Comprehensive Study on Machine Learning Applications for Heart Disease Risk Prediction

Lashyn Adiat[1], Aizhan Altaibek[1,2], Marat Nurtas[1,2] and Aigerim Altayeva[3]

[1] International Information Technology University, Manas St. 34/1, Almaty, 050000, Kazakhstan
[2] Institute of Ionosphere, Gardening community IONOSPHERE 117, Almaty, 050020, Kazakhstan
[3]Al-Farabi Kazakh National University, al-Farabi Avenue 71, Almaty, 050040, Kazakhstan

Abstract

This research investigates an analytical approach to machine learning applications for cardiovascular disease (CVD) risk prediction on a publicly accessible database. The dataset contains crucial information about patients' individual characteristics, such as age, blood pressure, ECG at rest, heart rate, and four types of chest pain. The purpose of this research is to choose the most suitable model for heart attack analysis. Descriptive analytics and exploratory data analysis based on various factors were done to predict risk by employing machine learning algorithms and techniques, including k-nearest neighbours, logistic regression, support vector machines (SVM) and random forests. The research involves thorough data analysis and rigorous model training processes.

Keywords

Heart disease prediction, comprehensive study, machine learning, cardiovascular risks, exploratory data analysis

1. Introduction

The World Health Organization (WHO) classifies and reports on various causes of death worldwide through its International Classification of Diseases (ICD) system [1].

According to WHO, Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. More than four out of five CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age [2].

Cardiovascular diseases (diseases of the heart or blood vessels) have become a significant public health concern in economically advanced countries. This is primarily due to the difficulties in making an early diagnosis and patients' unwillingness to seek medical assistance when the first symptoms occur. A fast-paced lifestyle, an unhealthy diet, a lack of physical activity, alcohol and tobacco addictions, and insufficient sleep all contribute to the harmful influence on the cardiovascular system [3].

According to epidemiological data, in 2018, cardiovascular disease was the leading cause of death in China. The number of patients with cardiovascular disease is 330 million in China, including 11 million stroke and more than 270 million diseases related to the heart [4].

In the United States, heart disease causes more than 600,000 deaths annually, accounting for approximately one in every four deaths [5].

Predicting and diagnosing heart disease is the biggest challenge in the medical industry and it is based on factors like physical examination, symptoms, and signs of the patient [6]. Body cholesterol levels, smoking habits, obesity, family history of diseases, blood pressure, and working environment are all factors that influence heart disease [10].

DTESI 2023: Proceedings of the 8th International Conference on Digital Technologies in Education, Science and Industry, December 06–07, 2023, Almaty, Kazakhstan

adiatlasyn@gmail.com (L. Adiat); aizhan.altaibek@yandex.ru (A. Altaibek); maratnurtas@gmail.com (M. Nurtas); aikosha1703@gmail.com (A. Altayeva)

0009-0006-2969-1695 (L. Adiat); 0000-0001-8431-7950 (A. Altaibek); 0000-0003-4351-0185 (M. Nurtas); 0000-0002-9802-9076 (A. Altayeva)

Early detection and prevention of heart attacks can improve patient outcomes and reduce the strain on healthcare systems.

Traditional risk assessment methods, such as the Framingham Risk Score [7], focus on clinical and demographic data that may not fully capture the complexity of the underlying causes of cardiovascular risk.

Machine learning has transformed disease detection by enabling the creation of predictive models [3] that analyze massive datasets to uncover subtle trends and anomalies, thereby assisting in early diagnosis and intervention. Machine learning techniques have been extensively used in recent years to forecast the likelihood of heart attacks based on these parameters [8].

Although traditional risk assessment models are effective, their scope and predictive power are frequently constrained. In contrast, machine learning offers a data-driven method that can more accurately forecast cardiac disease and find complex correlations between numerous risk factors.

As healthcare organizations strive to acquire patient records, it is estimated that one trillion bytes of data are generated every day. This information is of the utmost importance and must be properly extracted to yield valuable insights [13]. Patients may not always accurately describe their medical conditions, and laboratory test results can be subject to errors. Healthcare specialists may struggle to make informed decisions about a patient's illness because of their limited expertise in all areas [12]. To address this challenge, the development of a disease prediction system that integrates medical knowledge with a comprehensive system is necessary to produce the most effective results and benefit society [14]. Previous investigations have attempted to use patient laboratory tests [15-17] and medication [18] to predict disease onset. Some prototypes have also been used to identify unknown risk factors while simultaneously improving the sensitivity and specificity of detection. Recent studies have demonstrated success in predicting diseases through several methods, including support vector machines [19-21], logistic regression [22], random forests [23], neural networks [17], and time series modelling techniques [24].

Machine learning models can be flexible and tailored to fit the range of data sources that are becoming increasingly accessible in healthcare. The advancement in technology has greatly enhanced the capacity to forecast a wide range of diseases, such as cancer, cardiovascular conditions, and infectious outbreaks. This development has a two-fold impact: it empowers healthcare providers to identify high-risk individuals for early intervention, which could potentially save lives, and it supports public health agencies in proactive surveillance and resource allocation, helping to curb the spread of diseases on a larger scale. As machine learning continues to advance, its contribution to disease prediction is expected to improve healthcare outcomes and minimize the avoidable economic and human costs associated with preventable illnesses [9].

The goal of this research is to build and test a machine-learning model for forecasting the risk of heart disease. This will be accomplished using a dataset containing patient information and clinical measurements. A comparative analysis is performed to evaluate and contrast the efficacy of various machine learning algorithms that have been utilized in this context, including logistic regression and more advanced models and feature selection strategies in the domain of heart attack prediction. The selection of the most appropriate algorithm is of the utmost importance, as it has a direct impact on the model's ability to process the data, recognize complex relationships, and produce trustworthy predictions.

Considering the study's findings, it is critical to provide relevant insights and evidence-based recommendations to healthcare providers and policymakers. The scope of the investigation is limited to the assessment of a single dataset containing patient information and clinical measurements. The model developed in this study was built primarily for research purposes and should not be used in place of clinical diagnosis or therapy.

2. Methodology

The exploratory data analysis was performed using publicly accessible data on heart disease. The dataset comprised 303 records with 14 attributes, including age, blood pressure, blood glucose level, ECG at rest, heart rate, and four types of chest pain.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

303 rows × 14 columns

Figure 1: dataset for predicting cardiovascular disease risk

Table 1
Heart dataset features

age	Age of the patient
sex	Sex of the patient
cp	Four chest pain types: 0 = Typical Angina 1 = Atypical Angina 2 = Non-anginal Pain 3 = Asymptomatic
trtbps	Resting blood pressure (in mm Hg)
chol	Cholestoral in mg/dl fetched via BMI sensor
fbs	(fasting blood sugar > 120 mg/dl) ~ 1 = True, 0 = False
restecg	Resting electrocardiographic results ~ 0 = Normal, 1 = ST-T wave normality, 2 = Left ventricular hypertrophy
thalachh	Maximum heart rate achieved
exng	Exercise induced angina ~ 1 = Yes, 0 = No
oldpeak	Previous peak
slp	Slope
caa	Number of major vessels
thall	Thalium Stress Test result ~ (0,3)
output	Target variable

The nearly balanced data on the proportion of people experiencing heart attacks (54%) suggests that there is no need to further balance them.

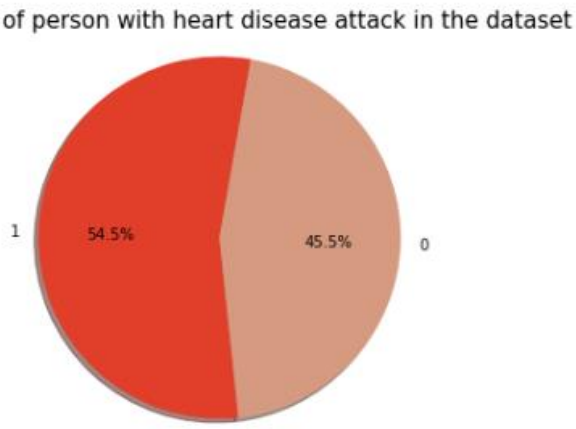


Figure 2: Percentage of people with heart disease attacks in the dataset

The majority of individuals fall within the age range of 50-60 years old, have relatively low chest pain, have blood pressure within the range of 120-140, have cholesterol levels between 200-300, have blood sugar levels below 120, and are male. The majority of these individuals had a heart rate within the range of 150-175.

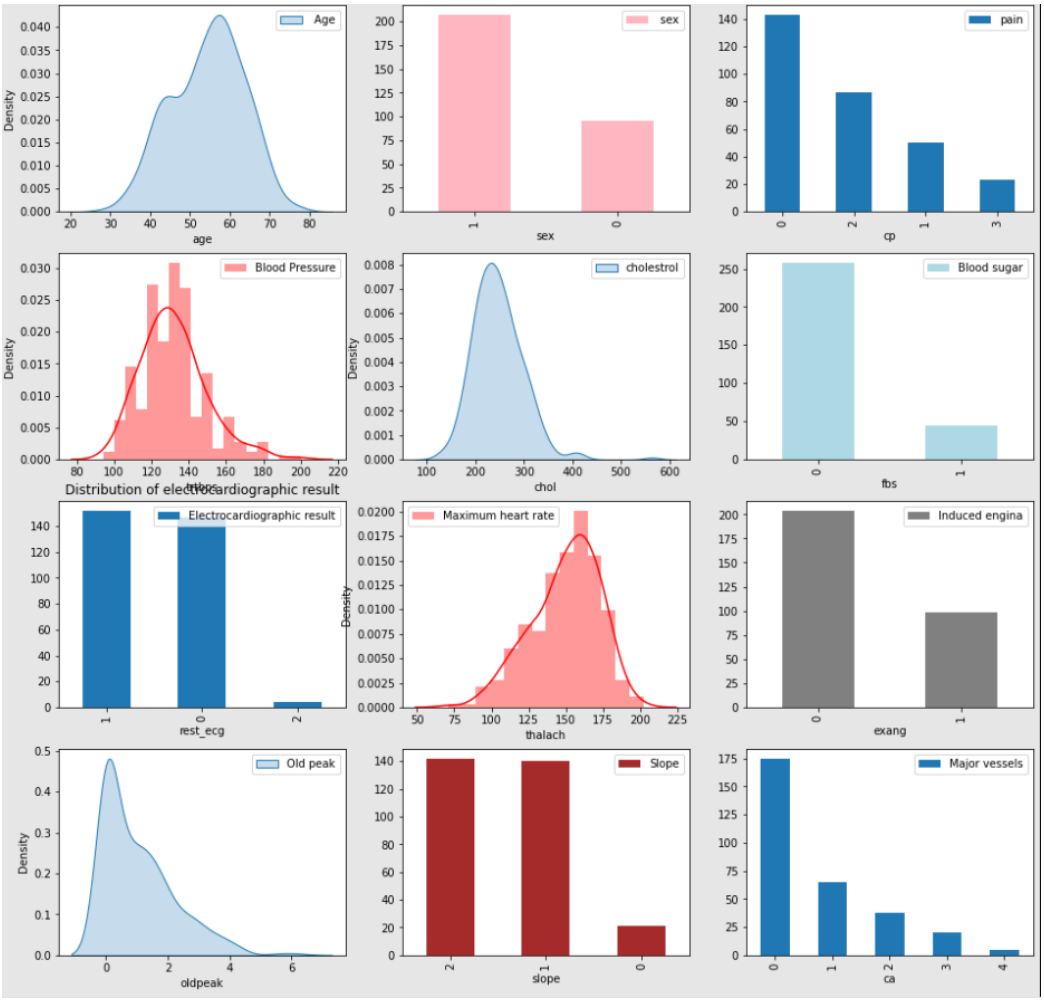


Figure 3: Distribution of features

Individuals aged 40-60 are more likely to have heart disease, whereas those with a higher resting heart rate are at a higher risk of experiencing a heart attack.

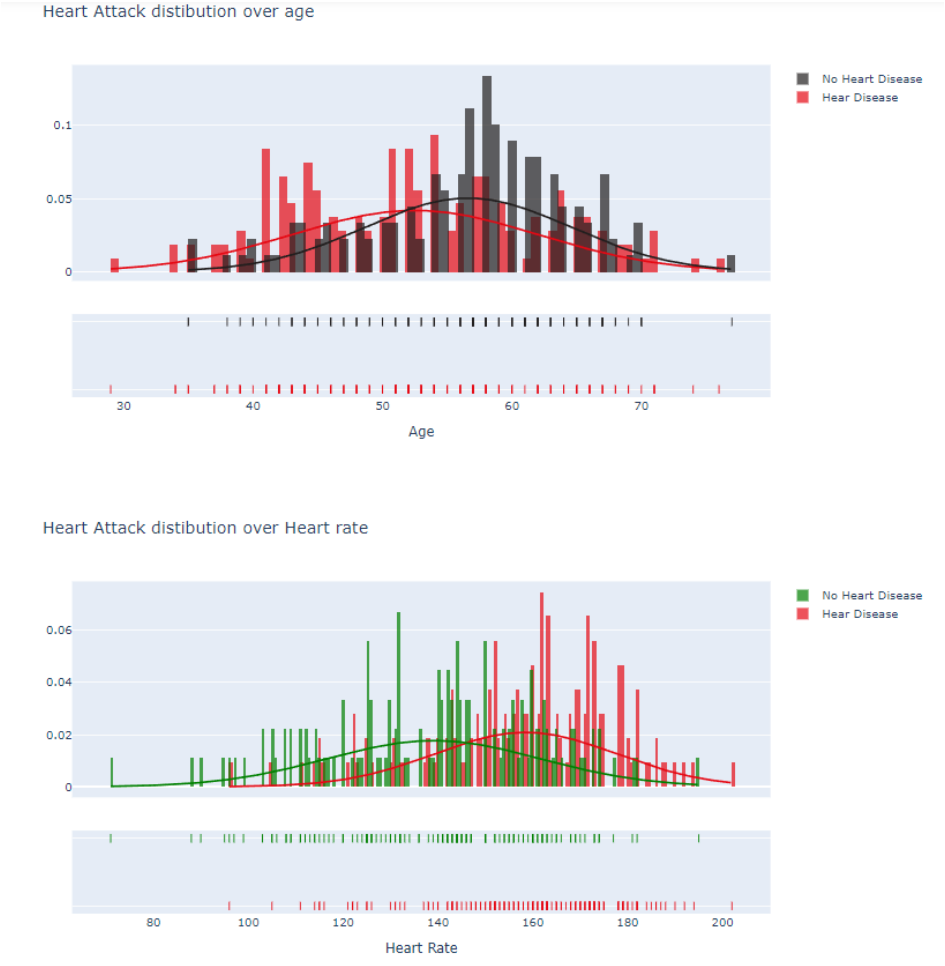


Figure 4: Heart Attack distribution over age and heart rate

People having cholesterol of 120-250 and blood between 110 to 140 are more likely to have a heart attack.

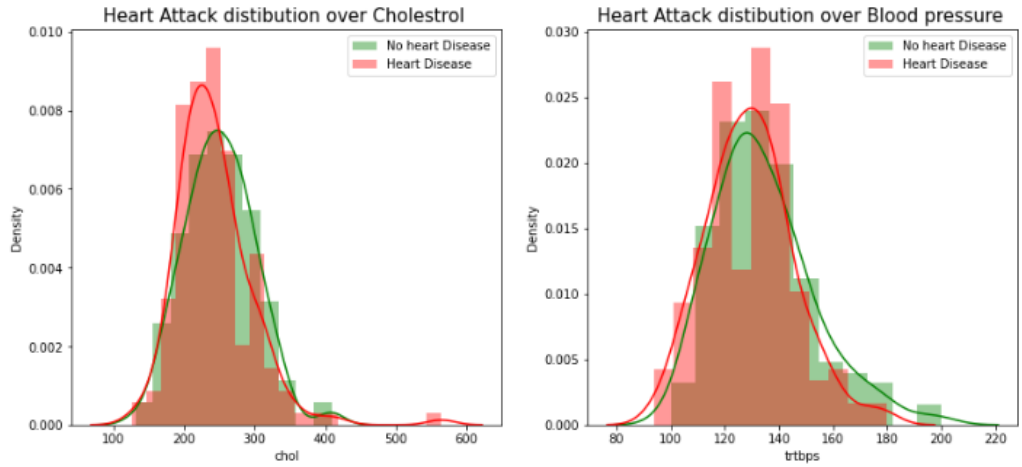


Figure 5: Heart Attack distribution over cholesterol and blood pressure

A large percentage of men are more likely to experience heart attacks than women, with 73% of men and 45% of women suffering from heart attacks.

If someone experiences chest pain, it is highly probable that they will suffer from a heart attack.

The impact of blood sugar level on the likelihood of a heart attack is relatively small. In other words, whether a person has high blood sugar levels does not necessarily determine whether they will have a heart attack.

People who do not regularly exercise their cardiovascular system are highly likely to suffer from heart attack.

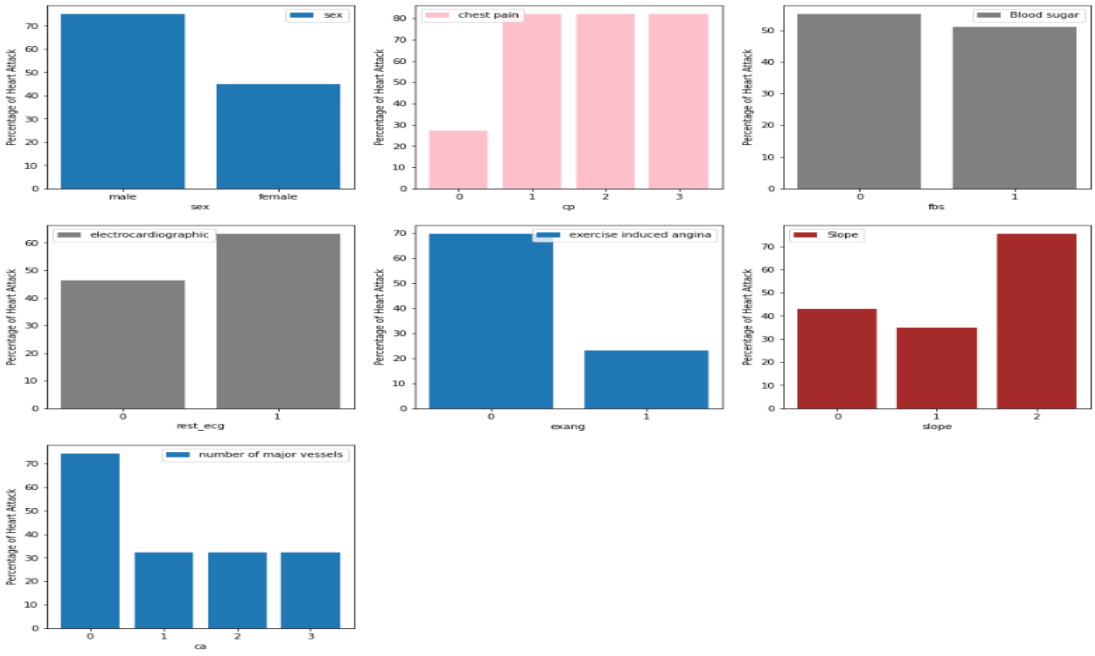


Figure 6: Distribution of features with reference to Heart Decease

The higher the chest pain and the higher the person's heart rate, the more likely they are to suffer a heart attack.

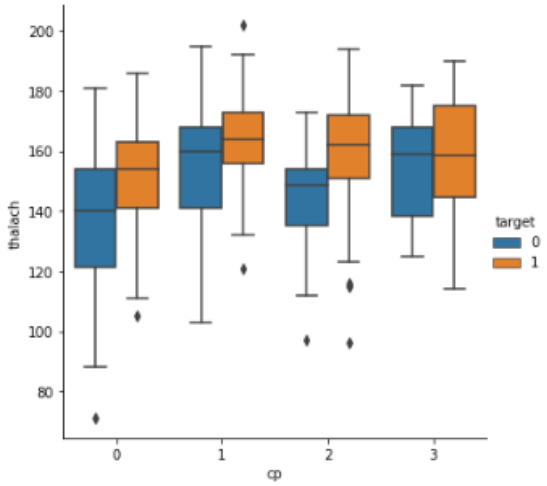


Figure 7: Distribution of attribute cp with reference to other attribute thalachh having hue= Heart Attack

Individuals with a low level of exercise-induced angina are more likely to experience heart disease, even though age does not significantly contribute to the risk of heart attack.

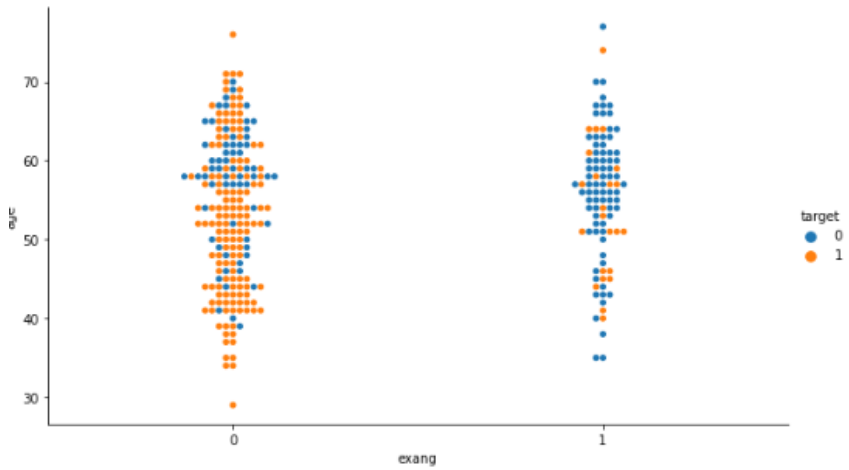
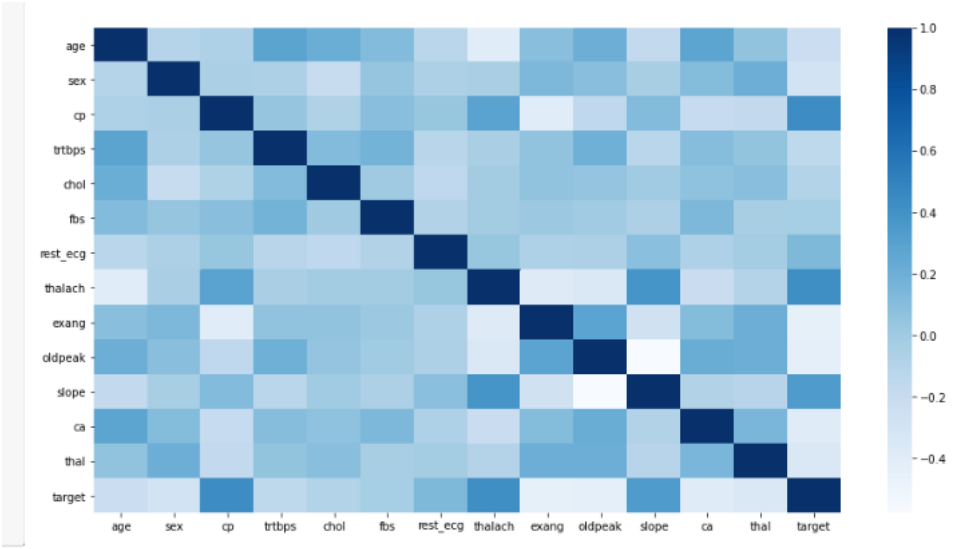


Figure 8: Distribution of attribute exng with reference to other attribute age having hue= Heart Attack

The graph and table below show that there is a positive correlation between heart attack and chest pain, heart rate, and slope. However, there was a negative correlation between heart attacks and age, induced angina, and major vessels.




```
]: df.corr()
```

]:

	age	sex	cp	trtbps	chol	fbs	rest_ecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.000000	-0.098447	-0.068653	0.279351	0.213678	0.121308	-0.116211	-0.398522	0.096801	0.210013	-0.168814	0.276326	0.068001	-0.225439
sex	-0.098447	1.000000	-0.049353	-0.056769	-0.197912	0.045032	-0.058196	-0.044020	0.141664	0.096093	-0.030711	0.118261	0.210041	-0.280937
cp	-0.068653	-0.049353	1.000000	0.047608	-0.076904	0.094444	0.044421	0.295762	-0.394280	-0.149230	0.119717	-0.181053	-0.161736	0.433798
trtbps	0.279351	-0.056769	0.047608	1.000000	0.123174	0.177531	-0.114103	-0.046698	0.067616	0.193216	-0.121475	0.101389	0.062210	-0.144931
chol	0.213678	-0.197912	-0.076904	0.123174	1.000000	0.013294	-0.151040	-0.009940	0.067023	0.053952	-0.004038	0.070511	0.098803	-0.085239
fbs	0.121308	0.045032	0.094444	0.177531	0.013294	1.000000	-0.084189	-0.008567	0.025665	0.005747	-0.059894	0.137979	-0.032019	-0.028046
rest_ecg	-0.116211	-0.058196	0.044421	-0.114103	-0.151040	-0.084189	1.000000	0.044123	-0.070733	-0.058770	0.093045	-0.072042	-0.011981	0.137230
thalach	-0.398522	-0.044020	0.295762	-0.046698	-0.009940	-0.008567	0.044123	1.000000	-0.378812	-0.344187	0.386784	-0.213177	-0.096439	0.421741
exang	0.096801	0.141664	-0.394280	0.067616	0.067023	0.025665	-0.070733	-0.378812	1.000000	0.288223	-0.257748	0.115739	0.206754	-0.436757
oldpeak	0.210013	0.096093	-0.149230	0.193216	0.053952	0.005747	-0.058770	-0.344187	0.288223	1.000000	-0.577537	0.222682	0.210244	-0.430696
slope	-0.168814	-0.030711	0.119717	-0.121475	-0.004038	-0.059894	0.093045	0.386784	-0.257748	-0.577537	1.000000	-0.080155	-0.104764	0.345877
ca	0.276326	0.118261	-0.181053	0.101389	0.070511	0.137979	-0.072042	-0.213177	0.115739	0.222682	-0.080155	1.000000	0.151832	-0.391724
thal	0.068001	0.210041	-0.161736	0.062210	0.098803	-0.032019	-0.011981	-0.096439	0.206754	0.210244	-0.104764	0.151832	1.000000	-0.344029
target	-0.225439	-0.280937	0.433798	-0.144931	-0.085239	-0.028046	0.137230	0.421741	-0.436757	-0.430696	0.345877	-0.391724	-0.344029	1.000000

Figure 9: Confusion matrix of correlation among attributes

3. Results

Among the aforementioned machine learning algorithms, the utilization of logistic regression, K-nearest neighbours (KNN) [11], support vector classifiers, and random forest classifiers are recommended because they demonstrate comparative accuracy with traditional methods. Our preliminary examination suggests that the search for optimal coefficients may be accommodated through iterative coefficient selection, specifically, Logistic Regression. However, the accuracy of the models can only be ascertained after a thorough evaluation.

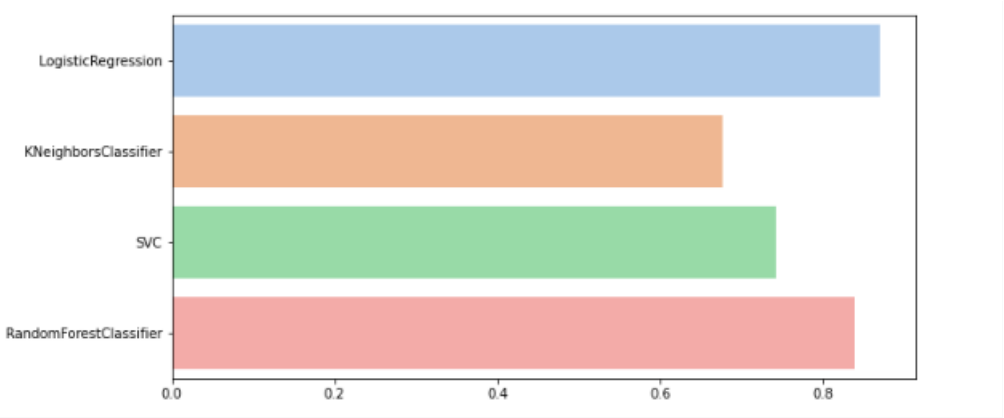


Figure 10: Diagram of accuracies by machine learning methods

Logistic Regression, a statistical method used for binary classification, was first introduced to address the problem of binary classification. It assumes that the data follow a Bernoulli distribution and solves for the optimal parameters through maximum likelihood estimation [4]. The first Logistic Regression model shows values of around 89 per cent accuracy, which is quite good for this model.

Figure 11: Roc Curve for logistic regression

KNN demonstrates an accuracy of around 70%, which unfortunately means that finding the optimal K requires additional calculations up to K=100(th neighbour). SVC is showing progress and is approaching 76% but it is still not the best model at the moment. Finally, Random Forest shows progress around 82%, which is clearly better than the previous ones, but still falls short of the best model.

We can conclude that the Logistic Regression model shows good progress in comparison with other machine learning algorithms, thereby suggesting that the neural network will demonstrate a more optimal result since Logistic Regression is based on gradient descent in finding the necessary parameters.

4. Acknowledgements

We would like to express our sincere gratitude to Rashik Rahman, the data provider on Kaggle for providing valuable data resources that were essential for the successful completion of this study. We extend our appreciation to the Kaggle community for their efforts to make diverse datasets accessible and to promote collaboration among data enthusiasts and researchers.

5. Conclusion

During our study, we conducted an exploratory data analysis and comparative analysis of machine learning algorithms' accuracies. After conducting a thorough analysis and applying several widely used machine learning algorithms for forecasting heart disease, we discovered that logistic regression demonstrated outstanding performance. In fact, it was the algorithm that attained the highest level of accuracy, allowing us to confidently classify patients with heart disease. The application of machine learning algorithms in predicting heart disease is an ongoing research area with significant promise. Integrating advanced machine learning methods in this domain is likely to significantly alleviate the burden on health care and improve the prognosis of diseases, leading to improved overall patient health.

Therefore, in future research, we aim to develop personalized forecasting methods that consider additional risk factors and model adaptations to changing conditions and patient needs. This approach will enhance the accuracy and effectiveness of heart disease forecasts, leading to earlier diagnosis and treatment, and ultimately improving patient health and quality of life.

6. References

[1] Rahman, A. U., Saeed, M., Saeed, M. H. (2023). A framework for susceptibility analysis of brain tumours based on uncertain analytical cum algorithmic modeling. Bioengineering, 10(2), 147. DOI:10.3390/bioengineering10020147.

[2] World Health Organization, "cardiovascular diseases (CVDs)", URL: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.

[3] Marat Nurtas**,** Baishemirov Zharasbek, Zhanabekov Zhandos, (2020), Applying Neural Network for predicting cardiovascular disease risk." News of the National Academy of sciences of the Republic of Kazakhstan, 4(332): 28–34. https://doi.org/10.32014/2020.2518-1726.62.

[4] Siyi Wang. (2023). Research on the heart attack prediction based on logistic regression. Highlights in Science, Engineering and Technology, Volume 65. DOI: 10.1016/j.tele.2018.11.007.

[5] D. for H. D. and S. P., National Center for Chronic Disease Prevention and Health Promotion, “Heart Disease Facts,” 2021. https://www.cdc.gov/heartdisease/facts.htm#:~:text=Coronary heart disease is the, killing 375%2C476 people in 2021.&text=About 1 in 20 adults, have CAD (about 5%25).&text=In 2021%2C about 2 in, less than 65 years old.

[6] V. Manikantan, S. Latha. (2013). Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods", International Journal on Advanced Computer Theory and Engineering, 2(2): 5-10.

[7] Tommy Pocana and Michael Fuchs. (2012). The Cardiovascular Link to Nonalcoholic Fatty Liver Disease: A Critical Analysis. Clinics in Liver Disease, 16(3): 599-613. DOI: 10.1016/j.cld.2012.05.008.

[8] B. Marqas, R., Mousa, A., Özyurt, F. and Salih, R. (2023). A Machine Learning Model for the Prediction of Heart Attack Risk in High-Risk Patients Utilizing Real-world Data. Academic Journal of Nawroz University, *12*(4), 286–301. DOI: 10.25007/ajnu.v12n4a1974Aighuraibawi.

[9] Aighuraibawi, A. H. B., Manickam, S., Abdullah, R. (2023). Feature Selection for Detecting ICMPv6-Based DDoS Attacks Using Binary Flower Pollination Algorithm." Computer Systems Science & Engineering, 47(1). DOI: 10.32604/csse.2023.037948.
[10] X. Su. (2020). Prediction for cardiovascular diseases based on laboratory data: An analysis of random forest model." J. Clin. Lab. Anal., 301. 34(9): 1–10, DOI: 10.1002/jcla.23421.
[11] Walid Sherif. (2018). Optimization of K-NN algorithm by clustering and reliability coefficients: application to breast-cancer diagnosis." Procedia Computer Science 127: 293–299. DOI: 10.1016/j.procs.2018.01.125.
[12] G. Saranya, A. Pravin. (2020). A comprehensive study on disease risk predictions in machine learning." International Journal of Electrical and Computer Engineering (IJECE), 10(4): 4217-4225. DOI: 10.11591/ijece.v10i4.pp4217-4225.
[13] R. Snyderman. (2012). Personalized health care: from theory to practice. Biotechnology Journal, 7(8): 973–979, Aug. 2012. DOI: 10.1002/biot.201100297.
[14] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu. (2011). A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries." J. Am Med Inform Assoc, 18(5): 601–606. DOI: 10.1136/amiajnl-2011-000163.
[15] N. Razavian and D. Sontag. (2015). Temporal convolutional neural networks for diagnosis from lab tests. arXiv:1511.07938v4.
[16] R. Ranganath, J. S. Hirsch, D. Blei, and N. Elhadad. (2015). Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis." Journal of the American Medical Informatics Association: JAMIA, 22(4): 872–880. DOI: 10.1093/jamia/ocv024.
[17] N. Tangri, L. A. Stevens, J. Griffith, H. Tighiouart, O. Djurdjev, D. Naimark, A. Levin, and A. S. Levey. (2011). A predictive model for progression of chronic kidney disease to kidney failure." JAMA. 305(15): 1553–1559. DOI: 10.1001/jama.2011.451.
[18] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun. (2016). Doctor AI: Predicting clinical events via recurrent neural networks." Proceedingsofthe1st Machine Learning for Healthcare Conference, ser. Proceedings of Machine Learning Research, pp. 301–318.
[19] N. Barakat, A. P. Bradley, and M. N. H. Barakat. (2010). Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus. IEEE Transactions on Information Technology in Biomedicine, 14(4): 1114– 1120. DOI: 10.1109/TITB.2009.2039485.
[20] Wu Jionglin M. S., et al. (2010). Prediction Modeling Using EHR Data: Challenges, and a Comparison of Machine Learning Approaches. Journal of the Medical Care section, American Public Health Association, 48(6): S106-S113.
[21] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes." BMC Medical Informatics and Decision Making, 10(16): 1-7. DOI: 10.1186/1472-6947-10-16.
[22] N. Razavian, S. Blecker, A. M. Schmidt, A. Smith-McLallen, S. Nigam, and D. Sontag. (2015). Population-Level Prediction of Type 2 Diabetes from Claims Data and Analysis of Risk Factors. Big Data, 3(4): 277–287. DOI: 10.1089/big.2015.0020.
[23] A. V. Lebedev, E. Westman, G. J. P. Van Westen, M. G. Kramberger, A. Lundervold, D. Aarsland, H. Soininen, I. Kłoszewska, P. Mecocci, M. Tsolaki, B. Vellas, S. Lovestone, and A. Simmons. (2014). Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness." NeuroImage: Clinical, 6: 115–125. DOI: 10.1016/j.nicl.2014.08.023.
[24] A. Perotte, R. Ranganath, J. S. Hirsch, D. Blei, and N. Elhadad. (2015). Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. Journal of the American Medical Informatics Association: JAMIA, 22(4): 872–880. DOI: 10.1093/jamia/ocv024.