Proceedings of the 2nd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/38/20240569

Logistic regression for cardiovascular diseases prediction by integrating PCA and K-means ++

Hancheng Miao

Tandon School of Engineering, New York University, NY, USA

hm3004@nyu.edu

Abstract. This research introduces a novel method for forecasting cardiovascular diseases using an advanced combination of K-means++ clustering, Principal Component Analysis (PCA), and Logistic Regression techniques. Given the global impact of cardiovascular diseases as a primary cause of death, this research utilizes a comprehensive dataset to tackle the prediction challenges associated with CVDs. Initially employing K-means++ for enhanced data quality, followed by PCA for dimensionality reduction, the study applies Logistic Regression for outcome prediction, achieving remarkable accuracy, specificity, and sensitivity. This methodological rigor offers a promising avenue for early and accurate CVDs detection, significantly outperforming traditional predictive models. By refining data through these steps, the study ensures the predictive model is built on a solid foundation, enhancing the reliability and generalizability of the predictions. The integration of these advanced analytical techniques marks a step forward in the pursuit of effective cardiovascular disease management, highlighting the importance of data preprocessing in predictive modeling.

Keywords: Cardiovascular diseases, PCA, K-means++, logistic regression.

1. Introduction

Worldwide, cardiovascular diseases persist as the leading cause of mortality, accounting for over 30% of all deaths. In the year 2019 alone, these conditions resulted in the deaths of 17.9 million individuals. Furthermore, there is a trend towards younger people being affected by cardiovascular diseases, especially rheumatic heart disease [1]. The complexity of cardiovascular diseases necessitates a multifaceted approach to understanding the interplay of various risk factors, including but not limited to age, blood pressure, cholesterol levels, and lifestyle choices such as smoking and physical activity. Traditional statistical methods have provided insights into the relationships between individual risk factors and heart disease. The advent of machine learning and data-driven methodologies offers a promising avenue to explore these associations further. This paper will utilize logistic regression, a binary classifier to predict the CVDs. However, dealing with datasets with many dimensions can be challenging due to the significant memory requirements and the risk of overfitting associated with analyzing numerous features [2]. By applying feature weighting, it can reduce redundancy in the data and cut down on processing time. This approach helps enhance the efficiency of the algorithm [3].

Principal component analysis (PCA), the technique employed in this article, is categorized as feature extraction. Hence, PCA's fundamental aim is to streamline the analysis by converting a broad array of variables into a more compact set that preserves the majority of the original data's insights. This method

DOI: 10.54254/2753-8818/38/20240569

is particularly useful in dealing with multicollinearity among variables, a common issue in epidemiological datasets where risk factors often exhibit intercorrelations.

In recent years, many researchers have combined PCA to build prediction models for heart disease. Gárate-Escamila et al. utilized chi-square (CHI) analysis and PCA in combination with machine learning techniques to determine if patients are afflicted with heart disease. The study gave an accuracy result of 98.7% by integrating CHI, PCA, and the RF classifier [4]. Zhu et al. combined PCA and K-means for diabetes prediction, and then used the logistic regression to do the classification. The result attained a 97% accuracy [5]. In a like manner, Rathore et al. developed a combined clustering and PCA framework for forecasting heart disease through logistic regression, attaining an accuracy of 98.82% [6]. Jhaldiyal et al. utilized PCA with support vector machines (SVM) to build a prediction model for diabetes. The model provided a 93.66% accuracy [7].

Before applying PCA to reduce the dimensionality of the dataset, this article will utilize a clustering method, K-means ++, to do the data cleaning. Since original datasets often contain noise, missing values, errors, or inconsistent records. Those inaccurate or low-quality data can lead to unreliable analytical results [8]. Through data cleaning, these issues can be identified and corrected, thereby enhancing the overall quality of the dataset to reduce misleading analyses and erroneous decisions [8].

Some researchers utilized clustering methods to do the data cleaning. Loureiro et al. employed hierarchical clustering techniques for identifying outliers, utilizing the dimensions of the resultant clusters as signals for outlier existence [9]. To address scalability challenges related to data observation and cleaning, Hu et al. employed a special cut-clustering technique to categorize keys into various groups. This classification is grounded on specific attributes, including age, cell line, and disease type. By organizing keys into these clusters, identifying duplicates and errors within each group becomes more straightforward, facilitating efficient data management and quality improvement [10]. Moreover, Guo et al. proposed a data-cleaning method based on improved K-means clustering and error feedback to achieve data cleaning [11].

Nevertheless, K-means chooses the initial centroids randomly, which might make a significant difference in the result. Therefore, a more accurate version of K- means—K-means ++ will be employed in the article. K-means++ improves upon K-means by carefully choosing initial centroids using a weighted probability distribution, which is more likely to spread out the centroids and avoid poor clustering. Due to the smarter initialization, K-means++ has a higher chance of converging to a better final solution closer to the global optimum. Moreover, with a better starting point, K-means++ often requires fewer iterations to converge, saving computational resources, especially in cases where convergence is slow. The k-means++ algorithm demonstrated a minimum improvement in accuracy of 10% over the traditional k-means method, frequently achieving significantly superior performance [12].

Therefore, this article will initially employ K-means++ to enhance the data quality, followed by the use of PCA for dimensionality reduction, and ultimately applying LR (Logistic Regression) for outcome prediction. Additionally, by refining the dataset through these steps, this study aims to address potential biases and ensure that the predictive model is built on a solid foundation, thus enhancing the reliability and accuracy of the predictions.

2. Methodology

2.1. Data source and description

The dataset “Cardiovascular Disease dataset”, as shown in Table 1, by SVETLANA ULIANOVA, obtained from Kaggle, was used for this project. This dataset encompasses a variety of input features of 70000 observations, categorized into three distinct groups: objective features (O), based on factual information; examination features (E), derived from the results of medical examinations; and subjective features (S), information provided by the patients. Furthermore, the dataset incorporates a Target Variable: The presence or absence of cardiovascular disease.

DOI: 10.54254/2753-8818/38/20240569

Table 1. Features of the Cardiovascular Disease dataset.

Serial	Group	Features	Features Descriptions
1	O	Age	Patient's age in days
2	O	Gender	Gender: 1 for female, 2 for male
3	O	Height	Patient's height in centimeters
4	O	Weight	Patient's weight in kilograms
5	E	Ap_hi	Systolic blood pressure
6		Ap_lo	Diastolic blood pressure
7	E	Cholesterol	Cholesterol level: 1 for normal, 2 for above normal, 3 for well above normal
8	E	Gluc	Glucose level: 1 for normal, 2 for above normal, 3 for well above normal
9	S	Smoke	Smoking status: 1 for yes, 0 for no
10	S	Alco	Alcohol consumption: 1 for yes, 0 for no
11	S	Active	Exercise engagement: 1 for yes, 0 for no
12	Target Variable	Cardio	Whether the patient has the disease or not. 1 - Yes, 0 - No.

2.2. Method Introduction

2.2.1. Standardization. To construct the model, initial feature standardization was performed with the following formula: $Z = \frac{(X - \mu)}{\sigma}$, where σ and μ represent the standard deviation and mean of each variable, and X is the original dataset. This process is to mitigate the sensitivity of clustering and PCA algorithms to feature scale.

2.2.2. K-means ++. K-means ++ enhances the initialization step of K-means clustering to improve cluster quality. Following the random selection of the initial cluster center, the ensuing centers are picked based on a probability directly related to the squared distance from the closest current center. This method continues until the establishment of K centers, after which the process proceeds with regular K-means clustering until it stabilizes.

For this model, the k-means++ clustering algorithm, set with K = 2, segmented Z, the standardized dataset, into two distinct clusters. Cluster 0 comprised observations indicative of a healthy classification within the dataset, whereas Cluster 1 included those associated with heart disease. Subsequently, data points that were incorrectly clustered were identified and excluded from both clusters. Following the exclusion of these inaccuracies, data points from both clusters were amalgamated and randomized to eliminate any potential sequence dependency present in the original dataset. In this way, the subsequent processes, like principal Component Analysis (PCA), would avoid the influence of the original data ordering.

2.2.3. PCA. Post shuffling and merging, PCA was utilized for dimensionality reduction, selecting principal components that accounted for at least 90% of the variance. Then, the data was then partitioned into training and test sets after the PCA application, preventing information leakage and validating model efficacy.

PCA begins with the computation of the covariance matrix with the formula: $\Sigma = \frac{1}{n-1} A^{\top} A$, A is the shuffling and merging dataset from Z. Then, eigenvalues and eigenvectors are computed from Σ, and eigenvectors are sorted by descending eigenvalues to capture the principal components. The dataset is

DOI: 10.54254/2753-8818/38/20240569

then transformed into a lower-dimensional space using the top k eigenvectors, resulting in a transformed dataset.

2.2.4. Logistic regression. The selected principal components served as features for training the logistic regression model and applied 0.5 as the Decision Threshold. The logistic regression model predicts the probability of a given input belonging to the class labeled "1" (as opposed to class "0") using the logistic function. For a set of features X (the principal components), the probability that Y=1 is expressed as:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}} \quad (1)$$

Where $P(Y = 1|X)$ is the probability that the outcome Y is 1 given the predictors X; $\beta_0, \beta_1, \ldots, \beta_k$, are the coefficients of the model, including the intercept β_0, and the slope coefficients $\beta_1, \ldots, \beta_k$ for each predictor $X_1, \ldots, X_k$; and e is the base of the natural logarithm.

Finally, the performance of the logistic regression model was evaluated using the test set, with accuracy, recall, and specificity as the metrics for assessment (Table 2).

Table 2. Confusion Matrix.

	Predicted: No	Predicted: Yes
Actual: No	True Negative (TN)	False Positive (FP)
Actual: Yes	False Negative (FN)	True Positive (TP)

The Confusion Matrix, as shown above, provides TN, FP, FN, and TP to calculate accuracy: the proportion of correctly made predictions, calculated as $\frac{TP+TN}{TP+TN+FP+FN}$; specificity: the proportion of negative identifications that were correct, calculated as $\frac{TN}{TN+FP}$; and recall (or Sensitivity): the proportion of actual positives correctly identified, calculated as $\frac{TP}{TP+FN}$. Higher values in accuracy, recall, and specificity generally indicate better performance of a logistic regression model.

3. Results and discussion

3.1. Feature contribution

The PCA graph, as shown in Figure 1, represents the variables from the dataset projected onto the first two principal components, which are the axes of the graph. The orientation and magnitude of the arrows show the extent of each variable's contribution to the two main components.

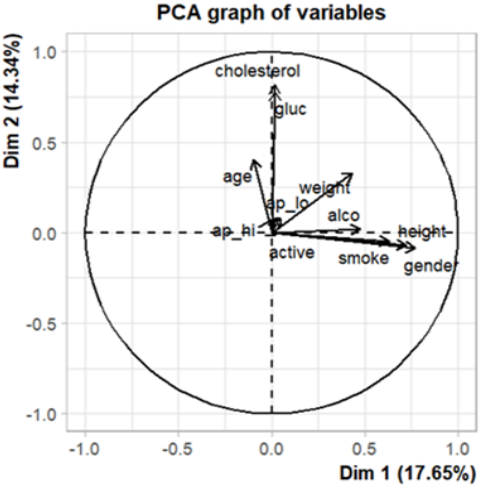

Figure 1. Features contribution to the first and second principal components.

DOI: 10.54254/2753-8818/38/20240569

The PCA biplot visualizes the variables' contributions to the primary axes of variance within the dataset. Notably, cholesterol and glucose levels align closely with the second principal component, suggesting these factors share a similar pattern of variance. Additionally, gender, height, smoking, and alcohol intake appear interrelated, exerting a substantial influence on the first principal component, while age had a moderate contribution to both components.

This alignment indicates a linkage between these lifestyle factors and the underlying principal components. The proximity of cholesterol to the axis of the second principal component signifies its strong association with this axis, paralleled by gender's alignment with the first principal component. These observations emphasize the relative importance of these variables in the dataset and their potential impact on cardiovascular health outcomes (Figure 2).

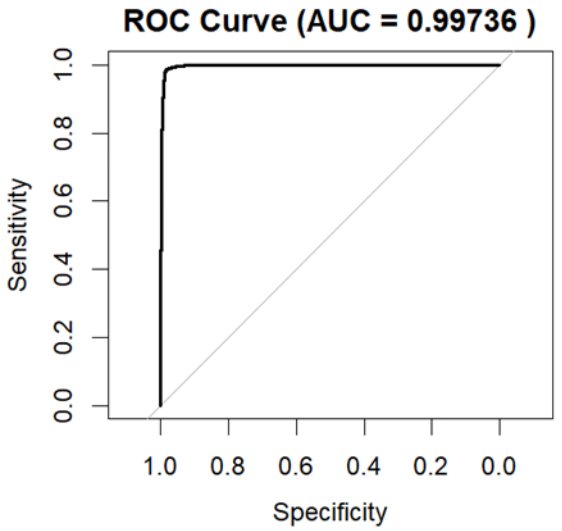

Figure 2. Receiver Operating Characteristic curve.

3.2. Performance improvement

Table 3 shows the changes in accuracy, specificity, and sensitivity for different datasets. Upon the deployment of K-means++, PCA, and logistic regression, the model demonstrated exceptional proficiency in discerning between the presence and absence of CVDs, registering an accuracy of 98.31%, a specificity of 98.00%, and a sensitivity of 98.49%. These metrics collectively indicate that the K-means++ clustering and PCA preprocessing significantly enhance the model's ability to predict cardiovascular diseases with high precision. This suggests that the advanced preprocessing steps contribute to the elimination of noise and reduction of irrelevant information, leading to a more refined and accurate prediction by the logistic regression model. The striking improvement from the raw to the K-means++ and PCA processed data underlines the importance of proper data preprocessing in predictive modeling.

Table 3. Performances of each dataset.

	Raw dataset	PCA processed	K-means++ & PCA processed
Accuracy	71.59%	72.92%	98.31%
Specificity	67.45%	69.25%	98.00%
Sensitivity	75.70%	76.61%	98.49%

The Receiver Operating Characteristic (ROC) curve supports these assertions, offering a detailed depiction of the predictive model's efficacy as illustrated in Fig 2. An AUC of 0.99736 signifies the model's superior ability to differentiate between individuals with and without cardiovascular conditions. This almost flawless AUC value signifies both a high true positive rate (sensitivity) and a minimal rate

DOI: 10.54254/2753-8818/38/20240569

of false positives (1-specificity), highlighting the accuracy of the model in forecasting cardiovascular events.

3.3. Comparison with other studies

To further assess the model, this study compares the accuracy with other recent studies using the same dataset by SVETLANA ULIANOVA, as shown in Table 4. Rana et al. used weight and height to determine the BMI as a new feature. Then they checked and removed the outlier by using the Interquartile Range (IQR). Finally, they used logistic regression as the classifier and got an average accuracy of 72.18% [13]. Comlan et al. utilized the CRISP-DM framework to develop the prediction model, starting with data selection and preparation. They applied several algorithms as classifiers, and the Decision Tree Classifier gave the highest accuracy of 85% [14]. Shorewala applied combined methods such as bagging, boosting, and stacking to enhance the efficacy of classic algorithms. By layering K-Nearest Neighbors, the random forest classifier, and the support vector machine atop logistic regression, they achieved a 75.1% accuracy [15]. Additionally, Theerthagiri and Vidya devised a Recursive Feature Elimination-Gradient Boosting (RFE-GB) strategy, beginning with the dataset's complete feature set and gradually removing the less significant features to isolate a set number of crucial ones [16]. They determined that blood pressure, cholesterol, and physical activity are the key predictors. Then, they used GB as a classifier to get an accuracy of 89.78%. By comparison, the model of this study shows an outstanding performance, but all these studies applied the data preprocessing techniques to achieve a significant improvement in model performance.

Table 4. Accuracy of different prediction models.

Author	Methodology	Accuracy
This study	K-means ++ + PCA + logistic regression	98.31%
Rana et al.	Removing the outlier + logistic regression	72.18%
Comlan et al.	CRISP-DM with Decision Tree	85.00%
Shorewala	Stacking of K-Nearest Neighbors, random forest classifier, and support vector machine with logistic regression	75.10%
Theerthagiri and Vidya	Recursive Feature Elimination-Gradient Boosting	89.78%

4. Conclusion

This study embarked on an integrative approach combining K-means++, PCA, and logistic regression to predict the presence of CVDs, leveraging a robust dataset reflective of key objective, examination, and subjective features. The findings underscore the efficacy of K-means++ in refining the dataset quality, which, when coupled with PCA for dimensionality reduction, significantly enhances the performance of the logistic regression model. This analysis yielded remarkable accuracy, specificity, and sensitivity, demonstrating the model's capability to discern between patients with and without CVDs effectively.

The strategic use of PCA facilitated the identification and retention of the most informative features, thereby mitigating the risk of multicollinearity and overfitting — common challenges in high-dimensional datasets, crucial for the subsequent logistic regression phase, ensuring the model was underpinned by data of the highest fidelity.

The convergence of these methods made a predictive performance that not only aligns with but also extends the current literature on CVDs risk prediction, signifying a step forward in the pursuit of early and accurate disease detection. The study's strength lies in its methodological rigor and the synergistic application of advanced analytical techniques that together enhance the model's reliability and generalizability.

Future work may explore the integration of additional data sources and the deployment of the model in clinical settings to further validate its practical utility. In striving for a model that performs with high

DOI: 10.54254/2753-8818/38/20240569

accuracy across diverse populations, acknowledge the continuous evolution of predictive analytics and its role in transforming cardiovascular disease management.

References

[1] Bose P 2023 Rising threat: cardiovascular disease on the rise among young adults. in News-Medical. Net.
[2] Domingos P 2012 A few useful things to know about machine learning. in Commun. ACM, 55, 78.
[3] Imani M and Ghassemian H 2015 Feature extraction using weighted training samples. in IEEE Geosci. Remote Sensing Lett, 12, 1387–1391.
[4] Gárate-Escamila A K, et al. 2020 Classification models for heart disease prediction using feature selection and PCA. in Informatics in Medicine Unlocked, 19.
[5] Zhu C, Idemudia C U and Feng W 2019 Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," in Informatics in Medicine Unlocked, 17, 100179.
[6] Rathore V S, et al. 2020 A Hybrid Cluster and PCA-Based Framework for Heart Disease Prediction Using Logistic Regression. in Advances in Intelligent Systems and Computing, Springer Singapore, 111–117.
[7] Jhaldiyal T and Mishra P K 2014 Analysis and prediction of diabetes mellitus using PCA, REP, and SVM. in Int. J. Eng. Tech. Res. (IJETR), 2, 164-166.
[8] Ilyas I F and Chu X 2019 Data Cleaning. Association for Computing Machinery.
[9] Loureiro A, Torgo L and Soares C 2004 Outlier Detection Using Clustering Methods: A Data Cleaning Application. in University of Porto, LIACC.
[10] Hu W, et al. 2017 Cleaning by clustering: methodology for addressing data quality issues in biomedical metadata. in BMC Bioinformatics, 18, 415.
[11] Guo X, et al. 2020 Study on Data Cleaning Based on Improved K-Means Clustering and Error Analysis. in 2020 IEEE 4th Conf. on Energy Internet and Energy System Integration (EI2), Wuhan, China, IEEE, 4243-4248.
[12] Arthur D and Vassilvitskii S 2007 K-means++: The Advantages of Careful Seeding. in Proc. of the eighteenth annual ACM-SIAM symp. on Discrete algorithms, 1027–1035.
[13] Rana J H, et al. 2022 Cardiac Abnormality Prediction Using Multiple Machine Learning Approaches. in Bangabandhu and Digital Bangladesh, ICBBDB 2021. Communications in Computer and Information Science, 1-12.
[14] Comlan M and Kpodohoun L 2023 Implementation of a Model for Risk Assessment of Cardiovascular Diseases Using Artificial Intelligence. in 2023 Int. Conf. on Electrical, Computer and Energy Technologies (ICECET), Cape Town, South Africa, 1-6.
[15] Shorewala V 2021 Early detection of coronary heart disease using ensemble techniques," in Informatics in Medicine Unlocked, 100655.
[16] Theerthagiri P and Vidya J 2022 Cardiovascular Disease Prediction Using Recursive Feature Elimination and Gradient Boosting Classification Techniques. in Expert Systems, 9.