American Research Journal of Cardiovascular Diseases

Volume 5, Issue 1, 01-06 Pages
Research Article | Open Access
ISSN (Online)- 2575-7601
DOI : 10.21694/2575-7601.24001

AMERICAN RESEARCH JOURNALS
An Academic Publishing House

Cardiovascular Disease Predictive Modeling with Machine Learning Feature Importance

Daniel Han

Bergen County Academies, Hackensack, NJ 07601.

ABSTRACT

Cardiovascular diseases are one of the leading factors of death around the world. To provide insights on the correlation between general factors and the existence of cardiovascular diseases, a dataset supplied by Svetlana Ulianova in Kaggle with 70,000 patient records with 11 features and a target was used to determine what attributes have the most influence on cardiovascular conditions. The research results suggest that 1) Smoking, height, and gender did not have a significant contribution to cardiovascular disease 2) Systolic and diastolic were shown to have a strong contribution to cardiovascular disease 3) Random Forest model performance yielded the highest metrics compared to both Gassausian Naive Bayes and the benchmark model Logistic Regression. This research can assist Doctors with determining patients who have a high susceptibility to cardiovascular disease.

INTRODUCTION

According to the World Heart Federation, cardiovascular disease has increased by 60% globally during the past 30 years. Cardiovascular disease, often abbreviated as CVD, is a classification of disorders in the heart or blood streams. These include "ischemic heart disease, stroke, heart failure, peripheral arterial disease, and a number of other cardiac and vascular conditions" (Mensah et al., 2019). The two most common CVD deaths are from heart attack and stroke. Out of these deaths, most of them were from less affluent countries. CVDs, along with ischemic heart disease and stroke, are few of the leading causes of global mortality and disability. In fact, CVD cases doubled from 1990 to 2019 with around 523 million cases (Roth et al., 2020).

These issues are critical in healthcare, and identifying and predicting such CVD early is an important factor in preventing further progression and reducing mortality rates. However, due to the challenge in the identification of such diseases may be challenging to do manually, more efficient methods (that utilize machine learning) have to be implemented (Ahsan, 2022). For example, EayanAlanzi (2022) used convolutional neural networks and K-nearest neighbor classification models to predict diabetes, CVD, and cancer, using both structured and unstructured data.

There have also been successful attempts at predicting CVD using AI that implemented models such as support vector machines (SVM) and logistic regression using data from the Internet of Things (Subramani et al., 2023). Additionally, other studies such as one from K. Arumugam and Mohd Naved (2023) focused on different ML models that included decision trees, Gaussian Naive Bayes, and SVM. This study concluded that the fine-tuned decision tree model with optimal performance yielded the best results in disease prediction (Arumugam et al., 2023). Overall, multiple ML models are used for various disease predictions, and standardly including "decision trees, support vector machines, neural networks, and ensemble methods" (Chotrani, 2022).

However, these ML models have their faults and limitations. The difficulty in obtaining data, time complexity, and difficulty to verify predictions can obscure the usage of ML in different contexts (Onyema et al., 2022). For example logistic regression training somewhat suffers from the time complexity of $O\ (mn^2+n^3)$ where n is the number of features and m is the sample size (Chu et al., 2006). In addition, the usage of ML models are limited within the technology available. Thus the space complexity of ML algorithms are another problem (Khanzode et al., 2020).

A brief analysis of the data with 70,000 data points of patient records including the features age, height, weight, gender, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake, and physical activity gives a rough idea of the impact of such factors in CVD. To combat the issue of time complexity, NB was used, and the performance of RF (ensemble of DT) and DT were compared. Evaluation metrics used included precision, recall, and f1 scores which is part of the standard evaluation metric for ML models (Naidu, 2023). Thus this research leverages several machine-learning techniques to seek out to determine which patient attributes have the most influence on cardiovascular disease.

Data

The research leveraged a dataset of 70000 data points from Kaggle. The features include age, height, weight, gender, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake, and physical activity. There is no null values for each feature. Some insights on the data may be obtained from graphs of the data after preprocessing. The dataset used: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

METHODOLOGY

Data Description & Pre-processing

Model/Method

This research leverages the dataset of patient attributes to predict the existence of cardiovascular disease in an individual. To remove outliers within the dataset, a z-score methodology was implemented on the height, systolic, and diastolic attributes. We trained the dataset with the Logistic Regression model as a baseline model and also leveraged Random Forest and Gaussian Naive Bayes.

The Logistical Regression model is a model that finds a "line of best fit," except it is a sigmoid curve, and the curve is in multi-dimensional space. The model is trained by trying to minimize the error it has from the predicted value to the actual value, with the equation $Y=b_0+b_1\ x_1+b_2\ x_2+...+b_n\ x_n$. Once the LR model is trained, the output is the probability that it is classified as CVD, thus if it is closer to 1 than it is closer to 0, the output of the data would indicate that the data most likely means a person has CVD. With this training, unnecessary features are given less importance, and vice versa.

On the other hand, the Random Forest model leverages Decision Tree models, feeding different subsets of data into the DT models. These DT models basically ask multiple questions about the dataset in series, in order to classify the data.A classification is made based on all of the decision tree models' predictions, and this is why RF is called Random Forest. Because RF utilizes multiple DT models, it works fine even with outliers, but for the same reason, it takes up a lot of memory and has a high computation time.

Lastly, the Gaussian Naive Bayes is a fast and simple model that calculates a combined probability based on the distribution of the features with separate classifications. For this research, the patients with CVD and without CVD would have the variance and average for each feature separately calculated by the NB model. With these calculations, the model will determine if the data inputted follow the data with CVD more closely or not, based on the Naive Bayes theorem which assumes that the data is approximately normal (NB models also assume that the features are independent).

To make sure we get models with accurate predictions, multiple parameters for the models were evaluated using parametric grid search optimization. Once optimized parameters were isolated, the models were retrained and model performance metrics were computed. The models chosen have transparency in determining which attributes are influencing the class label (0: no cardiovascular disease, 1: the presence of cardiovascular disease.

Finally, to compare the model performances, values such as precision, recall, and F1 scores will be compared **(explanation of these metrics are in later parts of this paper)**. Through this, the optimal model to predict the existence of CVD will be determined. The models will also be separately analyzed through the feature importance through the optimized coefficients of the models.

Summary

1. Age in days is converted to age in years, and then rounded to decrease the cardinality of the unique values in the dataset, to speed up computation.
2. Datapoints with outliers in quantitative features are removed; i.e. |z| is above 3.
3. To account for the illogical values in diastolic and systolic blood pressure, any values below 50 or above 250 are removed.
4. Machine learning models such as RF, NB, and LR are set up.
5. Optimized the models using 5 fold cross validation, allocating 20% of the dataset for testing. However the models' parameters were limited to a select few choices in order to reduce runtime.
6. Optimized models are tested, outputting precision, recall, and F1 scores
7. The optimized RF model also outputs the feature importances, using Gini importance.

Diagram 1. Methodology Flow Diagram

EVALUATION

Result

Table 1. Model Performance

	Precision	Recall	F1-score
RF	0.733832	0.732072	0.731377
NB	0.723955	0.714327	0.710733
LR	0.716578	0.715569	0.715064
DT	0.635322	0.635314	0.635191

For these CVD prediction models, precision, recall, and F1 scores can show how well these models perform in predicting the existence of CVD given multiple features. The values of these indicators range from 0 to 1, with higher values meaning better performance. Here is a confusion matrix diagram that helps understand the evaluation metrics.

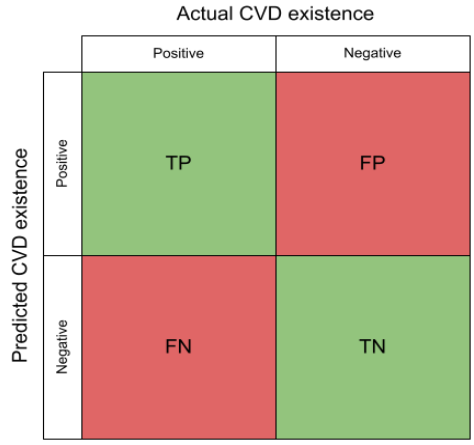


Diagram 2. Confusion Matrix

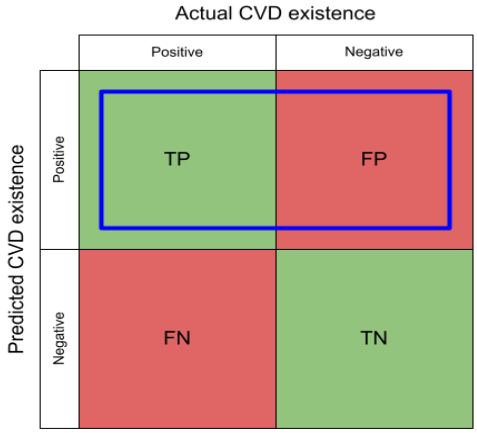


Diagram 3. Matrix for Precision

This table indicates the data that precision measures. Precision $\frac{TP}{(TP+FP)}$. In other words, precision measures the rate at which a positive predicted CVD is true. With Table 1, we can see that the RF model had the best precision of around .734, then NB with .724, LR with .717, and finally DT with .635.

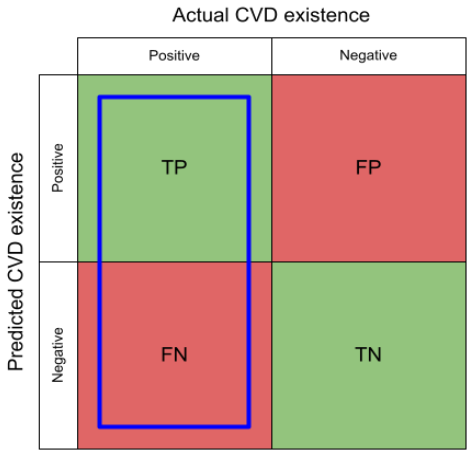


Diagram 4. Matrix for Recall

This table indicates what recall measures. Recall= $\frac{TP}{(TP+FP)}$. This means the rate at which a patient with CVD is predicted correctly. Again, with table 1, we can conclude that RF again has the highest performance with .732, but next was LR with .716, NB with .714, and DT with .635

The last indicator of performance on Table 1 was the F1 score. This combines the precision and recall values using their harmonic mean. F1= $\frac{2\times Precision\times Recall}{Precision+Recall}$. With table 1, it is shown that overall, RF is the best model for CVD, with LR and NB following and DT at the end.

Table 2. RF Feature importance

Feature	RF Importance	DT Importance
ap_hi (systolic blood pressure)	0.468428	0.195399
ap_lo (diastolic blood pressure)	0.194903	0.055061
age_years	0.127057	0.137403
cholesterol	0.091773	0.036322
weight	0.055170	0.250485
height	0.027109	0.226033
gluc (glucose level)	0.013491	0.024832
active (physical activity)	0.008646	0.018804
smoke (if patient smokes)	0.004864	0.012854
gender	0.004528	0.030881
alco (if patient consumes alcohol)	0.004033	0.011925

This table shows that ap_hi, which represents systolic blood pressure, has the highest feature importance for CVD prediction. In Random Forest classification models, feature importance is calculated using Gini importance, which is a calculation of reduction in the impurity of nodes when a particular feature is used for splitting in an individual decision tree.

In decision trees, as mentioned, data is split into subsets based on certain features at each node, which is the point in the tree where a decision is made based on the value of a certain feature. Impurity of nodes represent the degree of mixing of the different target values (in this research's case, CVD existence). This means that low impurity suggests that the node is mostly consisted of instances of a single target value, and high impurity suggests that the node has a mixed distribution of such target values.

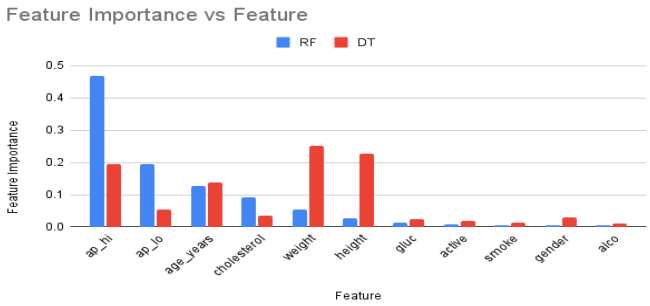
The Gini importance calculates the extent to which each feature reduces impurity in the nodes of each decision tree. When a feature is used to split the data at a node, resulting subsets of data should have a lower impurity. The Gini importance is calculated by summing up such reductions in impurity across all the decision trees in the RF, in proportion to how much data the nodes split.

These importance scores are then normalized such that their sum would total to 1. Because each Gini importance is the measure of the mean decrease in impurity, the higher the feature importance, the better the metric is for prediction. This means that according to Table 2, surprisingly smoking did not serve a big role in CVD prediction for the RF classifier.

CONCLUSION

Out of the three models that were tested, overall RF is the best, and then LR and NB follows due to the ranking of their

respective F1 scores. However, it is to be noted that the difference between the F1 scores were not vastly different, with just .016 difference between RF and LR. Thus, it may be more beneficial for people in the medical field with limited access to use NB instead of RF due to the similar performances. Decision Tree had the worst performance metrics in all respective scores.



Graph 1. Feature Importance vs Feature

1. Smoking, gender, and alcohol consumption did not have a significant contribution in predicting CVD according to the feature importance of the optimized RF model
    a. limitation is that the smoke feature did not indicate how often a person smoked, but rather if the person smokes
    b. another limitation lies in the alcohol consumption, as this feature is also binary, and dependent on if the person just consumes alcohol or not.
2. Systolic and Diastolic blood pressures were the most important features in determining if a person had CVD. Higher values of these blood pressures corresponds to a higher likelihood of having CVD.
3. The RF model outperformed NB and the baseline model LR in precision, recall, and F1 scores. LR did do better than NB in recall, but was inferior for precision and F1 scores.
    a. However NB assumes that all the features are independent, which is not true.
4. The feature importance ranking from greatest to least is as following: ap_hi, ap_lo, age_years, weight, height, cholesterol, gluc, active, smoke, gender, alco
5. The performance difference between RF, LR, and NB was small, and even though RF did outperform both ML models, NB may be used by individuals in the medical field with limited resources, in order to reduce memory usage and runtime.

A. Using more data with more extensive preprocessing

a. In the future, we can focus on eliminating the limitations mentioned in the conclusion, perhaps using more accurate data on alcohol/smoking levels with higher cardinality.

b. Filtering out features that are dependent on each other. This may potentially lead to biases in using NB, as the algorithm assumes every feature is independent of one another.

c. We may combine this dataset with other datasets to have a more extensive dataset. However, since not all datasets share the same features, there needs to be much more preprocessing.

B. Implementingmoremachinelearningmodels

a. We may implement more machine learning models to have a comprehensive ranking of machine learning models that are good predictors of CVD existence. These ML models include, but are not limited to Convolutional Neural Networks, K Nearest Neighbors, and Support Vector Machines.

C. Experimenting with more possible parameters of the ML models

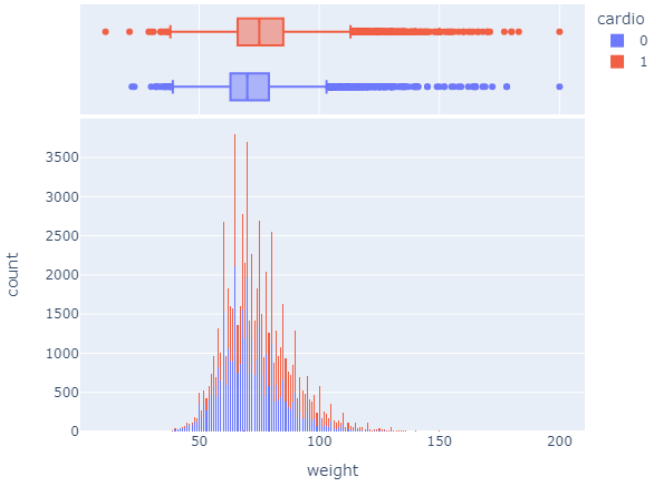
a. In this research, to reduce runtime, the process for training the models used limited parameter options. In the future, with higher computing power, a much wider range of parameters may be used, and this may yield better precision, recall, or f1 scores.

b. Instead of doing 5-fold validation on the dataset, we can use a higher number of folds as well, to balance the amount of unused data with the accuracy of the models (account for both overfitting and underfitting)

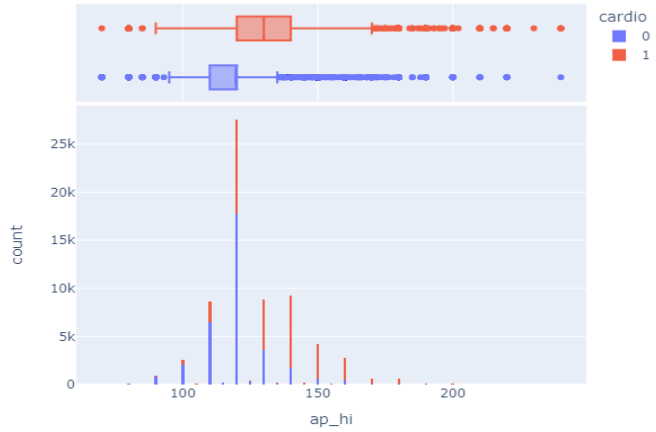
D. Analyze training time and prediction time of different ML Models

a. Using the known time complexities for different ML models and experimented training/prediction time, we can create a new metric that evaluates the time ML takes with its accuracy to evaluate tradeoffs between time and accuracy so that people with less developed technology (with less space for data) could know what ML model they should implement.

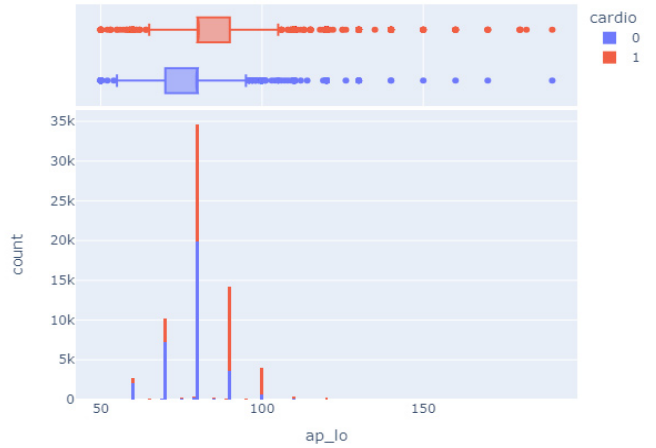
APPENDIX



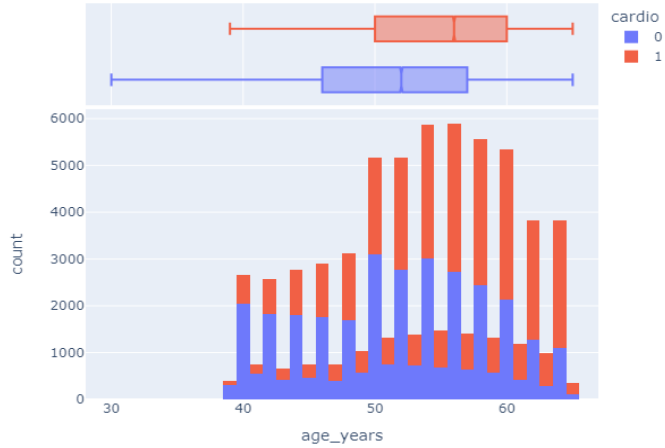
Graph 2-1. Histogram and Boxplot for weight



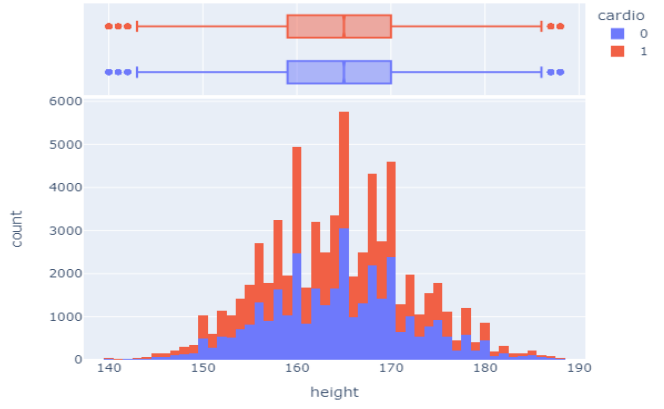
Graph 2-2. Histogram and Boxplot for ap_hi



Graph 2-3. Histogram and Boxplot for ap_lo

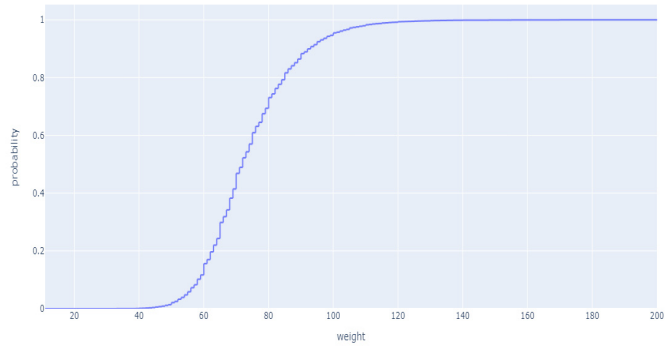


Graph 2-4. Histogram and Boxplot for age_years

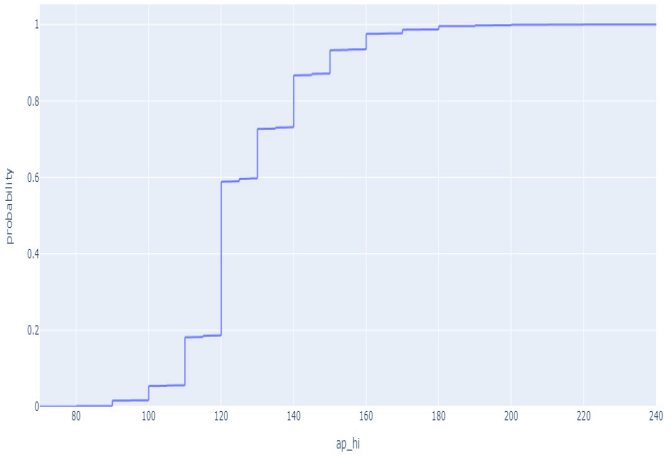


Graph 2-5. Histogram and Boxplot for height

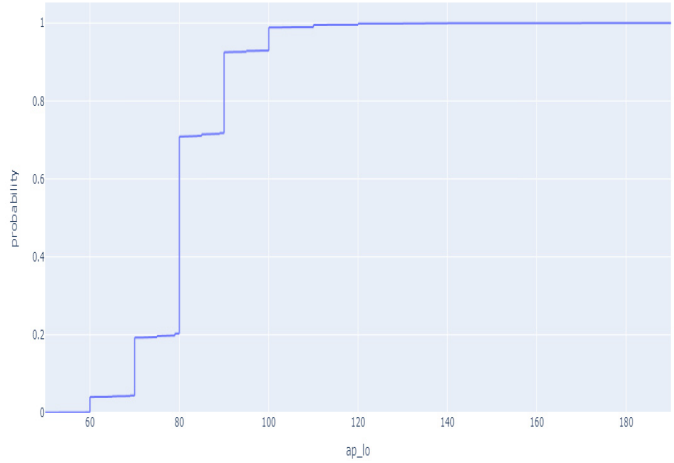
These graphs (2-1~5) each contain a box and whisker plot of each quantitative feature categorized by the existence of cardiovascular disease. The red indicates a 1 on the cardio target, which means that CVD is present. On the other hand, the blue indicates a 0 on the cardio target, which means that CVD is not present. The graph below is a histogram of the features, also categorized by the target. Utilizing the combination of these two graphs, it can be inference that age, systolic blood pressure, and diastolic blood pressure will have high feature importance. In addition, you can observe the shape of each graph, and observe that height is approximately normal and symmetric, weight is skewed to the right, etc. NB assumes that these features are normally distributed, and thus we can see potential sources of error with these graphs as well.



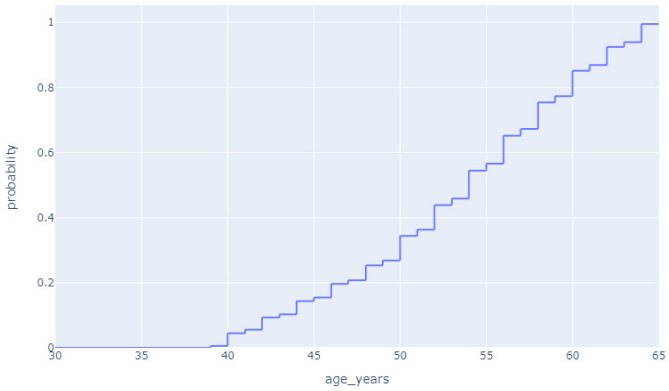
Graph 3-1. Cumulative Relative Frequency for weight



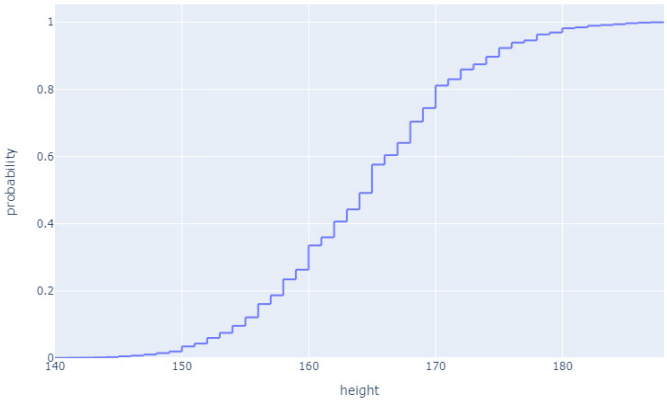
Graph 3-2. Cumulative Relative Frequency for ap_hi



Graph 3-3. Cumulative Relative Frequency for ap_lo



Graph 3-4. Cumulative Relative Frequency for age_years



Graph 3-5. Cumulative Relative Frequency for height

These graphs(3-1~5) indicate the cumulative relative frequency for each of the quantitative features, which are age, height, weight, systolic blood pressure, and diastolic blood pressure respectively.

REFERENCES

1. Gregory A. Roth. & George A Mensah et al. (2020 Dec). Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update From the GBD 2019 Study. *Journal of the American College of Cardiology, Volume 76,* 2982-3021. Retrieved from https://www.jacc.org/doi/10.1016/j.jacc.2020.11.010.
2. K. Arumugam., & Mohd Naved et al. (2023 April). Multiple disease prediction using Machine learning algorithms. *Materials Today: Proceedings, Volume 80,* 3682-3685. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S2214785321052202?via%3Dihub/
3. Sivakannan Subramani. & Neeraj Varshney et al. (2023 April). Cardiovascular diseases prediction by machine learning incorporation with deep learning. *Frontiers in Medicine, Volume 10,* Retrieved from https://www.frontiersin.org/articles/10.3389/fmed.2023.1150933/
4. Edeh Michael Onyema., & Khalid K. Almuzaini et al. (2022 June). Prospects and Challenges of Using Machine Learning for Academic Forecasting. *Computational Intelligence and Neuroscience, Volume 2022,* Retrieved from https://www.hindawi.com/journals/cin/2022/5624475/
5. Eayan Alanazi. (2022 Feb). Identification and Prediction of Chronic Diseases Using Machine Learning Approach. *Journal of Healthcare Engineering, Volume 2022,* Retrieved from https://www.hindawi.com/journals/jhe/2022/2826127/
6. Ahsan, M., Luna, S. A., & Siddique, Z. (2022). Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare*, *10*(3), 541. https://doi.org/10.3390/healthcare10030541
7. World Health Organization. (n.d.). *World Health Organization (WHO)*. World Health Organization. https://www.who.int/
8. Naidu, G., Zuva, T., & Sibanda, E. M. (2023). A review of evaluation metrics in machine learning algorithms. *Artificial Intelligence Application in Networks and Systems*, *724*, 15–25. https://doi.org/10.1007/978-3-031-35314-7_2
9. Chu, C.-T., Kim, S. K., Lin, Y.-A., Yu, Y., Bradski, G., Ng, A. Y., & Olukotun, K. (2007). Map-reduce for machine learning on multicore. *Advances in Neural Information Processing Systems 19*, 281–288. https://doi.org/10.7551/mitpress/7503.003.0040
10. Khanzode, Ku. C. A., & Sarode, R. D. (2020). ADVANTAGES AND DISADVANTAGES OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING: A LITERATURE REVIEW. *International Journal of Library and Information Science (IJLIS)*, *9*(1), 30–36. https://doi.org/https://doi.org/10.17605/OSF.IO/GV5T4

Citation: Daniel Han, "Cardiovascular Disease Predictive Modeling with Machine Learning Feature Importance", American Research Journal of Cardiovascular Diseases, Vol 5, no. 1, 2024, pp. 01-06.