_j|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\,\sigma_y^2}\right)$$

It is important to note that the denominator of the $P(y|x)$ equation is the same for all entries in the dataset and can be removed to form the following proportionality.

$$P(y|\mathbf{x}) \propto P(y)\prod_{j=1}^{D} P(x_j|y)$$

Finally, to determine the classification of a patient or sample we would find the class with the maximum probability for it.

$$y = argmax_y P(y)\prod_{j=1}^{D} P(x_j|y)$$

IV. RESULT AND ANALYSIS

Description of metrics:

1) Precision: How close measured values are to each other. Measurements can be accurate but not precise and vice versa. Precision can be calculated by dividing the true positive by the sum of the true positive and false negative as below:

$$\text{Pre}cision = \frac{TP}{TP+FP}$$

2) F-1 Score: F1 is the harmonic mean of precision and recall. F1 is useful in situations where the analysis views precision and recall as equally valuable. The formula for calculating f1 is as follows:

$$\text{F1} = \frac{2 * \text{Pr}\,e\,cision * Recall}{\text{Pr}\,e\,cision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

3) Recall(sensitivity): Recall is the number of values that were classified correctly. The recall values can be calculated by dividing the true positives by the sum of the true positives and false negatives as follows:

$$Recall = \frac{TP}{TP + FN}$$

4) Accuracy: Accuracy represents the number of correctly classified data instances over the total number of data instances. Accuracy may not be a good measure if the dataset is not balanced (both negative and positive classes have a different number of data instances) [7]. Hence, we calculated the other metrics as well in order to evaluate the performance of the classifiers. Accuracy can be represented as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Results:

Algorithm	Accuracy Score	precison Score	recall	F1
Random Forest	0.874	0.86	0.88	0.88
Decision Tree	0.771	0.79	0.72	0.72
Logistic Regression	0.816	0.80	0.84	0.84
Gradient-boosted	0.759	0.78	0.72	0.72
SVC	0.828	0.82	0.84	0.84
Naive Bayes	0.793	0.79	0.72	0.72

Fig. 11. Metrics Comparison Table

As we can see the result, with those 6 different models, the random forest model has 0.874 accuracy score which is relative higher than other models. For some

reason, the recall and F1 score are all same. Decision tree model has the lowest accuracy score.

V. CONCLUSION

In this research, various Supervised ML classifiers namely, Random Forest, Decision Tree, Gradient-boosted Tree, SVC, Naïve Bayes, and Logistic Regression have been used to deploy a model for Heart Attacks prediction. Based on our previous discussion and analysis, the results impersonate that the Logistic Regression classifier compared to other ML classifiers is achieving the highest accuracy score in such a way that prediction used by our model. Despite some limitations to be overcome to implement ML algorithms in medical practice, in general, ML algorithms show promising results. This research study delves deeply into machine learning techniques for heart attack identification.

The contribution of the classifier is essential in the healthcare industry because the findings can be used to forecast the diagnosis that can be given to patients. Existing techniques are studied and compared to develop efficient and accurate systems. Machine learning techniques improve the accuracy of disease risk prediction, allowing patients to be identified at an early stage of disease and benefit from preventive treatment. It can be concluded that machine learning algorithms have a large potential for predicting heart conditions or heart-related diseases.

VI. FUTURE WORK

On the dataset obtained from the Heart Attack database, we investigated six different types of classification techniques. In the future, we could indeed test different types of machine learning algorithms on about the same dataset whether there are any better solutions for predicting heart attacks. Artificial neural networks and k-nearest neighbors are two other machine learning classification techniques which can be used. It is also possible to investigate other data balancing techniques to determine which method produces the best results with the dataset.

Aside from experimenting with other machine learning techniques, different datasets can be analyzed and trained for comparison. Various datasets may contain multiple predictor variables of heart attack, resulting from different effects because appropriate elements have greater or lesser degrees of correlation with heart attack. Furthermore, analyzing different or additional groups of patient data can considerably reduce overfitting and producing better results.

VII. REFERENCES

[1] Rashik Rahman, 2021. Heart Attack Analysis & Prediction Dataset

[2] T. Yiu, “Understanding Random Forest: How the Algorithm Works and Why it Is So Effective,” *Towards Data Science*, 2019.

[3] N. Donges, “A complete guide to the random forest algorithm,” *Built In*, 2019.

[4] S. Moro, P. Cortez and P. Rita, ”A Data-Driven Approach to Predict the Success of Bank Telemarketing,” Decision Support Systems, Elsevier, vol. 62, pp. 22-31, Jun. 2014

[5] V. Singh and S. P. Lal. Digit recognition using single layer neural network with principal component analysis., 2014.

[6] Xgboost 1.5.1 Documentation. Introduction to boosted trees.

[7] Harikrishnan N B. Confusion matrix, accuracy, precision, recall, f1 score.