ISSN: 2508-7894 © 2023 KODISA
KJAI website: http://acoms.kisti.re.kr/kjai
doi: http://dx.doi.org/10.24225/kjai.2023.11.1.1

A Study on the Comparison of Predictive Models of Cardiovascular Disease Incidence Based on Machine Learning

Ji Woo SEOK[1], Won ro LEE[2], Min Soo KANG[3]

Received: January 20, 2023. Revised: January 27, 2023. Accepted: March 04, 2023.

Abstract

In this paper, a study was conducted to compare the prediction model of cardiovascular disease occurrence. It is the No.1 disease that accounts for 1/3 of the world's causes of death, and it is also the No. 2 cause of death in Korea. Primary prevention is the most important factor in preventing cardiovascular diseases before they occur. Early diagnosis and treatment are also more important, as they play a role in reducing mortality and morbidity. The Results of an experiment using Azure ML, Logistic Regression showed 88.6% accuracy, Decision Tree showed 86.4% accuracy, and Support Vector Machine (SVM) showed 83.7% accuracy. In addition to the accuracy of the ROC curve, AUC is 94.5%, 93%, and 92.4%, indicating that the performance of the machine learning algorithm model is suitable, and among them, the results of applying the logistic regression algorithm model are the most accurate. Through this paper, visualization by comparing the algorithms can serve as an objective assistant for diagnosis and guide the direction of diagnosis made by doctors in the actual medical field.

Keywords : Cardiovascular Disease, Prediction, Machine Learning

Major Classification Code : Artificial Intelligence

1. Introduction

Cardiovascular Diseases (CVD) is a comprehensive disease that refers to disorders of the heart and blood vessels. CVD includes Coronary Artery Disease (CAD) such as Angina and Myocardial Infarction. Other CVDs include Stroke, Heart Failure, Hypertensive Heart Disease, Cardiomyopathy, Arrhythmia, Congenital Heart Defect, Valvular Heart Disease, Carditis, Aortic Aneurysm, etc. According to the announcement of the World Health Organization, CVD is the number one disease accounting for about 32% of all deaths worldwide as of 2019 (WHO, 2019). According to the 2021 cause of death statistics released in September 2022 in Korea, among the 317,680 deaths in 2021, the three major causes of death are malignant neoplasia (cancer), heart disease, and pneumonia, accounting for 43.1%. Among them, heart disease was found to account for 9.9% of all deaths (National Statistical Office, 2022).

*This work was supported by the research grant of the KODISA Scholarship Foundation in 2023.
1 First Author, Dept. of Medical IT, Eulji University, Korea.
Email: wldn9020@daum.net
2. Second Author, CEO, NASCALLAB, Korea,
Email: nascalab@gmail.com
3. Corresponding Author, Professor Dept. of Medical IT, Eulji University, Korea. Email: mskang@eulji.ac.kr

Table 1: Ranking of causes of death, 2011-2021

(Unit : persons per 100,000 population, persons, %)

Ranking	2011 Cause of Death	2011 Death Rate	2020 Cause of Death	2020 Death Rate	2021 Cause of Death	2021 Mortality	2021 Component Ratio	2021 Death Rate	2021 '11 Rank Contrast	2021 '20 Rank Contrast
1	malignant neoplasm (cancer)	142.8	malignant neoplasm (cancer)	160.1	malignant neoplasm (cancer)	82,688	26.0	161.1	-	-
2	cerebrovascular disease	50.7	heart disease	63.0	heart disease	31,569	9.9	61.5	↑+1	-
3	heart disease	49.8	pneumonia	43.3	pneumonia	22,812	7.2	44.4	↑+3	-
4	Intentional self-harm (suicide)	31.7	cerebrovascular disease	42.6	cerebrovascular disease	22,607	7.1	44.0	↓-2	-
5	diabetes	21.5	Intentional self-harm (suicide)	25.7	Intentional self-harm (suicide)	13,352	4.2	26.0	↓-1	-
6	pneumonia	17.2	diabetes	16.5	diabetes	8,961	2.8	17.5	↓-1	-
7	chronic lower respiratory tract disease	13.9	Alzheimer's disease	14.7	Alzheimer's disease	7,993	2.5	15.6	↑+4	-
8	liver disease	13.5	liver disease	13.6	liver disease	7,129	2.2	13.9	-	-
9	transport accidents	12.6	hypertensive disease	11.9	blood poisoning	6,429	2.0	12.5	↑+5	↑+1
10	hypertensive disease	10.1	blood poisoning	11.9	hypertensive disease	6,223	2.0	12.1	-	↓-1

As shown in Table 1, heart disease accounts for 61.5 per 100,000 people, the second-highest cause of death after malignant neoplasms (161.1 per 100,000 people). Furthermore, looked at CVD, a comprehensive concept that includes not only heart disease but also vascular disorders. Circulatory system diseases include cerebrovascular disease (44.0 per 100,000 people) and hypertensive disease (12.1 per 100,000 people), totaling 121.5 per 100,000 people (National Statistical Office, 2022).

Table 2: Mortality rates of circulatory diseases by sex, 2011-2021

(Unit : persons per 100,000 population, %)

			circulatory system disease	hypertensive disease	heart disease	ischemic[1] heart disease	other[2] heart disease	cerebrovascular disease
Male and Female Totality	2011		113.5	10.1	49.8	27.1	22.7	50.7
	2020		121.1	11.9	63.0	27.4	35.6	42.6
	2021		**121.5**	**12.1**	**61.5**	**27.5**	**34.0**	**44.0**
	'20 Rank Contrast	Increase and Decrease	0.4	0.2	-1.5	0.1	-1.6	1.5
		Rate of Change	0.3	2.0	-2.4	0.5	-4.5	3.4
Male	2011		106.5	6.5	48.3	28.3	20.0	48.6
	2020		115.5	7.8	62.3	31.1	31.2	41.5
	2021		**116.2**	**8.3**	**60.3**	**31.0**	**29.3**	**43.4**
	'20 Rank Contrast	Increase and Decrease	0.7	0.5	-2.0	-0.1	-1.9	1.9
		Rate of Change	0.6	6.4	-3.3	-0.4	-6.1	4.6
Female	2011		120.5	13.6	51.3	26.0	25.3	52.8
	2020		126.7	15.9	63.7	23.7	40.0	43.6
	2021		**126.8**	**15.9**	**62.7**	**24.1**	**38.6**	**44.7**
	'20 Rank Contrast	Increase and Decrease	0.1	-0.0	-1.0	0.4	-1.3	1.0
		Rate of Change	0.0	-0.1	-1.5	1.6	-3.3	2.4
Mortality Sex Ratio (Male/Female)	2011		0.9	0.5	0.9	1.1	0.8	0.9
	2020		0.9	0.5	1.0	1.3	0.8	1.0
	2021		**0.9**	**0.5**	**1.0**	**1.3**	**0.8**	**1.0**

1) Ischemic heart disease includes myocardial infarction and angina pectoris.
2) Other heart disease includes heart failure and endocarditis.

In Table 2, heart disease (61.5 people), cerebrovascular disease (44.0 people), and hypertensive disease (12.1 people) are in order, and women (126.8 people) are 1.1 times higher than men (116.2 people). The mortality rate of hypertension and cerebrovascular disease was higher in women than in men, but the mortality rate of ischemic heart disease was higher in men (31.0 persons) than in women (24.1 persons) (National Statistical Office, 2022).

Table 3: Mortality rates of circulatory diseases by age, 2011-2021

(Unit : persons per 100,000 population, persons)

Age (years)	2020	2021	circulatory system disease				
			hypertensive disease	heart disease	ischemic[1)] heart disease	other[2)] heart disease	cerebrovascular disease
Total*	121.1	121.5	12.1	61.5	27.5	34.0	44.0
0	1.8	3.9	-	2.7	-	2.7	0.4
1-9	0.6	0.6	-	0.4	-	0.4	0.2
10-19	0.8	0.9	-	0.5	0.0	0.4	0.4
20-29	2.1	1.9	0.0	1.3	0.3	1.0	0.5
30-39	7.0	7.0	0.1	3.7	1.6	2.1	2.9
40-49	19.8	19.5	0.7	10.0	5.9	4.0	8.1
50-59	46.9	43.6	1.6	23.5	13.7	9.8	17.0
60-69	103.3	100.3	4.9	51.0	28.4	22.6	40.1
70-79	369.0	350.3	24.1	171.4	83.0	88.5	141.1
80 over	1,894.0	1,809.6	236.9	917.1	362.0	555.1	607.7

1) Ischemic heart disease includes myocardial infarction and angina pectoris.
2) Other heart disease includes heart failure and endocarditis.
* Including unknown age

Also, as shown in Table 3, the mortality rate tends to increase with increasing age. In particular, it has increased rapidly since the age of 70, and in the case of heart disease, the mortality rate of ischemic heart disease is high in the 40-60s, and the mortality rate of other heart diseases is high in the age of 70 or older (National Statistical Office, 2022). As described above, typical risk factors for CVD include not only aging but also blood pressure (systolic and diastolic), obesity, waist circumference, fasting blood sugar, triglyceride, high-density lipoprotein (HDL) cholesterol, and total cholesterol (Lee D.C 2012).

In addition to the mortality rate, the results of a study showing that the prevalence of cardiovascular diseases increases after infection with COVID-19 are drawing attention. A paper published in Nature Medicine in March 2022 (Long-term cardiovascular outcomes of COVID-19) contains the US National Healthcare Databases for a total of 153,760 patients with Long Covid Syndrome. As a result of the one-year study based on this data, the normal control group during the same period was 5,637,647. According to the results, the prevalence of cardiovascular disease increases 30 days after infection. In other words, the frequency of diseases such as ischemic heart disease such as cerebrovascular disease, arrhythmia and angina pectoris, non-ischemic heart disease, pericarditis, myocarditis, heart failure, and thrombosis has increased (Rapportian, April 2022). As mortality and morbidity rates increase, it is necessary to focus on prevention and management. People have a high level of awareness of cardiovascular disease, but a relatively low level of awareness of how to prevent it. CVD is a disease that must be prevented and treated throughout life above all, and primary prevention, which is presented before the disease occurs, is the most important disease (Seung Seok W., 2022). For this reason, research is needed to find a method for predicting the occurrence of potential cardiovascular diseases. Therefore, this paper aimed to find an accurate method for predicting the occurrence of cardiovascular disease. In order to find the most accurate method of predicting the occurrence of cardiovascular disease, three algorithm models suitable for disease prediction were performed and compared.

2. Related research

Pyung-Woo Park and Min-Koo Kim's “Comparative study of machine learning algorithms for diagnosing ischemic heart disease” sets ischemic heart disease, one of the cardiovascular diseases, as a research domain. It compares and analyzes and proposes usable algorithms and efficient approaches within the medical expert system for the diagnosis of the disease. The purpose of the study is to assist medical professionals and doctors based on data from previous patients' first visit records. It is meaningful in helping to explain the causal relationship to ischemic heart disease and minimizing unnecessary related tests. In addition, by configuring experimental data, medical experts and doctors can use it as a learning model, thereby efficiently maximizing experience and knowledge. The data provided in the thesis is a total of 4 data sets, which were created as standard data sets through medical experts and related books and papers. Based on a standard data set, listing the used attributes by their old names: Diagnosis Date, Sex, Age, Height, Weight, Body Mass Index, Body Temperature, Pulse, Smoking, Drinking, Pain, Duration, Month, Heartbeat, Diastolic Blood Pressure, Systolic Blood Pressure, Pulse Pressure, Dyslipidemia, Diabetes Mellitus, Family History, Chest Pain, Dizziness, Dyspnea On Exertion, Dyspnea, Headache, Orthopnea, Palpitation, Syncope, Hypertension, Pulmonary Tuberculosis, Hepatitis, Self Medicine, On Medication, Allergy, and Drug Allergy. Of a total of 35 attributes, the last class label, the Class attribute, indicates the presence or absence of confirmation. As a result of the study, it was found that when the multilayer perceptron classifier was applied to the data set in which the meaning of missing values was assigned, it showed good performance in diagnosing confirmed patients. In addition, it was found that the naive Bayesian and support vector machine classifiers were efficient in accurately diagnosing non-confirmed patients. Through this, it will be used in combination with the development of advanced medical services in the future to improve people's quality of life. In addition, it can play an objective auxiliary role in disease diagnosis, and can create an intelligent value-added service with high accuracy and reliability (Park & Kim, 2018).

Yujun Jeong’s “Development of a 1-year follow-up mortality prediction model for patients with acute myocardial infarction using machine learning methods” utilized KAMIR data. There are 72 attributes of ACS patients extracted from the KAMIR database. There are 22 continuous variables in the experimental data, including age, BMI, WHR, symptoms to balloon time, and arrival to balloon time. Among them, the values used for symptoms of balloon time and arrival to balloon time are unified in minutes. There are 43 categorical variables, including gender, pain, dyspnea, and previous angina before MI symptoms. Among them, the discharge medication variable is applicable only to discharged patients and is not included in the prediction of in-hospital mortality. There are six discrete variables, including variables that are graded according to the severity of symptoms, such as Killip class, lesion type, and TIMI flow. The purpose of this study is to develop a 1-year mortality prediction model for acute myocardial infarction and an in-hospital mortality prediction model including Korean characteristics. Based on the machine learning algorithms DNN, GBM, GLM, and RF, an in-hospital mortality prediction model and a mortality prediction model within one year after discharge were developed. The optimal performing prediction model was selected through 4-fold stratified cross validation and hold-out cross validation. The performance of the machine learning-based predictive model was evaluated by comparing and analyzing previous models GRACE, TIMI, ROC curve (AUC), F1-score, recall, precision, and confusion matrix. The AUC values of the in-hospital mortality prediction model were 0.983 (DNN), 0.982 (GBM), 0.978 (RF), 0.976 (GLM), 0.886 (GRACE), and 0.825 (TIMI), and DNN showed the best performance. The AUC values of the mortality prediction model within one year after discharge were 0.898 (GBM), 0.898 (DNN), 0.883 (RF), 0.873 (GLM), 0.810 (GRACE), and 0.764 (TIMI), with GBM showing the best performance. As a result, it was shown that a machine learning-based model for predicting in-hospital mortality and mortality within 1 year after discharge for Korean patients with acute myocardial infarction outperformed previous statistical models (Yu Jun J., 2018).

Sikandar Ali's “Deep neural network-based clinical decision support and diagnosis for cardiovascular disease in patients with acute myocardial infarction” aims to develop a deep neural network-based diagnosis system that provides a web-based platform for diagnosis and stratification of cardiovascular disease. For this study, the KAMIR-IV dataset collected from 52 domestic hospitals was used. A feature selection technique was applied to select the most suitable feature for the target label. Among the 550 features, the top 14 features with high impact and high correlation with the target feature were selected. Results: sex, age, Killip class, systolic blood pressure (SBP), diastolic blood (DBP), heart rate (HR), hypertension (HTN), diabetes mellitus, smoking, neutrophil, platelets, creatinine, high density lipoprotein (HDL) , Low density lipoprotein (LDL). As a result of the experiment, the proposed deep neural network-based diagnosis system achieved 98.35% accuracy and 99.43% AUC. These results were compared with five other machine learning models: Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression (LR), and Decision Tree (DT). Performance refers to the respective accuracy and AUC. RF (94.53%, 99.29%), SVM (94.4%, 99.0%), KNN (94.35%, 83.05%), LR (88.44%, 89.66%), and DT (94.55%, 91.04%). As a result, it can be seen that our deep neural network-based diagnostic system outperforms other machine learning models in acute coronary syndrome patients (Ali, 2021).

Joo Ki-hoon's "Prediction of cardiovascular disease using health insurance big data and machine learning" focused on three cardiovascular diseases: atrial fibrillation, coronary artery disease, and heart failure. Using the National Health Insurance Service health checkup cohort database, a model predicting the disease was compared with predictive performance of whether the subject would have the disease 2 years later using logistic regression, deep neural network, random forest, and Light GBM techniques. Based on the health checkup questionnaire results, basic information and exercise status of 12 patients including AGE, SEX, BLDS, BMI, BP_HIGH, BP_LWST, GAMMA_GTP, HMG, OLIG_PROTE_CD, SGOT_AST, SGPT_ALT, and TOT_CHOLE were analyzed. In addition, five characteristics were included: smoking status, annual smoking amount, average number of drinking days per week and amount of alcohol consumption, and presence or absence of cardiovascular disease-related diseases. Data were segmented into Medicine, Exclude, and Not_Medicine based on a total of 26 features including patient identification ID and expected correct label (Gi hoon, J., 2020).

Table 4: Experiment Result (Joo, 2020)

	Atrial Fibrillation(AF)			Coronary Artery Disease(CAD)			Heart Failure(HF)		
	Medicine	Exclude	Not Medicine	Medicine	Exclude	Not Medicine	Medicine	Exclude	Not Medicine
LR	0.804	0.777	0.729	0.784	0.778	0.770	0.831	0.822	0.803
DNN	0.806	0.779	0.728	0.790	0.782	0.774	0.838	0.828	0.804
RF	0.776	0.710	0.672	0.748	0.685	0.725	0.818	0.776	0.789
Light GBM	0.803	0.773	0.707	0.786	0.779	0.751	0.835	0.821	0.796

As a result, it was confirmed that the predictive performance of DNN was the best. As can be seen from the results, overall, the deep neural network shows good performance compared to other models. In addition, random forests show significantly lower performance than other models, which is thought to be due to the increased generalization performance, which is a characteristic of random forests. When tested with the same models, heart failure shows the best performance. Through this, it is thought that it will be possible to develop applications that can help doctors diagnose in actual medical settings in the future (Gi hoon, J., 2020).

According to Park Pyung-Woo's "K-Nearest Neighbor Algorithm for Heart Disease Diagnosis and Prediction", heart disease is set as a research domain. It proposes an efficient approach that benefits both doctors and patients by incorporating AI's data mining technology. The ultimate goal is to assist medical experts and doctors based on the accumulated information of patients with suspected heart disease, and to be used as a model for education and learning for medical personnel. The data sets used in the paper were provided by hospitals in Cleveland, Hungary, Switzerland, and Long Beach, and all four data sets were combined into one data set. It consists of 13 attributes: age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, and num, and one class label indicating the presence or absence of heart disease. For the existing patient information, missing values are replaced through the K-nearest neighbor algorithm, and the results are compared and analyzed by applying representative predictive classifiers such as naive Bayes, support vector machine, and multilayer perceptron. The experiments in this study included a K optimization process and were performed in a 10-fold cross-validation method, and comparison and analysis were determined through accuracy and kappa statistics. Through the experiments conducted, it was found that the multilayer perceptron classification method was the best in the data set in which missing values were replaced by applying the 3-nearest neighbor algorithm, and the accuracy was 90.9783% and the kappa statistic was 0.8179 (Park & Lee, 2017).

3. Data Set Implementation

3.1 Experimental Environment

In this paper, Microsoft Azure Machine Learning Studio was used to find the most accurate features when predicting the occurrence of cardiovascular disease. Azure Machine Learning is a cloud service that accelerates and simplifies the management of machine learning project lifecycles. You can create models in Azure Machine Learning or use models built on open-source platforms such as Pytorch, TensorFlow, or scikit-learn. MLOps tools can be used to monitor, retrain, and redeploy models. The Azure platform offers over 600 services, and we are using one of these services, Azure Machine Learning Studio. Use Azure Machine Learning Studio because it provides modules in a variety of formats to help you develop and deploy machine learning models in real-world environments (Kang & Choi, 2018).

3.2 Generate Dataset

After building the experiment environment of Azure Machine Learning Studio, you need to load the data set. Data sets can be created or downloaded from sites such as UCI, Kaggle, and Amazon AWS datasets. The clinical data used in this study were collected from Kaggle as a Heart Failure Prediction Data set (Kaggle, 2021). The data set consists of 12 columns: Age, Sex, Chest Pain Type, Resting BP, Cholesterol, Fasting BS, Resting ECG, Max HR, Exercise Angina, Old Peak, ST_Slope, and Heart Disease.

3.3 Data Preprocessing

Data preprocessing and cleaning is an important task that needs to be done before a dataset can be used for model training. Raw data is often noisy, unstable, and has missing values. Using this data for modeling can lead to misleading results. Building good predictive models requires good data. In order to avoid “garbage in and garbage out” and improve data quality and ultimately increase model performance, it is important to perform data health checks to detect data problems early and determine appropriate data processing and cleaning steps (Microsoft Azure ML, 2022). In this study, Azure Machine Learning Studio was used for clinical data to be utilized. Execute the preprocessing step of changing the name of the column from English to Korean, changing the attribute of the binary feature from Numeric to Categorical, and changing the heart disease to be predicted from Feature to Label. When the data set is cleaned up, it consists of 918 rows and 12 columns. After data preprocessing, the data set was divided into training data and test data at a ratio of 8 to 2, and the random seed was fixed at 3 so that the same value could be obtained when different algorithms were applied.

3.4 Algorithms

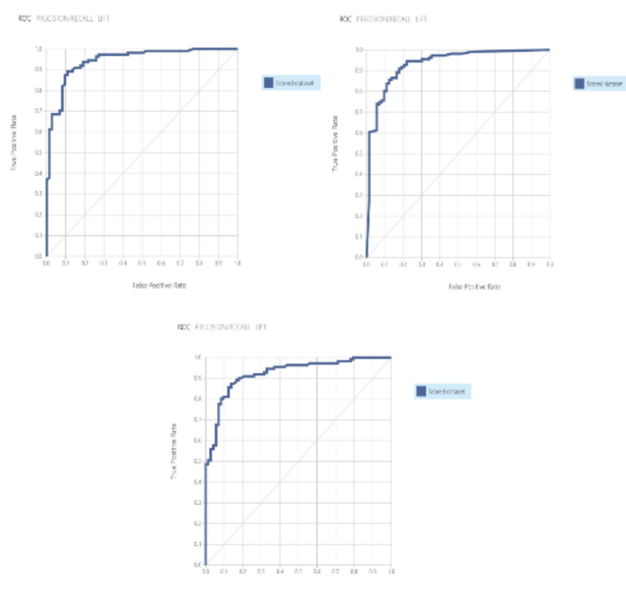
In order to find the most accurate model for predicting cardiovascular disease, two-class Logistic Regression, two-class Boosted Decision Tree, and two-class SVM algorithm models, which are classification algorithm models of supervised learning, were applied. Logistic Regression can analyze the relationship between an independent variable and a dependent variable when the dependent variable is measured on a dichotomous scale in machine learning, and predict the probability of an event occurring (Choi J.S, 2000). Decision Tree classifies a sample group into similar groups based on a specific criterion, and repeats the process of finding and classifying the classified subgroups again based on a specific criterion. It is an analysis method that finds patterns or relationships between dependent variables and independent variables or target variables and input variables through this process. Decision rules are tabulated to classify the group of interest into several subgroups, or make predictions (Kwon & Koo, 2007). SVM (Support Vector Machine) is known to have excellent predictive power along with artificial neural networks among machine learning methods, and research using SVM algorithm models is being conducted in many fields (Park & Hansen, 2012). After experimentation by applying three algorithms, the performance was evaluated through the accuracy, precision, recall, and f1 score of the ROC curve, and each was compared and analyzed.

4. Results

The first experiment obtained 0.886 accuracy, 0.925 precision, 0.883 recall, and 0.903 f1 scores with logistic regression. Figure 1 (a) visualizes how the Logical Regression algorithm model came out.

The second experiment obtained 0.864 accuracy, 0.913 precision, 0.856 recall, and 0.884 f1 scores as a decision tree. Figure 1 (b) visualizes how the Decision Tree algorithm model came out.

The last experiment obtained 0.837 accuracy, 0.909 precision, 0.811 recall, and 0.857 f1 scores with a support vector machine (SVM). Figure 1 (c) visualizes how the Support Vector Machine (SVM) algorithm model came out.



(a) Logistic Regression (b) Decision Tree (c) SVM

Figure 1: ROC curve for each algorithm

Table 5: Results of each algorithm

Results	Logistic Regression	Decision Tree	SVM
Accuracy	0.886	0.864	0.837
Precision	0.925	0.913	0.909
Recall	0.883	0.856	0.811
F1 score	0.903	0.884	0.857

5. Conclusion

In this paper, a total of three experiments were performed: two-class Logistic Regression, two-class Boosted Decision Tree, and two-class Support Vector Machine (SVM) to compare cardiovascular disease incidence prediction models. The performance of the machine learning model can be evaluated through the evaluation indicators that represent the accuracy, precision, recall, f1 score, and AUC of the ROC curve. When comparing the accuracy of the ROC curve as a result of the experiment, the difference in accuracy between the two-class Logistic Regression and the two-class Boosted Decision Tree was 2.2%. The accuracy difference between the two-class Logistic Regression and two-class Support Vector Machine (SVM) was 4.9%, and finally, the accuracy difference between the two-class Boosted Decision Tree and two-class Support Vector Machine (SVM) was 2.7%. Even when the remaining precision, recall, and f1 scores were compared, a difference could be seen. In addition, AUC was 94.5%, 93%, and 92.4% in the order of two-class Logistic Regression, two-class Boosted Decision Tree, and two-class Support Vector Machine. This shows that all three algorithms are closer to 1 between 0 and 1, and it can be seen that the performance of the model is also suitable according to the criterion that the closer to 1, the better the model's performance. However, since all three algorithms satisfy the conditions and the differences are slightly different, it is not certain which algorithm is the most accurate. When all conditions are compared, the result table of the logistic regression algorithm is the highest, so it can be assumed that the accuracy may be higher among the three. A diagnosis can have many decision criteria. It may already exist, or new diagnostic methods may be developed and added. Visualization, as in the paper, effectively makes each comparable. Depending on the effective visualization, the optimal standard can be derived and it can move in the best direction. Furthermore, it is thought that it can serve as an objective assistant in the diagnosis of cardiovascular diseases and guide the direction of diagnosis made by doctors in the actual medical field.

References

Choi. J.S. (2000). Modern statistical analysis using SPSS Ver 10. Bogdoo Publish. pp.514 (in Korean).

Gi-Hun J. (2020). Cardiovascular Disease Prediction using the NHIS Big Data and Machine Learning. Department of Computer Science. Graduate School. Kangwon National University. Korea.

Kaggle. Heart Failure Prediction Dataset. (2021). From https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Lee D. C. Sui X. Church T. S. Lavie C. J. Jackson A. S. & Blair S. N. (2012). Changes in fitness and fatness on the development of cardiovascular disease risk factors: hypertension, metabolic syndrome, and hypercholesterolemia. *Journal of the American College of Cardiology. 59*(7). pp.665-672.

Li X. and D. Lord and Y. Zhang and Y. Xie. (2008). “Predicting motor vehicle crashes using Support Vector Machine models”. *Accident Analysis & Prevention. 40*(4). pp.1611-1618.

Microsoft Azure Machine Learning. (2022). From https://docs.microsoft.com/ko-kr/azure/machine-learning/studio/what-is-ml-studio

Min Soo K. Eun Soo C. (2018). Getting started Machine Learning with MicroSoft AZURE ML. Hanti Media. pp.63-71

Microsoft Azure Machine Learning. (2022). From https://learn.microsoft.com/ko-kr/azure/architecture/data-science-process/prepare-data

National Statistical Office. (2022). 2021 Causes of Death Statistics Results. From

Pyoung-Woo P. Min-Koo K. Hong Seok L. Duk-Yong Y. Seok-Won L. (2018). A Comparative Study of Machine Learning Algorithms for Diagnosis of Ischemic Heart Disease. *Journal of KIISE, 45*(4). pp. 376-389.

Pyoung-Woo P. Seok-Won L. (2017). “Classification of Heart Disease Using K-Nearest Neighbor Imputation”. *Journal of the Korean Society for Information Processing. 24*(2). pp.742-745.

Park S. H. and B. Hansen. (2012). “Prediction of Protein-Protein Interaction Sites Based on 3D Surface Patches Using SVM”. *Korea Information Processing Society. 19*(1). pp.21-28.

Sikandar A. (2021). Deep neural network-based clinical decision support and diagnosis system for cardiovascular disease in patients with acute myocardial infarction. Graduate School of Chungbuk National University Cheongju. Korea

Seungseok, W. (2022). Relationship between Cardiovascular Disease Risk Factors & Rate and Physical Fitness. The 6th to 8th Korea National Health and Nutrition Examination Survey (2014-2019) subjects.

Tae-Won K. Yu-Hoe K. (2007). Comparative Analysis Comparative Analysis Analysis of Prediction Prediction Prediction Taekwondo Taekwondo Taekwondo Trainee`s Trainee`s Defection using Defection using Decision Decision Decision Tree and Logistic Logistic Logistic Regression. *The Korea Journal of Sports Science. 17*(2). pp.71-83.

Rapportian. (2022). The prevalence of cardiovascular disease increases after COVID-19 infection. From https://www.rapportian.com/.

Wikipedia. (2022). From https://en.wikipedia.org/wiki/Cardiovascular_disease

WHO. (2019). From https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Yu Jun, J. (2018). Development of mortality prediction model during admission and 1-year after discharge in acutemyocardial infarction patients using machine learning techniques. Graduate School of Chungbuk National University Cheongju. Korea.

Yun Seon R. Heung Sik C. Sun Woong K. (2016). Prediction of VKOSPI Changes and Application to Actual Option Trading Using SVM. *Korea Intelligent Information System Society. 22*(4). pp.177-192.