Predicting Heart Disease Through the Application of Machine Learning Techniques Using the Python

Zihang Fu *

KangChiao International School

* Corresponding Author Email: 1403082227@qq.com

Abstract. This paper is mainly about using machine learning methods to predict cardiovascular disease. Because cardiovascular disease now ranks first in mortality and the methods and costs of curing heart disease are very small and expensive, so why not take a step earlier? As for discovering diseases, there are not many projects that use machines to predict diseases, and it costs a lot of money to predict diseases in hospitals. Using machines can be faster and more accurate. Another point is that society has entered the era of artificial intelligence, and the medical industry It should also usher in an improvement and should keep pace with other industries, so I want people to experience predicting diseases, such as performing cardio-tracheal examinations faster and more accurately with less money. This helps a lot of families save a lot of money and get the same services as a hospital, which is not a good thing and that's why I did this research.

Keywords: Artificial intelligence, heart disease, convenience.

1. Introduction

Heart disease is one type of circulatory disease that regularly affects humans. The circulatory system is made constituted by the heart, blood arteries, and neurohumoral tissues that regulate blood flow. All of the aforementioned tissues and organs are modified by circulatory disorders, often known as cardiovascular diseases, which are frequent ailments in internal medicine, the most common of which is heart disease, which has a significant impact on a patient's ability to work and is now inextricably linked to the entire world. Heart disease is still the leading cause of mortality and disability worldwide from 2000 to 2019, according to the World Health Organization, which indicates that hospitals are still not addressing the condition. no thorough treatment, and it is quite challenging, using a machine to forecast heart disease would be tremendously beneficial for academic research. I recently looked up many ways to forecast heart disease on the Internet, such as the detection of C-reactive protein and carotid artery ultrasonography. The aforementioned techniques may be a little time-consuming for us and not very popular, but with the advancement of AI, we can now utilize a variety of memory models to assist more people in determining whether they have heart disease and lower the likelihood of misdiagnosis, such as logistic regression, random forest, etc.

FIRST PART:

First paper-LRLSLDA model, LRLSLDA idea of calculating similarities separately in different spaces leads to the parameters of the model More, a large number of parameters can only rely on empirical knowledge preset, which limits the model Predict further improvements in performance. In addition, matrix regularization operates the complexity of the calculation time is high, and it is difficult to adapt to the current rapid growth of people Genomics-like data.

The major components of the second paper-LightGBM are a decision tree method based on histograms, a depth-limited Leaf-Wise leaf growth strategy, unilateral gradient sampling, and straight Category feature support. The LightGBM model no longer divides the data when processing a large amount of data because the cardiovascular disease dataset has a relatively large amount of data. As a result, the amount of computation is reduced, and the algorithm can quickly and effectively obtain prediction results for cardiovascular disease prediction.

Third paper-In this paper, a decision tree model is constructed using RapidMiner data mining tools, and finally compared with the actual results for detection

The predictive power of decision trees. The prediction accuracy obtained after the decision tree model is run is72.26%, from the results, it is necessary to further optimize the model and improve the accuracy.

Fourth paper- In the field of computer-aided diagnosis, this paper mainly focuses on lung CT imaging for the diagnosis of benign and malignant lung cancer, using a single layer CNN and SDAE and DBN solved the problem of benign and malignant classification of lung nodes, with a classification accuracy of 89.9 %

Therefore, from these four papers, it can be seen that machine research can sometimes be a bit troublesome to operate, but the final result is relatively close.

SECOND PART:

First paper-uses a random forest algorithm to predict the outcome of emergency patients, and the model can obtain results accurately and quickly.

Second paper-uses logistic regression, gradient enhancement machines, and artificial neural networks to train a model to obtain 65 clinically meaningful ECG features to predict ACS.

Third paper-used the random forest method to take a patient with chest pain from 8 Clinical signs and 5 min HRV data were screened out with primary not

Systolic blood pressure, mean RR interval, and mean instantaneous heart rate were the three factors that were most important to benign cardiac events,

Fourth paper-take advantage of K-nearest neighbors and multilayer perceptron’s

Neural networks classify healthy people and people at risk of SCD.

The study found that the type and time point of the extracted features differed in their pairs the accuracy of SCD prediction should also differ, while the time-frequency characteristics and non-linear will be the combination of sexual characteristics can achieve higher precision.

So, in my opinion, machine learning is used in all four papers, and it is not difficult to see that the advantages of using machines to predict diseases are accurate, fast and save a lot of manpower Method.

2. Methodology

logistic regression:

Despite having the word "regression" in its name, logistic regression is actually a classification model that is frequently applied in many different domains. These classic approaches are still frequently utilized in many disciplines because to their distinct advantages, despite the fact that deep learning is now more popular than them.

$$\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{P}(\boldsymbol{X} \leq \boldsymbol{x}) = \frac{\mathbf{1}}{\mathbf{1}+e^{-(x-\mu)/\gamma}} \tag{1}$$

X is a continuous random variable, μ is a positional parameter and $\gamma > 0$is the shape parameter.

SVM:

SVM can not only realize classification problems, that is, the output is the type of label, such as handwritten digit recognition, Iris classification, but also realize the prediction of continuous values, that is, the output is a continuous value, that is, regression problems, such as Boston house price forecast.

$$w^T + b = 0 \tag{2}$$

Where w is the normal vector and b is the displacement term. w and b are actually the parameters of the model.

3. Experiment

Age	Sex	Chest pain	BP	Cholesterc	FBS over 1	EKG result	Max HR	Exercise ar	ST depress	Slope of S	Number o	Thallium	Heart Disease
70	1	4	130	322	0	2	109	0	2.4	2	3	3	Presence
67	0	3	115	564	0	2	160	0	1.6	2	0	7	Absence
57	1	2	124	261	0	0	141	0	0.3	1	0	7	Presence
64	1	4	128	263	0	0	105	1	0.2	2	1	7	Absence
74	0	2	120	269	0	2	121	1	0.2	1	1	3	Absence
65	1	4	120	177	0	0	140	0	0.4	1	0	7	Absence
56	1	3	130	256	1	2	142	1	0.6	2	1	6	Presence
59	1	4	110	239	0	2	142	1	1.2	2	1	7	Presence
60	1	4	140	293	0	2	170	0	1.2	2	2	7	Presence
63	0	4	150	407	0	2	154	0	4	2	3	7	Presence
59	1	4	135	234	0	0	161	0	0.5	2	0	7	Absence
53	1	4	142	226	0	2	111	1	0	1	0	7	Absence
44	1	3	140	235	0	2	180	0	0	1	0	3	Absence
61	1	1	134	234	0	0	145	0	2.6	2	2	3	Presence
57	0	4	128	303	0	2	159	0	0	1	1	3	Absence
71	0	4	112	149	0	0	125	0	1.6	2	0	3	Absence
46	1	4	140	311	0	0	120	1	1.8	2	2	7	Presence
53	1	4	140	203	1	2	155	1	3.1	3	0	7	Presence
64	1	1	110	211	0	2	144	1	1.8	2	0	3	Absence
40	1	1	140	199	0	0	178	1	1.4	1	0	7	Absence
67	1	4	120	229	0	2	129	1	2.6	2	2	7	Presence
48	1	2	130	245	0	2	180	0	0.2	2	0	3	Absence
43	1	4	115	303	0	0	181	0	1.2	2	0	3	Absence
47	1	4	112	204	0	0	143	0	0.1	1	0	3	Absence
54	0	2	132	288	1	2	159	1	0	1	1	3	Absence
48	0	3	130	275	0	0	139	0	0.2	1	0	3	Absence
46	0	4	138	243	0	2	152	1	0	2	0	3	Absence
51	0	3	120	295	0	2	157	0	0.6	1	0	3	Absence
58	1	3	112	230	0	2	165	0	2.5	2	1	7	Presence

Figure 1. The dataset (Pi*cture credit:* Heart Disease Prediction | Kaggle**)**

These three are very normal so I don't explain the definition of them - Age, Sex, Heart Disease, Number of vessels Fluro.

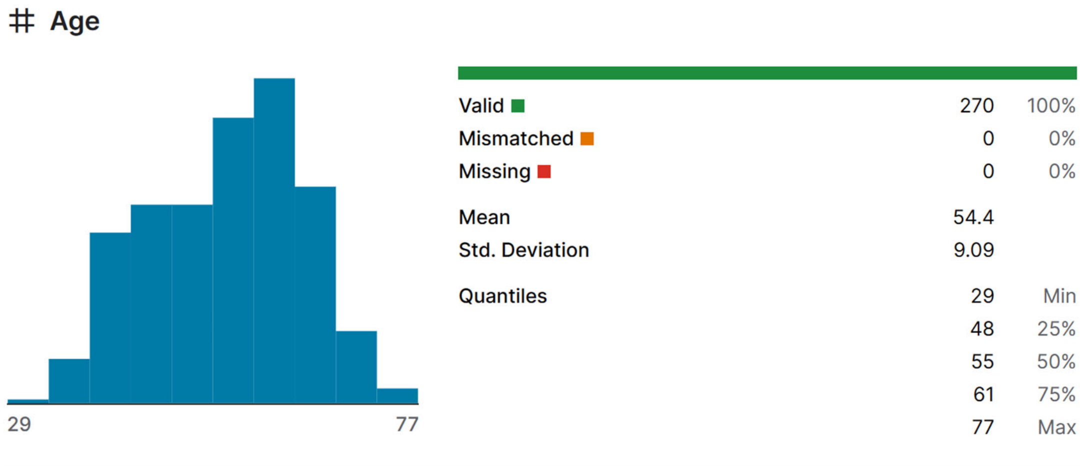

Figure 2. Age (Pi*cture credit:* Heart Disease Prediction | Kaggle**)**

After the test, this data is 100% usable. After the test, there is no mismatch in this data, and nothing is missing. The average value between 29 and 77 is 54.4, and the error of the data is 9.09.

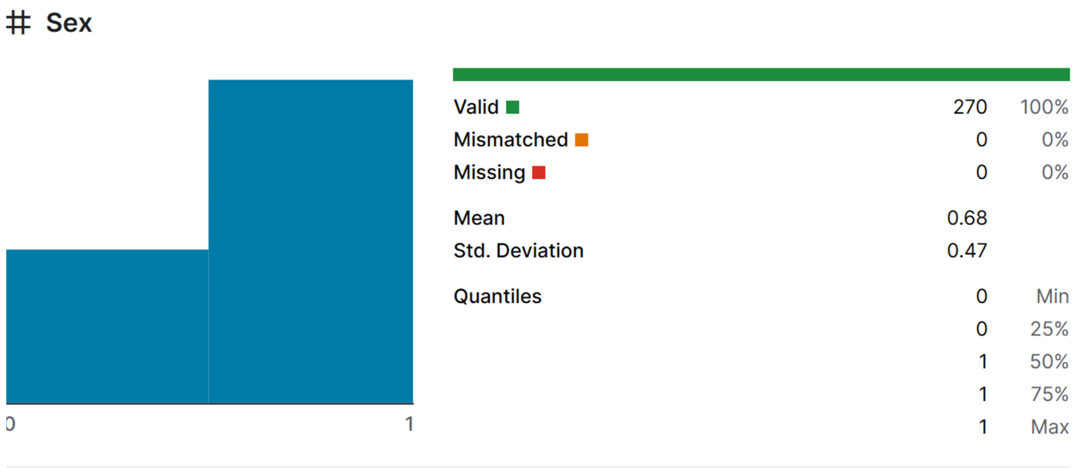

Figure 3. Sex (Pi*cture credit:* Heart Disease Prediction | Kaggle**)**

This data is 100% usable after testing, there is no mismatch in this data after testing, nothing is lost either, the average between 0 and 1 is 0.68, the error of the data is 0.47.

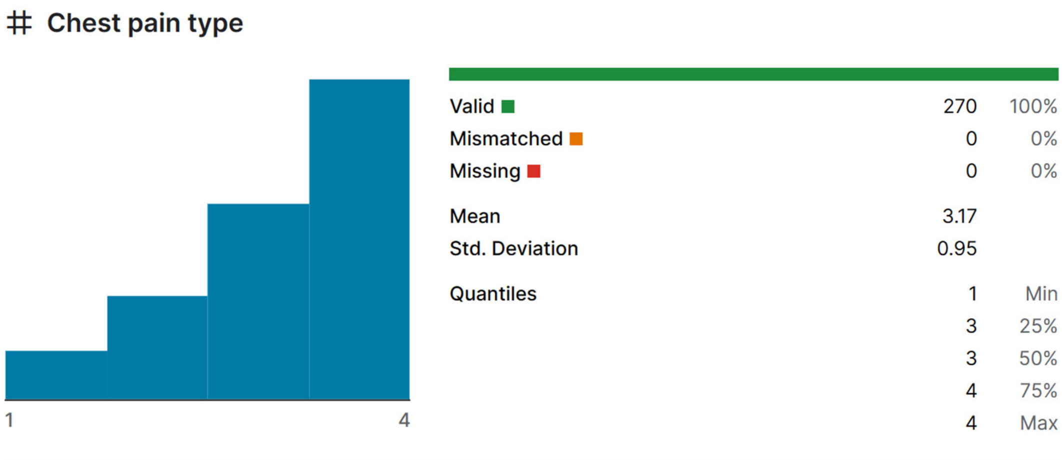

Figure 4. Chest pain type (Pi*cture credit:* Heart Disease Prediction | Kaggle**)**

This data is 100% usable after testing, there is no mismatch in this data after testing, nothing is lost either, the average between 1 and 4 is 3.14, the error of the data is 0.95.The Fig 4 shows that chest pain type is

Chest discomfort might feel like anything from a subtle aching to a violent stabbing. Chest pain can occasionally feel crushing or scorching. Sometimes the pain starts in the jaw and moves up the neck before spreading to the back or down one or both limbs.

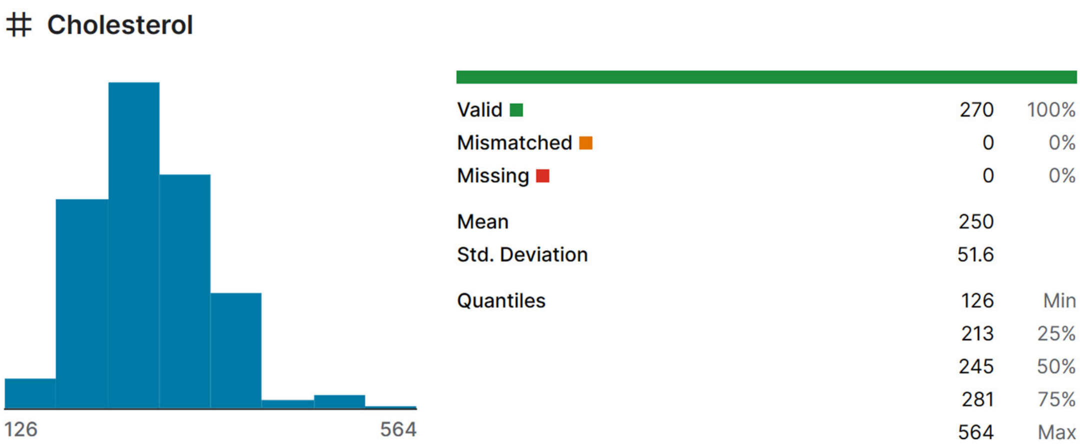

Figure 5. Cholesterol (Pi*cture credit:* Heart Disease Prediction | Kaggle**)**

This data is 100% usable after testing, there is no mismatch in this data after testing, nothing is lost either, the average between 126 and 564 is 250, the error of the data is 51.6. The Fig 5 shows that cholesterol is it exists a waxy, fat-like substance called cholesterol present in every cell of your body. Your body utilizes cholesterol to make vitamins, testosterone, and lipids valuable assistance accelerate food digestion. Your body synthesizes all the cholesterol its requirements are. Additionally, there are foods made from animals like cheese, meat, and egg yolks.

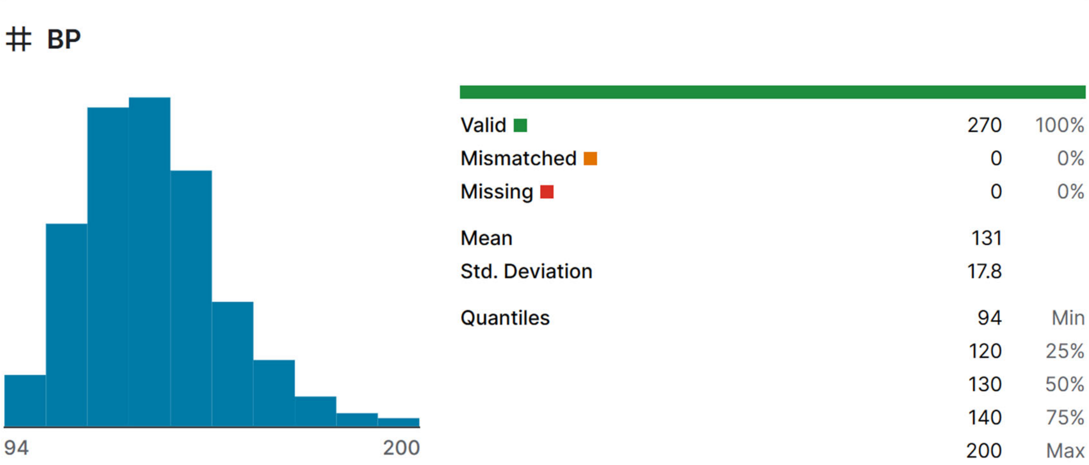

Figure 6. BP (Pi*cture credit:* Heart Disease Prediction | Kaggle**)**

This data is 100% usable after testing, there is no mismatch in this data after testing, nothing is lost either, the average between 94 and 200 is 131, the error of the data is 17.8. The Fig 6 shows that BP is blood pressure is the driving force that causes blood to flow inside blood vessels. It is the lateral pressure that acts on the wall of blood vessels per unit area when blood flows in blood vessels. It is referred to as arterial blood pressure in the systemic circulation and is also known as capillary pressure, venous blood pressure, and arterial blood pressure in various blood vessels.

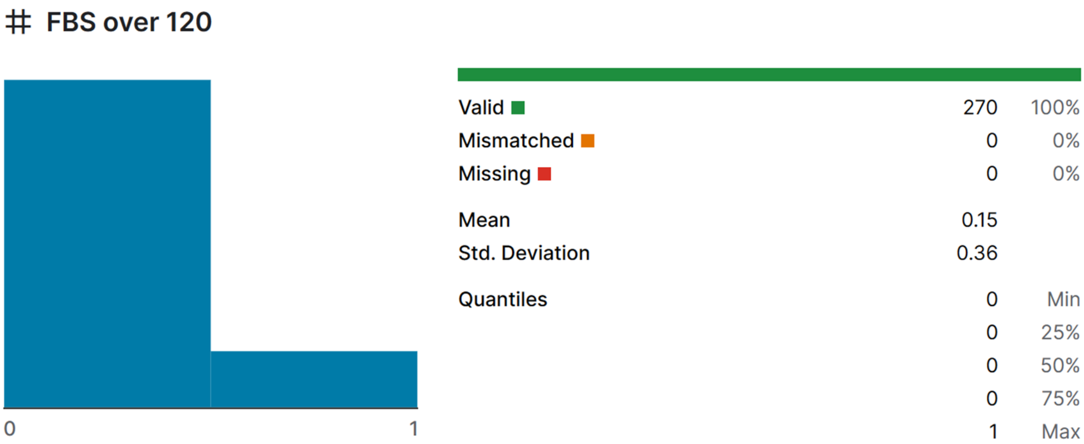

Figure 7. FBS over 120(Pi*cture credit:* Heart Disease Prediction | Kaggle**)**

This data is 100% usable after testing, there is no mismatch in this data after testing, nothing is lost either, the average between 0 and 1 is 0.15, the error of the data is 0.36. The Fig 7 shows that FBS over 120 is fasting blood glucose levels should be between 70 and 100 mg/dL. According to American Diabetes Association guidelines, prediabetes may be indicated by a fasting blood sugar level between 100 and 125 mg/dL (5.6 and 6.9 mmol/L) that is higher than usual. This demonstrates an elevated risk of Type 2 diabetes.

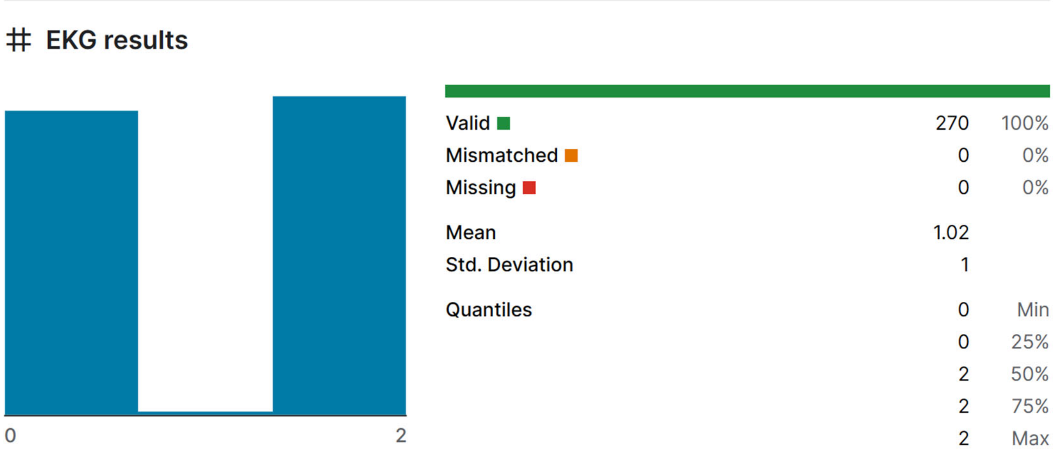

Figure 8. EKG results (Pi*cture credit:* Heart Disease Prediction | Kaggle**)**

This data is 100% usable after testing, there is no mismatch in this data after testing, nothing is lost either

The average between 0 and 2 is 1.02, the error of the data is 1. The Fig 8 shows that EKG results is A sinus rhythm (ECG) monitors the heart's electrical activity. It is a frequently performed, painless test that is used to immediately identify cardiac issues and monitor the condition of the heart.

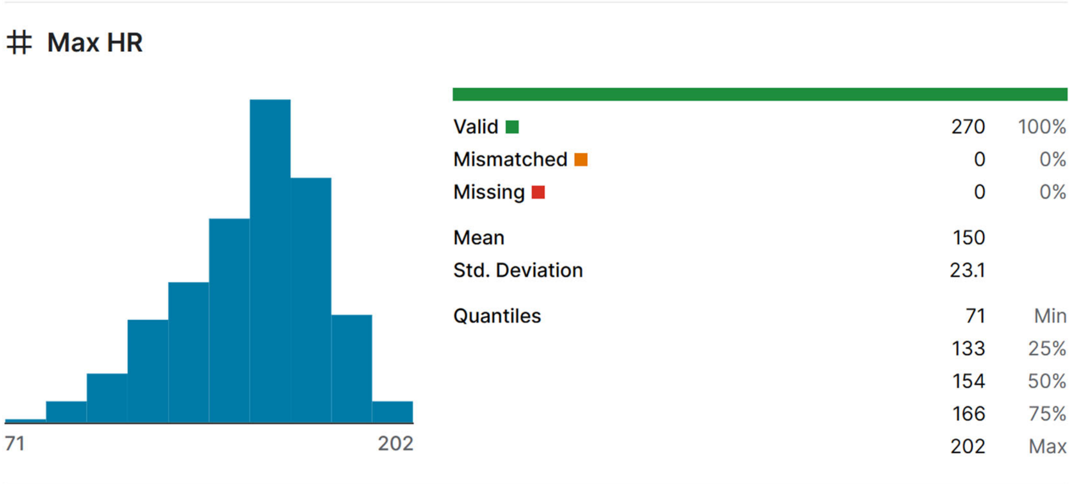

Figure 9. Max HR (Pi*cture credit:* Heart Disease Prediction | Kaggle**)**

This data is 100% usable after testing, there is no mismatch in this data after testing, nothing is lost either, the average between 71 and 202 is 150, the error of the data is 23.1. The Fig 9 shows that Max HR is Your age less 220 is your maximum heart rate. 3 Find your goal heart rates by reading across in the age group that is closest to you. Target heart rate ranges between 50 and 70 percent of maximal heart rate for moderate-intensity exercises and between 70 and 85 percent for strenuous physical activity.

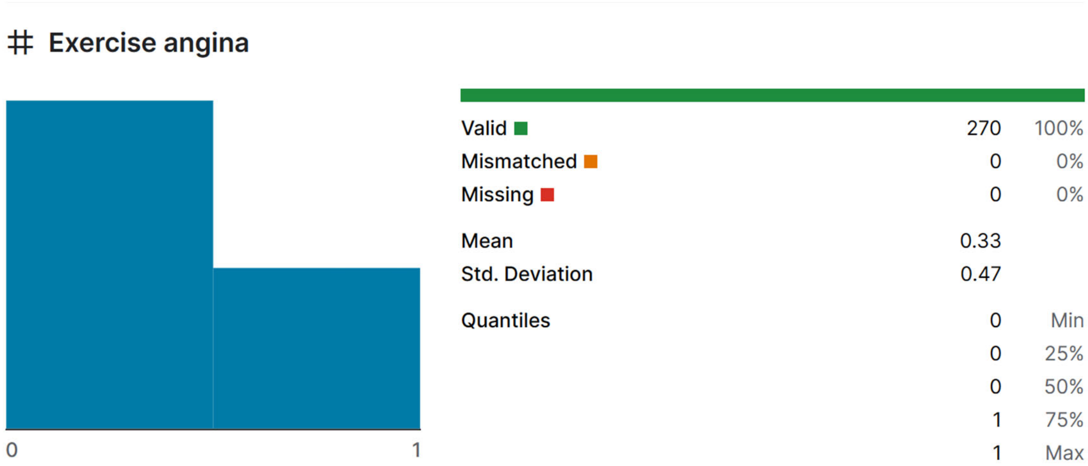

Figure 10. Exercise angina (Pi*cture credit:* Heart Disease Prediction | Kaggle**)**

This data is 100% usable after testing, there is no mismatch in this data after testing, nothing is lost either, the average between 0 and 1 is 0.33, the error of the data is 0.47. The Fig 10 shows that Exercise angina is Chronic heart condition known as sports angina is linked to ill health and higher mortality.

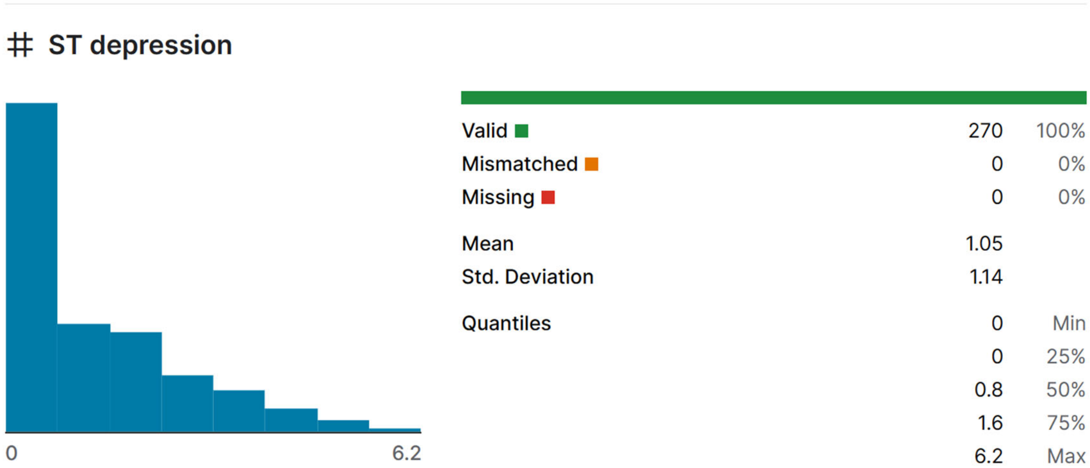

Figure 11. ST depression (Pi*cture credit:* Heart Disease Prediction | Kaggle**)**

This data is 100% usable after testing, there is no mismatch in this data after testing, nothing is lost either, the average between 0 and 6.2 is 1.05, the error of the data is 1.14. The Fig 11 shows that ST depression is A down sloping or horizontal depressive symptoms of 1 mm or more that is measured 80 mms after the inflection point between the QRS complex and the left ventricle segment (J point) for at least two adjacent leads for at least three consecutive beats (with baseline stability) is the most reliable ECG indicator of myocardial ischemia.

Thallium:

Thallium (Tl), a chemical element and metal from Periodic Table's Main Group 13 (IIIa, or the boron group), is toxic and has little industrial use. Thorium is a soft, slow-melting element with low tensile strength, similar to lead. When thallium is first cut, it has a metallic luster that fades to a bluish gray color when exposed to air. Long-term exposure to air causes the metal to continue to oxidize, producing a thick non-protective oxide crust. In diluted sulfuric acid, nitric acid, and hydrochloric acid, thallium dissolves slowly.

Slope of ST:

ST elevation refers to the elevation of the ST segment on the ECG compared to the normal level, which is common in acute myocardial ischemia, acute myocardial infarction, variant angina, acute pericarditis and other diseases.

4. Conclusion

With two findings of 0.7578947368421053 and 0.7684210526315789, I was able to forecast heart disease using two distinct approaches (logistic regression and Svm). It is clear that there is still a small difference between the two approaches.

The primary goal of my research is to avoid heart disease, which is the leading cause of death worldwide and the number one killer of human life. Every year, heart disease claims the lives of hundreds of thousands of people in our nation. It will be extremely important to prevent heart disease if you can extract human body-related physical indicators and study the effects of many characteristics on heart disease through data mining.

The advantage of the whole project is that we are able to predict whether you are at risk of heart disease from different angles through different methods in Python.

The downside of the whole project is that the way to predict by Python is not 100% correct, and this method is not so common.

References

[1] Ren Shanshan: Research on disease prediction based on CNN disease data processing technology, Middle School Affiliated to Shaanxi Normal University, Xi'an 710000.

[2] Shi Shengyuan, Zhu Lei, Ye Lin and Luo Tieqing: Research on cardiovascular disease prediction based on random forest algorithm, School of Information Science and Engineering, Hunan University of Traditional Chinese Medicine, Changsha 410208.

[3] Lan Xin1 Wei Rong1 Cai Hongwei1 Guo Youmin2 Hou Mengwei1 Department of Imaging, The First Affiliated Hospital of Jiaotong University, Xi'an, 710061.

[4] Liao Hualong 1 Zeng Xiaoqian 2 Li Huafeng 3 Yu Yang 4 Zhao Can 1 Chen Yu 1: Application of machine learning in disease prediction, 1. Provincial Key Laboratory of Biomechanical Engineering, Sichuan University, Chengdu 6100652. West China Big Data Center, West China Hospital of Sichuan University, Chengdu 6100413. Department of Anesthesiology, West China Second Hospital of Sichuan University, Chengdu 6100414. Department of Nephrology, West China Hospital of Sichuan University, Chengdu 610041.

[5] Liu Jimin 1 Zhang Kai 2 Wen Long 1 Jia Quanqiu 3 Xie Chuangsen 1 Faye Wong 2: Improved algorithm model for cardiovascular disease prediction, 1. School of Intelligent Equipment, Shandong University of Science and Technology, Tai'an, Shandong 2710002. School of Computer Science and Engineering, Shandong University of Science and Technology, Shandong Qingdao 2660003. Big Data College, Taishan University of Science and Technology, Tai'an, Shandong 271000.

[6] Wang Zengwu and Wu Yangfeng: Methods for predicting cardiovascular disease, 100037, Institute of Cardiovascular Diseases, Fuwai Cardiovascular Hospital, Chinese Academy of Medical Sciences, Peking Union Medical College, 100037, Chinese Academy of Medical Sciences, Peking Union Medical College, China Institute of Vascular Diseases, Fuwai Cardiovascular Hospital, Epidemiology Research Department.

[7] Zhu Xiaotong, Pang Chunying, Zhu Han: Cardiovascular disease prediction model based on deep learning, College of Life Science and Technology, Changchun University of Science and Technology, Changchun 130022.

[8] Liu Yan, Fu Jianfeng and Hu Jiakai: Research on cardiovascular disease prediction based on classification model, School of Information Management, Shanghai Lixin University of Accounting and Finance, Shanghai 201209.

[9] Wang Zengwu and Wu Yangfeng: Methods for predicting cardiovascular disease, 100037, Institute of Cardiovascular Diseases, Fuwai Cardiovascular Hospital, Chinese Academy of Medical Sciences, Peking Union Medical College, 100037, Chinese Academy of Medical Sciences, Peking Union Medical College, China Institute of Vascular Diseases, Fuwai Cardiovascular Hospital, Epidemiology Research Department.

[10] Hu Manman 1 Chen Xu 1 Sun Yuzhong 2 Shen Xi 2 Wang Xiaoqing 3 Yu Tianyang 4 Mei Yudong 1 Xiao Li 2 Cheng Wei 5 Yang Jie 6 Yang Yan 7: Disease prediction model based on dynamic sampling and

transfer learning, 1. Computing Technology Research, Chinese Academy of Sciences Institute Beijing 1000802. University of Chinese Academy of Sciences Beijing 1000493. Institute of Computing Technology, Chinese Academy of Sciences Beijing 1000804. Beijing Chaoyang Hospital Affiliated to Capital Medical University Beijing 1000205. Nanchang University Nanchang 3300006. China Academy of Chinese Medical Sciences Xiyuan Hospital Beijing 1000917. China Academy of Chinese Medical Sciences Traditional Chinese Medicine Data Center Beijing 1007008. Information Department, No. 983 Hospital of the Joint Logistics Support Force of the Chinese People's Liberation Army, Tianjin 300142.