Cardiovascular Disease Classification Project

Introduction/Background

Our project focuses on classifying cardiovascular disease using medical data, including biometrics like gender, height, weight, and blood pressure. A study that used machine learning to predict whether a patient had heart disease tested several machine learning algorithms, and “the results showed that RF (91.80%) had the highest accuracy in predicting heart disease, followed by NB (88.52%) and SVM (88.52%)” [1]. Given these results, we will prioritize RF and SVM in our supervised learning approach. Our dataset includes 70,000 entries, each with 12 features—11 biometrics and 1 indicating cardiovascular disease.

Cardiovascular Disease Dataset

Problem Definition

According to the WHO, “cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year” [2]. By creating a model to classify whether a person has cardiovascular disease, preventative actions can be taken to mitigate the devastating effects of cardiovascular diseases.

Methods

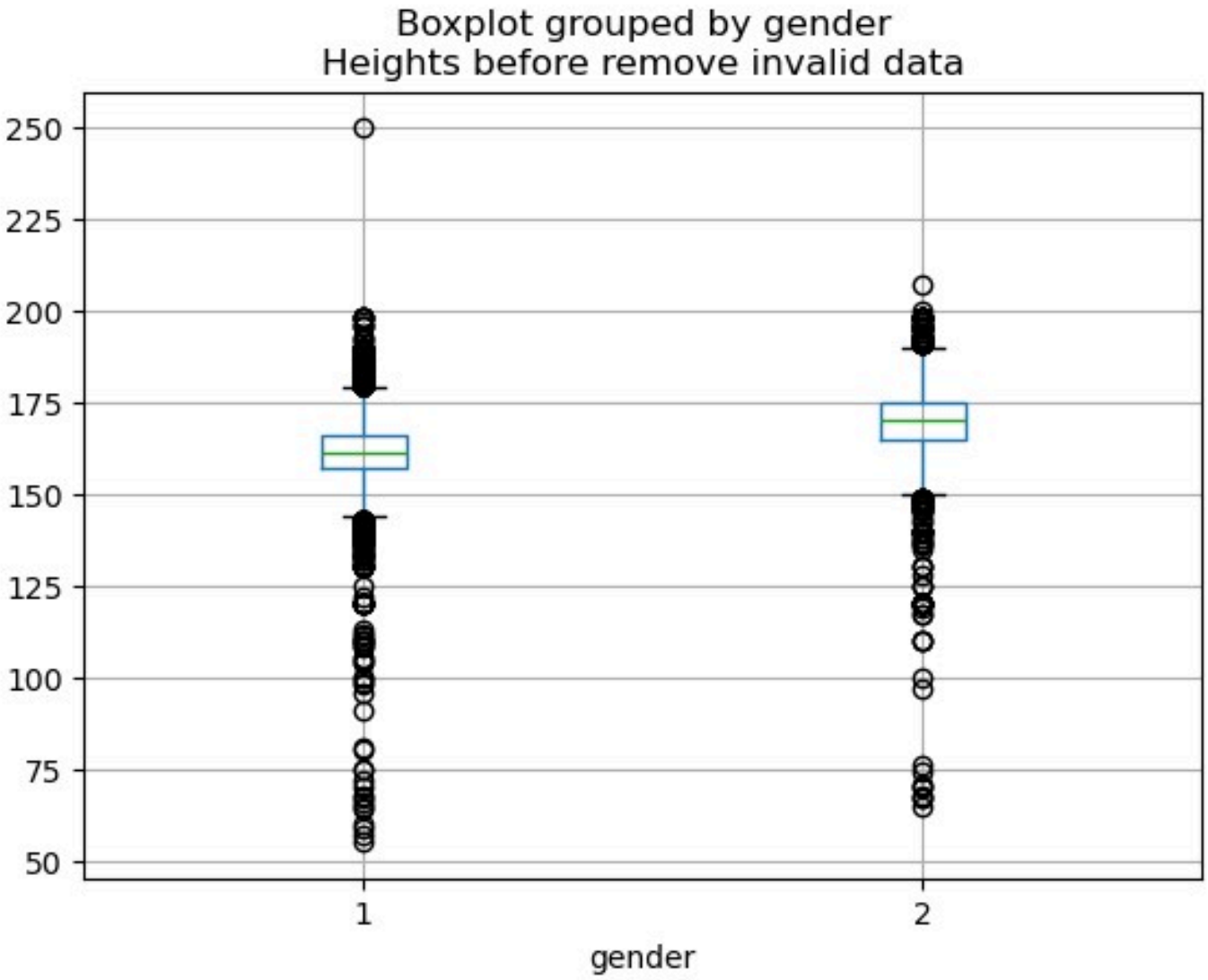
Upon initial inspection, our dataset is primarily composed of numerical values, so we decided to use Pandas and Numpy library.

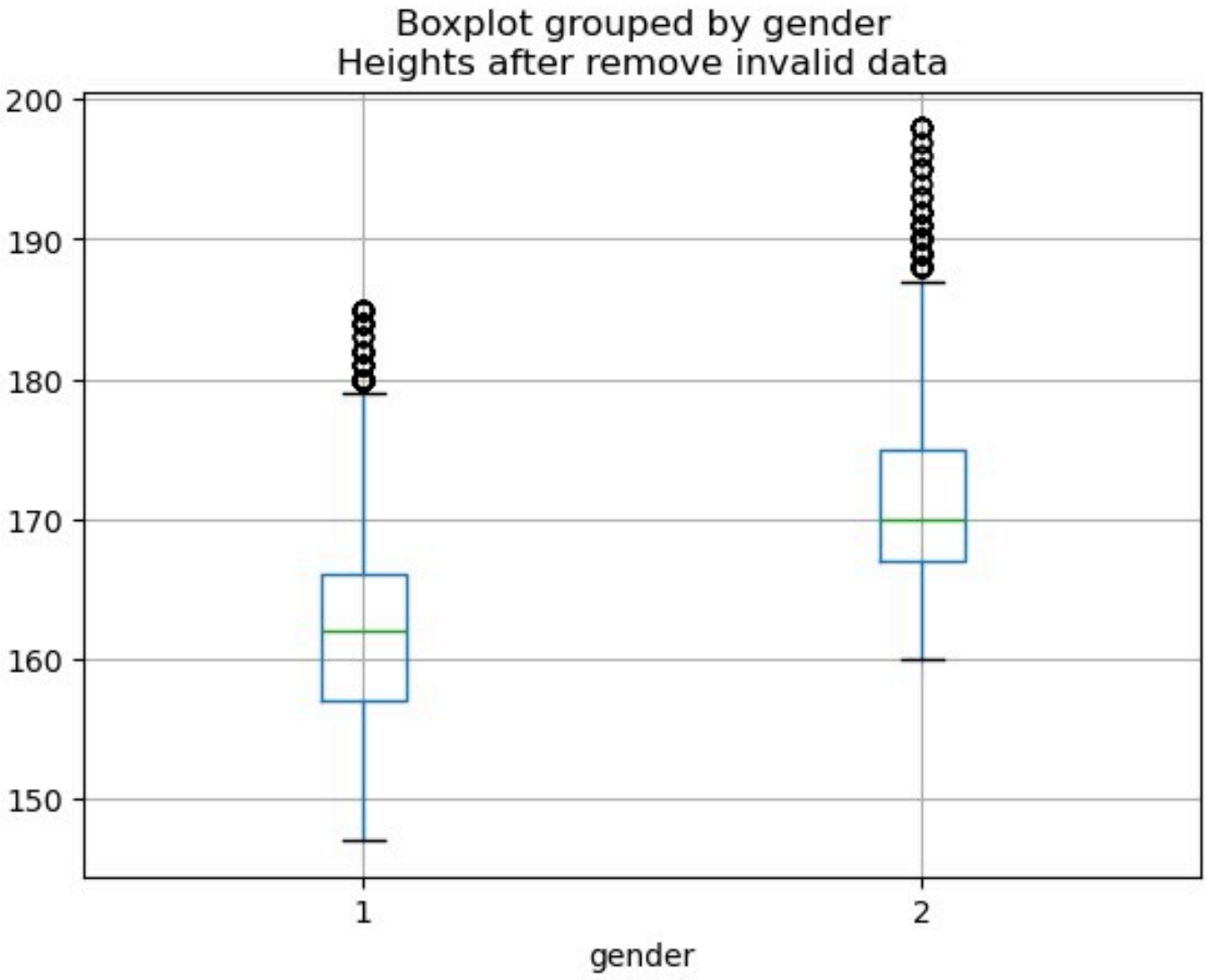
Our 3 data preprocessing methods are as follows:

1. ***Data Cleaning***
    - Duplicate Values: 24 data points removed
    - Missing Values: No missing values in our dataset

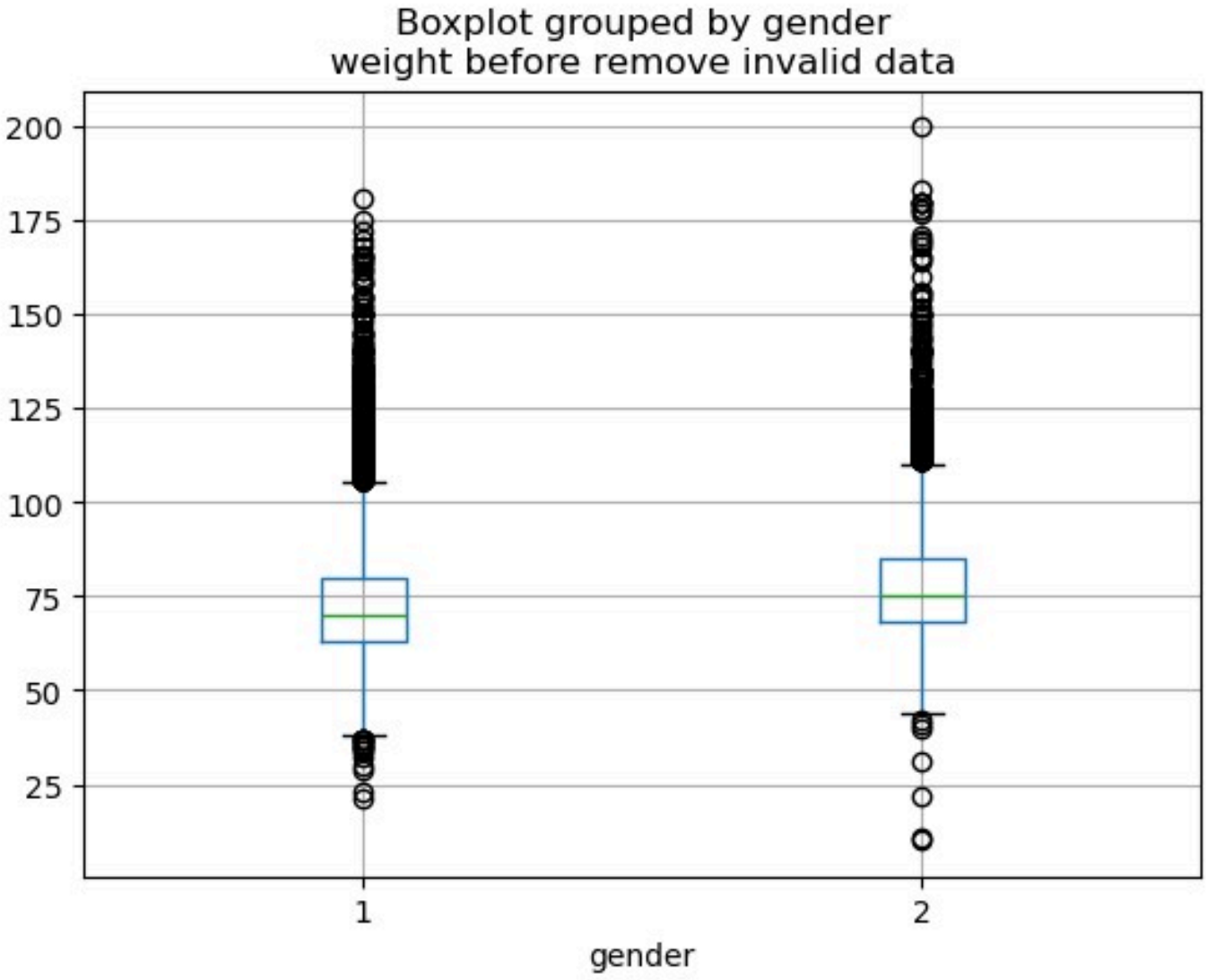
- **Height:** -Since our data focuses on individuals in their middle age, spanning a height range from 5’3" (160 cm) to 6’6" (198 cm) for men and the 4’10’’ (147cm) to 6’1’’ (185cm) for women, according to the 2007-2008 census data [3]. It's worth noting that the cumulative percentage of the population falling

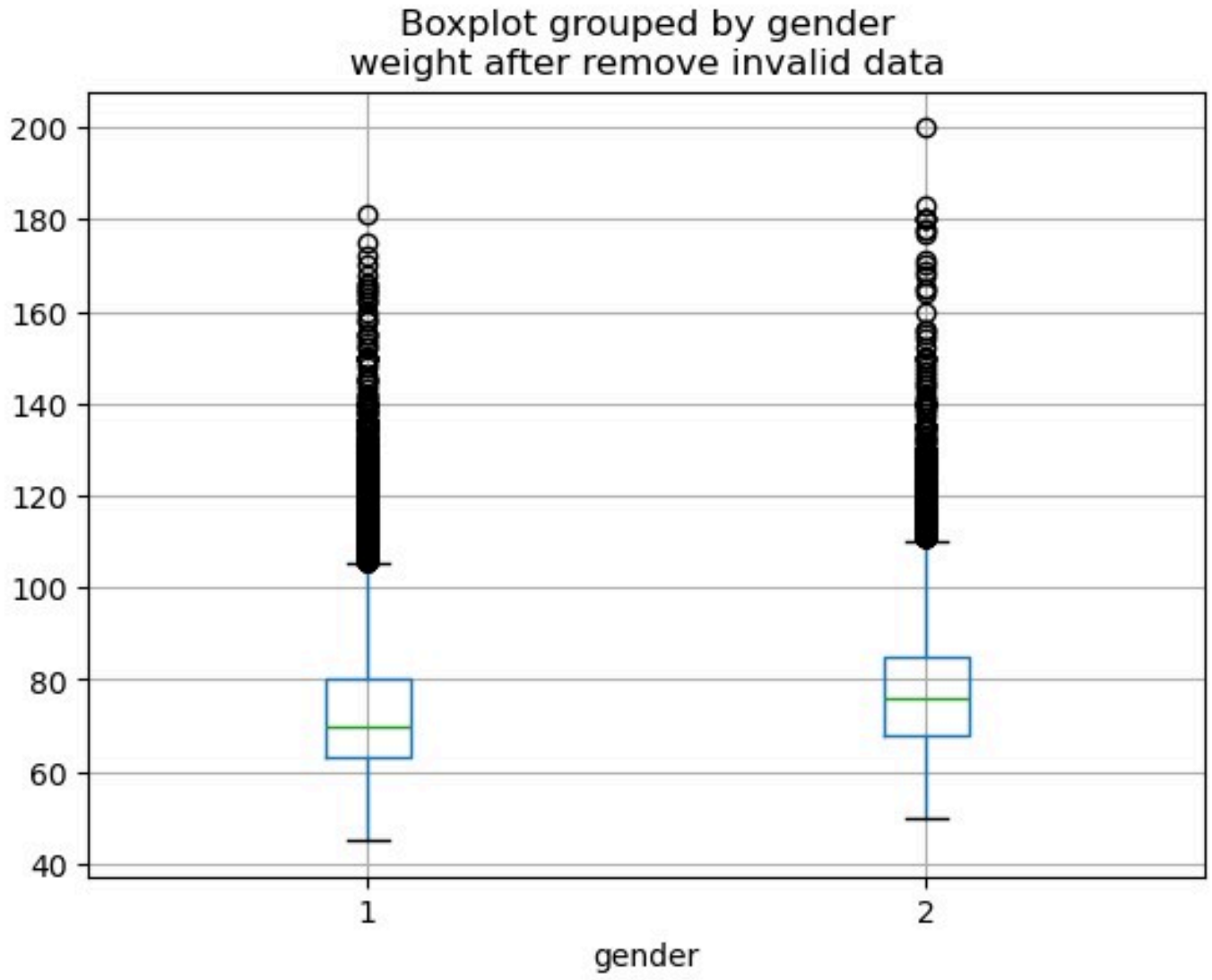
outside this range is roughly 1%. Consequently, we classify these data points as outliers and exclude them from our dataset. The box plot of data after removal shows lower variability with a few tall individuals.



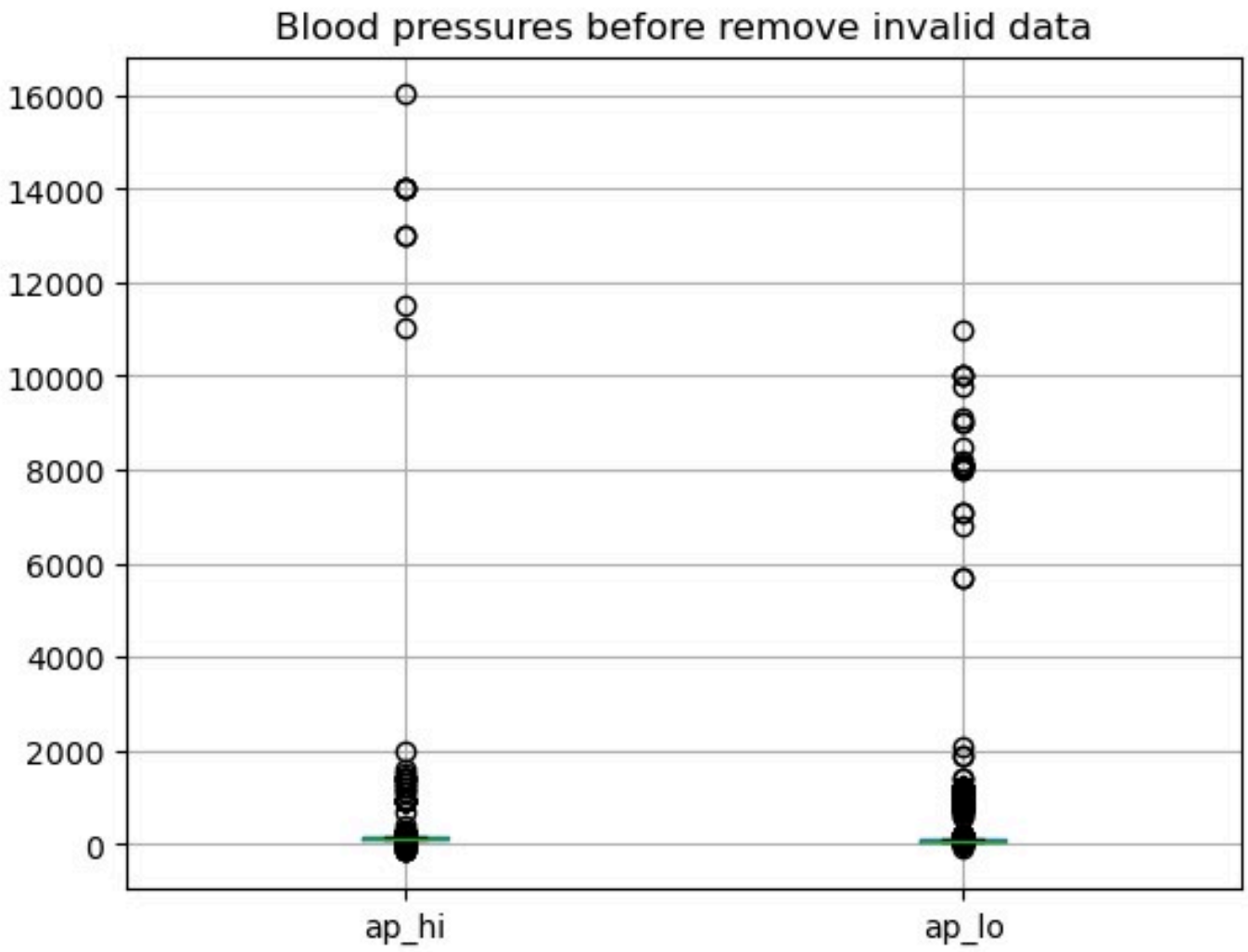


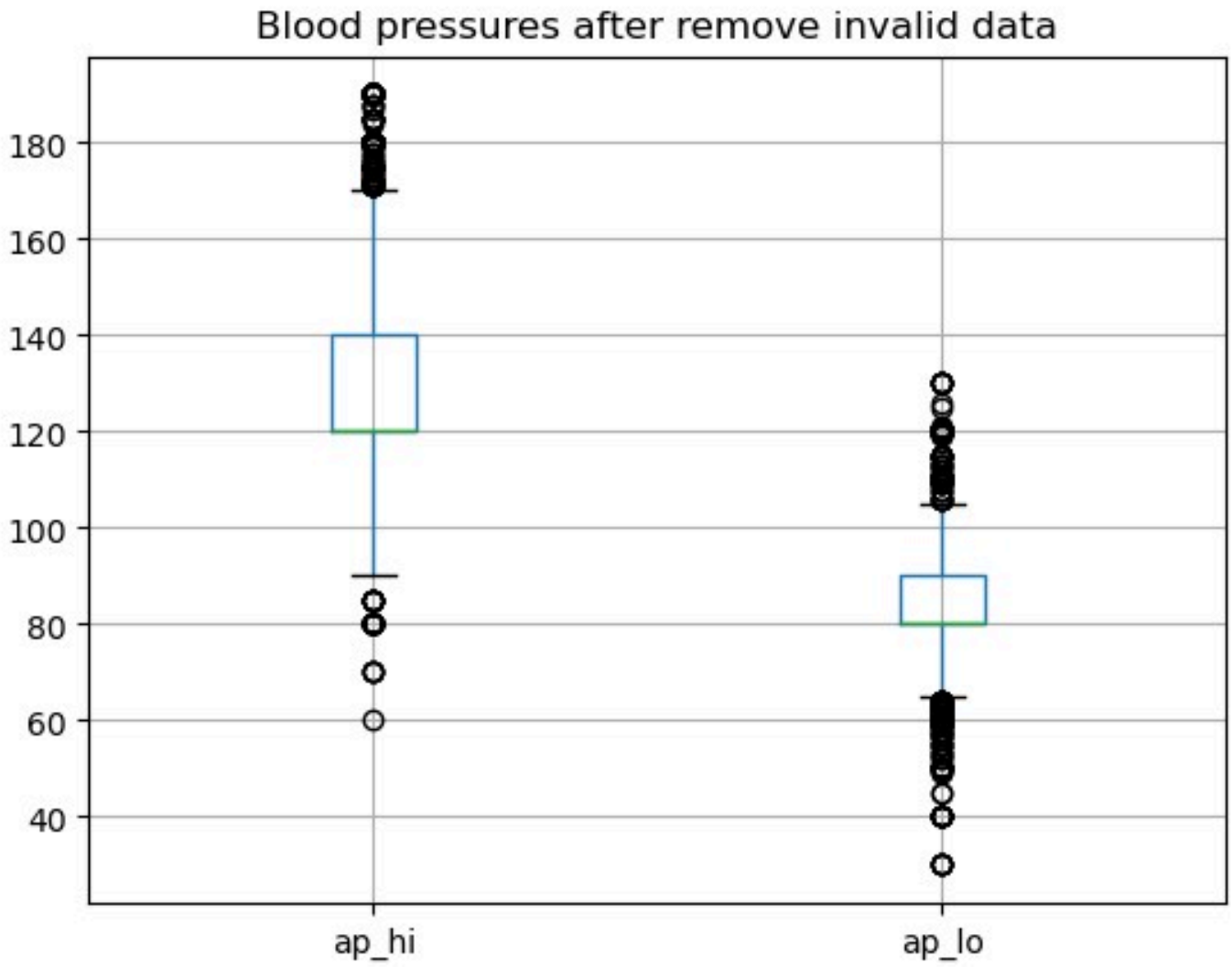
- **Weight:** -Similarly, we removed weights that are under 45kg (99lbs) for women and 50kg (110lbs) for men since the cumulative percentage of those are less than cumulatively 1% based on census. However, we did not remove weights on the higher side since overweight directly contributes to cardiovascular risk factors [4]


Boxplot grouped by gender
weight before remove invalid data
200
175
150
125
100
75
50
25
1
2
gender

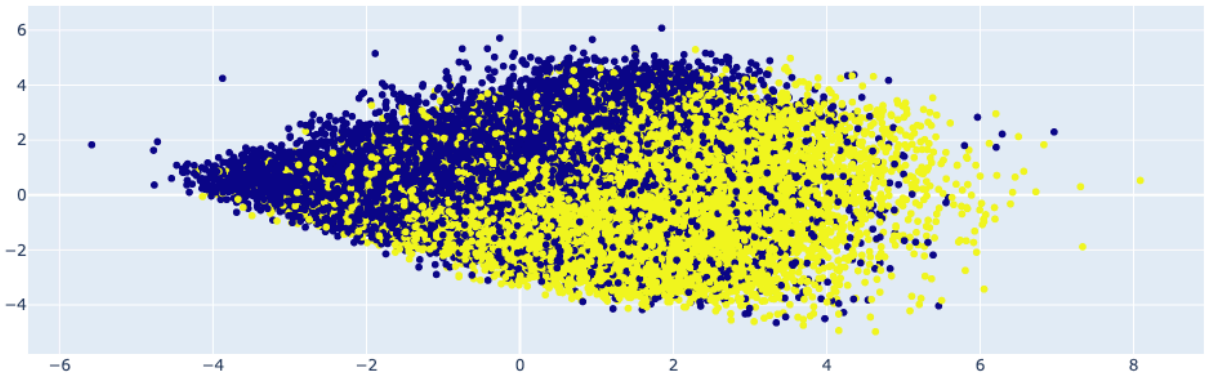


- **Blood Pressure:** -The systolic and diastolic blood pressures range from normal 90/60 mm Hg to stage 2 hypertension 160/100 mm Hg [5]. Any pressure outside of this range is considered life threatening emergency. So, we excluded those datapoints with buffer of +/- 30mm Hg.


Blood pressures before remove invalid data
16000
14000
12000
10000
8000
6000
4000
2000
0
ap_hi
ap_lo

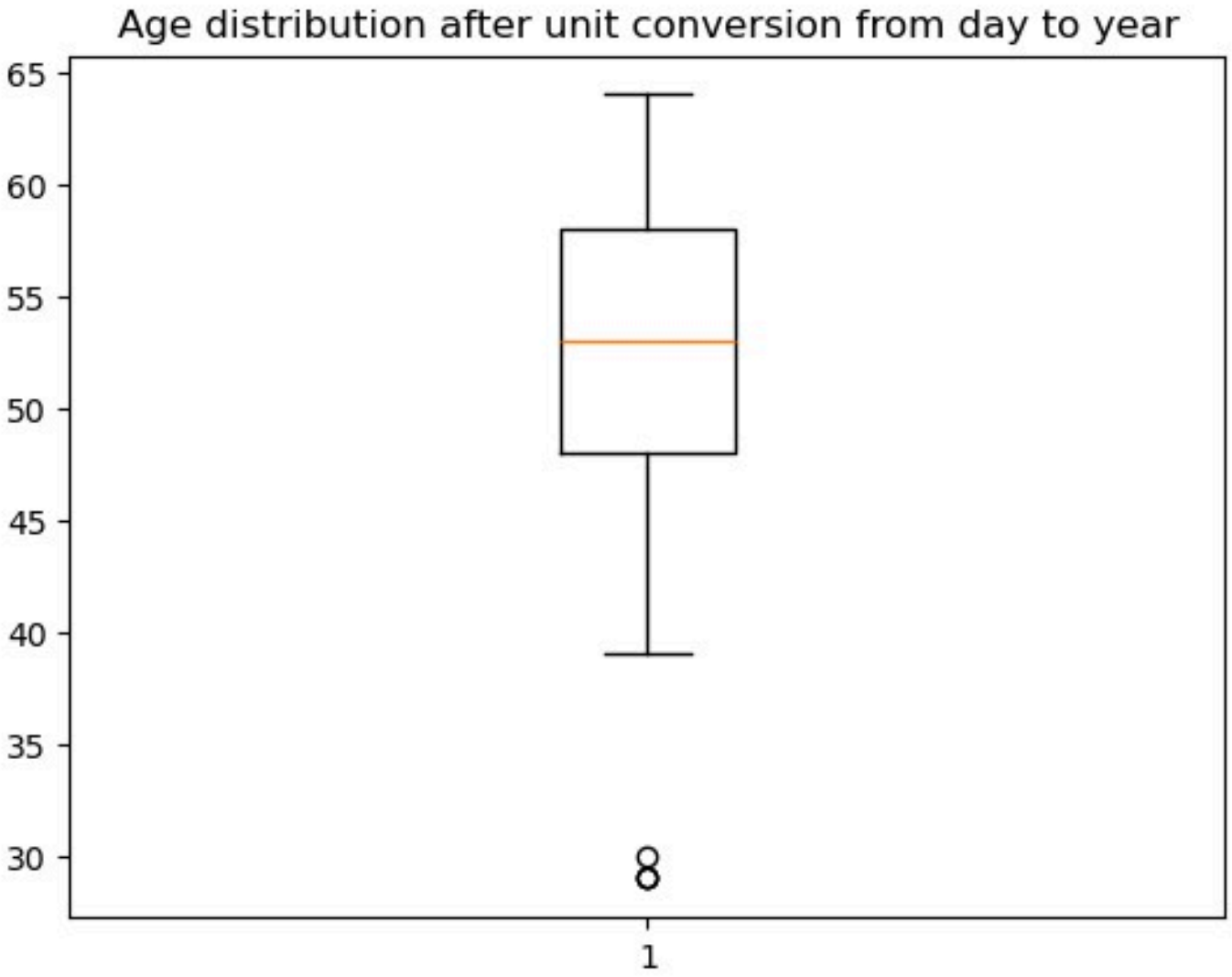


We considered applying Principal Component Analysis to the original and the validated datasets to visualize any possible outlier groups. Based on plot PCA plot using 1st and 2nd components, there is no big cluster of outliers so all remaining datapoints are kept. Additionally, from the projected PCA, there is a high degree of overlap between the two outcomes of cardiovascular diseases.



2. ***Feature Engineering***

-In the original dataset, ages were recorded in days, leading to unnecessary variation. To address this, we converted the unit from days to years. The boxplot reveals that our data primarily focuses on individuals in their middle age.

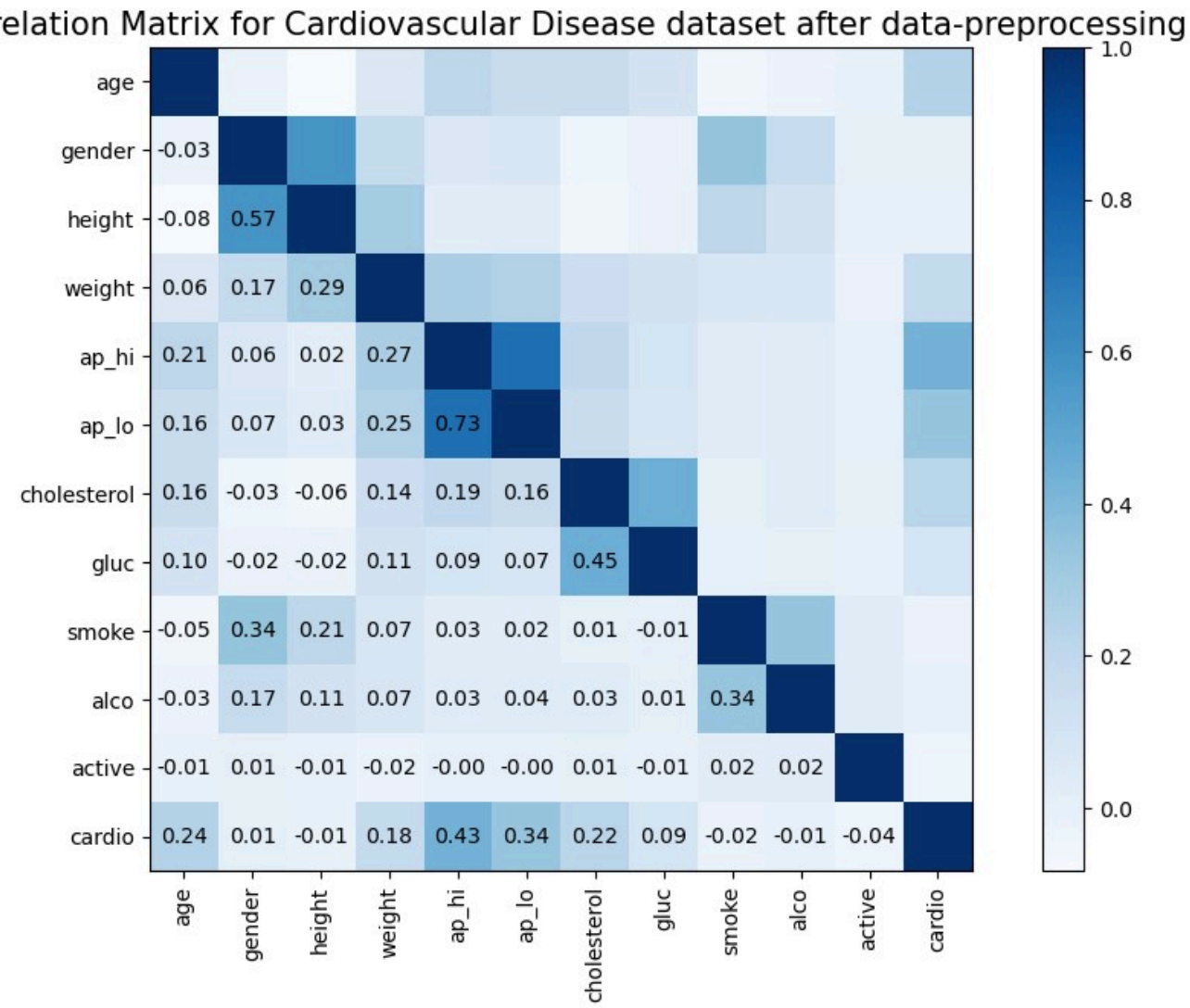


-We combined height and weight to calculate BMI to further evaluate the validity of our dataset. Most BMI reports don't record BMI that is higher than 60 so we excluded those data from as well.

3. ***Standardization of Values inluding BP, Height, Weight***

-Certain features exhibit a broad range of values, such as blood pressures, which could dominate the classification algorithm. To balance their impacts on distance calculations, we plan to scale these features with mean of 0 and stdev of 1, utilizing the scikit-learn library

After data-processing, the correlation matrix is as below. Since there are only 12 features, we decided to keep them all.



Upon examining the correlation matrix between features, it shows that the majority do not exhibit high correlation, suggesting our dataset is complex and potentially contains both linear and non-linear relationships. Given the dataset's size of 70,000 entries and its complexity, the Random Forest algorithm initially seemed like the ideal choice for our project. However, we still want to explore other, simpler algorithms to verify whether our initial assumption about the dataset's large size and complexity is accurate. Two algorithms under consideration are logistic regression, which can serve as a basic benchmark but assumes the linear relationship between features, and SVM, which is effective for high-dimensional spaces but slow on large datasets. We chose to implement logistic regression first to achieve a general baseline for performance that we could compare other models to. We then proceeded to have working implementations of our other two proposed supervised learning models: SVM and Random Forest. We found that SVM currently has a much poorer performance than both logistic regression and random forest, and we plan to improve the performance of the SVM algorithm by tuning parameters and experimenting with different kernels. We found that logistic regression and random forest return very similar results, which will be discussed further in the next section. We chose our supervised learning models carefully, specifically choosing logistic regression to obtain a good baseline for performance, choosing SVM because we have a high-dimensional feature space, and choosing random forest because of

its flexibility and ability to model nonlinear relationships. We have also brainstormed which unsupervised models we will use. We have already used PCA to reduce the dimensionality of the feature space, but we have used this solely for visualization up to this point, rather than using it to simplify our data. Therefore, we are planning to employ PCA in the future more directly. We are also considering using k-means clustering, specifically with two clusters, since this is a binary classification problem.

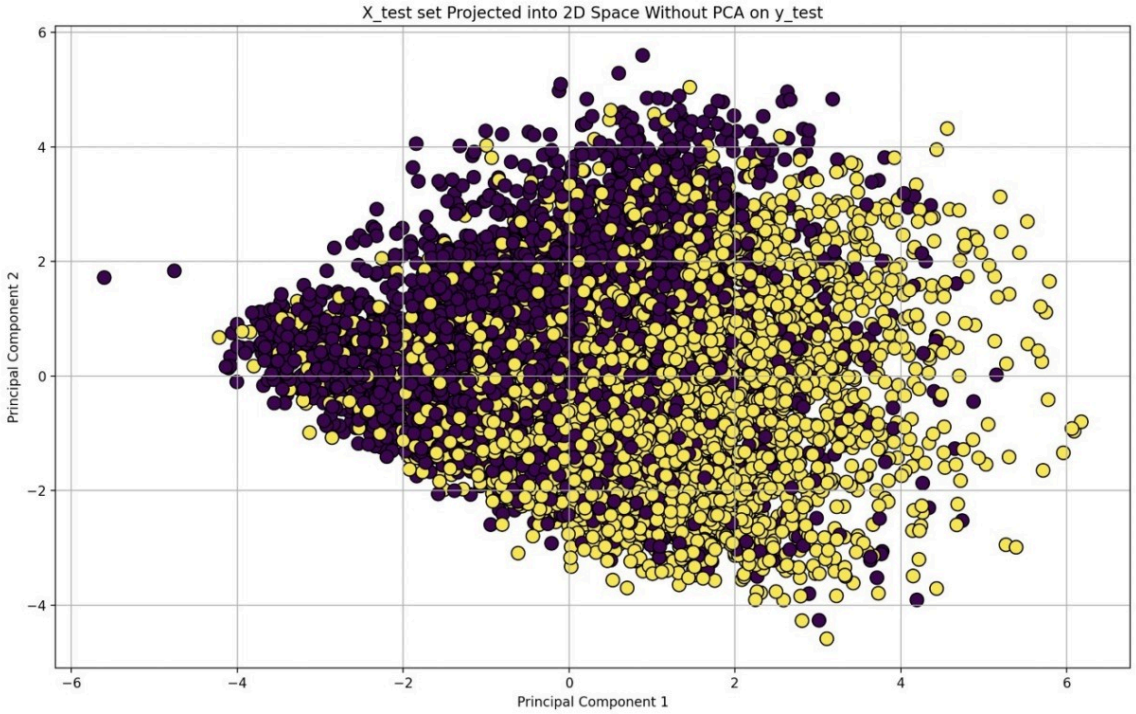
Results and Discussion

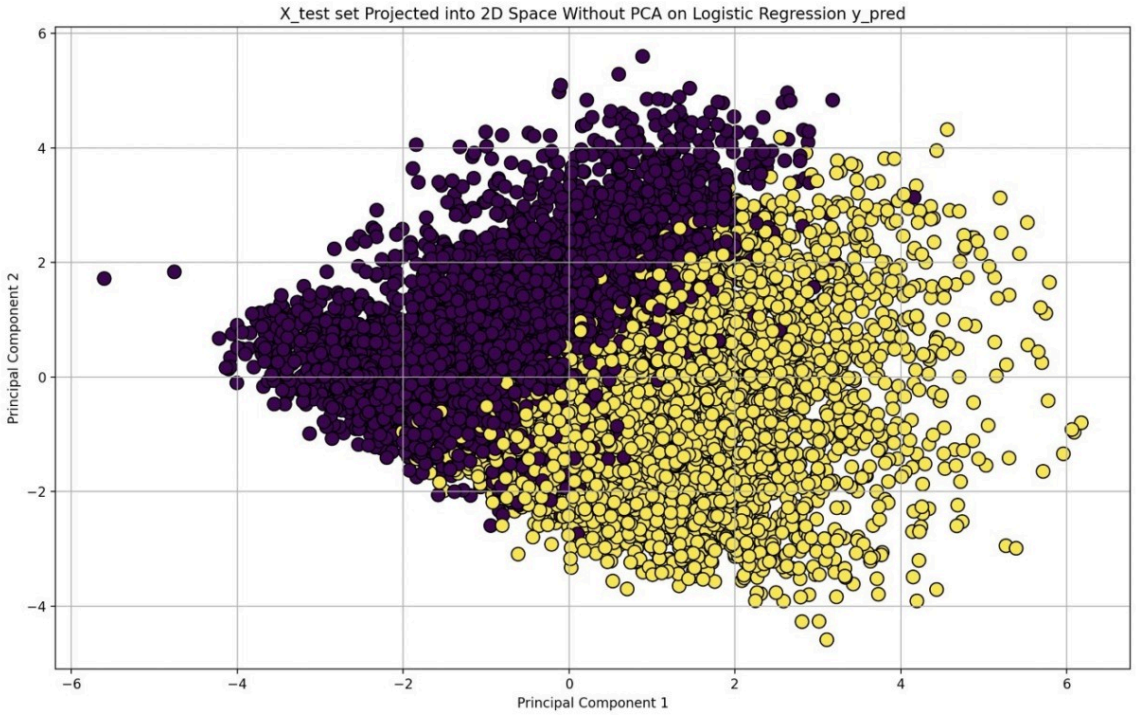
For this project, we will evaluate our ML model using several quantitative metrics.

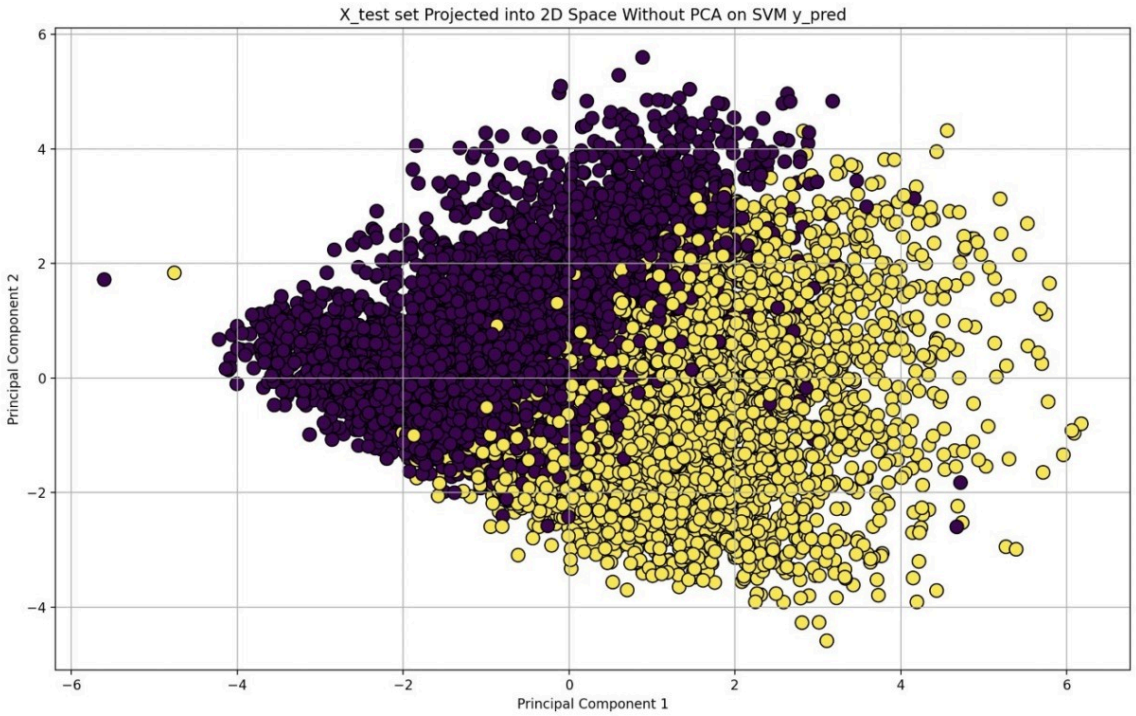
- **Accuracy** -To measure the proportion of true results (true positives & true negatives) among the total cases. Our goal is to achieve accuracy exceeding 85%, which would indicate strong model performance. Furthermore, we are looking to measure false positive and false negative counts, since those values are crucial when providing a medical diagnosis.
- **F1 Score** -To balance precision and recall. Using the F1 score would help us minimize false positives while accurately identifying patients with cardiovascular disease. We aim for an F1 Score of at least 0.75.
- **Receiver Operating Characteristic**
  -To assess the model's ability to accurately distinguish between patients with and without cardiovascular disease. We aim for our mode to have a ROC-AUC score above 0.80.

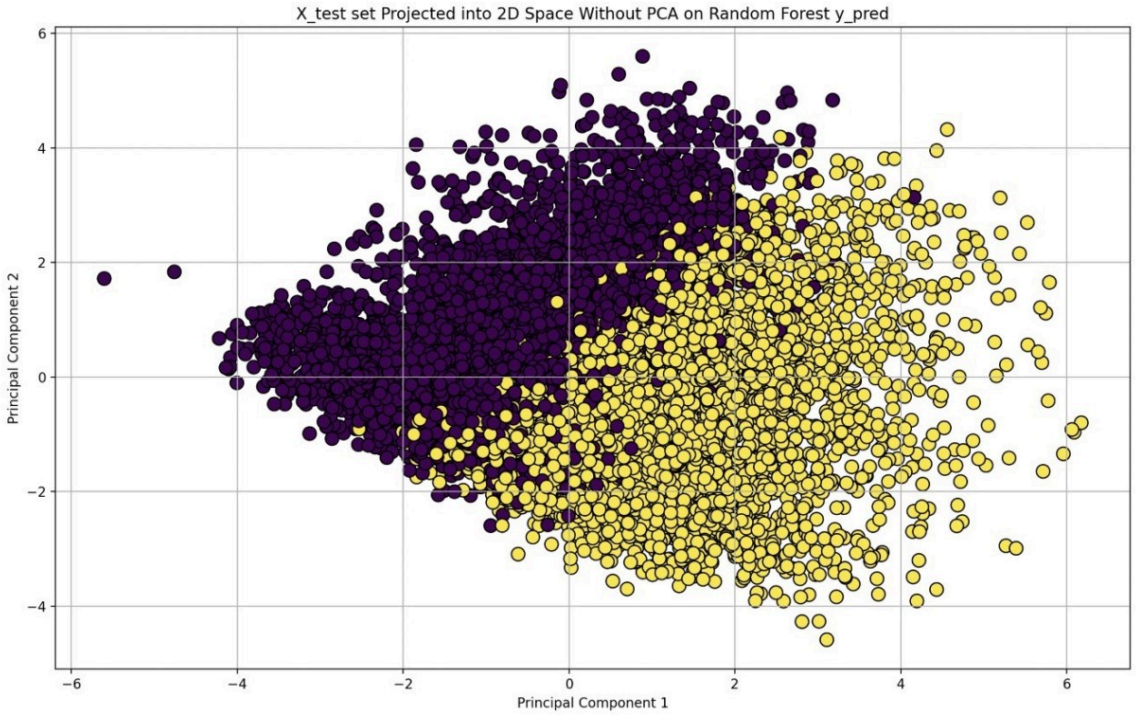
Model Results

	Models	Accuracy	F1	ROC-AUC
0	Logistic Regression	0.72	0.71	0.78
1	Random Forest	0.73	0.71	0.79
2	SVM	0.72	0.69	0.5


X_test set Projected into 2D Space Without PCA on y_test
Principal Component 2
Principal Component 1
6
4
2
0
−2
−4
−6
−4
−2
0
2
4
6


X_test set Projected into 2D Space Without PCA on Logistic Regression y_pred
Principal Component 2
Principal Component 1
6
4
2
0
−2
−4
−6
−4
−2
0
2
4
6

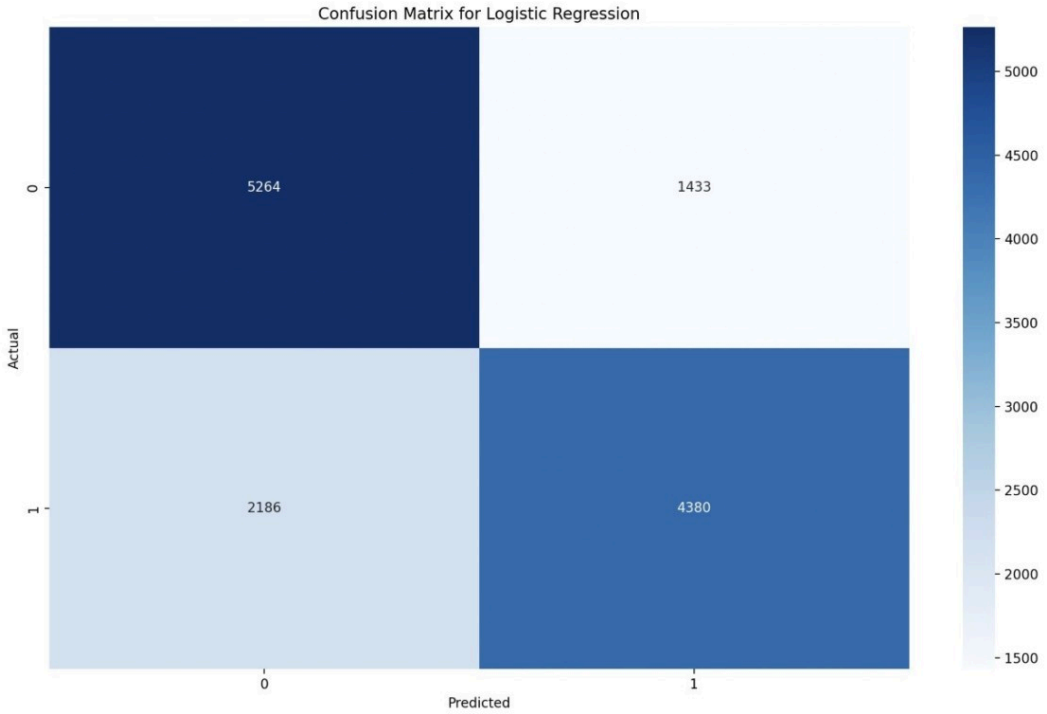


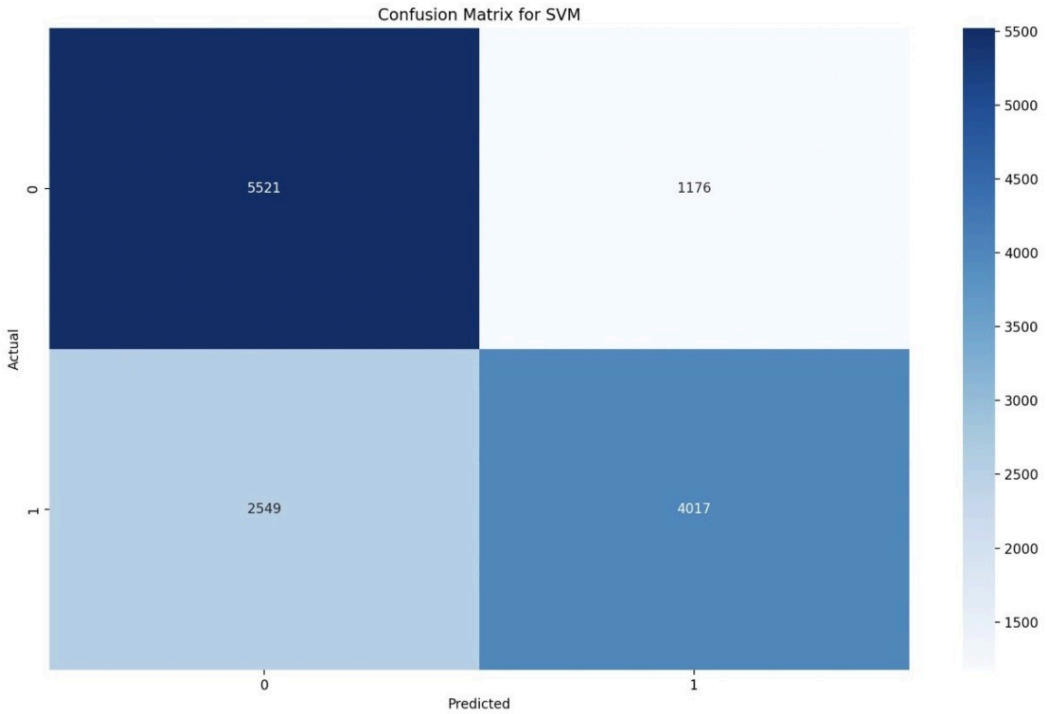


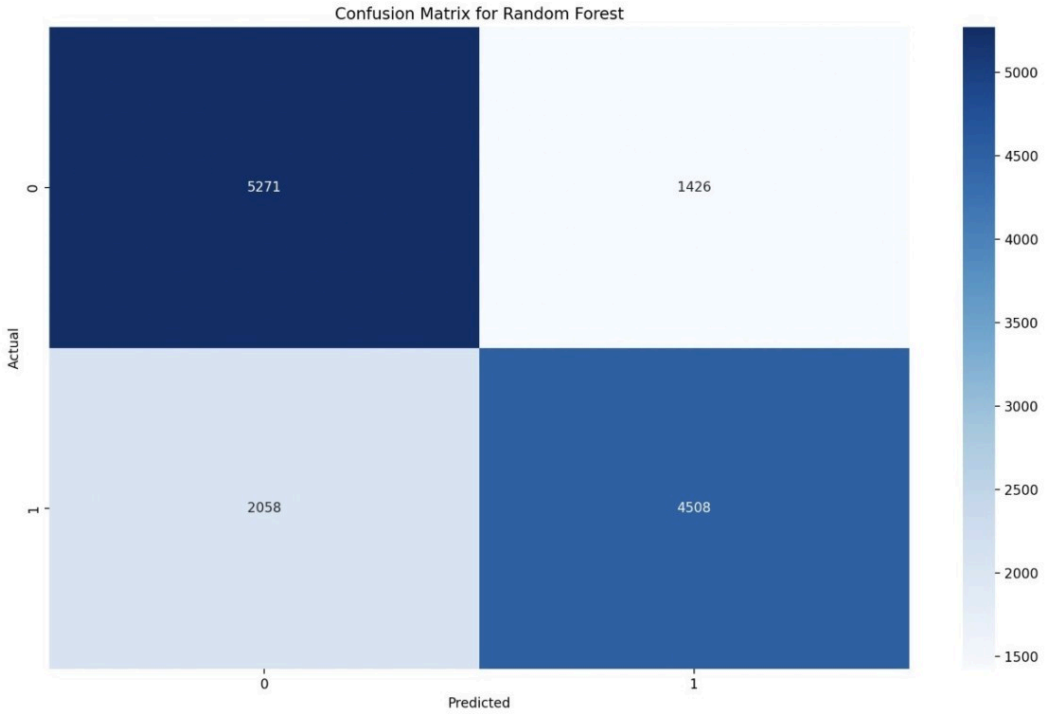
For the supervised algorithms we have implemented thus far, we have unfortunately not been able to reach the goals we listed above. On our midterm report, our SVM implementation did perform far worse than the other two models, with an accuracy score of 0.53, an F1 score of 0.64, and an ROC-AUC score of 0.59. We believed that these results were to the implementation of our SVM algorithm that has not quite been optimized yet. So, we went ahead and optimized it. We changed the kernel to be polynomial,

because the earlier kernel was not capturing any non-linear relationships in the data. We experimented with differing degrees but found 3 to be the best. The accuracy improved to an accuracy score of 0.72, an F1 score of 0.69, and an ROC-AUC score of 0.5.

Our logistic regression and random forest models also returned encouraging results. Our logistic regression model returned an accuracy score of 0.72, an F1 score of 0.71, and an ROC-AUC score of 0.78. We tried various methods to increase these scores such as trying different regularization constants and changing the maximum number of iterations, but the accuracy wouldn't budge. Our random forest model returned very similar results, specifically an accuracy score of 0.73, an F1 score of 0.71, and an ROC-AUC score of 0.79, and again tweaking hyperparameters such as the maximum depth and the number of decision trees didn't improve performance. We believe this is because even though each model can create decision boundaries, there is always a section of overlapping values for each model, which can be seen in the photos above. This puts a ceiling on our accuracy, which is further evidenced by the fact that all three models returned almost the exact same accuracy.





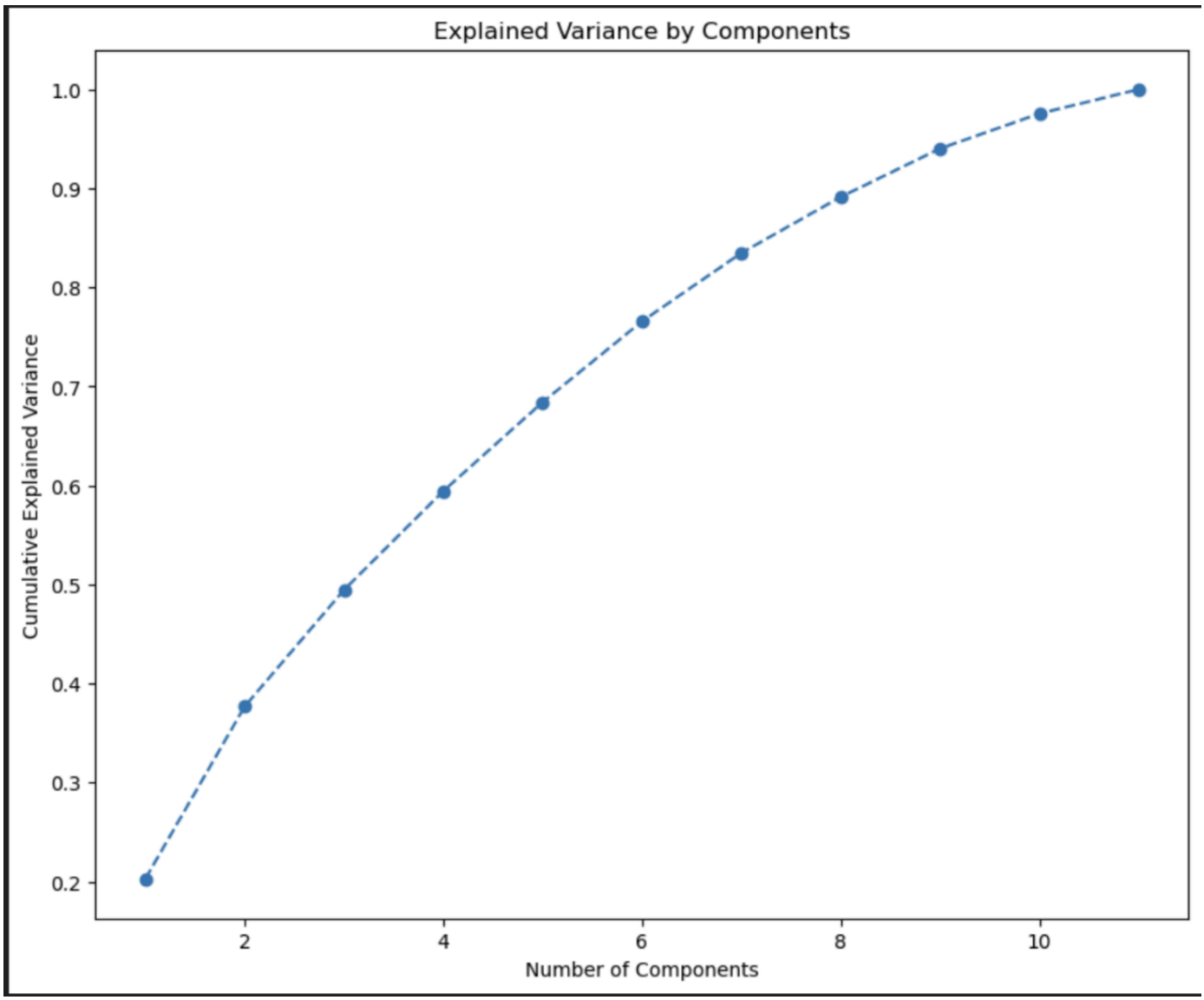


The above photos represent the confusion matrices for each of the three supervised learning models. As can be seen, the results are quite similar across the three models, particularly between logistic regression and random forest. Once again, this is likely due to the fact that the decision boundaries are very similar for each of the three models, and the same data points seem to cause confusion. SVM seems to be more

willing to classify data points as positive since it has more correct true positive judgments, but also more false positive judgments.

Integrating PCA with each classifier

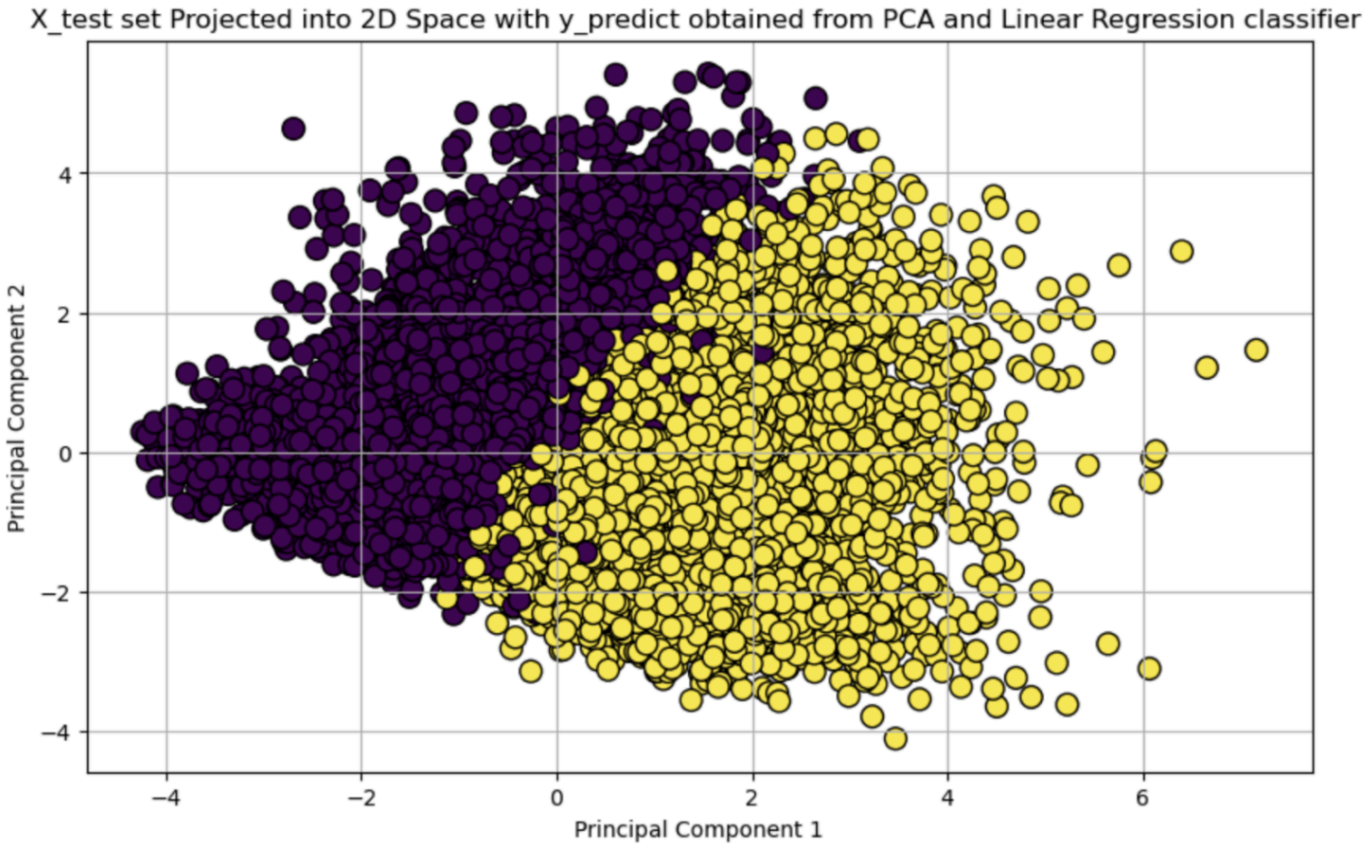
We tested PCA integrated with each classifier for the purpose of denoising. Based on the variance graph for each component, we chose to retain six principal components, as this selection accounts for 80% of the variance.

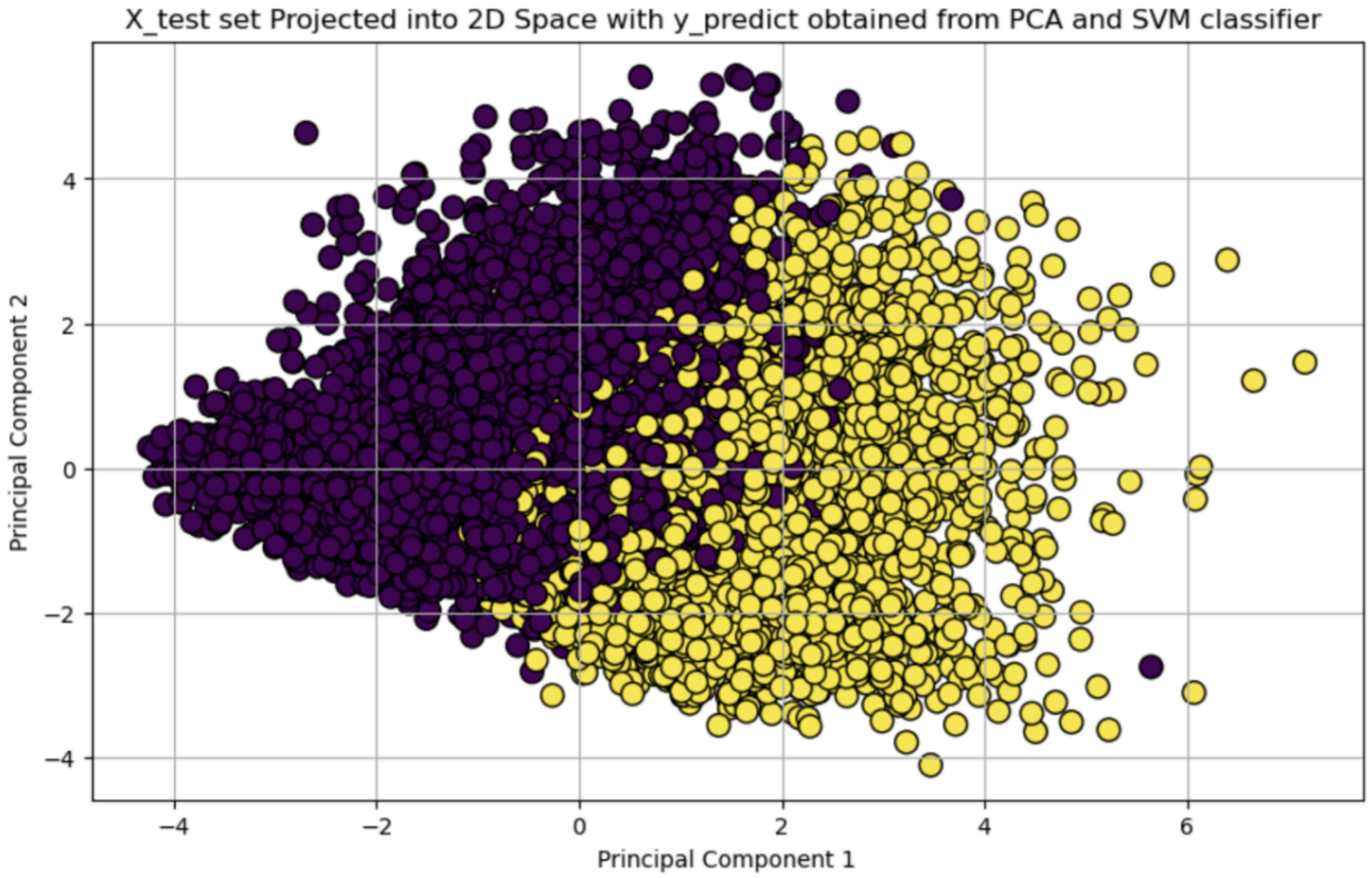


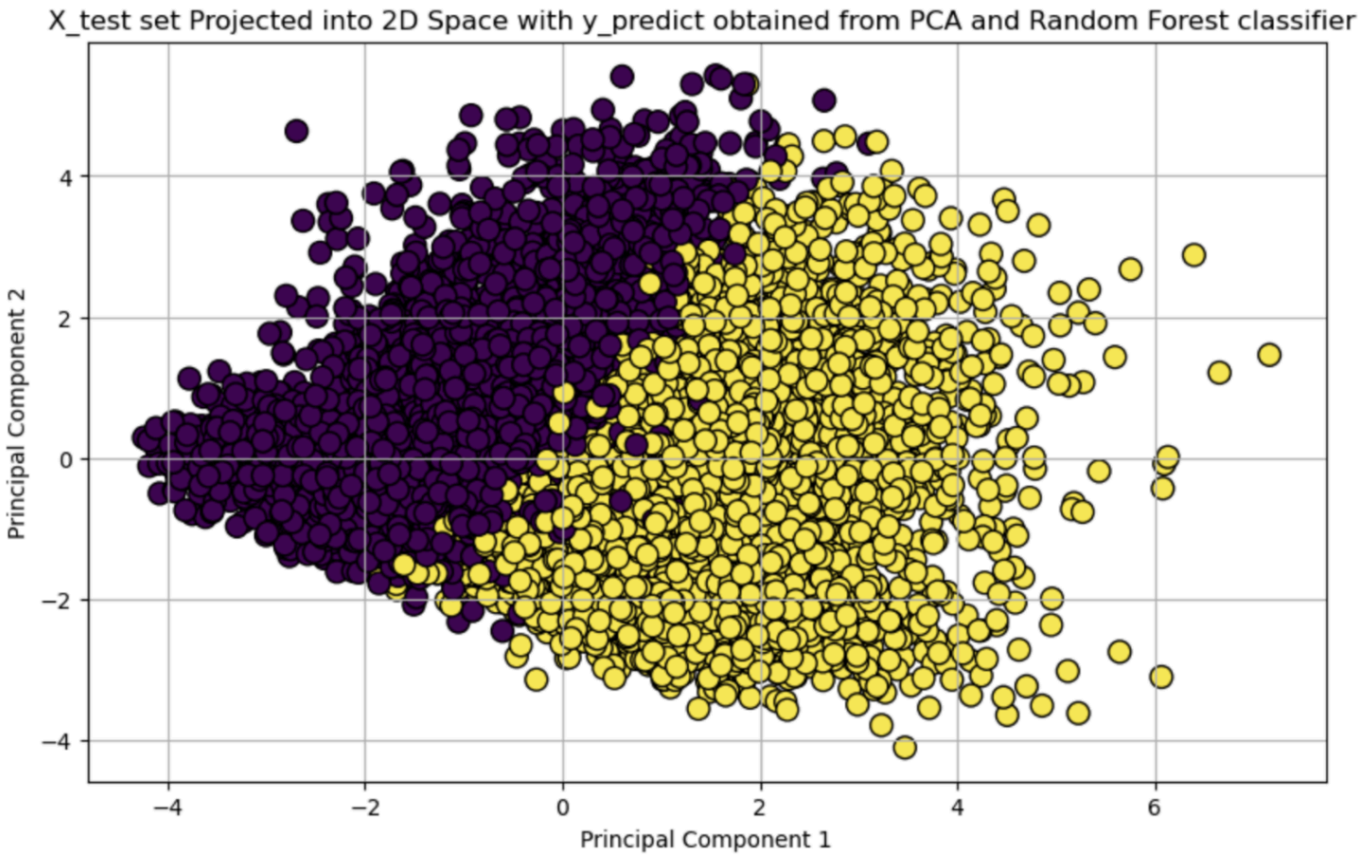
Model Results with Integrated PCA

	Models	Accuracy	F1	ROC-AUC
0	Logistic Regression	0.72	0.7	0.78
1	Random Forest	0.67	0.6	0.5
2	SVM	0.72	0.7	0.78

However, using six components with PCA, our models did not improve and even performed slightly worse than the trained models without PCA (as shown in the table above). We believe this is because, when visualizing the original data in a 2D space using PCA (Figure 1), there are not many points identified as outliers. No distinct clusters of points are evident as outliers. Therefore, when using PCA to remove noise, there would not be much change in the data. Below are the graphs of each model when integrated with PCA. Compared to Figure 2, 3 and 4, the decision line is consistent across all figures, further confirming the same results obtained.




X_test set Projected into 2D Space with y_predict obtained from PCA and SVM classifier
Principal Component 2
Principal Component 1
4
2
0
−2
−4
−4
−2
0
2
4
6


X_test set Projected into 2D Space with y_predict obtained from PCA and Random Forest classifier
Principal Component 2
Principal Component 1
4
2
0
−2
−4
−4
−2
0
2
4
6

References

1. **A. A. Ahmad and H. Polat**, “Prediction of heart disease based on machine learning using jellyfish optimization algorithm,” *Diagnostics (Basel, Switzerland)*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10378171/#::text=The%20results%20showed%20that%20RF,patients%20more%20accurately%20than%20other%20algorithms (accessed Oct. 3, 2024).

2. **World Health Organization**, “Cardiovascular diseases,” *World Health Organisation*, 2024. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

3. **U.S. Census Bureau**, “Health and Nutrition 135 Table 205. Cumulative Percent Distribution of Population by Height and Sex: 2007 to 2008,” 2010. Available: https://www2.census.gov/library/publications/2010/compendia/statab/130ed/tables/11s0205.pdf

4. **American Heart Association Obesity Committee**, “Obesity and cardiovascular disease: A scientific statement from the American Heart Association,” *Circulation, 143*(21), e984–e1010. https://doi.org/10.1161/CIR.0000000000000973

Gantt Chart

Here is the Gantt Chart: Group 39 Gantt Chart

Contribution Table

	Name	Proposal Contributions
0	Quyen Tran	Data Cleaning/PCA
1	Varun Chandrashekhar	Model Development/Optimization
2	Aryan Shah	Model Development/Discussion
3	Aria	Data Cleaning/Deployment
4	Rugved	Model Development/Optimization
5	Everyone	Research/Model Dev

Developed with ❤️ using Streamlit