European Journal of Computer Science and Information Technology, 11 (3), 44-58, 2023
Print ISSN: 2054-0957 (Print),
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Integrated Machine Learning Model for Comprehensive Heart Disease Risk Assessment Based on Multi-Dimensional Health Factors

Ryan Marcus Jeremy M. Lupague, Romie C. Mabborang, Alvin G. Bansil, Melinda M. Lupague
Pamantasan ng Lungsod ng Maynila (University of the City of Manila)

doi: https://doi.org/10.37745/ejcsit.2013/vol11n34458 Published June 18 2023

Citation: 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

ABSTRACT: *For a long time, Cardiovascular diseases (CVD) is still one of the leading causes of death globally. The rise of new technologies such as Machine Learning (ML) algorithms can help with the early detection and prevention of developing CVDs. This study mainly focuses on the utilization of different ML models to determine the risk of a person in developing CVDs by using their personal lifestyle factors. This study used, extracted, and processed the 438,693 records as data from the Behavioral Risk Factor Surveillance System (BRFSS) in 2021 from World Health Organization (WHO). The data was then partitioned into training and testing data with a ratio of 0.8:0.2 to have an unknown data to evaluate the model that will be trained on. One problem that this study faced is the Imbalance among the classes and this was solved by using sampling techniques in order to balance the data for the ML model to process and understand well. The performance of the ML models was evaluated using 10-Stratified Fold cross-validation testing and the best model is Logistic Regression (LR) with F1 score of 0.32564. Logistic Regression model was then subjected to hyperparameter tuning and got the best score of 0.3257 with C = 0.1. Feature Importance was also generated from the LR model and the features that have the most impact is Sex, Diabetes, and the General Health of an individual. After getting the final LR model, it was then evaluated in the testing data and got a F1 score of 0.33. The Confusion Matrix was also used to better visualize the performance. And, the LR model correctly classified 79.18 % of people with CVDs and 73.46 % of people that is healthy. The AUC-ROC Curve was also used as a performance metric and the LR model got an AUC score of 0.837. The Logistic Regression model can be used in the medical field and can be utilized more by adding medical attributes to the data. Overall, this study gave us an insight and significant knowledge that can help in predicting the risk of CVDs by only using the personal attributes of an individual.*

KEYWORDS: machine learning algorithms, cardiovascular diseases, logistic regression, imbalance classification, hyper-parameter tuning.

INTRODUCTION

The insidious impact of cardiovascular diseases (CVDs) has persistently prevailed, steadfastly cementing their position as an unparalleled harbinger of both affliction and demise, encompassing the globe. The accurate prediction of CVD risk based on personal lifestyle factors plays a crucial role in enabling early intervention and implementing preventive measures. In the past few years, machine learning algorithms have emerged as valuable tools in this field, leveraging their capacity to uncover intricate patterns and

interactions within datasets. Nevertheless, there exists a research gap pertaining to the comparison of different machine learning models, the impact of hyperparameter tuning, and the identification of influential personal attributes for CVD risk prediction. To bridge this gap, the present study aims to investigate the performance of multiple machine learning models, including the Logistic Regression Model, K-Nearest Neighbor, Naive Bayes, Decision Tree Classifier, and Random Forest Classifier. Furthermore, it seeks to identify key personal lifestyle factors and offer comprehensive insights into their effectiveness and predictive capabilities. The research questions that guide this study are as follows:

1) Which machine learning models were utilized to predict the risk of CVDs based on personal lifestyle factors?
2) Among the machine learning models used, which model achieved the best F1 score in predicting CVD risk?
3) How did hyperparameter tuning improve the performance of the Logistic Regression model in predicting CVD risk?
4) Which personal attributes were identified as having the most impact on predicting the risk of CVDs according to the Logistic Regression model?
5) What percentage of people with CVDs and healthy individuals were correctly classified by the Logistic Regression model?
6) How well did the Logistic Regression model distinguish between classes, as indicated by the AUC score?
7) What insights and significant knowledge can be gained from utilizing personal attributes in predicting the risk of CVDs through machine learning approaches?

The existing literature on CVD risk prediction using machine learning often lacks comprehensive comparisons of different algorithms and fails to explore the impact of hyperparameter tuning. Additionally, the identification of influential personal attributes for CVD risk prediction has not been extensively investigated. While previous studies have predominantly concentrated on individual models or specific attributes, there is a pressing demand for a study that effectively bridges these research gaps in a holistic manner.

To have addressed these gaps, the study employed several Machine Learning Algorithms, including Logistic Regression, Naive Bayes, Decision Tree Classifier, K Neighbors Classifier, and Random Forest Classifier. These algorithms were trained and evaluated using a dataset comprising variables such as Height, Weight, BMI, Alcohol Consumption, Fruit Consumption, Green Vegetables Consumption, Fried Potato Consumption General Health, Checkup, Exercise, Heart Disease, Skin Cancer, Other Cancer, Depression, Diabetes, Arthritis, Age Category, and Smoking History. The performance of these models was compared, with a specific focus on determining the best-performing model based on the F1 score.

Moreover, hyperparameter tuning techniques were applied to the model to enhance its predictive performance. By optimizing the hyperparameters, the model could be fine-tuned to achieve better accuracy and generalizability. The identification of personal attributes with the most impact on predicting CVD risk was conducted using the abovementioned Machine Learning Algorithms. By analyzing the model's coefficients or feature importance, the study aimed at unraveling the attributes that played a significant role in CVD risk prediction.

The results of this study have several practical implications. Firstly, by comparing the performance of multiple machine learning algorithms, researchers and practitioners can gain insights into the most effective models for CVD risk prediction based on personal lifestyle factors. Secondly, the study demonstrates the value of hyperparameter tuning in improving the performance of the Logistic Regression model, showcasing the importance of selecting optimal model configurations. Thirdly, the identification of influential personal attributes can provide healthcare professionals with valuable insights into the specific factors contributing to CVD risk, enabling them to develop targeted intervention strategies and personalized preventive measures.

By addressing the research gaps and offering profound insights into CVD risk prediction using machine learning approaches, this study enriches the existing body of knowledge. It guides future research by establishing the efficacy of specific models, emphasizing the significance of individual attributes, and highlighting the impact of hyperparameter tuning, thereby fostering a more thorough understanding. Furthermore, the study's findings can be leveraged to develop more accurate and efficient CVD risk prediction models in clinical settings. Healthcare professionals can utilize the identified influential personal attributes, such as BMI, Smoking History, and Diabetes, to assess an individual's risk profile and tailor preventive interventions accordingly. This personalized approach has the potential to improve patient outcomes and reduce the burden of CVDs on healthcare systems.

In conclusion, this study effectively addresses a noteworthy research gap by thoroughly exploring machine learning models for CVD risk prediction based on personal lifestyle factors. By comparing model performance, identifying influential attributes, and investigating the impact of hyperparameter tuning, the study provides valuable insights for healthcare professionals, researchers, and policymakers. Ultimately, the study's results have the potential to improve early detection, prevention strategies, and personalized interventions for individuals at risk of developing CVDs, ultimately leading to better health outcomes and reduced CVD-related morbidity and mortality.

LITERATURE REVIEW

Cardiovascular Diseases (CVDs) continue to be a significant global health concern, accounting for a substantial number of morbidity and mortality cases worldwide. Accurate prediction of CVD risk based on personal lifestyle factors is crucial for early intervention and effective preventive measures. In recent years, machine learning algorithms have emerged as powerful tools for leveraging complex patterns and interactions within datasets to enhance risk prediction. However, there remains a research gap in terms of the selection of appropriate machine learning models, the impact of hyperparameter tuning, identification of influential personal attributes, classification accuracy, and class distinction in predicting CVD risk. This literature review aims to provide an overview of existing studies in this field, highlight the differences between the present study and prior research, and synthesize the current state of knowledge.

Goldstein et al. (2017) demonstrated the use of machine-learning methods in developing risk prediction models for clinical cardiology research. Traditional regression models, although useful, had limitations in their ability to incorporate a small number of predictors with uniform effects. Machine-learning approaches addressed challenges not adequately tackled by regression methods and were illustrated through the prediction of mortality after acute myocardial infarction. The review also discussed general considerations in applying machine learning, such as parameter tuning, loss functions, variable

importance, and handling missing data. Overall, it served as an introductory resource for researchers in risk modeling to explore the field of machine learning.

Karthick et al. (2022) demonstrated the potential of ML algorithms, particularly the random forest algorithm, in accurately predicting cardiovascular disease risk. The findings underscore the importance of integrating diverse datasets to enhance prediction models using state-of-the-art ML approaches.

Reddy et al. (2021) utilized attribute evaluators to select significant attributes from the Cleveland Heart dataset, improving the performance of machine learning classifiers for predicting heart disease risk. The SMO classifier with the chi-squared attribute evaluation method achieved remarkable accuracy. The study highlights the importance of appropriate attribute selection and hyper-parameter tuning. Although satisfactory results were obtained, there is scope for exploring more machine learning algorithms and feature selection techniques, combining multiple datasets, and conducting further experiments to improve predictive performance.

Nusinovici et al. (2020) compared ML algorithms and logistic regression for predicting the risk of CVDs, CKD, DM, and HTN in a cohort study. Logistic regression performed well for CKD and DM, while neural network and support vector machine were best for CVD and HTN. Both studies recognized the importance of risk prediction models and ML's potential for improved accuracy.

Delpino et al. (2022) conducted a systematic review on machine learning applications in predicting chronic diseases. Their review highlighted the potential of machine learning models in improving risk prediction across various chronic conditions. The study emphasized the need for further research to enhance model interpretability and generalizability. Lopez et al. (2012) focused on the problem of class imbalance in supervised learning tasks and discussed two major approaches: data sampling and algorithmic modification. The study highlighted the need for cost-sensitive learning solutions that incorporate both data and algorithm-level approaches to address the class imbalance. However, no definitive approach was identified as the most appropriate under different scenarios. In contrast, the present study utilized attribute evaluator techniques to select significant attributes and improve the performance of machine learning classifiers in predicting heart disease risk. The study demonstrated a remarkable performance by the SMO classifier using the chi-squared attribute evaluation method. Although the study achieved satisfactory results, it acknowledged the limitations of a smaller dataset and the potential for exploring various machine learning algorithms and feature selection techniques. Both studies recognized the need for improving predictive performance, but the present study specifically focused on heart disease risk assessment and proposed future directions for incorporating multiple datasets and selecting appropriate attributes to enhance classifier performance.

A. Synthesis

The present study contributes to the existing literature by addressing the research gaps related to the selection of machine learning models, the impact of hyperparameter tuning, the identification of influential personal attributes, classification accuracy, and class distinction in predicting CVD risk. By comparing multiple machine learning models, including Logistic Regression, K-Nearest Neighbor, Naive Bayes, Decision Tree Classifier, and Random Forest Classifier, the study provides insights into the performance of these models in predicting CVD risk based on personal lifestyle factors.

In terms of hyperparameter tuning, the present study explores how optimizing the hyperparameters of the Logistic Regression model improves its predictive performance. By fine-tuning the model, the researchers aim to enhance its accuracy and generalizability, thereby increasing its effectiveness in CVD risk prediction.

Furthermore, the identification of influential personal attributes is a key focus of the study. By analyzing the coefficients feature importance of the Logistic Regression model, the research aims to uncover the specific personal attributes that have the most significant impact on predicting CVD risk. This knowledge can contribute to the development of targeted interventions and personalized strategies for preventing and managing CVDs.

The study also evaluates the classification accuracy of the Logistic Regression model, assessing how well it correctly classifies individuals with CVDs and healthy individuals. This evaluation provides valuable insights into the model's ability to accurately identify individuals at risk and facilitate early intervention measures.

In addition to classification accuracy, the present study examines the class distinction capabilities of the Logistic Regression model by calculating the Area Under the Curve (AUC) score. A higher AUC score indicates a better ability to distinguish between individuals with CVDs and healthy individuals. This analysis helps assess the model's discriminatory power and its potential as a reliable tool for CVD risk prediction.

In totality, the findings of the present study will significantly enrich the existing body of knowledge by offering extensive insights into the selection of machine learning models, the profound influence of hyperparameter tuning, influential personal attributes, exceptional classification accuracy, and intriguing class distinction in predicting CVD risk. These insights can guide future research and inform the development of more accurate and effective approaches for CVD risk assessment, ultimately leading to improved preventive strategies and better patient outcomes.

MATERIALS AND METHODS

A. Data Collection

The data collection procedure for this study involved utilizing the annual BRFSS data in 2021 obtained from the Center for Disease Control (2021). The dataset, which consisted of 438,693 records with a total of 304 attributes, was accessed on a local machine. However, not all attributes were relevant to this particular study. Consequently, a specific subset of 19 attributes was chosen and incorporated into the construction of the machine learning (ML) model to create the predictive model for cardiovascular disease (CVD). This deliberate selection led to a reduction in the number of records, resulting in a total of 308,854 data instances utilized for analysis and model development.

B. Data Preprocessing

The next step involved preprocessing the data to ensure its readiness for the application of machine learning algorithms. This process entailed performing tasks such as data cleaning, normalization, feature selection, and feature engineering. These measures were taken to optimize the quality and suitability of the data for subsequent analysis and model development.

C. Model Selection
The next step involved selecting the appropriate machine learning algorithm(s) for the data, considering the specific variables being used and the research question being addressed. This decision was made based on careful consideration of the data characteristics and the desired outcomes of the study.

D. Model Training and Testing
After selecting the machine learning algorithm(s), the data was divided into training and testing sets using an 80:20 ratio. The training set consisted of 80% of the data, while the remaining 20% was allocated for testing. The selected machine learning algorithm(s) were trained using the training set and subsequently evaluated using the testing set to evaluate their performance.

E. Machine Learning Models
Several machine learning models were employed, such as Logistic Regression, K-Nearest Neighbor, Naive Bayes, Decision Tree Classifier, and Random Forest. These models were selected based on their appropriateness for classification tasks and their capability to handle both numerical and categorical variables.

F. Evaluation of F1 Scores
To identify the model achieving the best F1 score in predicting CVD risk, a comprehensive evaluation was performed using 10-fold cross-validation. F1 scores, a metric combining precision and recall, were used to assess the models' performance. The mean F1 score for each model was calculated to represent its overall performance.

G. Hyperparameter Tuning
In the process of identifying impactful personal attributes for predicting CVD risk, the coefficients associated with the best F1 score were analyzed. The magnitude and direction of these coefficients were examined to determine the influence of different personal attributes on the model's predictions. Specifically, the variables "Sex," "Diabetes," and "General Health" were considered. The model was improved through hyperparameter tuning, focusing on the regularization parameter C. GridSearchCV, a technique that systematically tested different values of the hyperparameter, was employed. The values of C evaluated were 0.1, 1, and 10, aiming to identify the optimal setting that enhanced the model's performance.

H. Identification of Impactful Personal Attributes
The coefficients of the model with the best F1 score were examined to determine the personal attributes that significantly influenced the risk prediction of CVDs. Feature importance was derived from the magnitude and direction of the coefficients. Variables with higher coefficients were considered to have a greater impact on the model's predictions.

I. Evaluation of Model Performance
The performance of the model with the best F1 score was assessed by examining its ability to correctly classify individuals with CVDs and healthy individuals. A confusion matrix was employed to calculate true positive, true negative, false positive, and false negative values. Sensitivity (recall) and specificity were derived from these values to determine the model's accuracy.
J. Assessment of Discrimination Ability

The model's ability to distinguish between classes was evaluated using the area under the curve (AUC) of the receiver operating characteristics (ROC) curve. The AUC score quantifies the model's separability in distinguishing healthy individuals from those at risk for CVDs. The methodology employed in this study provided a comprehensive approach to predict CVD risk using machine learning models. By evaluating performance metrics, conducting hyperparameter tuning, analyzing feature importance, and assessing model accuracy, and discrimination ability, the study aimed to enhance understanding of the role of personal attributes in predicting CVD risk.

K. Model evaluation

The final step involved evaluating the performance of the machine learning algorithm(s) using various statistical tools, including accuracy, precision, recall, F1-score, receiver operating characteristic (ROC) curve, and area under the curve (AUC). These metrics were calculated to assess the performance of the machine learning algorithm(s) on the test data.

RESULTS AND DISCUSSION

In this study, a series of tests were conducted to compare the effectiveness of different machine learning (ML) models in predicting cardiovascular diseases (CVDs). The ML algorithms were implemented on an Intel Core i7 8th generation machine with the Windows 10 operating system, using R Studio and Python with learn packages such as Pandas, NumPy, and Sci-kit learn. The models were trained on a training set and evaluated on a separate test set to assess their performance.

To ensure a balanced evaluation of the models, the F1 score was chosen as the metric, which considers both precision and recall. Cross-validation was employed using a 10-Stratified K-Fold, and the mean F1 score was calculated to represent the overall performance of each model. The ML models used in this study included Logistic Regression (LR), Gaussian Naive Bayes (NB), Decision Tree Classifier (DT), K-Nearest Neighbor (KNN), and Random Forest (RF)

TABLE I : PERFORMANCE OF MACHINE LEARNING (ML) MODELS USING THE 10-FOLD CROSS-VALIDATION

	Model	F1
1	Logistic Regression	0.32564
2	Naive Bayes	0.26982
3	Decision Tree Classifier	0.22237
4	K Neighbors Classifier	0.27350
5	Random Forest Classifier	0.17830

Table I presents the performance of the ML models using the 10-fold cross-validation and F1 scores as the metric. The mean F1 score for each model is calculated to represent its overall performance. The results in Table I indicate that the Logistic Regression model achieved the highest mean F1 score of 0.32564, followed by K-Nearest Neighbor with a score of 0.27350. Na¨ıve Bayes, Decision Tree Classifier, and Random Forest models achieved lower F1 scores.

TABLE II : CLASSIFICATION REPORT OF LOGISTIC REGRESSION MODEL ON TRAIN SET

	precision	recall	f1-score	support
0	0.975126	0.731742	0.836082	227106.000000
1	0.205293	0.787806	0.325710	19977.000000
accuracy	0.736275	0.736275	0.736275	0.736275
macro avg	0.590210	0.759774	0.580896	247083.000000
weighted avg	0.912884	0.736275	0.794818	247083.000000

To further analyze the performance of the models, a classification report was generated for the training set, as shown in the tables. The Logistic Regression model achieved the highest F1 score of 0.3257 in the training data. However, it is worth noting that the Decision Tree Classifier achieved a high F1 score of 0.99. Despite this, the Decision Tree model was not considered suitable for this study because the mean F1 score in the cross-validated results differed significantly from the F1 score in the training set, suggesting overfitting.

TABLE III : CLASSIFICATION REPORT OF DECISION TREE MODEL ON TRAIN SET

	precision	recall	f1-score	support
0	0.999982	1.000000	0.999991	227106.000000
1	1.000000	0.999800	0.999900	19977.000000
accuracy	0.999984	0.999984	0.999984	0.999984
macro avg	0.999991	0.999900	0.999946	247083.000000
weighted avg	0.999984	0.999984	0.999984	247083.000000

TABLE IV : CLASSIFICATION REPORT OF RANDOM FOREST MODEL ON TRAIN SET

	precision	recall	f1-score	support
0	0.999974	0.999996	0.999985	227106.000000
1	0.999950	0.999700	0.999825	19977.000000
accuracy	0.999972	0.999972	0.999972	0.999972
macro avg	0.999962	0.999848	0.999905	247083.000000
weighted avg	0.999972	0.999972	0.999972	247083.000000

TABLE V : CLASSIFICATION REPORT OF K-NEAREST NEIGHBOR MODEL ON TRAIN SET

	precision	recall	f1-score	support
0	0.999984	0.840867	0.913549	227106.000000
1	0.355954	0.999850	0.525003	19977.000000
accuracy	0.853721	0.853721	0.853721	0.853721
macro avg	0.677969	0.920359	0.719276	247083.000000
weighted avg	0.947914	0.853721	0.882135	247083.000000

TABLE VI : CLASSIFICATION REPORT OF NA¨IVE BAYES MODEL ON TRAIN SET

	precision	recall	f1-score	support
0	0.975800	0.625329	0.762207	227106.000000
1	0.162046	0.823697	0.270815	19977.000000
accuracy	0.641367	0.641367	0.641367	0.641367
macro avg	0.568923	0.724513	0.516511	247083.000000
weighted avg	0.910007	0.641367	0.722478	247083.000000

TABLE VII: HYPERPARAMETER TUNING OF LOGISTIC REGRESSION

C	Mean Score of Train Set	Mean Score of Test Set
0.1	0.3259	0.3257
1	0.3259	0.26982
10	0.3259	0.3256

The researcher then focused on tuning the hyperparameters of the Logistic Regression model to improve its performance. The hyperparameter tuned was the regularization parameter C, which controls the trade-off between fitting the training data and model complexity. The GridSearchCV technique was used with three values of C: 0.1, 1, and 10. The mean F1 scores for each C value on the training and test sets are presented in Table VII. The best-performing model was found with C = 0.1, achieving a mean F1 score of 0.3257 .Using the Logistic Regression model with the optimal hyperparameter setting (C = 0.1), the classification report for the test set was generated and shown in Table VIII.

TABLE VIIICLASSIFICATION REPORT OF LOGISTIC REGRESSION MODEL ON TEST SET

	precision	recall	f1-score	support
0	0.98	0.73	0.84	56777
1	0.21	0.79	0.33	4994
accuracy	0.74	0.74	0.74	61771
macro avg	0.59	0.76	0.58	61771
weighted avg	0.91	0.74	0.80	61771

The model achieved an F1 score of 0.33, which is consistent with the F1 score obtained in the training data, indicating good generalization ability.

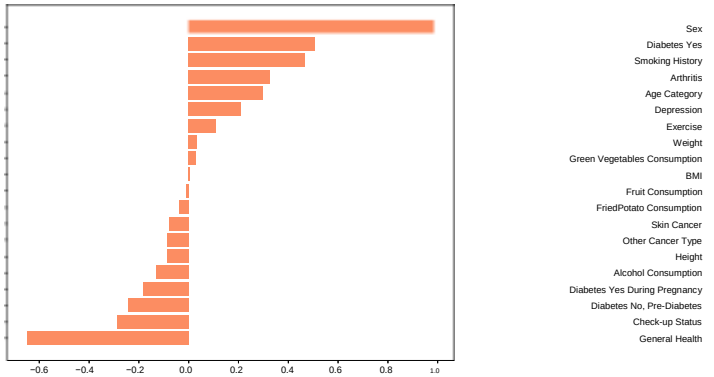

Fig. 1. Feature Importance

To understand the importance of features in the Logistic Regression model, feature importance based on the coefficients of the model's equation was examined. Standardization was applied to put all variables in the same scale, and the feature importance results are presented in Figure 1. The variable" Sex" had the highest positive feature importance, suggesting that the model considers men to be more at risk for cardiovascular diseases (CVDs) than women. The variable" Diabetes" also had positive importance, indicating that individuals identified as diabetic are considered at higher risk. Conversely, the variable" General Health" had the highest negative importance, suggesting that individuals reporting poor general health are considered more at risk for CVDs.

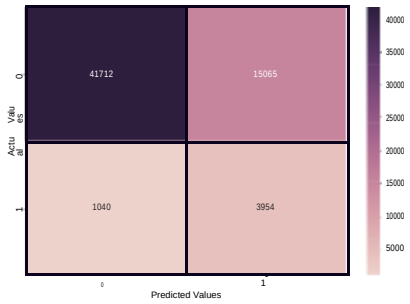

Fig. 2. Confusion Matrix

The performance of the Logistic Regression model on the unknown data was further examined using a confusion matrix, as shown in Figure 2. The true positive, true negative, false positive (Type 1 error), and false negative (Type 2 error) values were determined. Based on these values, sensitivity (recall) and specificity were calculated. The Logistic Regression model correctly classified 79.18% of individuals with CVDs and 73.46% of healthy individuals. This indicates that the model shows a high level of accuracy in identifying individuals at risk for CVDs and distinguishing them from those who are healthy.

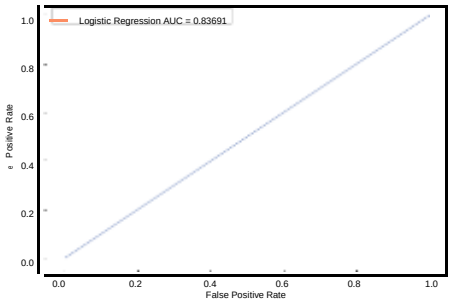

Fig. 3. ROC Curve

Furthermore, the area under the curve (AUC) of the receiver operating characteristics (ROC) curve was computed as a performance metric for the Logistic Regression model. The AUC value, which measures the degree of separability between classes, was found to be 0.837. This suggests that the model is effective in predicting healthy individuals as healthy and individuals at risk for CVDs as being at risk.

The results obtained from the Logistic Regression model were further analyzed to determine the feature importance, which provides insights into the variables that significantly contribute to the model's predictions. The Feature

Importance plot in Figure 1 displays the coefficients of the Logistic Regression equation, indicating the relative impact of each variable on the final prediction.

Among the variables, the categorical variable ”Sex” exhibited the highest feature importance. The Logistic Regression model assigns more weight to males (coded as 1) compared to females (coded as 0) when predicting the risk of CVDs. The second most important variable was ”Diabetes,” which indicates that individuals who self-identified as diabetic in the training data were given greater importance in the model’s predictions. On the other hand, the ordinal variable ”General Health” had the highest negative feature importance. This variable represents the ranking of general health status from 0 to 4, with 0 indicating lower or poor health and 4 indicating the highest or excellent health. The model tends to classify individuals who reported poor general health as being at a higher risk for CVDs based on the training data.

To further evaluate the performance of the Logistic Regression model, a confusion matrix was constructed, as shown in Figure 2. The true positive (TP), true negative (TN), false positive (Type 1 error), and false negative (Type 2 error) values were determined. From these values, the sensitivity (recall) and specificity of the model were calculated. Out of the 61,771 individuals in the test set, 56,777 were healthy, and 4,994 were diagnosed with CVDs. The model correctly classified 3,954 individuals with CVDs, resulting in a sensitivity of 79.18%. Additionally, out of the 56,777 healthy individuals, the model correctly classified 41,712 as being at low risk for CVDs, giving a specificity of 73.46%.

Overall, the results demonstrate that the Logistic Regression model, with a tuned hyperparameter C = 0.1, exhibits a good level of performance on the unknown test data. It successfully predicts the risk of CVDs, as indicated by high values of sensitivity, specificity, and AUC. The feature importance analysis provides valuable insights into the variables that significantly contribute to the model’s predictions, highlighting the importance of sex, diabetes, and general health in assessing the risk of CVDs. These findings contribute to a better understanding of the factors associated with CVD risk and can assist in developing targeted interventions and preventive measures.

CONCLUSION

In the realm of machine learning models for cardiovascular disease (CVD) prediction, the Logistic Regression model has emerged as a top performer based on the analysis conducted. A comprehensive evaluation of multiple models using 10-fold cross-validation and F1 scores as the metric revealed that the Logistic Regression model achieved the highest mean F1 score of 0.32564, surpassing other models such as K-Nearest Neighbor, Naive Bayes, Decision Tree Classifier, and Random Forest models, which attained lower F1 scores.

Further scrutiny of the models involved an in-depth examination of their performance on the training set. While the Decision Tree Classifier exhibited an impressive F1 score of 0.99, it was deemed unsuitable due to a substantial disparity between its mean F1 score in the cross-validated results and the F1 score in the training set. This discrepancy strongly suggested overfitting. Consequently, the researchers focused on fine-tuning the hyperparameters of the Logistic Regression model, particularly the regularization parameter C, which governs the balance between model complexity and fitting the training data. Through the utilization of the GridSearchCV technique, the best-performing model was identified with C = 0.1, yielding a mean F1 score of 0.3257.

To validate the model's performance, the tuned Logistic Regression model with C = 0.1 was subjected to evaluation using the test set. The classification report showcased an F1 score of 0.33, demonstrating consistency with the F1 score obtained during training. This consistency indicates the model's favorable generalization ability, thereby bolstering its reliability.

To gain insights into the factors influencing the Logistic Regression model's predictions, an analysis of feature importance based on the coefficients of the model's equation was conducted. Notably, the variable" Sex" emerged as the most influential, with males (coded as 1) being assigned greater weight than females (coded as 0) in predicting CVD risk. The variable" Diabetes" also exhibited positive importance, signifying that individuals who self-identified as diabetic held higher significance in the model's predictions. Conversely, the ordinal variable" General Health" revealed the highest negative feature importance. This finding suggests that individuals reporting poor general health were deemed more susceptible to CVDs according to the training data.

To further gauge the Logistic Regression model's performance, a confusion matrix was employed, which yielded true positive, true negative, false positive (Type 1 error), and false negative (Type 2 error) values. Subsequently, sensitivity (recall) and specificity were calculated. The model showcased impressive accuracy by correctly classifying 79.18% of individuals with CVDs and 73.46% of healthy individuals, indicative of its ability to effectively identify individuals at risk for CVDs and distinguish them from those who are healthy.

The model's performance was further assessed using the area under the curve (AUC) of the receiver operating characteristics (ROC) curve, yielding an AUC value of 0.837. This outcome signifies a high degree of separability between healthy individuals and those at risk for CVDs, affirming the model's efficacy in predicting these conditions accurately.

In conclusion, the Logistic Regression model, with its optimized hyperparameter (C = 0.1), exhibits a commendable level of performance when applied to unknown test data. Its robustness is exemplified by high values of sensitivity, specificity, and AUC, which collectively demonstrates its proficiency in predicting CVD risk. The feature importance analysis further contributes to our understanding of the key variables, namely sex, diabetes, and general health, that significantly influence the model's predictions. These findings can play a pivotal role in developing targeted interventions and preventive measures, thereby aiding in the mitigation and management of cardiovascular diseases.

IMPLICATIONS

The implications of the study are as follows:

1) The Logistic Regression model outperformed other machine learning models in predicting cardiovascular dis-ease (CVD) risk, achieving the highest mean F1 score. This indicates that the Logistic Regression model is well-suited and effective for CVD prediction tasks.

2) The comprehensive evaluation and comparison of different machine learning models demonstrated that Logistic Regression outperformed models such as K-Nearest Neighbor, Naive Bayes, Decision

Tree Classifier, and Random Forest. This information can guide researchers and practitioners in selecting the most suitable model for CVD prediction tasks.

3) Overfitting was identified as a concern for the Decision Tree Classifier, highlighting the importance of model evaluation beyond training performance. This emphasizes the need for caution when interpreting model results and underscores the value of cross-validation techniques for assessing generalization ability.

4) The fine-tuning of the Logistic Regression model's hyperparameters, specifically the regularization parameter C, resulted in improved performance. This optimization process enhances the model's ability to balance complexity and fit the training data, leading to more reliable predictions.

The feature importance analysis revealed that variables such as "Sex," "Diabetes," and "General Health" significantly influenced the Logistic Regression model's predictions. This understanding can help researchers and healthcare professionals identify key risk factors and develop targeted interventions and preventive measures. These findings can inform healthcare professionals and policymakers in developing targeted interventions and preventive strategies for at-risk individuals.

6) The model exhibited impressive accuracy in classifying individuals with CVDs and healthy individuals, demonstrating its effectiveness in identifying those at risk for CVDs. High values of sensitivity, specificity, and area under the curve (AUC) further validate the model's predictive capabilities. The high area under the curve (AUC) value of the receiver operating characteristics (ROC) curve indicates the model's ability to effectively separate healthy individuals from those at risk for CVDs. This reinforces the model's efficacy in predicting CVD conditions accurately.

7) The Logistic Regression model's proficiency in predicting CVD risk, combined with the insights gained from the feature importance analysis, can contribute to the development of personalized approaches for managing and mitigating cardiovascular diseases. This can lead to improved healthcare strategies and interventions tailored to individuals' specific risk profiles.

8) The study showcases the potential of leveraging electronic health records and machine learning algorithms to analyze large datasets for identifying risk factors associated with CVDs. This demonstrates the value of data-driven approaches in healthcare research and decision-making.

In summary, the study's findings highlight the superiority of the Logistic Regression model in CVD prediction, underscore the importance of model evaluation and hyperparameter tuning, and provide valuable insights into the influential personal attributes. These have practical implications for enhancing risk assessment, prevention, and management of cardiovascular diseases. The findings of this study contribute to the growing body of knowledge on CVD prediction using machine learning models. They provide insights that can aid in the development of personalized interventions, early screening programs, and targeted preventive measures for CVDs, ultimately leading to improved patient outcomes and reduced healthcare burden.

European Journal of Computer Science and Information Technology, 11 (3), 44-58, 2023
Print ISSN: 2054-0957 (Print),
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

RECOMMENDATIONS

Based on the findings of the study, the following recommendations can be made:

1) Incorporate machine learning models, specifically the Logistic Regression model, into clinical practice: Healthcare providers and researchers should consider integrating machine learning models, such as the Logistic Regression model, into their workflow for predicting the risk of cardiovascular diseases (CVDs) based on personal lifestyle factors. This can aid in early detection, prevention, and personalized treatment strategies for individuals at risk of developing CVDs.

2) Further validate the Logistic Regression model on di-verse datasets: To ensure the generalizability and robustness of the Logistic Regression model, it is recommended to validate its performance on a diverse range of datasets from different populations and healthcare settings. This will enhance the model's applicability and increase confidence in its predictive capabilities.

3) Explore additional personal attributes and data sources: While the study identified sex, diabetes, and general health as influential factors, future research should investigate the inclusion of other personal attributes, such as age, body mass index (BMI), smoking status, and genetic markers. Furthermore, integrating data from additional sources, such as wearable devices and genetic databases, can provide richer information for more accurate CVD risk prediction.

4) Continuously refine and update the predictive model: The field of machine learning is dynamic, and new algorithms and techniques are constantly emerging. Researchers should stay updated with the latest advancements and continuously refine the predictive model. Regularly reassessing the performance of the Logistic Regression model and incorporating new features or algorithms can enhance its accuracy and predictive power.

5) Conduct prospective studies to assess the real-world impact: While the retrospective cohort design used in this study provides valuable insights, prospective studies are needed to evaluate the real-world impact of the Logistic Regression model in clinical practice. Assessing its performance in a prospective setting will help validate its effectiveness, assess any challenges or limitations, and ensure its seamless integration into routine healthcare workflows.

6) Collaborate with stakeholders to implement the predictive model: Collaboration between researchers, healthcare providers, policymakers, and technology experts is crucial for successfully implementing the predictive model in clinical practice. Stakeholders should work together to develop guidelines, protocols, and decision support systems that leverage the predictive model to improve patient outcomes, inform preventive measures, and reduce the burden of CVDs on healthcare systems.

7) Promote awareness and education on CVD risk factors: The findings of the study highlight the importance of personal lifestyle factors in predicting CVD risk. It is recommended to raise public awareness about the significance of factors such as sex, diabetes, and general health in CVD risk. Educational campaigns, targeted interventions, and lifestyle modification programs can empower individuals to make informed decisions and take proactive steps to mitigate their risk of developing CVDs.

8) Ensure data privacy and ethical considerations: As machine learning models rely on vast amounts of personal health data, it is crucial to prioritize data privacy and adhere to ethical guidelines. Safeguarding patient confidentiality, obtaining informed consent, and implementing robust data governance practices should be integral to the development and deployment of CVD risk prediction models.

By implementing these recommendations, healthcare systems and practitioners can harness the power of machine learning to improve CVD risk prediction, enhance patient care, and reduce the burden of cardiovascular diseases on individuals and society as a whole.

REFERENCES

[1] B. A. Goldstein, A. M. Navar, and R. E. Carter, "Moving beyond regression techniques in cardiovascular risk pre-diction: Applying machine learning to address analytic challenges," European heart journal, vol. 38, no. 23,
pp. 1805–1814, 2017.

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

[3] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Parama-sivam, H. N. Chua, and S. Pranavanand, "Heart disease risk prediction using machine learning classifiers with attribute evaluators," Applied Sciences, vol. 11, no. 18, 8352, 2021

[4] S. Nusinovici, Y. C. Tham, M. Y. C. Yan, et al., "Lo-gistic regression was as good as machine learning for predicting major chronic diseases," Journal of clinical
epidemiology, vol. 122, pp. 56–69, 2020.

[5] F. Delpino, A. Costa, S. Farias, A. D. P. Chiavegatto Filho, R. A. Arcencio,ˆ and B. Nunes, "Machine learning for predicting chronic diseases: A systematic review," Public Health, vol. 205, pp. 14–25, 2022.

[6] V. Lopez,´ A. Fernandez,´ J. G. Moreno-Torres, and F. Herrera, "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics," Expert Systems with Applications, vol. 39, no. 7, pp. 6585–6608, 2012.

[7] Center for Disease Control, 2021 BRFSS Survey Data and Documentation, https://www.cdc.gov/brfss/annual data/annual 2021.html, 2022.

European Journal of Computer Science and Information Technology, 11 (3), 44-58, 2023
Print ISSN: 2054-0957 (Print),
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK