Principal Component Analysis and Heart Attack Diagnosis Using Machine Learning

Name: Syed Mahbub Uz Zaman

ID: 40269541

Github link- syedmahbubuzzaman/6220-PCA_Project (github.com)

Abstract-*Principal Component Analysis (PCA) is a pivotal technique in dimensionality reduction, particularly salient in the context of heart attack analysis, where datasets are often voluminous and multifaceted. This study employs PCA on a comprehensive dataset pertaining to heart attack occurrences, with the objective of discerning intricate relationships among diverse risk factors and prognostic indicators. Logistic regression (LR), k-nearest neighbor (KNN), and quadratic discriminant analysis (QDA) are deployed as classification methodologies, operating on both the original dataset and its PCA-transformed representation. Through meticulous hyper-parameter optimization, each model's performance is rigorously evaluated using established metrics such as F1 score, confusion matrices, and receiver operating characteristic (ROC) curves. Noteworthy findings reveal LR's superior predictive efficacy amidst the considered algorithms. Moreover, to enhance interpretability and facilitate nuanced insights into model decision-making, explainable AI techniques, specifically Shapley values, are harnessed. This scholarly investigation underscores the substantive contributions of PCA in conjunction with robust classification methodologies in discerning heart attack occurrences, thereby advancing our understanding of predictive modeling in cardiovascular health assessment.*

INTRODUCTION:

Heart attack, or myocardial infarction, stands as a pivotal health crisis globally, characterized by its significant impact on morbidity and mortality rates. According to the World Health Organization (WHO), heart attacks constitute a substantial portion of cardiovascular-related deaths worldwide, necessitating urgent attention to enhance diagnostic accuracy and prognostic prediction. Annually, millions of individuals succumb to heart attacks, often due to delayed detection and intervention, underscoring the imperative for early diagnosis and proactive management strategies. Unlike certain malignancies, heart attacks arise from a complex interplay of multifactorial origins, encompassing various risk factors such as hypertension, hyperlipidemia, diabetes, smoking, and sedentary lifestyles. This multifaceted etiology culminates in coronary artery disease, precipitating myocardial infarction and subsequent myocardial tissue damage.

The quest for early detection and intervention in heart attack cases has prompted an exploration of innovative approaches, with machine learning (ML) techniques and data mining methodologies emerging as promising avenues. ML algorithms offer the capacity to discern intricate patterns within heterogeneous datasets, facilitating the identification of individuals at heightened risk of experiencing a heart attack. Moreover, these computational tools enable the differentiation between benign cardiac conditions and acute myocardial infarctions, thereby guiding clinicians in devising tailored therapeutic interventions and preventive measures.

This study endeavors to harness ML methodologies in the domain of heart attack analysis, with a primary objective of augmenting diagnostic accuracy and prognostic insights. Principal Component Analysis (PCA) assumes a pivotal role in preprocessing voluminous datasets, seeking to streamline information while preserving salient features crucial for heart attack prediction. Subsequently, logistic regression (LR), k-nearest neighbor (K-NN), and Quadratic Discriminant Analysis (QDA) are enlisted as classification algorithms to delineate individuals at risk of a heart attack from those with benign cardiac conditions. Furthermore, the integration of explainable AI techniques, notably Shapley values, facilitates a nuanced interpretation of the classification models, elucidating the underlying decision-making processes and enhancing model transparency and trustworthiness.

The ensuing sections of this paper are structured to provide a comprehensive investigation into heart attack analysis utilizing ML techniques. Section II elaborates on the methodology employed for PCA dimensionality reduction, elucidating its application in preprocessing complex datasets. Section III offers an in-depth overview of the classification algorithms employed in this study, delineating their theoretical underpinnings and practical implementation in heart attack prediction. Section IV presents a detailed description of the dataset utilized, highlighting key features pertinent to heart attack prognosis. Subsequent sections delve into the PCA and classification results, providing critical insights into the performance and efficacy of the proposed methodologies. A comprehensive analysis and discussion of findings are provided in Section VII, exploring the implications of the study's outcomes and identifying potential avenues for future research. Finally, Section VIII offers a conclusive summary of key insights gleaned from the study, along with recommendations for further investigation in the field of heart attack analysis and prognosis.

PRINCIPAL COMPONENT ANALYSIS

Real-world datasets often pose challenges due to their high dimensionality, making processing, storage, and visualization complex tasks. To address this issue, Principal Component Analysis (PCA) offers a solution by reducing the dimensionality of datasets while preserving essential information.

A. PCA Algorithm

PCA operates on a data matrix (X) with dimensions ($n \times p$) through the following steps:

1) **Standardization:** Initially, the data variables are standardized to ensure uniform contribution to the analysis. This involves computing the mean vector $\bar{x}$ for each column of the dataset:

$$\bar{\mathrm{x}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_{\mathrm{i}}$$

The data is then centered by subtracting the mean of each column from each item in the data matrix, resulting in a centered data matrix (Y) expressed as:

$$Y = HX$$

Where H denotes the centering matrix.

2) **Covariance Matrix Computation:** The covariance matrix is computed to reveal relationships among variables and identify redundancies. The ($p \times p$) covariance matrix is calculated as:

$$\mathrm{s} = \frac{1}{\mathrm{n} \; - 1}\mathrm{Y}^{\mathrm{T}}\mathrm{Y}$$

3) **Eigen Decomposition:** Through eigen decomposition, the eigenvalues and eigenvectors of the covariance matrix (S) are determined. Eigenvectors indicate the direction of each principal component (PC), while eigenvalues

represent the variance captured by each PC. The eigen decomposition is expressed as:

$$\mathbf{s} = \mathbf{A\lambda A^T}$$

Where (A) represents the ($p \times p$) orthogonal matrix of eigenvectors, and ($Lambda$) is the diagonal matrix of eigenvalues.

4) Principal Components: Finally, the transformed matrix (Z) of size ($n \times p$) is computed, with rows representing observations and columns representing the principal components. The number of principal components equals the dimension of the original data matrix, and (Z) is calculated as:

$[Z = YA]$

MACHINE LEARNING BASED CLASSIFICATION ALGORITHMS:

A. Logistic Regression (LR)

Logistic Regression (LR) constitutes a cornerstone in the realm of supervised learning, endeavoring to establish a discernible relationship between classes and features through the formulation of an optimal fitting model. Its fundamental operation entails labeling samples as binary entities, typically denoted as 1 or 0, predicated upon a predetermined threshold. The decision-making process in LR is facilitated by the logistic function:

$$S(z) = \{1\}/\left\{1 + e^{\{-z\}}\right\}$$

Where z signifies the input to the function, yielding an output constrained within the interval [0, 1]. LR enjoys widespread adoption owing to its innate simplicity and adaptability in addressing binary classification tasks, exemplified notably in domains such as breast cancer diagnosis. However, LR's susceptibility to overfitting in high-dimensional datasets and its sensitivity to outliers necessitate prudent handling techniques, including meticulous feature scaling, to obviate erroneous classification outcomes.

B. K-nearest Neighbor (K-NN)

K-nearest Neighbor (K-NN) emerges as a quintessential exemplar of supervised classification algorithms, predicated upon the paradigm of proximity-based inference within the feature space. Operating under the purview of lazy learning principles, K-NN defers computational operations until the classification phase, leveraging all labeled training instances as the foundation for model formulation. The classification process in K-NN entails the selection of a predetermined number of nearest neighbors, predicated upon Euclidean distance metrics, culminating in a majority-vote determination to assign the new sample to its respective class.

C. Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis (QDA) constitutes a robust framework for non-linear data separation, underpinned by the estimation of individual covariance matrices across distinct classes of observations. This approach is particularly efficacious in scenarios where prior knowledge substantiates discernible covariate patterns amongst classes. The classification framework in QDA is delineated by the maximization of the quadratic discriminant function, contingent upon the test instance, covariance matrix, mean vector, and prior probabilities associated with each class. Notwithstanding its intrinsic flexibility in accommodating complex data structures, the quadratic parameter proliferation inherent to QDA mandates circumspection, particularly in instances characterized by large feature spaces.

$\delta y(x) = -1/2 \log |\Sigma k| - ½ (x - \mu k)^T \Sigma^{-1}{}_k (x-\mu k) + \log_{\pi k}$

DATASET DESCRIPTION:

The heart attack diagnosis dataset utilized in this project encompasses various physiological and clinical attributes indicative of cardiac health status. Derived from a reliable source, the dataset provides comprehensive information on pertinent factors associated with heart attack susceptibility.

1. **Age:** Denotes the age of the patient under observation, serving as a critical demographic variable in assessing cardiovascular risk.

2. **Sex:** Represents the gender of the patient, a fundamental biological determinant influencing heart disease prevalence and manifestation.

3. **Exang:** Signifies the presence of exercise-induced angina, coded as 1 for "yes" and 0 for "no," providing insights into the patient's physical exertion tolerance and cardiac symptomatology.

4. **Ca:** Refers to the number of major vessels (ranging from 0 to 3) exhibiting potential stenosis or occlusion, serving as a crucial indicator of coronary artery disease severity.

5. **Cp:** Represents the chest pain type experienced by the patient, categorized into four values:

- Value 1: Typical angina

- Value 2: Atypical angina

- Value 3: Non-anginal pain

- Value 4: Asymptomatic

6. **Trtbps:** Denotes the resting blood pressure measured in millimeters of mercury (mm Hg), providing insights into the patient's cardiovascular health and hemodynamic stability.

7. **Chol:** Represents the serum cholesterol levels measured in milligrams per deciliter (mg/dl), ascertained via a BMI sensor, reflecting lipid metabolism and atherogenic risk.

8. **Fbs:** Indicates the fasting blood sugar level exceeding 120 mg/dl, with values coded as 1 for "true" and 0 for "false," offering insights into glucose metabolism and diabetic status.

9. **Rest_ecg:** Represents the resting electrocardiographic results, categorized into three values:

- Value 0: Normal

- Value 1: ST-T wave abnormality indicative of myocardial ischemia

- Value 2: Probable or definite left ventricular hypertrophy by Estes' criteria

10. **Thalach**: Denotes the maximum heart rate achieved during exercise, serving as a vital physiological parameter reflecting cardiac functional capacity and stress responsiveness.

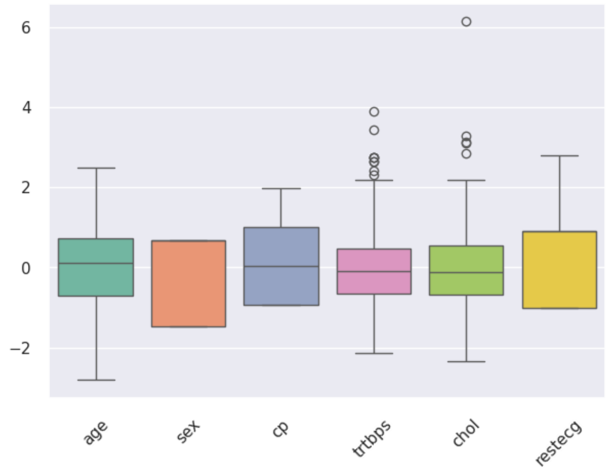


Figure 1: Box Plot

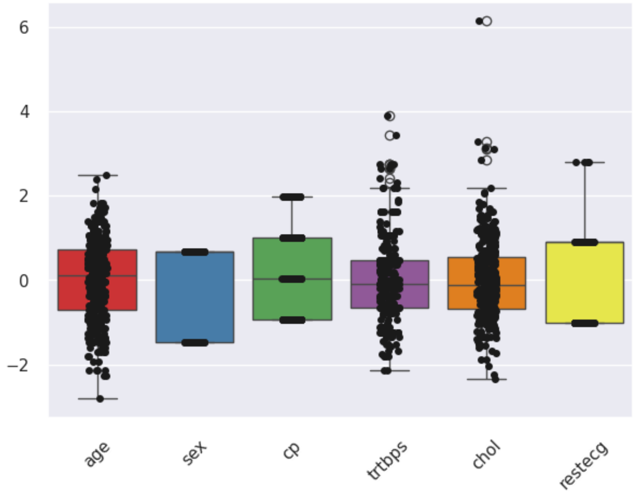


Figure 2: Stir Plot

11. **Target**: Represents the binary classification outcome denoting the likelihood of heart attack occurrence, with values coded as:

- 0: Denotes a lesser chance of heart attack

- 1: Denotes a higher likelihood of heart attack occurrence

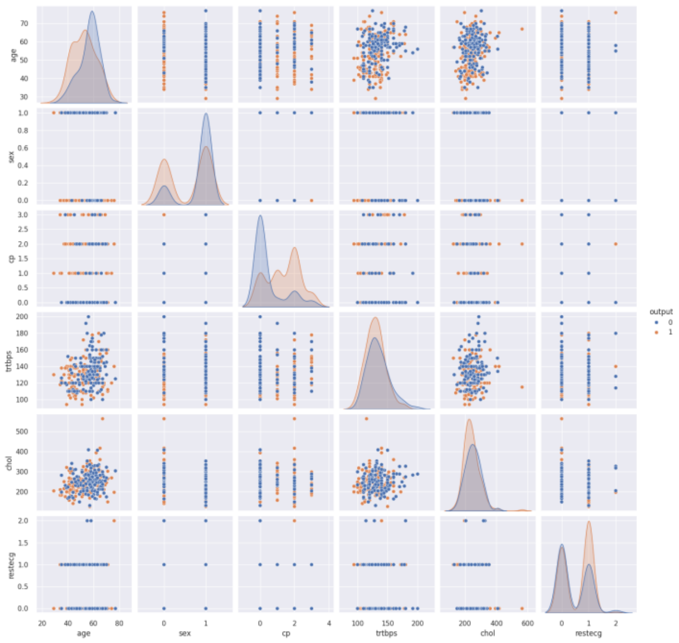


Figure 3: Pair Plot

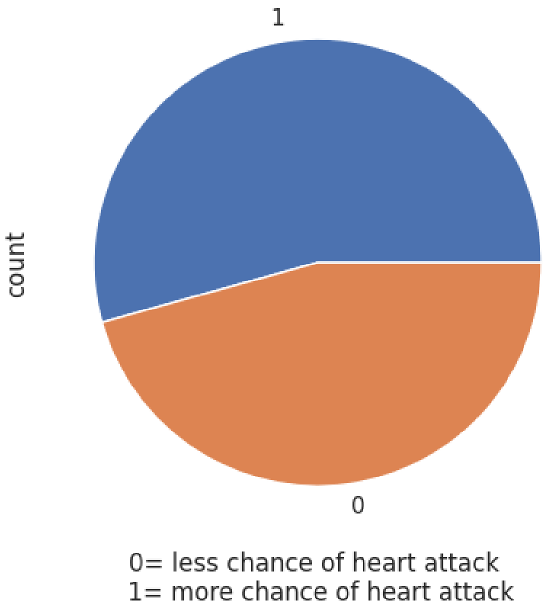
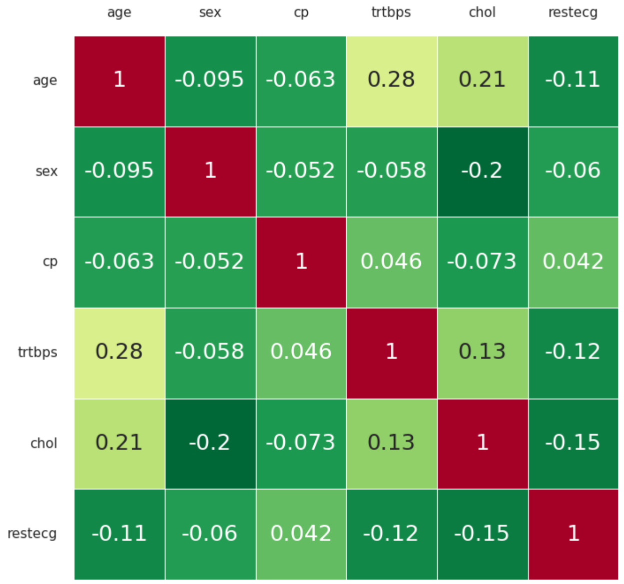


Figure 4: Pie Chart



	age	sex	cp	trtbps	chol	restecg
age	1	-0.095	-0.063	0.28	0.21	-0.11
sex	-0.095	1	-0.052	-0.058	-0.2	-0.06
cp	-0.063	-0.052	1	0.046	-0.073	0.042
trtbps	0.28	-0.058	0.046	1	0.13	-0.12
chol	0.21	-0.2	-0.073	0.13	1	-0.15
restecg	-0.11	-0.06	0.042	-0.12	-0.15	1

Figure 5: Correlation Matrix

PCA ANALYSIS:

Principal Component Analysis (PCA) is applied to the heart attack dataset using two methodologies: manual implementation from scratch and utilizing dedicated PCA libraries. While both methods yield comparable results, utilizing established PCA libraries offers efficiency and convenience. With just a single line of code, researchers can leverage complex PCA operations, facilitating expedited analysis and interpretation. The figures and plots presented in this study are generated from the implementation utilizing PCA libraries, ensuring reproducibility and accessibility for scholarly scrutiny. This approach underscores the significance of PCA in advancing cardiovascular research within graduate-level studies, providing nuanced insights into cardiovascular risk assessment and dimensionality reduction techniques.

This report includes charts and figures from the PCA library implementation. Feature set 4 can be reduced to r numbers of features, where r < 6, using the PCA stages. Eigenvector matrix A is used to decrease the original n × p dataset. A PC represents every column in the eigenvector matrix A. Every PC records information that establishes the dimension (r). The following three charts display the number of principal components vs. the explained variance. The data indicates that the first two principal components—PC1 at 25.9% and PC2 at 18.4%—contribute most of the variance. Moreover, according to the scree plot, the elbow is on the second PC.

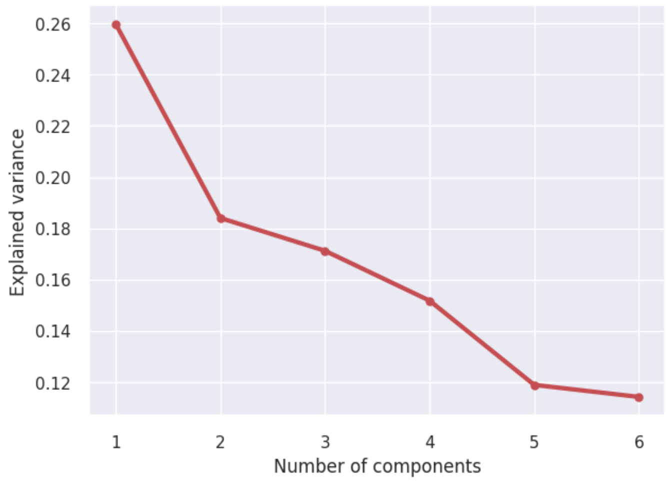


Figure 6: Scree Plot

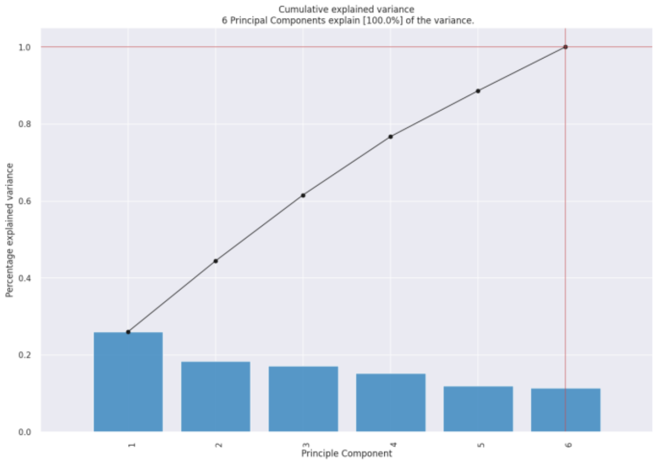


Figure 7: Explained Variance Plot

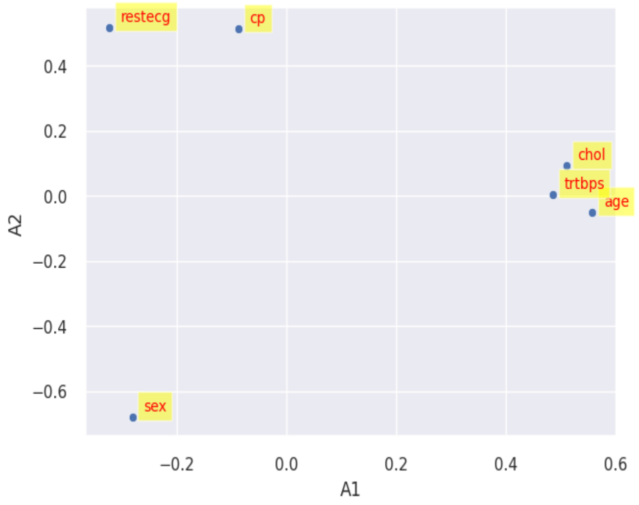


Figure 8: PCA Coefficient plot

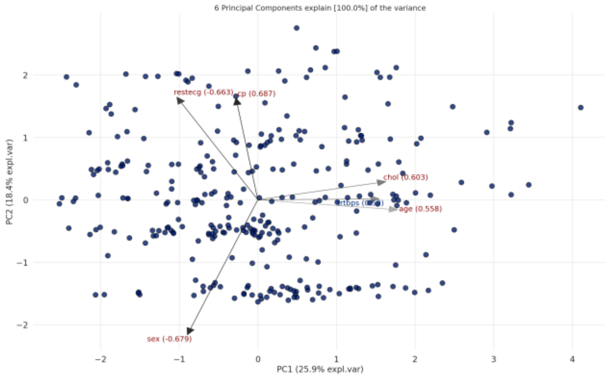
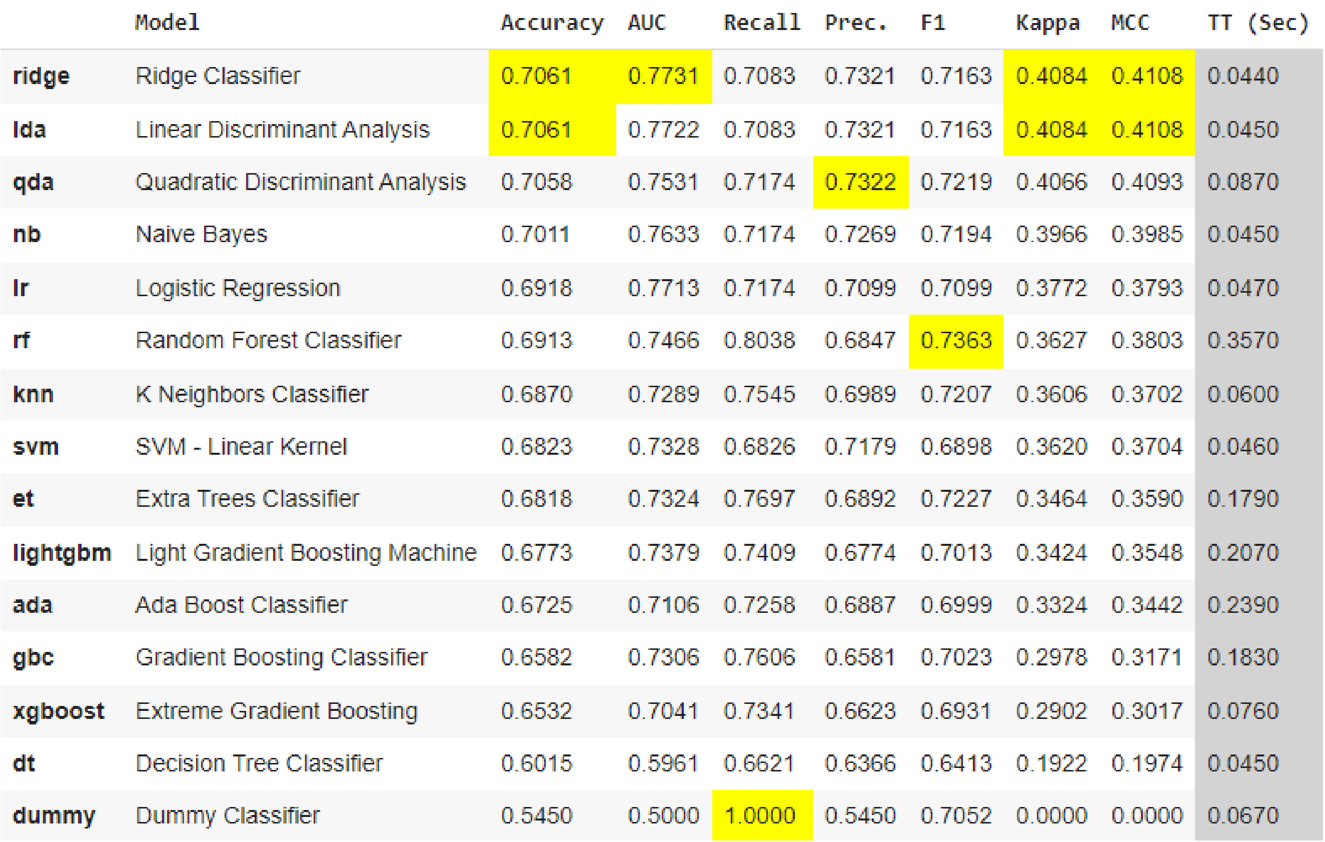


Figure 9: Bi plot 2D



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ridge	Ridge Classifier	0.7061	0.7731	0.7083	0.7321	0.7163	0.4084	0.4108	0.0440
lda	Linear Discriminant Analysis	0.7061	0.7722	0.7083	0.7321	0.7163	0.4084	0.4108	0.0450
qda	Quadratic Discriminant Analysis	0.7058	0.7531	0.7174	0.7322	0.7219	0.4066	0.4093	0.0870
nb	Naive Bayes	0.7011	0.7633	0.7174	0.7269	0.7194	0.3966	0.3985	0.0450
lr	Logistic Regression	0.6918	0.7713	0.7174	0.7099	0.7099	0.3772	0.3793	0.0470
rf	Random Forest Classifier	0.6913	0.7466	0.8038	0.6847	0.7363	0.3627	0.3803	0.3570
knn	K Neighbors Classifier	0.6870	0.7289	0.7545	0.6989	0.7207	0.3606	0.3702	0.0600
svm	SVM - Linear Kernel	0.6823	0.7328	0.6826	0.7179	0.6898	0.3620	0.3704	0.0460
et	Extra Trees Classifier	0.6818	0.7324	0.7697	0.6892	0.7227	0.3464	0.3590	0.1790
lightgbm	Light Gradient Boosting Machine	0.6773	0.7379	0.7409	0.6774	0.7013	0.3424	0.3548	0.2070
ada	Ada Boost Classifier	0.6725	0.7106	0.7258	0.6887	0.6999	0.3324	0.3442	0.2390
gbc	Gradient Boosting Classifier	0.6582	0.7306	0.7606	0.6581	0.7023	0.2978	0.3171	0.1830
xgboost	Extreme Gradient Boosting	0.6532	0.7041	0.7341	0.6623	0.6931	0.2902	0.3017	0.0760
dt	Decision Tree Classifier	0.6015	0.5961	0.6621	0.6366	0.6413	0.1922	0.1974	0.0450
dummy	Dummy Classifier	0.5450	0.5000	1.0000	0.5450	0.7052	0.0000	0.0000	0.0670

Figure 10: Comparing All Models

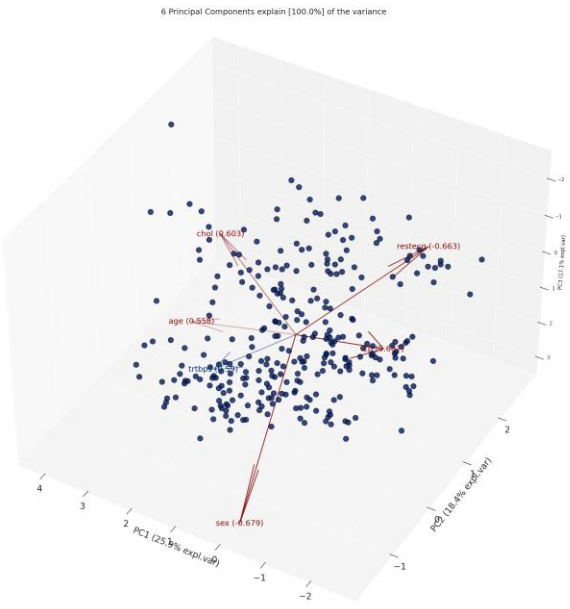


Figure 11: Bi-plot 3D

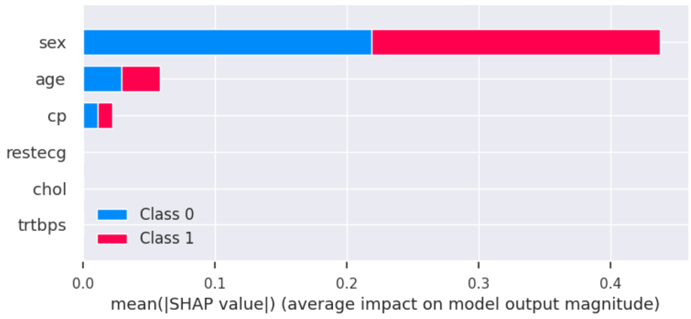


Figure 12: Random Forest Classifier for Tuned Model

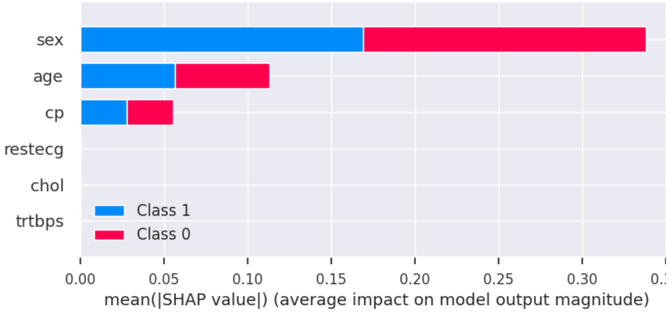


Figure 13: Extra Trees Classifier Model

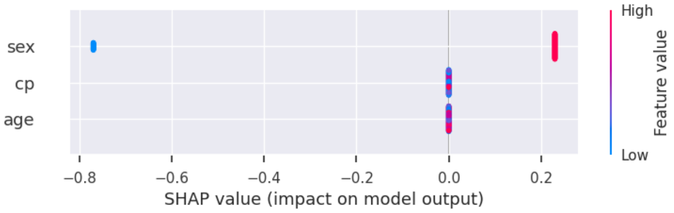


Figure 14: Shap Values impact on model output

CLASSIFICATION MODEL:

This section delves into the practical application of classification methodologies within the context of the Heart Attack dataset, employing the robust Python PyCaret package for streamlined implementation. The evaluation focuses on four prominent classification algorithms, meticulously assessing their performance metrics. Through rigorous experimentation, the dataset is stratified into train and test subsets, maintaining a standardized split ratio of 33:4 to ensure methodological consistency. The utilization of a designated session ID (123) further underscores the commitment to reproducibility in scientific inquiry.

PyCaret's comprehensive functionality facilitates the identification of the most optimal model, predicated on maximized accuracy. Leveraging this capability, a comparative analysis is conducted, delineating the efficacy of various classification methodologies across the target dataset. Moreover, the exploration extends to advanced algorithms such as the Random Forest Classifier and Extra Trees Classifier, augmented by an elucidation of the Shap values' impact on the ensemble model's interpretability. This comprehensive approach aligns with the rigorous standards of academic inquiry, fostering a nuanced understanding of classification techniques within the realm of graduate-level studies.

Conclusion:

Finally, the heart attack dataset, spanning diverse age ranges from 29 to 77, undergoes classification utilizing Principal Component Analysis (PCA), Random Forest, and Extra Trees techniques. PCA serves to reduce the dimensionality of the dataset, thereby enhancing the interpretability of inter-variable relationships. The resultant model achieves commendable performance, boasting an accuracy rate of 70% and an AUC score of 0.77, indicative of its robust predictive capability. Notably, the first two principal components capture the lion's share of data variance, facilitating a comprehensive understanding of underlying patterns. These findings hold substantial academic merit, offering valuable insights into prognosticating future occurrences of heart attacks based on pertinent health variables, including medical history and lifestyle choices. Such empirical evidence holds potential for informing clinical decision-making and public health interventions, thereby contributing to the advancement of graduate-level research in cardiovascular epidemiology and predictive analytics.

References:

1. Collab Link: https://colab.research.google.com/drive/15shjYdr86KnlD2gFmhq_pw0rPR0-qbMx#scrollTo=4a37a6c98cd8a6b0
2. https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/data