2023

Factors affecting Cardiovascular Disease

DS 861 – DATA MINING & ADVANCED STATISTICAL METHODS FOR BUSINESS ANALYSTS

PROF. MINH PHAM

PREPARED BY:

DHARA PARMANAND KHUBCHANDANI

PRAMOD MANTYA RAJU

Table of Contents

Dataset and Project Objective

We examined a health-related dataset containing respondent information such as ID, age, gender, height, weight, blood pressure readings (systolic and diastolic), cholesterol levels, glucose levels, smoking habits, alcohol intake, physical activity, and cardiovascular status for this study. The collection contains 70,000 instances, each with a unique ID ranging from 0 to 69999. Age, height, weight, and blood pressure are numerical qualities, whereas categorical factors such as gender, cholesterol level, glucose level, smoking, alcohol intake, physical activity, and cardiovascular status are encoded with unique numerical codes.

The goal of this investigation is to look for patterns and linkages in the data in order to get insight into potential factors influencing cardiovascular health. The richness of the dataset, which includes both numerical and categorical variables, provides a solid foundation for training and testing machine learning models. Notably, the cardio variable in the dataset acts as the response variable in a binary classification context, where 0 represents the lack of a heart problem and 1 represents the presence of a heart problem.

Data preprocessing, exploratory research to understand the distribution of variables and potential relationships, and the application of various machine learning models for classification tasks are all part of our technique. This complete approach enables us to evaluate the forecasting capacity of various models and make meaningful comparisons between them.

Dataset Description

Each entry in the dataset contains the following information about an individual.

Type	Feature	Description	Values/Ranges
Numerical	ID	Respondents ID	0 to 69999
Numerical	Age	Age of the individual	Positive integers
Categorical	Gender	Gender of the individual	1 for Male, 2 for Female

Numerical	Height	Height of the individual	In centimeters
Numerical	Weight	Weight of the individual	In kilograms
Numerical	Ap_hi	Systolic blood pressure	Integer values like 110, 140
Numerical	Ap_lo	Diastolic blood pressure	Integer values like 70, 80
Categorical	Cholesterol	Cholesterol level	1, 2, or 3 (1=normal, 2=above normal, 3=way above normal)
Categorical	Gluco	Glucose level	1, 2, or 3 (1=normal, 2=above normal, 3=way above normal)
Binary	Smoking	Smoking habits	0 = does not smoke, 1 = does smoke
Binary	Alcohol	Alcohol consumption	0 = does not consume alcohol, 1 = consumes alcohol
Binary	Active	Physical activity	0 = is not physically active, 1 = is physically active
Binary	Cardio	Cardiovascular status	0 = does not have a heart problem, 1 = does have a heart problem

Purpose: (Why is it interesting? How is it relevant to our class?)

The "Cardio Train" dataset is an exciting focus for our class since it provides a compelling exploration into the world of cardiovascular health prediction. This dataset is very interesting because it includes the use of many categorization algorithms. Logistic Regression was chosen to model the binary target variable Cardio (0 = no heart problem, 1 = has a heart problem) because it provides a clear and interpretable understanding of the correlations between selected parameters and cardiovascular health.

Ensemble approaches such as Random Forest and Gradient Boosting emerge from the dataset's rich feature landscape. Random Forest, which is adept at dealing with the high dimensionality and intricate feature interactions found in health and demographic data, provides useful insights.

The diverse data types in the dataset highlight the importance of decision trees. Decision trees, which are well-known for managing numerical and categorical data, provide interpretable results and can accept non-linear relationships. Because of this versatility, decision trees are an effective tool for building a model that not only predicts cardiovascular health but also provides critical insights into major attributes and decision routes that influence these predictions.

In conclusion, the strategic use of decision trees and ensemble methods on the "Cardio Train" dataset is perfectly aligned with our class objectives. This approach enables us to obtain interpretable and meaningful results from the examination of demographic and health data, so contributing to a thorough grasp of cardiovascular health prediction within the framework of our coursework.

Literature Review:

As a part of the literature review, we went over three articles from different researchers. These researchers have been doing the research in medical domain for a long time.

The first article we went over was about the research done by Author William Kannel, who is an MD. He has done 36 years, almost 4 decades, of research on Blood pressure as a cardiovascular disease factor. He says Hypertension is one of the most prevalent and powerful contributors to cardiovascular diseases, the leading cause of death in the United States. He has also mentioned in his paper that there is, on average, a 20 mm Hg systolic and 10 mm Hg diastolic increment increase in blood pressure from age 30 to 65 years. (William B. Kannel, 1996)

The second paper we went through was from Edward Lakatta who is also an MD. He has done 16 years of research on 'Age-Associated Cardiovascular changes in Health and its Impact on Older persons'. In finding the conclusions, he has used linear regression models. According to his research, there is an age-associated increase in vascular afterload on the heart which is due to arterial stiffening and is reflected in the age-associated modest increase in systolic blood pressure at rest. (MD, 2002)

The third research paper we went through was on 'Obesity and Cardiovascular disease' by multiple researchers – Carl, Richard, and Hector. As a part of research, they studied approx. 29000 individuals for 5 years. Their research says that "Obesity has a major impact on CV diseases, such as heart failure (HF), coronary heart disease (CHD), sudden cardiac death, and atrial fibrillation, and is associated with reduced overall survival" (Carl J. Lavie, 2009)

Data Preprocessing:

I. Missing values:

Once we imported the dataset, we tried to check if our dataset had any missing values to work on. After checking, we observed from the output that our dataset does not have any missing values.

II. Feature Engineering:

We added two attributes to the dataset – Age_Years and BMI. Age was in the form of days or integers in the dataset. That format did not help in doing the analysis. To do the analysis in a better way, we added age in years by dividing the given age by 365. Also, we added BMI as we believe it affects the decision for someone having or not having a cardiovascular disease. BMI was calculated by dividing the weight by the square of height.

III. Dealing with the outliers:

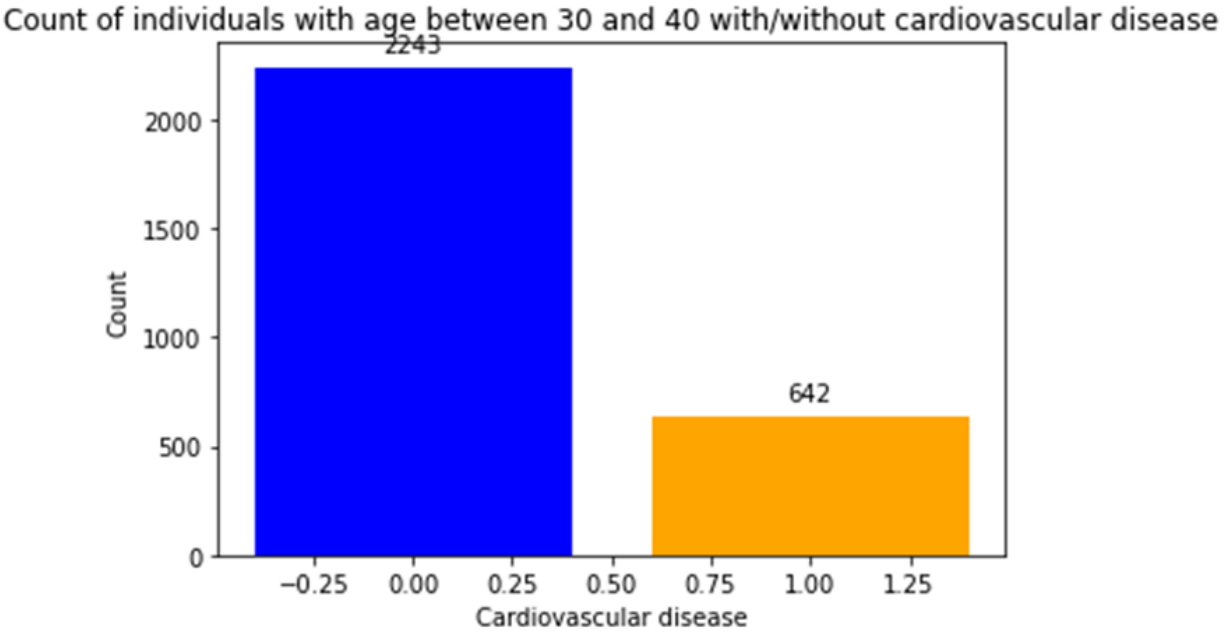
We initially tried to plot box and whisker plots for some attributes where we observed the graphs were not showing up properly due to outliers in some of the attributes. So, we removed the outliers from our dataset from the attributes – Height, Weight, Systolic blood pressure and Diastolic blood pressure. We removed the samples that were beyond the 97.5^{th} or 2.5^{th} percentile. After removing the outliers, our dataset was reduced from 70000 samples to 60142.

IV. Creating Dummy Variables:

For the categorical attribute cholesterol level, we created dummies by using the get_dummies function from the pandas library. So, we created 3 columns for each cholesterol level. After creating the dummy variables, we dropped the original 'cholesterol' column to replace it with its binary representations, making it suitable for machine learning algorithms.

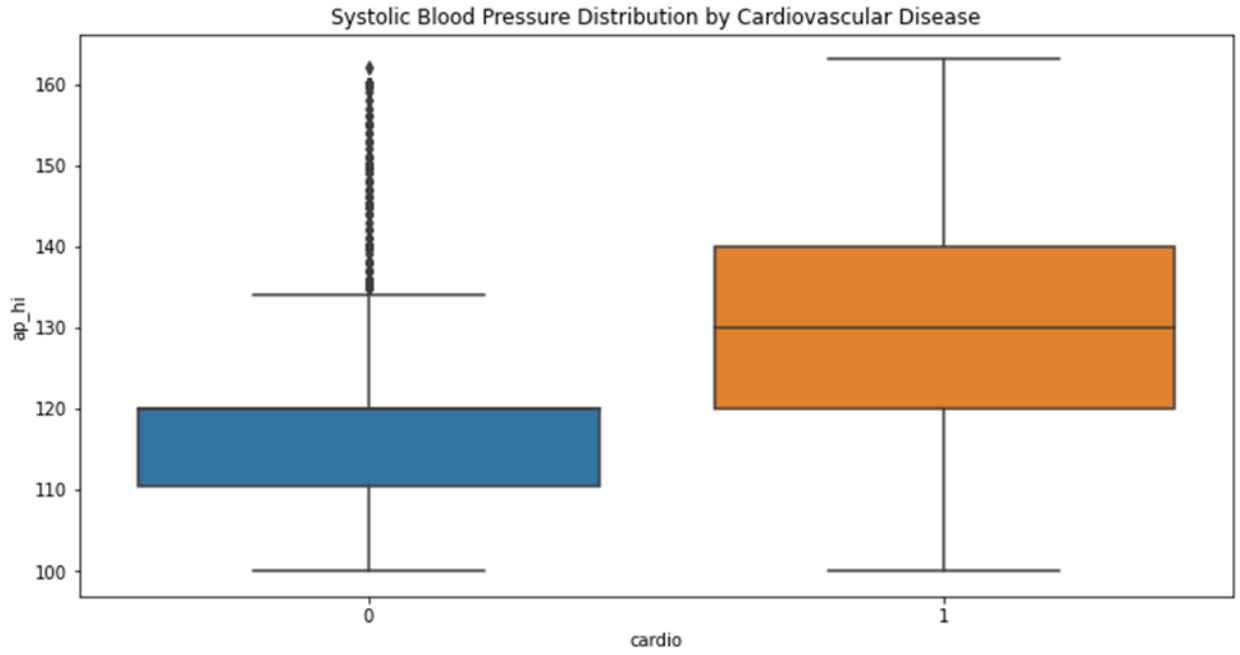
Exploratory Data Analysis:

1. Individuals between 30 to 40 years who have or do not have cardiovascular disease.



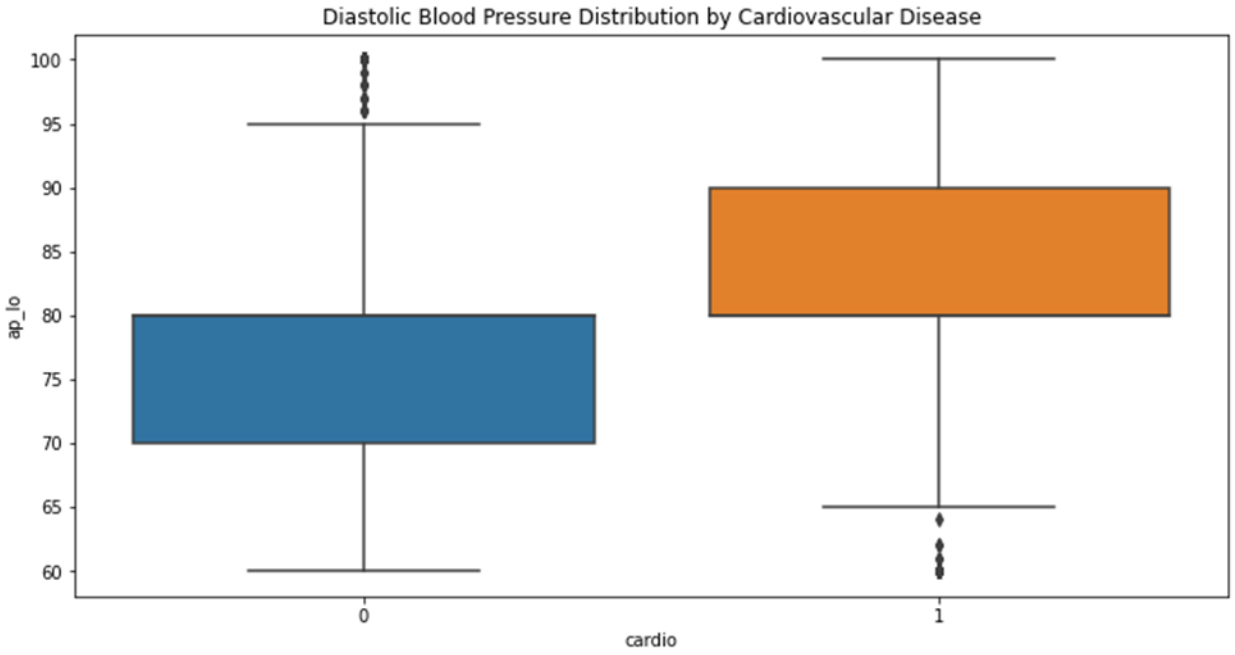
In our dataset, there are 642 individuals between the age of 30 to 40 years who have cardiovascular disease. Out of a total 2885 (2243 + 642) individuals, 642 individuals had cardiovascular disease. This means approximately 1 out of 5 individuals between the age group of 30 to 40 years have a cardiovascular disease.

2. Range of systolic blood pressure reading for individuals with or without cardiovascular disease.



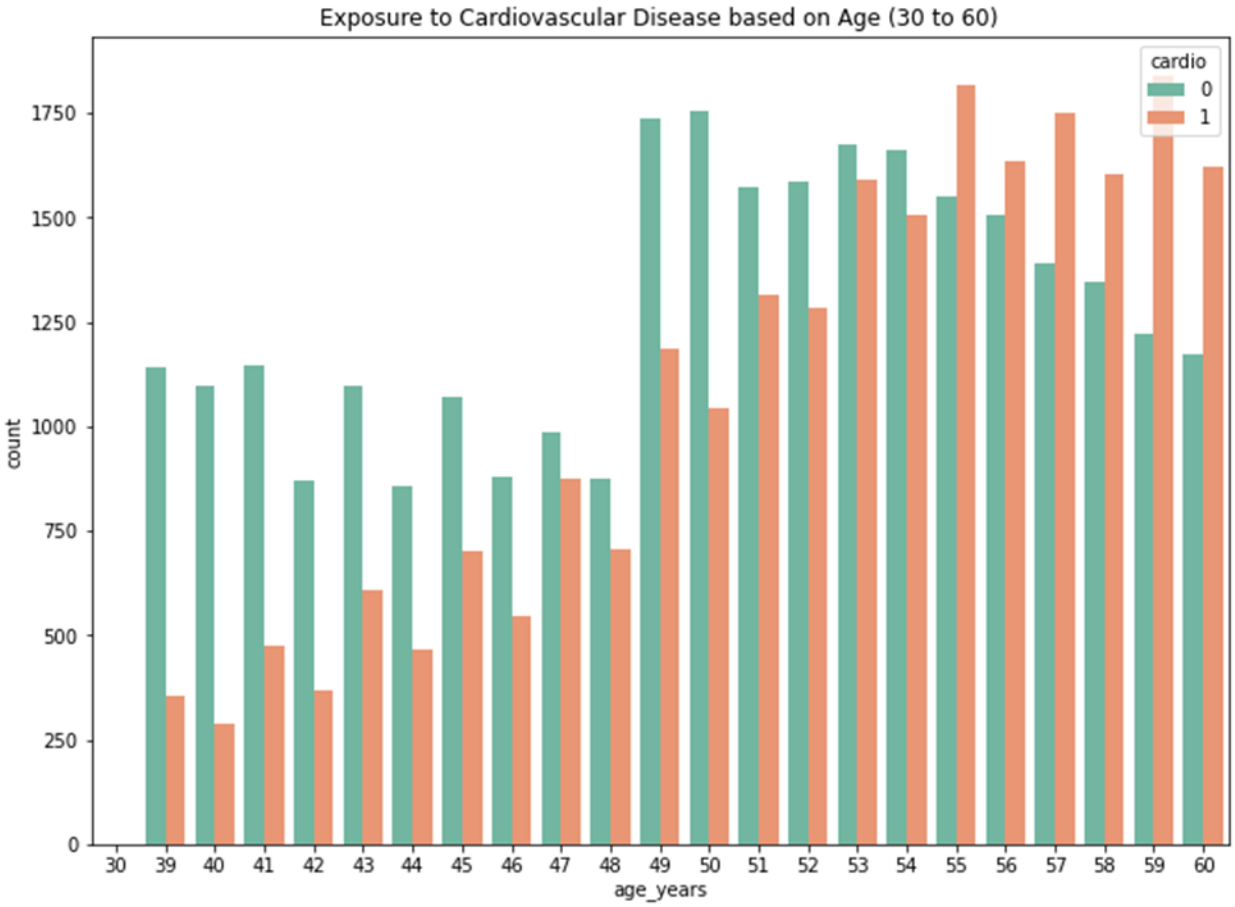
The box and whisker plot shows that the systolic blood pressure reading for individuals who have cardiovascular disease is more than the idea reading of 120. Their reading is between 120 to 140 which is much higher. Another observation is the normal readings for the individuals who do not have cardiovascular disease – their readings are between 110 to 120. For the box and whiskers of the individuals who do not have cardiovascular disease, there are some outliers. These might be the individuals who might have some early cardiovascular diseases which are not yet detected or they could be in the early age group.

3. Range of diastolic blood pressure reading for individuals with or without cardiovascular disease.



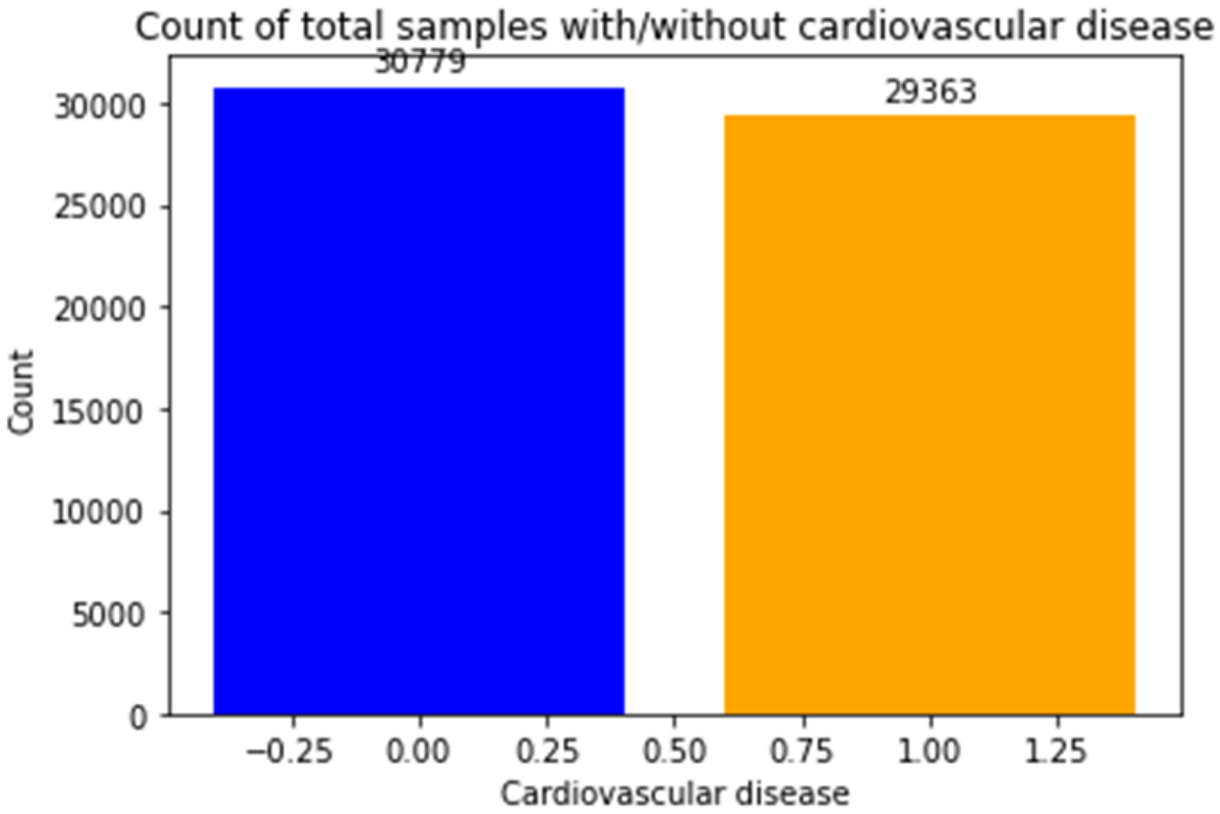
The box and whisker plot shows that the diastolic blood pressure reading for individuals who have cardiovascular disease is more than the idea reading of 80. Their reading is between 80 to 90 which is much higher. Another observation is the normal readings for the individuals who do not have cardiovascular disease – their readings are between 70 to 80. For the box and whiskers of the individuals who do not have cardiovascular disease, there are some outliers. These might be the individuals who might have some early cardiovascular diseases which are not yet detected, or they could be in the early age group.

4. Exposure to Cardiovascular disease between age 30 to 60



We can observe from the bar graph that after 55 years of age, there are more individuals with cardiovascular disease than the individuals without cardiovascular disease (orange bars are taller than green). We can also observe the trend that there is an overall increase in orange bars from 30 years to 60 years. If we also think about the dataset from this graph, we can say that the dataset contains of more samples who are in the age group between 49 – 60 years.

5. Is the target variable balanced?



Once we dropped the outliers, we wanted to check if in the dataset, the target attribute is balanced. The bar graph shows that in our final dataset, we have almost the same number of samples for having/not having cardiovascular disease.

Models:

Logistics Regression:

The first model we decided to train our dataset on was Logistics regression. Since our target variable (cardio) is binary, we thought to start off with logistics regression.

For all the models that we have trained, we cross-validated by splitting the dataset into training and testing in the ratio of 70:30, and then standardizing the attributes. Once the model was trained, we tested the model and evaluated the performance of the model.

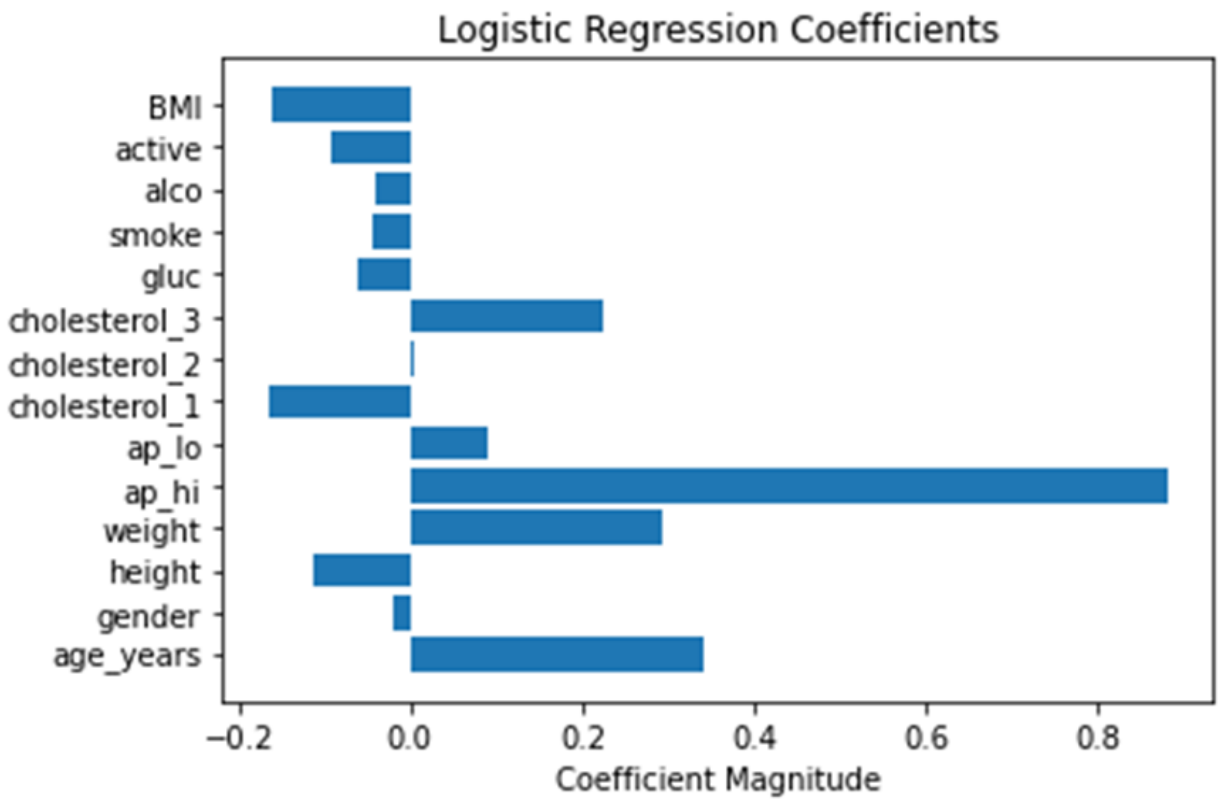
i) For Logistics regression, we are using statsmodel and sklearn (train_test_split, StandardScaler, and Logit). We wanted to train the model and see the coefficients. We also wanted to see which are the attributes which are insignificant as per Logistics Regression.

Logit Regression Results

Dep. Variable:	cardio	No. Observations:	42099
Model:	Logit	Df Residuals:	42085
Method:	MLE	Df Model:	13
Date:	Tue, 05 Dec 2023	Pseudo R-squ.:	0.1854
Time:	22:18:24	Log-Likelihood:	-23761.
converged:	True	LL-Null:	-29167.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0173	0.011	-1.540	0.124	-0.039	0.005
x1	0.3419	0.012	29.364	0.000	0.319	0.365
x2	-0.0188	0.013	-1.406	0.160	-0.045	0.007
x3	-0.1122	0.076	-1.485	0.138	-0.260	0.036
x4	0.2917	0.142	2.054	0.040	0.013	0.570
x5	0.8818	0.018	50.055	0.000	0.847	0.916
x6	0.0895	0.016	5.534	0.000	0.058	0.121
x7	-0.1659	4.07e+05	-4.08e-07	1.000	-7.98e+05	7.98e+05
x8	0.0027	3.21e+05	8.29e-09	1.000	-6.29e+05	6.29e+05
x9	0.2245	2.97e+05	7.56e-07	1.000	-5.82e+05	5.82e+05
x10	-0.0606	0.013	-4.631	0.000	-0.086	-0.035
x11	-0.0430	0.013	-3.429	0.001	-0.068	-0.018
x12	-0.0393	0.012	-3.253	0.001	-0.063	-0.016
x13	-0.0927	0.011	-8.377	0.000	-0.114	-0.071
x14	-0.1615	0.143	-1.127	0.260	-0.442	0.119

Since, the attributes that were printed does not have the attribute name, we plotted them on the bar graph:



We can see from the coefficients, for a unit increase in systolic blood pressure, the log-odds of the outcome to increase is 0.8818. For a unit increase in age, the log-odds of the outcome (presumably having cardiovascular disease) increases by 0.3419. For a unit increase in weight, the log-odds of the outcome increase by 0.2917. Thus the top 3 attributes having significant impact on having or not having a cardiovascular disease are – systolic blood pressure, age and weight.

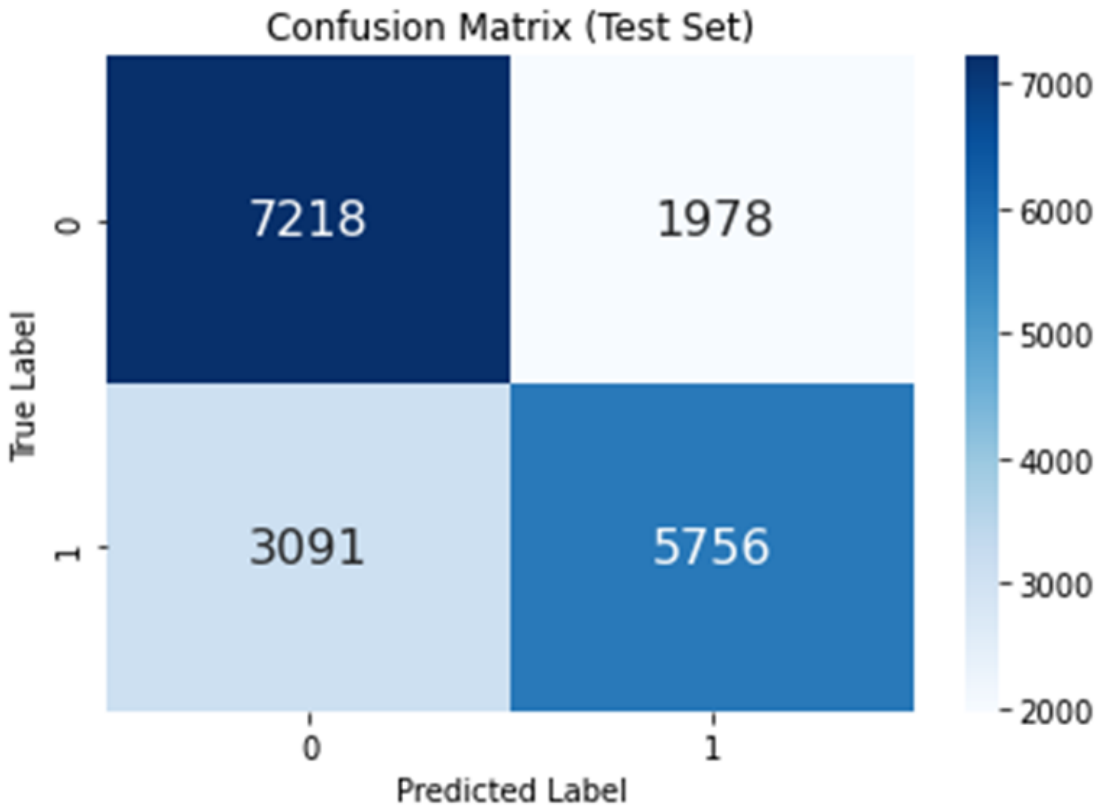
ii) We also printed the insignificant attribute:

```
Insignificant Feature Names:
['gender', 'height', 'cholesterol_1', 'cholesterol_2', 'cholesterol_3', 'BMI']
```

The insignificant features are the attributes with p-value greater than 0.05. It was interesting to see BMI as an insignificant feature and weight as a significant feature.

iii) Once the model was trained, we standardized the features for the final training data and then refitted the model. To evaluate the model's performance, we created a confusion matrix, calculated accuracy ratio, AUC and ROC visualization.

```
Confusion Matrix (Test Set):
[[7218 1978]
 [3091 5756]]
Accuracy: 0.7190600232777254
AUC (Test Set): 0.7830550426311134
```



· The top-left cell (7218) represents the number of true negatives (correctly predicted non-cardiovascular cases).

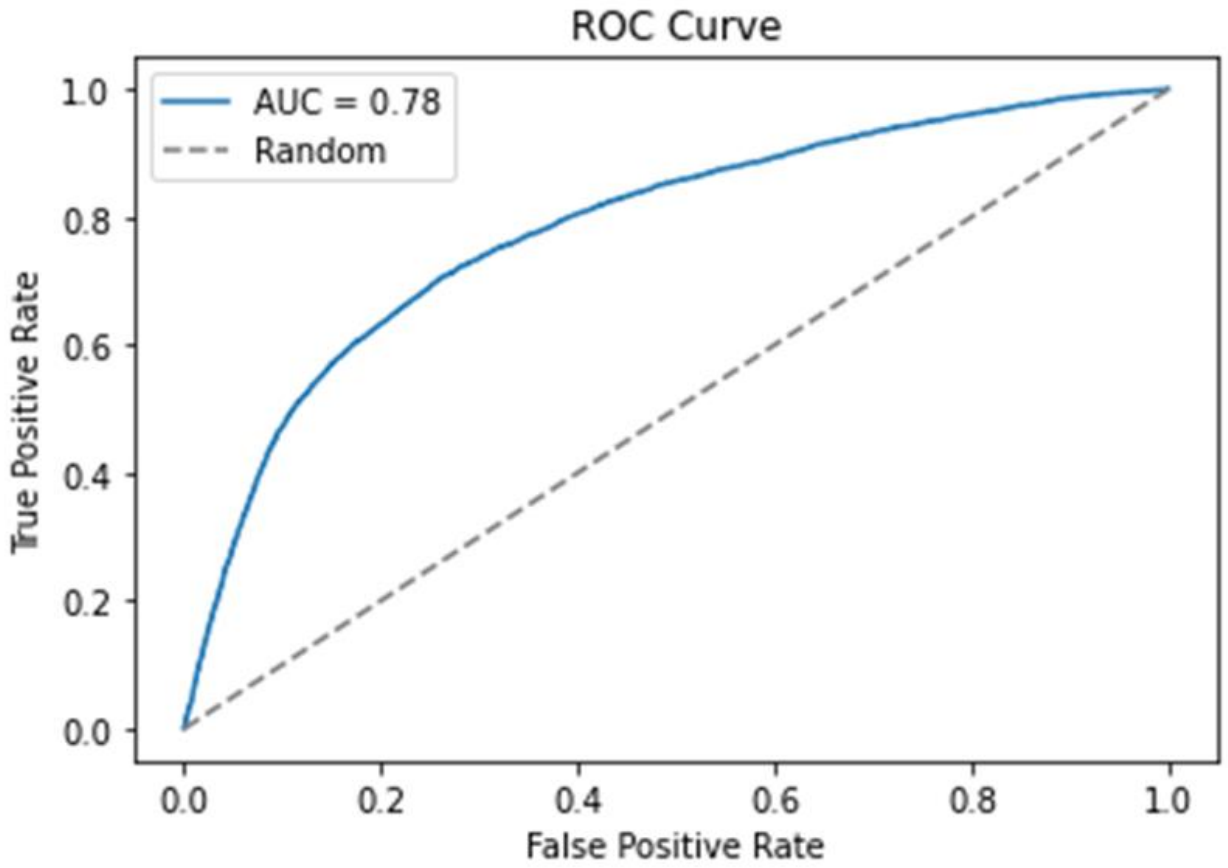
· The bottom-right cell (5756) represents true positives (correctly predicted cardiovascular cases).

· The top-right cell (1978) represents false positives (predicted as cardiovascular disease but actually non-cardiovascular).

· The bottom-left cell (3091) represents false negatives (predicted as non-cardiovascular but actually cardiovascular).

If we calculate Sensitivity (5756 / 5756+3091), it comes to 0.651. This means approximately 65.1% of the actual positive instances were correctly predicted.

The accuracy rate is 0.7190 meaning approximately 71.9% of all instances (both positive and negative) were correctly classified.

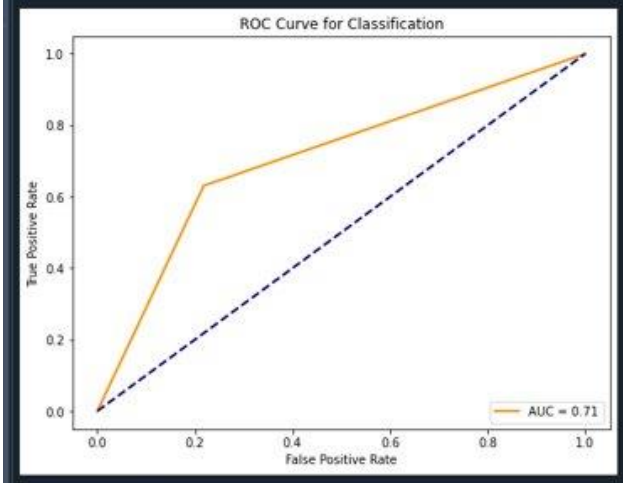


We can see the ROC curve is close to perfect classification or close to the top left corner which is a good sign. AUC is 0.78 which indicates a reasonably good ability of the model to discriminate between two classes.

KNN Model

```
Best K for Classification: 10

Misclassification Rate for Classification on Test Set: 0.29224630050435074
```




```
KNN Classification Model:
KNeighborsClassifier(n_neighbors=10)
```

Best K for Classification: 10

- To anticipate cardiovascular illness, our model, acting as a health counselor, compares it to 10 comparable cases (nearest neighbors). This optimal K value ensures a balanced approach that is neither either particular (overfitting) nor overly general (underfitting).

Misclassification Rate: 29.22%

- Our advice may be incorrect about 29.22% of the time when predicting whether someone has cardiovascular disease or not. The fraction of inaccurate predictions on the test set is represented by this misclassification rate.

In Simple Terms:

- The advisor does its best, but it may make a mistake in forecasting heart health 29.22% of the time. It's a useful tool, but it has a little margin for error.

ROC Curve and AUC Analysis

Understanding the ROC Curve and AUC:

- The ROC curve shows how well our counselor identifies between those with and without cardiovascular illness.

- AUC (Area Under the ROC Curve) quantifies the model's overall performance. Our advisor has a moderate capacity to differentiate between positive and negative cases, with an AUC of 0.71.

Model Performance:

- AUC: 0.71 - An AUC of 0.71 indicates that our advisor has a moderate ability to distinguish between people who have and do not have cardiovascular disease.

ROC Curve Interpretation: - The curve depicts the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) at different categorization thresholds.

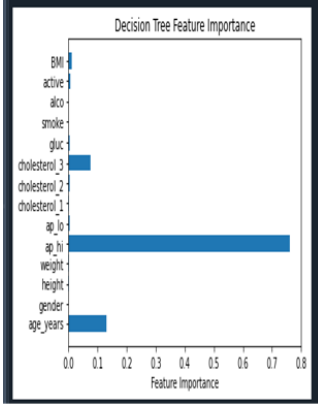
- The model performs better when the curve is closer to the upper-left corner.

Takeaway:

- Our adviser has a moderate ability to identify between patients with and without cardiovascular illness, as seen by the ROC curve and AUC of 0.71. While not particularly high, it nonetheless provides useful information for predicting heart health.

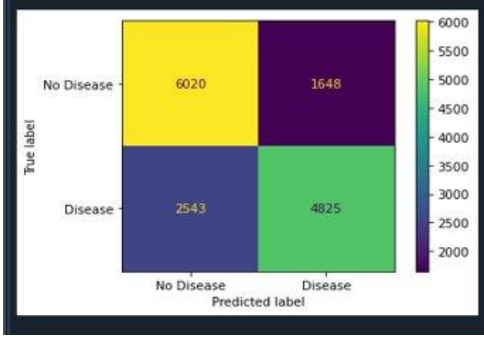
Decision Tree

```
Best Parameters for Classification: {'max_depth': 10, 'max_leaf_nodes': 30}
Accuracy on Test Set for Classification: 0.7212689545091779
```



	precision	recall	f1-score	support
0	0.70	0.79	0.74	7668
1	0.75	0.65	0.70	7368
accuracy			0.72	15036
macro avg	0.72	0.72	0.72	15036
weighted avg	0.72	0.72	0.72	15036

```
Confusion Matrix:
[[6020 1648]
 [2543 4825]]
```



1. Best Parameters for Classification: After experimenting with various combinations, the decision tree model discovered that it works best when it examines up to 10 levels of information and divides it into up to 30 separate groups to produce predictions.

2. Accuracy on Test Set for Classification: When you tested this decision tree on a portion of your data you hadn't shown it before, it correctly predicted whether individuals have cardiovascular disease or not about 72.13% of the time.

3. Classification Report:

- Precision: When the model predicts that someone has cardiovascular illness, it is correct roughly 75% of the time. When it predicts that someone does not have cardiovascular disease, it is approximately 70% correct.

- Recall : that the model detects approximately 65% of persons who have cardiovascular disease. It appropriately detects approximately 79% of persons who do not have cardiovascular disease.

- F1-score: The model receives an overall score of roughly 72 out of 100, taking into account both its accuracy and its capacity to identify persons with cardiovascular disease.

- Support: This is equivalent to stating how many people the model assisted for each category in your dataset.

4. Confusion Matrix:

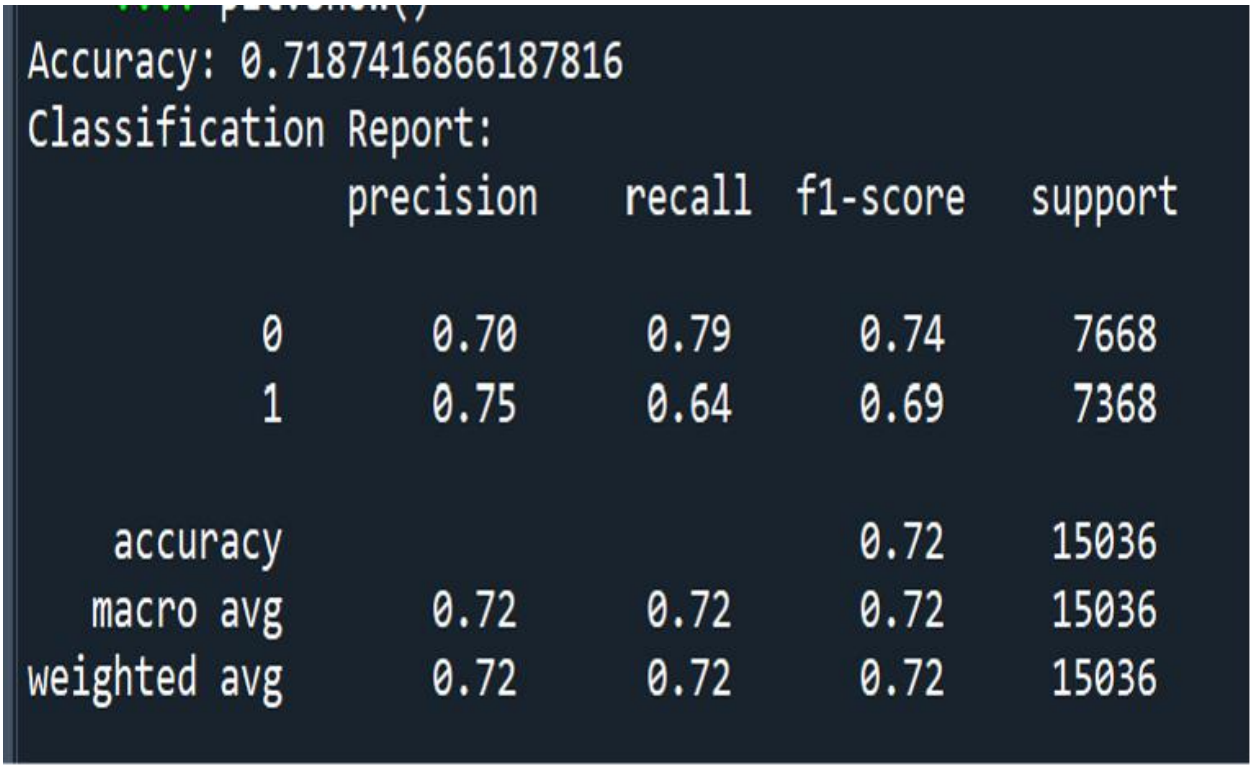
- True Negatives (6020): The model correctly identified people without cardiovascular disease.

- False Positives (1648): The model mistakenly thought some people had cardiovascular disease when they didn't.

- False Negatives (2543): The model missed some people who actually had cardiovascular disease.

- True Positives (4825): The model correctly identified people with cardiovascular disease.

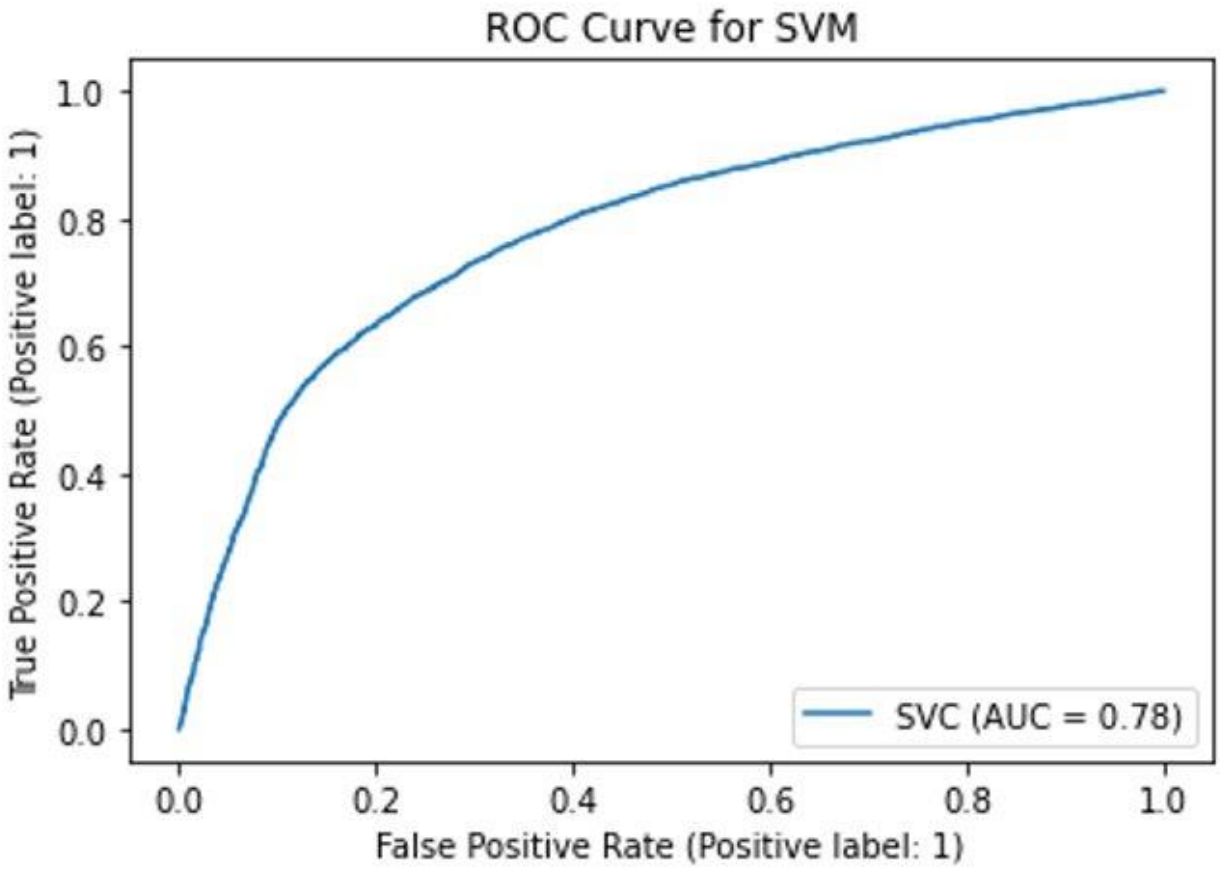
Support Vector Machine (SVM)



```
Accuracy: 0.7187416866187816
Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.79      0.74      7668
           1       0.75      0.64      0.69      7368

    accuracy                           0.72     15036
   macro avg       0.72      0.72      0.72     15036
weighted avg       0.72      0.72      0.72     15036
```



1. Accuracy (72%):

- This indicator reveals how accurate your model is overall. Your model successfully predicts the disease condition of 72 out of 100 people.

2. Precision for "No Disease" (70%):

Precision is a measure of the accuracy of optimistic forecasts. In this context, it means that 70% of individuals expected to be disease-free are indeed disease-free.

3. Precision for "Disease" (75%):

- This statistic, like the "No Disease" precision, focuses on the accuracy of positive forecasts. In this area, 75% of people projected to develop the condition really do.

4. Recall for "No Disease" (79%):

- Recall, also known as sensitivity, measures the model's ability to recognize positive cases properly. In this scenario, it means that your model properly detects 79% of all disease-free individuals.

5. Recall for "Disease" (64%):

- Recall shows the model's ability to properly identify individuals with the disease in the "Disease" class. The model properly recognizes 64% of them in this scenario.

6. F1-Score (Balanced Performance):

- The F1-Score is a balanced statistic that takes precision and recall into account. It enables a consistent evaluation of the model's performance. It is approximately 74% for "No Disease" and approximately 69% for "Disease."

7. Macro Avg (Overall Performance: 72%):

- The Macro Average computes the average performance across both illness and no disease classes. When both are averaged, the performance is roughly 72%.

8. Weighted Avg (Considering More Common Cases: 72%):

- The Weighted Average takes into account class imbalance, giving greater weight to more prevalent occurrences. In this scenario, even if more persons in your dataset are disease-free, you still get an overall accuracy of 72%.

Random Forest (with 5-fold cross validation)

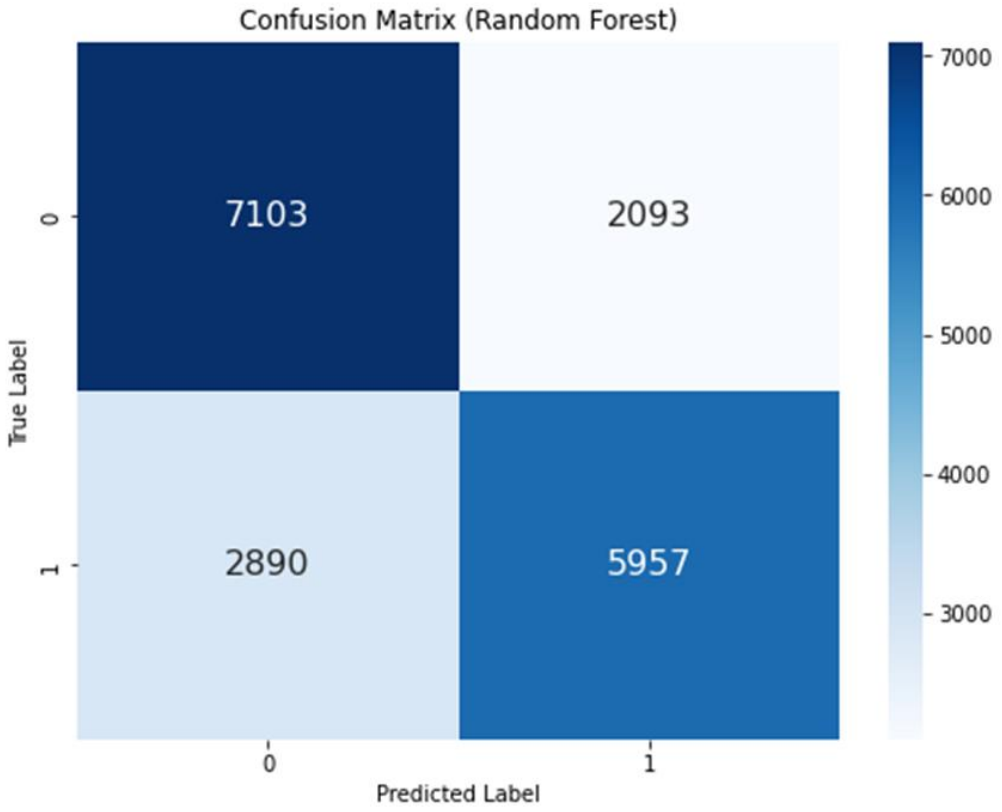
i) For Random forest, we primarily used the scikit-learn library. We split the dataset into training, validation and test sets using train_test_split. For training and tuning the random forest, we setup parameter grid for ‘n_estimators’ and used GridSearchCV in finding the best hyperparameters.

```
...: # Train and tune your Random Forest
...: rf_grid = {'n_estimators': np.linspace(100, 1000, 10, dtype=int)}
...: RF = GridSearchCV(RandomForestClassifier(min_samples_leaf=10, random_state=10,
max_features='sqrt'),
...:                   param_grid=rf_grid, cv=5, n_jobs=-1, scoring='f1')
...: RF.fit(X_train_classification, y_train_classification)
...:
```

ii) To evaluate the model, we are using the f1 score, Confusion matrix, ROC curve and the AUC score for the random forest model.

```
Best Parameters for Random Forest: {'n_estimators': 100}
F1 Score for Random Forest on Test Set: 0.705095579096881
AUC Score for Random Forest: 0.7889018035716455
Confusion Matrix for Random Forest:
 [[7103 2093]
 [2890 5957]]
```

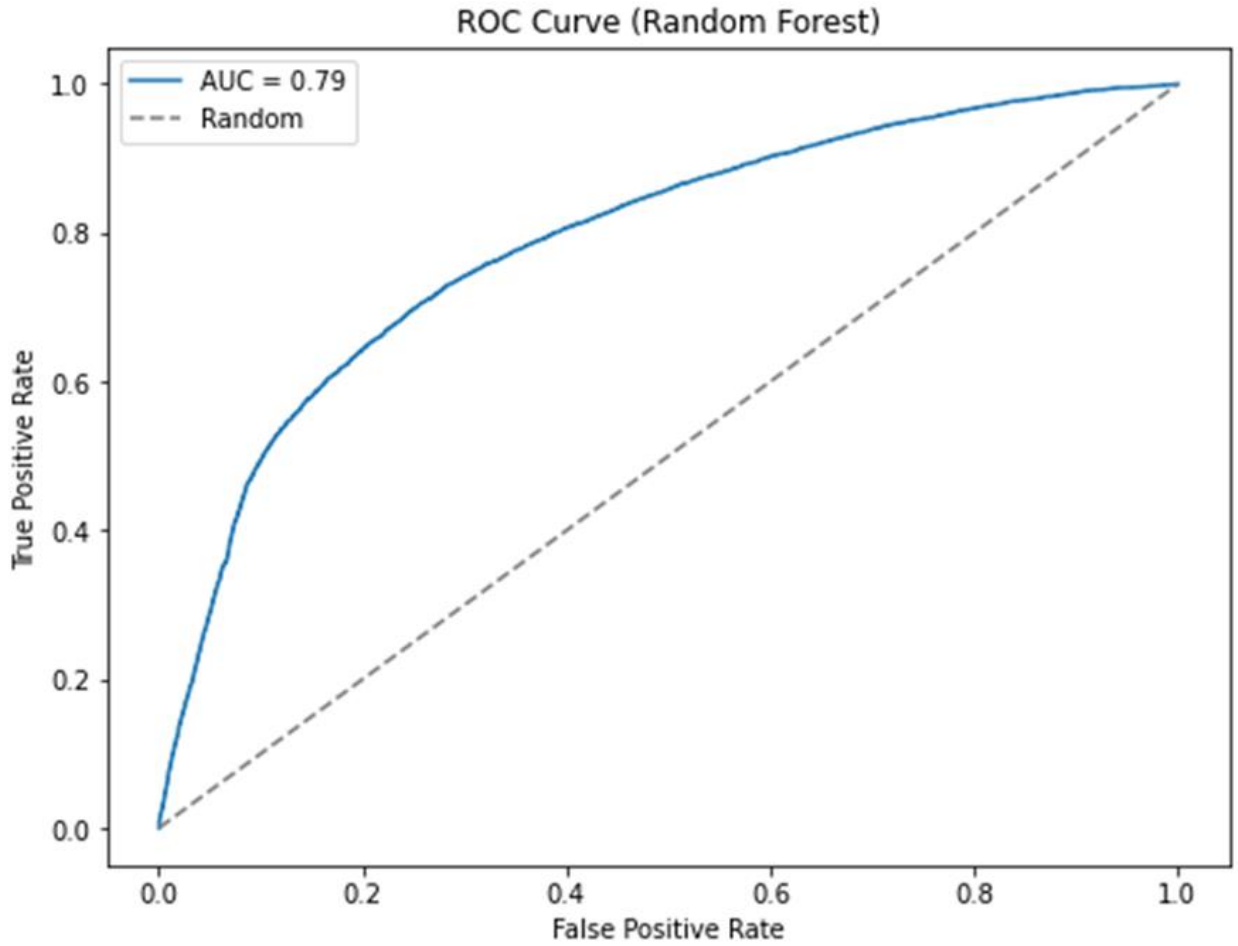
We can see from the output, the best-performing random forest model, based on grid search had 100 trees and when evaluated on the test set, it achieved an F1 score of approximately 0.705.



- True Negatives (TN): 7103 instances correctly predicted as not having cardiovascular disease.
- False Positives (FP): 2093 instances incorrectly predicted as having cardiovascular disease.
- False Negatives (FN): 2890 instances incorrectly predicted as not having cardiovascular disease.
- True Positives (TP): 5957 instances correctly predicted as having cardiovascular disease.

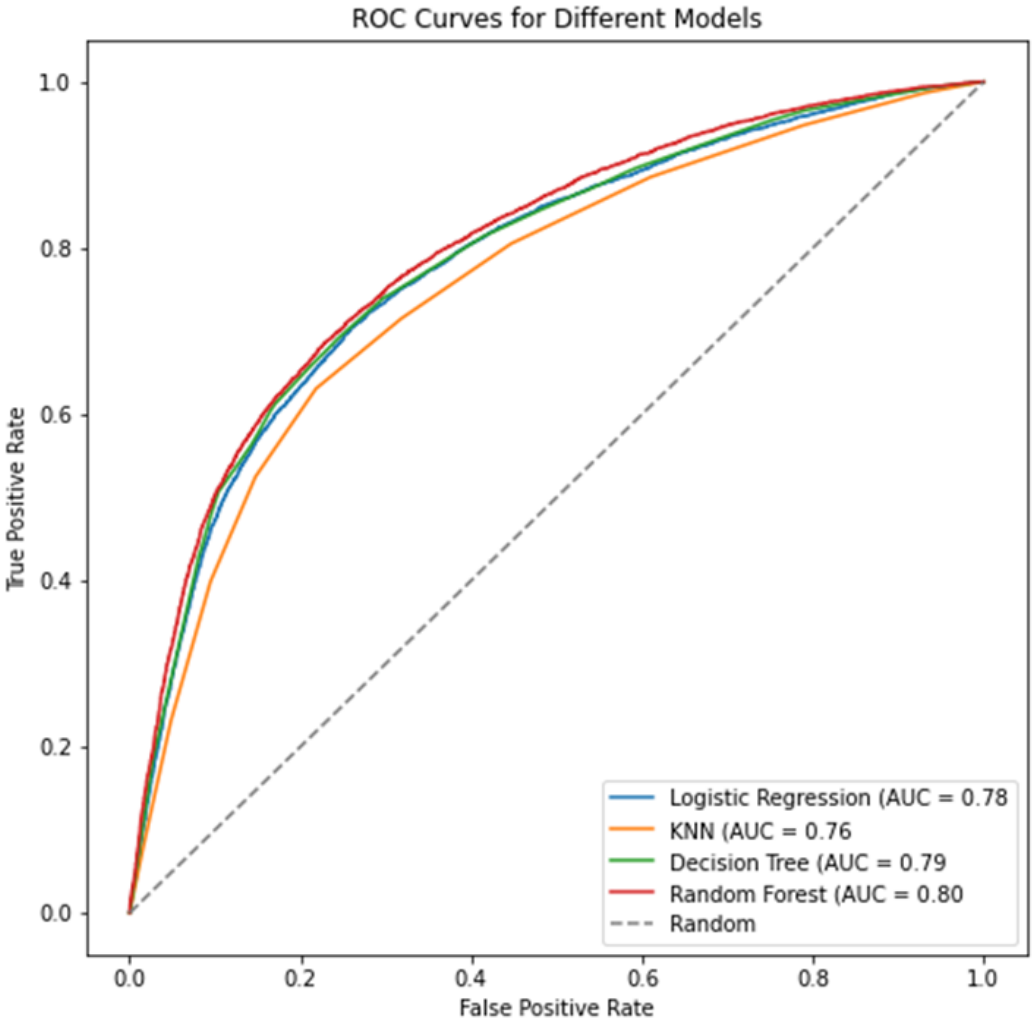
The accuracy of the Random Forest model is approximately 72.0%, representing the overall correctness of predictions.

The sensitivity of the Random Forest model is approximately 67.6%, indicating the proportion of actual positive instances correctly identified by the model.



The ROC curve for random forest is better than the previous models. It is slightly more towards the upper left corner. Also, the AUC is highest over here which is 0.79. So, it shows that this model has the higher ability to discriminate between instances with and without cardiovascular diseases.

Model Comparison:



We compared the 4 models (except support vector machine) by plotting the ROC curve. The above graph shows that random forest has the highest ROC curve which is close to the upper left corner of the graph. Also, the AUC score of random forest is highest at 0.80. This means out of the 4 models, Random Forest has the highest ability to discriminate between instances with and without cardiovascular diseases.

References

Carl J. Lavie, R. V. (2009). Obesity and Cardiovascular Disease: Risk Factor, Paradox, and Impact of Weight Loss. *JACC Journals*.

MD, E. G. (2002, Jan). *Age-associated Cardiovascular Changes in Health: Impact on Cardiovascular Disease in Older Persons.* Retrieved from SPRINGER LINK: https://link.springer.com/article/10.1023/A:1013797722156

ULIANOVA, S. (n.d.). *Cardiovascular Disease dataset*. Retrieved from Kaggle.com: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

William B. Kannel, M. (1996, May 22). *JAMA Network.com.* Retrieved from Blood Pressure as a Cardiovascular Risk Factor: https://jamanetwork.com/journals/jama/article-abstract/402826