EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Data Analysis and Predictive Modelling on Heart Disease based on People's Lifestyle

Edward Leonardo[1*], Dr. Murugananthan Velayutham[2], and Justin Gilbert[3]

[1]School of Computing, Asia Pacific University of Technology & Innovation (APU), Kuala Lumpur, Malaysia
[2]School of Computing, Asia Pacific University of Technology & Innovation (APU), Kuala Lumpur, Malaysia
[3]School of Computing, Asia Pacific University of Technology & Innovation (APU), Kuala Lumpur, Malaysia

Abstract

Coronary Artery Disease (CAD) is a form of heart disease primarily influenced by lifestyle choices. Despite preventative measures available to mitigate CAD risks, a significant proportion of the population remains unaware of its severity and consequently neglects necessary precautions. As a result, the influence of CAD continues to rise. This project aims to curb CAD cases by developing an early warning detection and educational accessible to the general population, leveraging Machine Learning and Data Visualization technologies. Research indicates that while Coronary Artery Disease can be mitigated through a lifestyle shift towards healthier living, the risk remains due to factors such as age and natural health deterioration.

Keywords: Coronary Artery Disease, Heart Disease, Machine Learning, Predictive Modelling, Data Analysis

Received on 11 02 2024, accepted on 30 05 2024, published on 25 06 2024

doi: 10.4108/eetpht.10.6432

1. Introduction

Heart disease encompasses conditions affecting the heart or blood vessels [1]. In 2019, prior to the COVID-19 pandemic, it was the leading cause of death globally, accounting for 19 million deaths or 34% of all fatalities [2]. Globally, 1 in 14 individuals lives with some form of heart disease, marking a 93% increase from the 1990s to the 2010s [2]. Unfortunately, the number of deaths from heart diseases is projected to continue rising.

Coronary heart disease (CHD), also known as coronary artery disease (CAD), is the most prevalent type of heart disease [3]. It is primarily caused by poor lifestyle choices such as smoking, unhealthy diet, excessive alcohol consumption, obesity, diabetes, and physical inactivity [3]. These unhealthy behaviors are often seen as Personal Key Indicators (PKIs). However, many individuals fail to recognize that these PKIs significantly increase the risk of heart disease, particularly during early adulthood when health may not be a top priority due to the "You Only Live Once" mentality [4]. As people age, they may mistakenly believe it's too late to adopt a healthy lifestyle, despite evidence to the contrary—healthy habits can benefit anyone at any age [5]. Establishing healthy habits is easier in childhood due to greater adaptability compared to adulthood [6]. Thus, increasing awareness of the risks of coronary artery disease across all age groups is crucial to prevent its onset later in life.

Machine Learning (ML) is a branch of Artificial Intelligence (AI) focused on using data and algorithms to enable machines to intelligently solve problems [7]. ML models to some degree mimic how a human would learn, improving accuracy with increased data quantity and quality, without requiring explicit programming for each task [8]. The most common use case of Machine Learning is in making prediction, classification, data mining insights and finding patterns via data mining. In real world application, ML is used in a wide range of applications

*Corresponding author. Email: tp058284@mail.apu.edu.my

which include utilized within recommendation engines, prediction models, spam filtering, malware detection detection [9]. As there are so many different use cases, choosing the correct suitable machine learning algorithm is critical as each model has specific strengths and applications and model performance is dependent on both data quantity and quality. [10]. A robust ML model continually learns from new data to enhance accuracy.

This paper aims to develop a machine learning-based heart disease prediction model to assess individuals' risk based on their Personal Key Indicators (PKIs). It will highlight crucial PKIs to educate the general public about lifestyle choices influencing heart disease risk. The prediction will be based on their Personal Key Indicators (PKIs).

2. Materials and Methods

2.1. Dataset

The dataset used in this paper originates from two distinct sources. The first source is the CDC's 2020 Annual Survey related to American Adults' health status. The dataset initially contains 279 columns/attributes, shortened into 18 columns/attributes with 17 features and 1 target variable. The attributes chosen related solely to general health conditions, such as difficulty in walking, BMI, and age. [11]. The target variable "HeartDisease" is a binary variable indicating whether or not the observed person suffered from coronary artery disease (CAD). The first data set collected a total of 319,795 observations.

The second source involves a survey questionnaire conducted beforehand, to gain additional data which could supplement the main dataset. The data gathered mostly matched the existing data, with some changes to allow easier data collection from participants, including participants' Body Weight and Body Height rather than Body Mass Index (BMI), since BMI can be derived from both. The survey questionnaire was performed using two languages, English Language and Bahasa Indonesia, to allow a higher variance of responses. Combining both versions, the data collected 99 observations across 33 columns/attributes, which was reduced to 19 columns during pre-processing, as the other 14 columns contained non-relevant questionnaire responses for the research process. The target variable of the questionnaire mirrored the target variable of the first source.

The inclusion of general health conditions and lifestyle choices ensured the outcomes, including the developed Machine Learning models, could be accessible to the general population without the need for any professional medical assistance.

2.2. Data Pre-Processing

Before utilizing the data to construct the Machine Learning model, data pre-processing is essential to ensure the data is suitable for Machine Learning applications. The exact steps of data pre-processing can vary widely depending on the initial condition of the data [12]. Through data pre-processing, the data undergoes cleaning, imputation, normalization, transformation, and encoding to optimize its Machine Learning result. In the case of this project, Data Cleaning is divided into two main phases: first, pre-processing of the survey data to assimilate it smoothly into the main data first; second, comprehensive pre-processing of the combined dataset.

First, the pre-processing of the survey data is conducted to allow the survey data to be added seamlessly into the main data. The first step is to fix all issues present in the data; this includes attribute renaming so that attributes match names in the main dataset. The next step involves standardizing inconsistencies such as capitalization, spacing, and units of measurement to ensure uniformity and reduce the variance of unique values in the dataset. Following this, data transformation is applied to transform the Body Weight data and the Body Height data into Body Mass Index (BMI) to calculate if a person's mass is within the healthy zone using the following formula:

$$BMI = \frac{BodyWeight}{BodyHeight^2} / BMI = \frac{kg}{m^2}$$

Once BMI is calculated, both Body Weight and Body Height attributes are discarded. Finally, the survey dataset is rearranged to align its structure with the main dataset, facilitating seamless concatenation.

The second step encompasses comprehensive pre-processing of the entire dataset. Initially, any duplicates entities from inside the dataset are removed to prevent data leakage during model development, where identical data appears in both training and testing datasets [13]. Subsequently, further inconsistency fixing is performed to standardize any remaining outliers or unique variance. The next pre-processing step covers missing data imputation by imputing missing numerical data with the mean (average) and imputing missing categorical data with the mode (most frequent value). Finally, all categorical attributes need to be encoded, to prepare them to fit into the machine learning models during the model building process [14]. For attributes with two unique values, Label Encoder assigns each unique value a numerical label ("0" for the first unique value and "1" for the second). For the attributes with more than two unique values, One Hot Encoding is utilized. One Hot Encoding create a new binary attribute for each unique category. The attribute corresponding to each category is marked with "1", while the others are marked with "0".

2.3. Model Building

Once the data has undergone pre-processing, it is ready for the Machine Learning Model Building phase. Prior to fitting the data into the ML models, several preparation steps are necessary. Firstly, the dataset is separated based

on feature-target split, resulting in two datasets, "x" containing all feature attributes, and "y" containing the target attribute. Following this, the data is further separated into a train-test split, with a 70:30 data distribution ratio, with 70% allocated to training, and 30% to testing. This means that the data will be split into four, "x_train", "x_test", "y_train", "y_test". Next, the data is normalized to allow better optimization of the data, by using a common scale for the numerical data. In this case, Min Max Scaler is used for data normalization. The data normalization is only applied to both "x_train" and "x_test" datasets. Finally, since the target class distribution is heavily imbalanced – 274,551 instances of "0" (Majority) against only 27,261 instances of "1" (Minority) - various data sampling techniques are applied to achieve a balanced class distribution. For testing purposes, three different data sampling techniques are utilized: Oversampling (where the Minority class data is synthetically increased to match the Majority), Undersampling (reducing the Majority class to match the Minority), and Combined Sampling (utilizing both oversampling and undersampling to achieve a balanced dataset).

With the data prepared, the Model building phase commences by evaluating eight different ML algorithms to see which best suits the data. These models are Multi-Layer Perceptron (MLP), Linear SVC (Support Vector Classification), Random Forest, XGBoost, Decision Tree, Ada Boost, K-Nearest Neighbors, and Gradient Boost. These algorithms are chosen based on their prevalence in prior research. Each model is tested with each of the data sampling techniques to determine the optimal combination of ML algorithm and data sampling approach for the dataset.

Once the most suitable ML model is identified, Hyperparameter Tuning (HP Tuning) is performed to fine-tune its parameters for enhanced prediction performance. HP Tuning employs three techniques: Keras Tuner, Randomized Search CV, and Grid Search CV, to identify the best parameters for the chosen model.

3. Results and Discussion

The model performance is evaluated by comparing all models with different data sampling techniques to determine the best configuration for the project's objective.

First, it is essential to explain the performance metrics. In Machine Learning, the model's performance is often measured using a confusion matrix. A confusion matrix is a 2x2 table that displays the outcome of a prediction: True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). It is used to visualize the results of the prediction [15]. In the case of this project, the target variable "HeartDisease" has two unique values, zero (0), indicating a person has a low risk of coronary artery disease (CAD), and one (1), indicating a high risk or a prior diagnosis with coronary artery disease. The confusion matrix evaluates the model's results based on the following:

- True Positive (TP): The prediction outcome is 1, the actual data is 1.
- False Positive (FP): The prediction outcome is 0, the actual data is 1.
- False Negative (FN): The prediction outcome is 1, the actual data is 0.
- True Negative (TN): The prediction outcome is 0, the actual data is 0.

Based on these four results, four performance metrics can be created [16]:

- Accuracy: The number of correct predictions over the total data amount.
- Precision: Positive predictive value, which covers the amount of true positive over the true positive and false positive. Thus, precision value become higher when False Positive decreases.
- Recall: Sensitivity, which covers the amount of true positive over true positive and False Negative. Thus, Recall value become higher when False Negative decreases.
- F-1 Score: Metrics that covers both precision and recall.

All these performance metrics have the formula as below [16]:

- $Accuracy = \frac{TN+TP}{TN+FN+FP+TP}$
- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$
- $F-1\,Score = 2 * \frac{Precision * Recall}{Precision + Recall}$

Since this project's main objective is to be a warning detection tool for detecting coronary artery disease (CAD), Recall is the important metric to be measured. Recall covers the False Negative (FN) situation, where the individual is predicted to be CAD-free despite having a high risk or an existing diagnosis of CAD. A high recall score indicates a lower number of False Negatives, making it a crucial measure for ensuring that at-risk individuals are accurately identified.

The next performance metric to evaluate is the accuracy of the predictions. Accuracy measures how well the Machine Learning model can correctly predict the outcomes. Precision, while important, is less critical in this case. A high false positive rate means the model predicts someone has a high risk of CAD when they do not, but this still prompts individuals to take preventive measures and make lifestyle changes to reduce their risk. Due to the considerations above, Recall and Accuracy are the primary metrics for determining the most suitable Machine Learning model for this project.

model	accuracy_train_cv	accuracy_test	precision_score	recall_score	f1_score
XGBClassifier	75.336702	72.729281	21.863286	78.454390	34.196781
GradientBoostingClassifier	76.154174	73.801688	22.570854	78.197603	35.030539
LinearSVC	76.143699	73.817150	22.539346	77.928589	34.965572
AdaBoostClassifier	76.033656	74.623388	22.924366	76.607973	35.288816
RandomForestClassifier	73.701733	70.849532	20.188531	75.421864	31.851278
MLPClassifier	74.715721	74.077796	22.215763	74.761555	34.253060
KNeighborsClassifier	72.352366	71.895432	20.404456	72.792859	31.874281
DecisionTreeClassifier	66.333930	66.507996	16.457153	66.434336	26.379549

Figure 1. Undersampling Performance's Results

model	accuracy_train_cv	accuracy_test	precision_score	recall_score	f1_score
LinearSVC	76.276504	73.640440	22.404756	77.879677	34.798525
AdaBoostClassifier	82.973697	80.523281	25.759984	61.445341	36.301246
MLPClassifier	82.723678	79.390131	23.486266	56.774272	33.227180
KNeighborsClassifier	86.470588	78.747349	21.687733	51.821961	30.578304
GradientBoostingClassifier	86.283269	85.462317	30.715667	48.544876	37.624982
DecisionTreeClassifier	88.219164	83.245715	21.011609	30.985571	25.042000
RandomForestClassifier	92.199964	86.571170	26.667448	27.818538	27.230834
XGBClassifier	90.992533	88.854038	33.823529	24.468085	28.395062

Figure 2. Oversampling Performance's Results

model	accuracy_train_cv	accuracy_test	precision_score	recall_score	f1_score
LinearSVC	76.108750	73.607307	22.400534	78.001956	34.805620
MLPClassifier	78.520510	72.733699	20.803565	71.924676	32.272578
AdaBoostClassifier	79.591536	78.192923	24.785281	69.515774	36.541861
KNeighborsClassifier	79.787529	73.583009	20.368209	66.153094	31.146550
GradientBoostingClassifier	81.257480	80.496775	26.537986	65.566153	37.783180
RandomForestClassifier	84.467956	80.868970	24.881881	55.380289	34.336619
XGBClassifier	82.727430	84.145830	29.427829	54.022989	38.100987
DecisionTreeClassifier	78.119851	75.811760	18.578559	49.608706	27.033150

Figure 3. Combined Sampling Performance's Results

Based on the results, considering performance metrics like Accuracy, Recall, and the time required for the Model Fitting Process, the Gradient Boosting algorithm with undersampled data emerges as the best ML model for this project.

During the Model Building phase, combining undersampled data using RandomUnderSampler with Gradient Boosting algorithms and tuning using Keras Tuner produced the best results. This approach aligns well with the established criteria for evaluating model performance for this dataset and the paper's objective.

```
Performance Metrics after Keras Tuning for GradientBoostingClassifier
              precision    recall  f1-score   support

           0       0.97      0.73      0.84     82366
           1       0.23      0.78      0.35      8178

    accuracy                           0.74     90544
   macro avg       0.60      0.76      0.59     90544
weighted avg       0.90      0.74      0.79     90544

[[60325 22041]
 [ 1759  6419]]
```

Figure 4. Final Model Performance's Metrics

The above figure shows the results of the Gradient Boosting with Undersampled data tuned using Keras Tuner. The model achieved an the accuracy of 74%, 23% precision, and 78% recall. This indicates that the model correctly classifies the data 74% of the time. The low precision suggests that the model tends to classify individuals as high risk for CAD even when they are actually low risk, which is problematic. However, the relatively high recall means that the model is less likely to produce False Negatives, where individuals with a high risk of CAD are predicted to be low risk. This is crucial given the project's objective to minimize such errors. These results are considered satisfactory, especially given the dataset's challenges, such as low correlation and heavily imbalanced target distributions.

Finally, comparing the project's Machine Learning model results with those from other developers' works can further validate the results. Such a comparison can demonstrate that this project's outcomes meet its objectives and are suitable for the target users. For benchmarking, three other developers' works, posted on Kaggle in the Code section of the main data page, will be used.

Figure 5. Comparison between All Developers' Final Results

Project	Model Name	Accuracy	Precision	Recall	F-1
This Project	Gradient Boosting	0.74	0.23	0.78	0.35
Elsayed (2022)	K-Nearest Neighbors	0.90	0.35	0.14	0.20
	Decision Tree	0.86	0.23	0.25	0.24
Mohaimin (2022)	Random Forest	0.59	0.83	0.22	0.35
Hossen (2023)	Extra tree	0.97	0.94	0.99	0.97

Compared with other developers' results, this project's outcomes show the best balance between Accuracy and Recall. Specifically, compared to Elsayed, the project's Recall score is significantly higher, although the accuracy

was not as high. Compared to Mohaimin, this project's performance metrics were superior across all scores. Finally, compared to Hossen, the performance metrics might not have been as high as Hossen's, but there is a major flaw in Hossen's methodology. Their process likely caused data leakage during data sampling, leading to unrealistically high results. Overall, this project's results are sufficient and valid, especially when benchmarked against other developers' work. While the results achieved were not optimal, they are still viable due to the nature of the project, the chosen Personal Key Indicators (PKIs), and the gathered data.

4. Conclusions

In conclusion, the primary objective of this project was to enhance public awareness about the risk factors influencing Coronary Artery Disease (CAD) based on lifestyle choices. The main outcome is an early warning detection tool that utilizes Machine Learning to predict a user's risk of CAD based on their current lifestyle. The results of the project indicate that the model is valid and viable for deployment and use by the general population.

The project faced limitations primarily related to the dataset and the resulting Machine Learning models. While the results are acceptable and align with the project's objectives, there is room for improvement in accuracy. These issues stem from the dataset's lack of variance and highly imbalanced class distribution. Improved variance and balanced classes in the dataset could lead to better results.

For future steps to progress the method further, the developer could enhance the Machine Learning models by conducting a larger-scale survey with more complex questions that correlate more strongly with CAD.

Overall, the developer feels that satisfactory results have been achieved, with all steps carefully executed to ensure the project's outcomes align with its objectives. Care was taken with the procedure and processes to ensure the successful aim of the project, which was to ensure that such models are useful for the general population, particularly in raising public awareness about the risk of coronary artery disease (CAD) based on lifestyle factors.

Acknowledgements

I would like to extend my deepest gratitude to several parties who have provided invaluable help and support throughout the development of this paper. First and foremost, I would like to thank my supervisor, Dr. Murugananthan Velayutham, for his continuous guidance and support during the whole process, pointing out mistakes or room for improvement. I also would like to express my sincere appreciation for all the time and attention he has dedicated to assisting me in this endeavour. An honourable mention goes to my second marker, Mr. Justin Gilbert A/L Alexius Silverster, who provided some insights that helped my paper. Next, I would like to show my gratitude towards my family who supported my whole undergraduate journey, as without them, I would not have had the chance to experience studying and living overseas. I would also like to express my gratitude to all the lecturers who have guided me during my undergraduate studies, my classmates who worked collaboratively with me during our learning process, and all my friends who have supported me. All the above mentioned have supported me to do my best and become the best version of myself.

References

[1] National Cancer Institute, "NCI Dictionary of Cancer Terms | Heart Disease," National Cancer Institute. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/heart-disease (accessed Feb. 08, 2022).

[2] British Heart Foundation, "Global Heart & Circulatory Diseases [Fact sheet]," British Heart Foundation, Feb. 2023. Accessed: Feb. 08, 2023. [Online]. Available: https://www.bhf.org.uk/-/media/files/research/heart-statistics/bhf-cvd-statistics-global-factsheet.pdf

[3] CDC, "Heart Disease Resources | CDC," Centers for Disease Control and Prevention, Jul. 12, 2022. https://www.cdc.gov/heartdisease/about.htm (accessed Feb. 08, 2023).

[4] S. Zhou, "Knowing the Importance of Being Healthy in your Early 20s," Flexispot, Jun. 21, 2021. https://www.flexispot.ca/spine-care-center/knowing-the-importance-of-being-healthy-in-your-early-20s/ (accessed Feb. 08, 2023).

[5] Johns Hopkins Medicine, "It's Never Too Late: Five Healthy Steps at Any Age," Johns Hopkins Medicine, Nov. 01, 2021. https://www.hopkinsmedicine.org/health/wellness-and-prevention/its-never-too-late-five-healthy-steps-at-any-age (accessed Feb. 08, 2023).

[6] J. Mock, "Is It Ever Too Late to Start Being Healthy?," Discover Magazine, Dec. 26, 2019. https://www.discovermagazine.com/health/is-it-ever-too-late-to-start-being-healthy (accessed Feb. 08, 2023).

[7] IBM, "What is Machine Learning? | IBM," IBM. https://www.ibm.com/my-en/topics/machine-learning (accessed Feb. 08, 2023).

[8] V. Kanade, "What Is Machine Learning? Definition, Types, Applications, and Trends for 2022," Spiceworks, Aug. 30, 2022. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ml/ (accessed Feb. 08, 2023).

[9] E. Burns, "Machine Learning," TechTarget, Mar. 30, 2021. https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML (accessed Feb. 08, 2023).

[10] E. Miah, "Key Factors in The Successful Use of Machine Learning," Data Science Central, Nov. 07, 2017. https://www.datasciencecentral.com/key-factors-in-the-successful-use-of-machine-learning/ (accessed Feb. 08, 2023).

[11] K. Pytlak, "Personal key indicators of heart disease," Kaggle, Feb. 16, 2022. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease (accessed Jul. 18, 2023).

[12] Tableau, “What Is Data Visualization? Definition, Examples, And Learning Resources,” Tableau, 2023. https://www.tableau.com/learn/articles/data-visualization (accessed Jul. 11, 2023).
[13] S. Chorev, “A practical guide to data cleaning,” Deepchecks, Mar. 24, 2023. https://deepchecks.com/what-is-data-cleaning/#:~:text=Duplicate%20entries%20can%20ruin%20the,disappointing%20the%20model%20in%20production. (accessed Jul. 18, 2023).
[14] T. Khan, “Different types of Encoding - AI ML Analytics,” AI ML Analytics, Jan. 02, 2022. https://ai-ml-analytics.com/encoding/ (accessed Jul. 15, 2023).
[15] Simplilearn, “What is a Confusion Matrix in Machine Learning?,” Simplilearn.com, Feb. 2023, [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/confusion-matrix-machine-learning#:~:text=A%20confusion%20matrix%20presents%20a,actual%20values%20of%20a%20classifier.
[16] B. Harikrishnan, “Confusion matrix, accuracy, precision, recall, F1 score,” Medium, Dec. 12, 2021. Accessed: Jul. 18, 2023. [Online]. Available: https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd