Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 7, S. 42-49, 10, 2023
© Telif hakkı IJANSER'e aittir
Araştırma Makalesi

IJANSER
https://alls-academy.com/index.php/ijanser
ISSN: 2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 7, pp. 42-49, 10, 2023
Copyright © 2023 IJANSER
Research Article

Comparison of Machine Learning Algorithms to Predict Cardiovascular Heart Disease Risk Level

Hakan Güler[*], Yunus Santur [2] and Mustafa Ulaş [2]

[1]Software Engineering, Fırat University, Türkiye
[1,2]Artificial Intelligence and Data Engineering, Fırat University, Türkiye

[*](hakanguler@firat.edu.tr) Email of the corresponding author

(Received: 13 October 2023, Accepted: 22 October 2023)

(2nd International Conference on Recent Academic Studies ICRAS 2023, October 19-20, 2023)

ATIF/REFERENCE: Güler, H., Santur, Y. & Ulaş, M. (2023). Comparison of Machine Learning Algorithms to Predict Cardiovascular Heart Disease Risk Level. *International Journal of Advanced Natural Sciences and Engineering Researches*, 7(10), 42-49.

Abstract – Cardiovascular diseases can pose a potential risk for almost every individual since they are associated with multiple parameters such as chronic disease, lifestyle, especially genetic factors. For this purpose, within the scope of the study, machine learning-based models were developed to predict the cardiovascular disease risk level and the metric performances of the algorithms were compared. For this purpose, the performances of the algorithms of the models developed using a data set accessible to all researchers were analyzed in a versatile way. In the study, the results obtained using Logistic Regression, Decision Trees, Random Forests, K-Nearest Neighbors, Gaussian Naive Bayes and LightGBM algorithms were compared. The results present the performance of each algorithm by evaluating it on metrics such as accuracy, precision, sensitivity and F1 score. The study aims to illuminate in which situations different algorithms are more effective and which variables are more determinant in terms of risk estimation. The results of this study can be used as an auxiliary diagnostic method for healthcare professionals working in the cardiovascular field. It can also be used as a predictive model for individuals who want to use artificial intelligence to determine the level of risk.

Keywords – Artifical Intelligence, Diagnosis, E-health, Ensemble Learning, Machine Learning

I. INTRODUCTION

Cardiovascular diseases are a serious concern in the field of health and have an important place among the causes of death worldwide. These diseases are often associated with various parameters such as lifestyle factors, genetic predisposition and chronic health conditions. Being able to accurately predict cardiovascular disease risk is vital for early detection and effective treatment. Cardiovascular Disease protection as a process that begins with the presence of the patient's risk and progresses through progressive vascular disease, targeting organ infection, end-organ treatment and death [1].

Machine learning is founded on the concept that computers can acquire the ability to carry out specific tasks through the analysis of data, without requiring explicit manual programming [2]. Machine Learning focuses on algorithms that improve their performance through experience. They are able to find non-linear relationships and patterns in datasets without being explicitly programmed to do so . The process of analytical modeling building to turn ML algorithms into concrete ML models for the use in intelligent systems is a four step process comprising data input, feature extraction, model building, and model assessment [3].

Karthick et al. have demonstrated the potential of ML algorithms, particularly the random forest algorithm, in accurately predicting cardiovascular disease risk. These findings emphasize the importance of integrating diverse datasets to enhance prediction models using state-of-the-art ML approaches [4].

Reddy et al. utilized attribute evaluators to select significant attributes from the Cleveland Heart dataset, improving the performance of machine learning classifiers for predicting heart disease risk. Using the chi-squared attribute evaluation method, the SMO classifier achieved remarkable accuracy. The study highlights the importance of appropriate attribute selection and hyper-parameter tuning. While satisfactory results have been achieved, there is potential for further experimentation to explore additional machine learning algorithms and feature selection techniques, combining multiple datasets, and enhancing predictive performance [5].

Delpino et al., conducted a systematic study on machine learning applications for predicting chronic diseases. Their research highlights the potential of machine learning models in predicting the risk of various chronic conditions. The study emphasizes the need for further research to enhance model interpretability and generalizability [6].

In their studies, Lupague et al. focused mainly on the use of different models to determine the risk of developing cardiovascular disease using a person's personal lifestyle factors. In their study, they used, extracted and processed records taken from the World Health Organization (WHO) Behavioral Risk Factor Surveillance System (BRFSS) in 2021 as data. But they encountered the problem of imbalance between classes [7].

In their study, Ramesh and his colleagues tried to discover methods that would help protect against heart diseases and locomotor disorders. Here they aim to help people gain valuable information on how to benefit from diagnosis and treatment for a particular patient. They used supervised learning methods such as Naive Bayes, SVM, Logistic regression, Decision Tree Classifier, Random Forest and K- in their studies [8].

In this context, this study addressed the development of machine learning-based models to predict cardiovascular disease risk. The research titled "Comparison of Risk Prediction Level of Cardiovascular Diseases with Machine Learning Algorithms" was concluded by evaluating the performance of different machine learning algorithms to predict the risk of this disease.

The main purpose of this study is to predict cardiovascular disease risk using different machine learning algorithms and compare the prediction performances of these algorithms. As machine learning algorithms were aimed in this study, machine learning approaches have been successfully applied to many different fields [9-13]. Within the scope of the study, how each algorithm performs with metrics such as accuracy, precision, sensitivity and F1 score was examined in detail.

The study includes a series of experiments using 19 different variables in the data set. These variables include lifestyle factors, genetics, and chronic disease states. The obtained results reveal in which situations different algorithms perform better and which variables are more effective in risk estimation.

II. Materials and Method

A. *Data Set*

In this study, a data set prepared by the Behavioral Risk Factor Surveillance System (BRFSS) was used. The BRFSS is a health survey system used to collect information about the health status of US residents, their chronic health conditions, and the use of preventive services. This data set includes 304 different variables in total, and 19 variables were selected for the purpose of presentation. This is 19 variables; 12 categorical variables and 7 digital variables. The data set size is the target variable, which consists of 308,854 observations and 19 variables, and the focus is presented as "Heart Disease" [14].

Table 1. Numerical Variable Description

Variable	Mean	Standart Deviation	%25	%75
Height (cm)	170.6	10.66	163	178
Weight (kg)	83.59	21.34	68.04	95.2
BMI	28.63	6.52	24.21	31.8
Alcohol Cons.	5.10	8.20	0.00	6.00
Fruit Cons.	29.84	24.88	12.00	30.0
Green Vegatables Cons.	15.11	14.93	4.00	20.0
Fried Potato Cons.	6.30	8.58	2.00	8.00

Table 2. Categorical Variable Description

Variable	Unique	Top	Frequency
General Health	5	Very Good	110395
Checkup	5	Within the past year	239371
Exercise	2	Yes	239381
Heart Disease	2	No	283883
Skin Cancer	2	No	278860
Other Cancer	2	No	278976
Depression	2	No	246953
Diabetes	4	No	259141
Arthritis	2	No	207783
Sex	2	Female	160196
Age Category	13	65-69	33434
Smoking History	2	No	183590

B. *Exploratory Data Analysis (EDA)*

In this study, an exploratory data analysis (EDA) analysis was performed in order to better understand the data set and to select the features suitable for the model.

Univariate analysis was performed for the target variable in the data set. When the target variable in the data set was examined, it was found to be unbalanced.

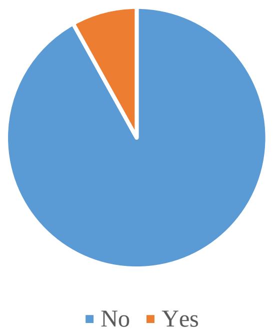

Fig. 1 Heart Disease Univariate Analysis

Bivariate analysis was performed for classification. This analysis allows understanding which features are important for classification and the differences between classes. It also better addresses the relationship between each feature and the target variable. A categorical comparison was made with each category. As an example, let's show its comparison with a categorical variable.

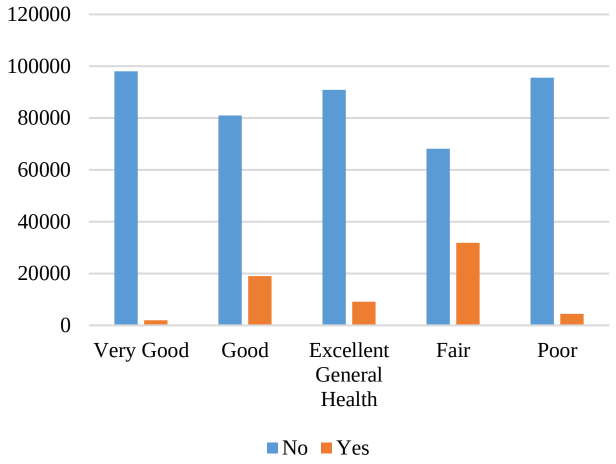

Fig. 2 General Health - Heart Disease

The relationship of numerical variables with the classification target variable (Heart Disease) was examined. The relationship between Body mass index (BMI), based on a person's height and weight [15] and the target variable, which is one of the more than one numerical variables, is visualized and analyzed.

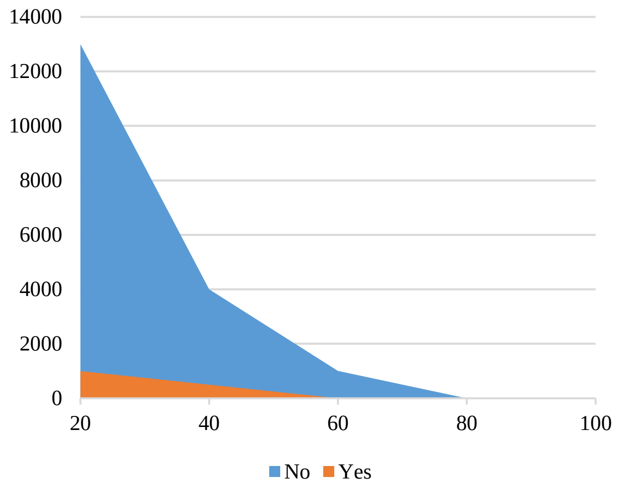

Fig. 3 BMI - Heart Disease

The multivariate analysis method, which is a data analysis method in which more than one variable is examined together, was used. This analysis is used to understand relationships between variables, identify patterns, and understand more complex data structures. With Multivariate Analysis, multidimensional data is made sense by using statistical analysis, graphics and other data visualization techniques.

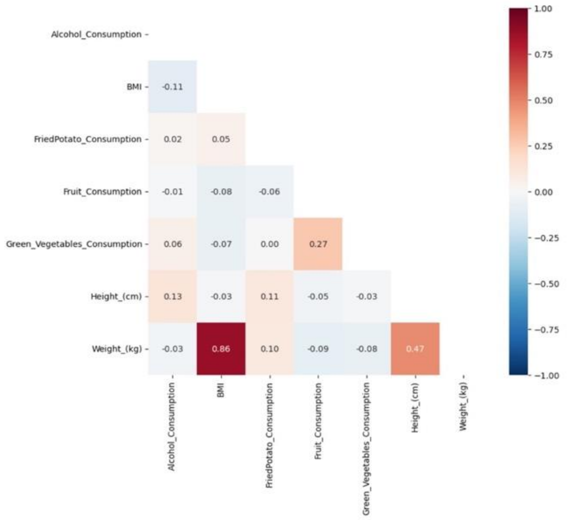

Fig. 4 Multivariate Analysis

This EDA phase provided basic information on understanding the structure and content of the dataset. This information will be observed to be the main target for model development and performance evaluation in the next steps.

C. *Data Preprocessing*

In the next stage of the work, the data pre-processes and the power that runs the machine is cut off from being ready. This process consists of stages such as validation, normalization, feature selection. The quality of this data has been increased and content optimized. In our data set, the Yes and No values of the target variable 'Heart Disease' are mathematically converted to 0 and 1. After the conversion process, 283883 numbers resulted as 0s and 24971 numbers as 1s. Afterwards, the data set was divided into two subsets as training and test data. The aim here is to train and test the machine learning model. There are a total of 247,083 examples and 19 features in the training dataset, while there are 61,771 examples and 19 features in the test dataset. This is a necessary step for training the model and evaluating its performance on an independent dataset.

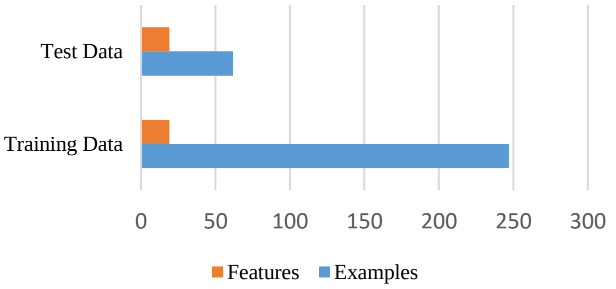

Fig. 5 Training and Test Data

The rates of the "Heart Disease" classes, which are the target variable in the training and test data sets, were analyzed separately for the training and test data sets. This analysis was performed to assess data imbalance by examining the proportions of individuals with and without heart disease in the training and testing datasets. The results obtained provided an important reference for model training and interpretation of the results.

Afterwards, only OneHotEncoder was implemented for the categorical pipeline. In this way, the data set was cleaned and it was ensured that there were no missing values. Logarithmic transformation and standardization of numerical data was achieved. In this way, numerical variables were processed and made ready for model training. For ordinal variables, variables were transformed according to their order. The lowest ranked values started at 0 and were incremented by 1.

After all these operations were carried out in order, there were changes in the data set. Data preprocessing means that the pipeline adds new columns to the data set or transforms existing columns. That is, some changes were made to the data preprocessing pipeline to make the dataset suitable for the model.

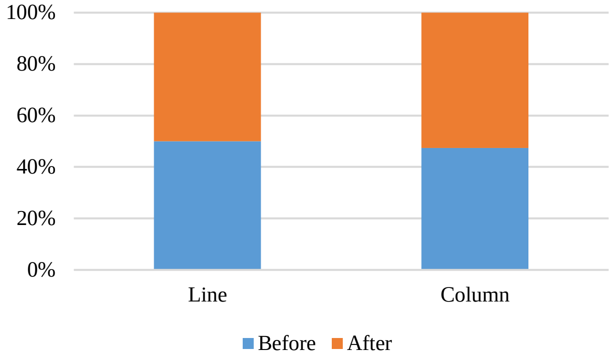

Fig. 6 Shape Before-After the Preprocessing

D. *Machine Learning Models*

Various machine learning models were used. These models were selected based on their suitability for classification tasks and their ability to handle both numerical and categorical variables. The models used include LightGBM, Logistic Regression, K-Nearest Neighbors, Naïve Bayes, Decision Tree Classifier and Random Forest.

LightGBM. is an algorithm designed by Microsoft Research Asia using the GBDT framework. It aims to increase computational efficiency so that prediction problems related to big data can be solved more effectively. In the GBDT algorithm, a pre-sorting approach is used to select and split indicators. Although this method can pinpoint the split point, it requires more time and memory. LightGBM is histogram-based algorithm that increases training speed and reduces memory consumption [16].

Logistic Regression, It is called logistic regression of the results of a generalized linear regression. Themed logic is to predict the probability of an event by using a logical function. The probability value is estimated and the output range is between 0 and 1 [17].

K-Nearest Neighbors, It is called the growth of classical non-parametric types in pattern recognition. It is widely used in many fields due to its simplicity, effectiveness and preservation. However, the performance of the results of the kNN process is negatively affected by the selection of a fixed and single value for all queries in the search phase and the existence of the simple density voting rule in the decision phase [18].

Naive Bayes is one of the most popular data mining cleaners. Its efficiency comes from feature independence, but this can be violated in many real-world datasets. Since feature selection is an important approach, many efforts have been made to mitigate it. However, traditional efforts to achieve feature selection in naive Bayes suffer from heavy programming overhead [19].

Decision Tree Classifier is a supervised machine learning algorithm suitable for solving classification and regression problems. Decision trees are a type of algorithm that iteratively builds training records by applying split conditions at each node, which divides them into subsets with output variables of the same class [20].

Random Forest, is a predictive statistical or machine learning algorithm [21].

These models offer different approaches to the complexity of the dataset and are chosen based on their appropriateness for the problem at hand. This diversity is important for model selection and comparing results. It allows for a more robust exploration of the dataset's patterns and behaviors. This comprehensive strategy enhances the reliability and depth of the analysis, ultimately aiding in the identification of the most effective model for predicting CVD risk.

E. *Evaluation of F1 Scores*

A comprehensive evaluation was conducted using 10-fold cross-validation to determine the model that achieved the best F1 score in predicting CVD risk. The models were assessed using F1 scores, a measure that combines precision and recall, to evaluate their performance. The average F1 score for each model was calculated to represent its overall performance.

III. Results And Discussion

This study evaluated the performance of different machine learning models for predicting cardiovascular disease risk. Models include Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Gaussian Naive Bayes and LightGBM. Models were selected considering their ability to handle both numerical and categorical variables.

According to the results obtained, it was observed that Logistic Regression and Decision Tree models performed better than others, based on the evaluation using the F1 score. While the Logistic Regression model provided a high precision value, the Decision Tree model was found to have a high recall value. However, both of these models showed low F1 score on the unbalanced dataset.

The results are as follows:

Decision Tree: The Decision Tree model exhibits a low performance in terms of F1 score of 0.222. Its precision is 1.000 and its recall value is 0.999. While it is successful in accurately detecting positive disease state, it appears to have a tendency to classify negative states as false positives.

Table 3. Decision Tree Report

Label	**Precision**	***Recall***	***F1-Score***	***Support***
0	0.9999	1.0000	0.9999	227106.00
1	1.0000	0.9998	0.9999	19977.000
Accuracy	0.9999	0.9999	0.9999	0.9999
Macro Avg	0.9999	0.9999	0.9999	247083.00
Weighted Avg	0.9999	0.9999	0.9999	247083.00

Logistic Regression: The Logistic Regression model stands out with a high F1 score of 0.326. Its precision is 0.205 and its recall value is 0.788. While it stands out for its ability to accurately detect positive disease states, it has been observed that it tends to classify negative states as false positives.

Table 4. Logistic Regression Report

Label	**Precision**	***Recall***	***F1-Score***	***Support***
0	0.9751	0.7317	0.8360	227106.00
1	0.2052	0.7879	0.3257	19977.00
Accuracy	0.7362	0.7362	0.7362	0.7362
Macro Avg	0.5902	0.7598	0.5808	247083.00
Weighted Avg	0.9128	0.7362	0.7948	247083.00

Gaussian Naive Bayes: The Gaussian Naive Bayes model has the lowest performance with a low F1 score of 0.270. Its precision is 0.162 and recall value is 0.824. Low levels of both precision and recall were observed.

Table 5. Gaussian Naïve Bayes Report

Label	**Precision**	***Recall***	***F1-Score***	***Support***
0	0.9758	0.6253	0.7622	227106.00
1	0.1620	0.8236	0.2708	19977.00
Accuracy	0.6413	0.6413	0.6413	0.6413
Macro Avg	0.5689	0.7245	0.5165	247083.00
Weighted Avg	0.9100	0.6413	0.7224	247083.00

K-Nearest Neighbor: The K-Nearest Neighbor model showed a high performance in terms of F1 score with 0.276. Its precision is 0.356 and its recall value is 0.999. However, in terms of sensitivity, it appeared to tend to classify negative cases as false positives.

Table 6. K-Nearest Neighbor Report

Label	**Precision**	***Recall***	***F1-Score***	***Support***
0	0.9999	0.8408	0.9135	227106.00
1	0.3559	0.9998	0.5250	19977.00
Accuracy	0.8537	0.8537	0.8537	0.8537
Macro Avg	0.6779	0.9203	0.7192	247083.00
Weighted Avg	0.9479	0.8537	0.8821	247083.00

Random Forest: The Random Forest model attracts attention with the highest F1 score of 0.325. Both precision (0.253) and recall (0.789) values are high. It offers a good balance in an unbalanced dataset and has been shown to accurately detect positive disease status.

Table 7. Random Forest Report

Label	**Precision**	***Recall***	***F1-Score***	***Support***
0	0.9999	0.9999	0.9999	227106.00
1	0.9999	0.9999	0.9999	19977.00
Accuracy	0.9999	0.9999	0.9999	0.9999
Macro Avg	0.9999	0.9999	0.9999	247083.00
Weighted Avg	0.9999	0.9999	0.9999	247083.00

LightGBM: The LightGBM model exhibits a low performance in terms of F1 score with 0.111. Its precision is 0.448 and its recall value is 0.063.

He appeared to have difficulty identifying negative situations correct.

Table 8. LightGBM Report

Label	**Precision**	***Recall***	***F1-Score***	***Support***
0	0.9233	0.9932	0.9570	227106.00
1	0.4481	0.0626	0.1099	19977.00
Accuracy	0.9179	0.9179	0.9179	0.9179
Macro Avg	0.6857	0.5279	0.5334	247083.00
Weighted Avg	0.8849	0.9179	0.8885	247083.00

These results suggest that models used to predict cardiovascular disease risk may often underperform on unbalanced data sets. More data collection or different feature engineering approaches can be considered to improve model performance. Additionally, taking a more thoughtful approach to model selection could be a potential strategy to achieve better results. The study also provides a basis for comparing the performance of different machine learning models used to assess cardiovascular disease risk. However, further research and model refinement may be required. It has been observed that future studies in this field play an important role to obtain more precise and reliable results.

IV. CONCLUSION

This study was conducted to evaluate the performance of different machine learning models to predict cardiovascular disease risk. Based on the results of the study, some important findings were reached.

Six different machine learning models were used in the study: Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Gaussian Naive Bayes and LightGBM. The abilities of each of these models to predict cardiovascular disease risk were compared using the F1 score. The results show that Logistic Regression and Decision Tree models perform better than others. While the Logistic Regression model attracts attention with a high F1 score (0.326), the Decision Tree model offers a high recall value (0.999). However, both models show low F1 score on the unbalanced dataset.

A key finding of the study suggests that models used to predict cardiovascular disease risk may underperform in unbalanced data sets. In particular, it has been shown to perform well at detecting positive disease states, while tending to classify negative states as false positives. This is predicted to significantly impact model performance in real-world applications.

The results obtained show some improvement ways to increase model performance. It is anticipated that collecting more data, applying feature engineering approaches, or using different class balance techniques will improve model performance. Additionally, to obtain a better result, it is necessary to adopt a more careful approach during model selection.

This study provided a basis for comparing the performance of different machine learning models used to assess cardiovascular disease risk. However, further research and model refinement may be required. Future studies may make further progress in this field to obtain more precise and reliable results.

V. ACKNOWLEDGEMENT

This work was supported by the Republic of Türkiye Ministry of Industry and Technology Attratcion Centers Supporting Program under Grant No: TRB1/22/CMDP-E1/0001.

REFERENCES

[1] Dahlöf, B. (2010). Cardiovascular disease risk factors: epidemiology and risk assessment. The American journal of cardiology, 105(1), 3A-9A.

[2] de Moraes Batista, A. F., Chiavegatto Filho, A. D. P. (2019). Machine learning aplicado à Saúde. Sociedade Brasileira de Computação.

[3] Herm, L. V., Heinrich, K., Wanner, J., Janiesch, C. (2023). Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. International Journal of Information Management, 69, 102538.

[4] Karthick, K., Aruna, S. K., Samikannu, R., Kuppusamy, R., Teekaraman, Y., Thelkar, A. R. (2022). Implementation of a heart disease risk prediction model using machine learning. Computational and Mathematical Methods in Medicine, 2022.

[5] Reddy, K. V. V., Elamvazuthi, I., Aziz, A. A., Paramasivam, S., Chua, H. N., & Pranavanand, S. (2021). Heart disease risk prediction using machine learning classifiers with attribute evaluators. Applied Sciences, 11(18), 8352.

[6] Delpino, F. M., Costa, Â. K., Farias, S. R., Chiavegatto Filho, A. D. P., Arcêncio, R. A., Nunes, B. P. (2022). Machine learning for predicting chronic diseases: a systematic review. Public Health, 205, 14-25.

[7] Lupague, R. M. J. M., Mabborang, R. C., Bansil, A. G., Lupague, M. M. (2023). Integrated Machine Learning Model for Comprehensive Heart Disease Risk Assessment Based on Multi-Dimensional Health Factors. European Journal of Computer Science and Information Technology, 11(3), 44-58.

[8] Ramesh, T. R., Lilhore, U. K., Poongodi, M., Simaiya, S., Kaur, A., Hamdi, M. (2022). Predictive analysis of heart diseases with machine learning approaches. Malaysian Journal of Computer Science, 132-148.

[9] Beam, A. L., Kohane, I. S. (2018). Big data and machine learning in health care. Jama, 319(13), 1317-1318.

[10] Panch, T., Szolovits, P., Atun, R. (2018). Artificial intelligence, machine learning and health systems. Journal of global health, 8(2).

[11] Yadav, B. P., Ghate, S., Harshavardhan, A., Jhansi, G., Kumar, K. S., Sudarshan, E. (2020, December). Text categorization Performance examination Using Machine Learning Algorithms. In IOP Conference Series: Materials Science and Engineering (Vol. 981, No. 2, p. 022044). IOP Publishing.

[12] Sun, H., Ramuhalli, P., Jacob, R. E. (2023). Machine learning for ultrasonic nondestructive examination of welding defects: A systematic review. Ultrasonics, 127, 106854.

[13] Kolk, M. Z., Deb, B., Ruipérez-Campillo, S., Bhatia, N. K., Clopton, P., Wilde, A. A., ... & Tjong, F. V. (2023). Machine learning of electrophysiological signals for the prediction of ventricular arrhythmias: systematic review and examination of heterogeneity between studies. EBioMedicine, 89.

[14] Online: https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset

[15] Khanna, D., Peltzer, C., Kahar, P., & Parmar, M. S. (2022). Body mass index (BMI): a screening tool analysis. Cureus, 14(2).

[16] Liang, W., Luo, S., Zhao, G., & Wu, H. (2020). Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms. Mathematics, 8(5), 765.

[17] Deng, S., Wei, M., Xu, M., & Cai, W. (2021). Prediction of the rate of penetration using logistic regression algorithm of machine learning model. Arabian Journal of Geosciences, 14, 1-13.

[18] Pan, Z., Wang, Y., & Pan, Y. (2020). A new locally adaptive k-nearest neighbor algorithm based on discrimination class. Knowledge-Based Systems, 204, 106185.

[19] Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. Knowledge-Based Systems, 192, 105361.

[20] Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. International Journal of Advanced Computer Science and Applications, 11(2), 612-619.

[21] Schonlau, M., Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal, 20(1), 3-29.