Group C

Pei-Yu Jheng, Yifan Wang, Jiahui Yang

ALY6040: Data Mining Applications

Winter Sec 01

Instructor: Justin Grosz

Module 3 Final Project — Modeling

Data Mining Analysis: Unveiling Insights into Cardiovascular Disease Risk Factors

03/17/24

Introduction

The analysis was conducted with the primary goal of predicting heart disease risk based on various health-related and demographic factors. The report outlines the steps taken, including data cleansing, exploratory data analysis, data processing, model building, and model evaluation.

The insights gained from the analysis can inform the development of a heart disease risk assessment tool or predictive model. Such a tool could be integrated into healthcare systems, enabling healthcare providers to identify high-risk individuals and tailor preventive interventions accordingly.

Analysis

The primary goal of this analysis was to predict the occurrence of heart disease based on various health-related and demographic factors. To achieve this, we followed a structured approach encompassing data cleansing, exploratory data analysis (EDA), data processing, model building, and model evaluation.

Tools and Techniques

The analysis was carried out using Python and several popular data science libraries, including Pandas, NumPy, Scikit-learn, and Matplotlib. The following techniques were employed:

1. Data Cleansing: This involved handling missing values, encoding categorical variables,

handling outliers, checking duplicate values, and splitting the dataset into train and test sets.

2. Exploratory Data Analysis (EDA): EDA was conducted to reveal the patterns within the dataset and understand the correlations between different variables and the target variable (heart disease). It has been presented in Appendix C.
3. Data Processing: Principal Component Analysis (PCA) was used to identify the most important features contributing to heart disease.
4. Model Building: Several machine learning models were trained and evaluated, including Logistic Regression, Decision Trees, Random Forests, K-Nearest Neighbors, and Gaussian Naive Bayes.
5. Model Evaluation: The models were compared using evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score. Cross-validation techniques were employed to ensure the robustness of the models.
6. Feature Importance: The coefficients of the Logistic Regression model and the feature importances of the Decision Tree and Random Forest models were analyzed to identify the most significant predictors of heart disease.

Results and Insights (The model evaluation table and variables significant figures have been stored in Appendix C)

1. Logistic Regression with PCA: The Logistic Regression model trained with the principal

components obtained from PCA achieved an accuracy of 93.88% on the test set. The model's coefficients revealed that Principal Components 1, 3, 6, and 10 had the highest impact on predicting heart disease.

2. Decision Tree and Random Forest: The Decision Tree and Random Forest models demonstrated high accuracy on the training set but exhibited signs of overfitting on the test set. The feature importance analysis for these models highlighted variables like BMI, height, alcohol consumption, and age category as crucial predictors of heart disease.
3. Logistic Regression without PCA: The Logistic Regression model without PCA achieved an accuracy of 73.56% on the test set. The model's coefficients indicated that variables like arthritis, age category, skin cancer, and general health were significant predictors of heart disease, which are proven through hypothesis testing in Appendix C.

According to Table D1, we can see that the Evaluation Metrics for Logistic Regression with PCA have the highest F-1 score and a smaller number of False Positives and False Negatives. This suggests that the combination of dimensionality reduction through PCA and the logistic regression algorithm was effective in capturing the underlying patterns in the data and making accurate predictions. Overall, the combination of logistic regression and PCA provided a balance between model interpretability, performance, and computational efficiency, making it a suitable choice for this heart disease prediction task.

Interpretations

The analysis revealed several significant findings:

1. Demographic and Lifestyle Factors: Variables such as age, BMI, alcohol consumption, smoking history, and physical activity levels emerged as crucial determinants of heart disease risk. Individuals with higher BMI, increased alcohol consumption, and sedentary lifestyles were more likely to develop heart disease.
2. Comorbidities: The presence of comorbidities like arthritis, diabetes, and other cancers was associated with an increased risk of heart disease. These conditions may share underlying biological mechanisms or contribute to lifestyle factors that exacerbate cardiovascular risk.
3. Socioeconomic Factors: Variables like general health and frequency of checkups, which may be influenced by socioeconomic status, played a role in predicting heart disease. Individuals with poorer general health and infrequent checkups were at higher risk.

These findings highlight the multifaceted nature of heart disease risk and underscore the importance of adopting a holistic approach to prevention and management. By addressing modifiable risk factors and promoting healthy lifestyles, healthcare providers can potentially reduce the burden of heart disease in the population.

Conclusion

In order to input relevant responder data and obtain real-time risk assessments within a user interface that we are going to build, we conducted the predictive model, This interface

could also provide personalized recommendations for lifestyle modifications, regular screenings, and appropriate medical interventions based on the individual's risk profile.

By leveraging the findings of this analysis and implementing a user-friendly risk assessment tool, responders can take a proactive approach to heart disease prevention, ultimately improving their health conditions and reducing the chances of getting cardiovascular disease.

References

Cardiovascular Diseases Risk Prediction Dataset. (2023, July 3). Kaggle. https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset/data

CDC - 2021 BRFSS Survey Data and Documentation. (n.d.). https://www.cdc.gov/brfss/annual_data/annual_2021.html

Appendix A

Table A1

Description of Variables

	Variable	Description	Type
1	General Health	Respondents considered their health conditions in general	object
2	Checkup	how long has it been since respondents last visited a doctor for a routine checkup?	object
3	Exercise [Yes or No]	Any participation in physical activities or exercise?	object
4	Heart Disease [Yes or No]	Respondents reported having coronary heart disease or mycardialinfarction.	object
5	Skin Cancer [Yes or No]	Respondents reported having skin cancer	object
6	Other Cancer [Yes or No]	Respondents reported having any other types of cancer	object
7	Depression [Yes or No]	Respondents reported having a depressive disorder (including depression, major depression, dysthymia, or minor depression)	object
8	Diabetes [Yes or No]	Respondents reported having a diabetes. If yes, what type of diabetes it is/was.	object
9	Arthritis [Yes or No]	Respondents reported having an Arthritis.	object
10	Sex	Gender of Respondents	object
11	Age Category	Age group of Respondents	object
12	Height (cm)	The height of the patients seems to follow a normal distribution, with the majority of patients having heights around 160 to 180 cm.	int64
13	Weight (kg)	The weight of the patients also appears to be normally distributed, with most patients weighing between approximately 60 and 100 kg.	float64
14	BMI	The distribution of Body Mass Index is somewhat right-skewed. A large number	float64

		of patients have a BMI between 20 and 30, which falls within the normal to overweight range. However, there are also a significant number of patients with a BMI in the obese range (>30).	
15	Smoking History		object
16	Alcohol Consumption	This feature is heavily right-skewed. Most patients have low alcohol consumption, but there are a few patients with high consumption.	int64
17	Fruit Consumption	This feature is also right-skewed. A lot of patients consume fruits regularly, but a significant number consume them less frequently.	int64
18	Green Vegetables Consumption	This feature appears to be normally distributed, with most patients consuming green vegetables moderately.	float64
19	FriedPotato Consumption	This feature is right-skewed. Many patients consume fried potatoes less frequently, while a few consume them more often.	int64

Appendix B

Data Cleaning Process

Figure B1

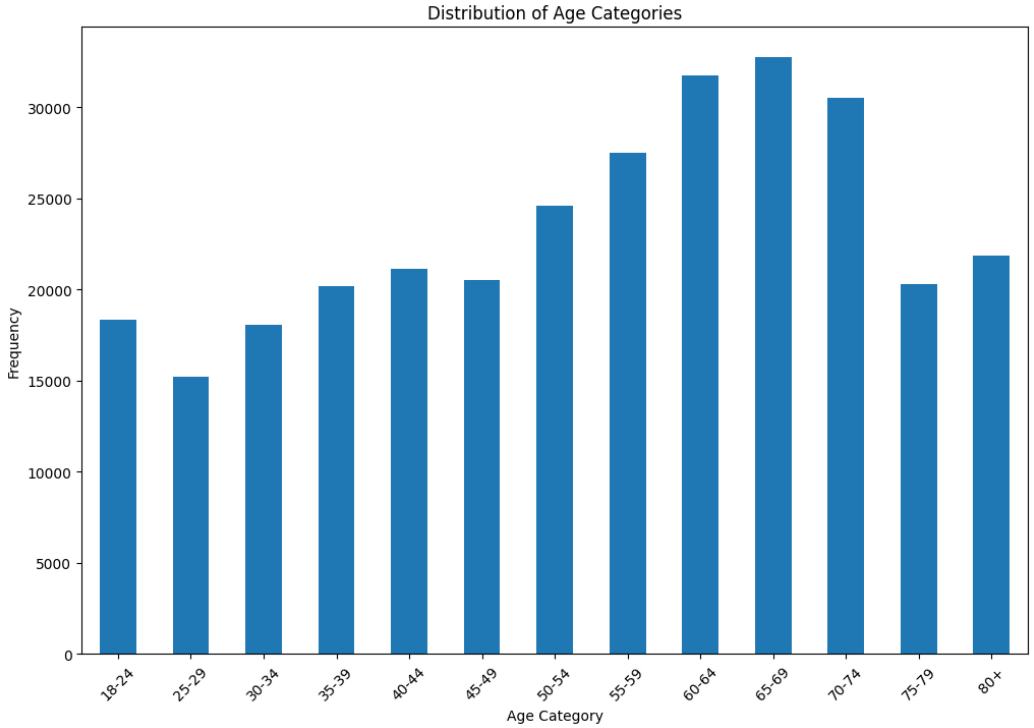
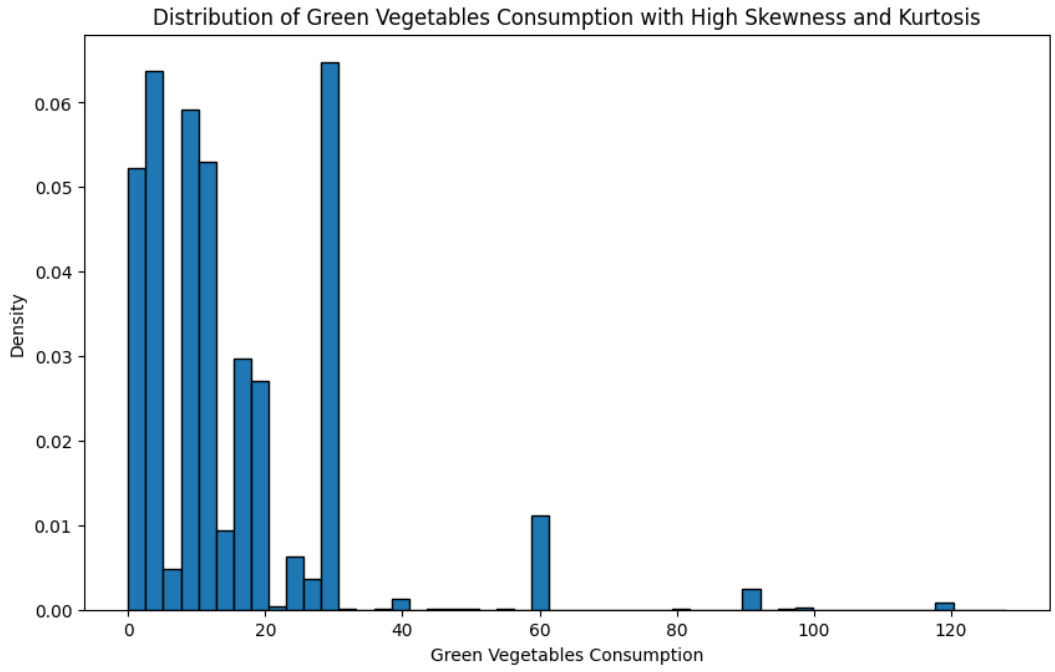


Figure B2



Appendix C

Predictor Selection EDA

Figure C1 – Correlation Metrix

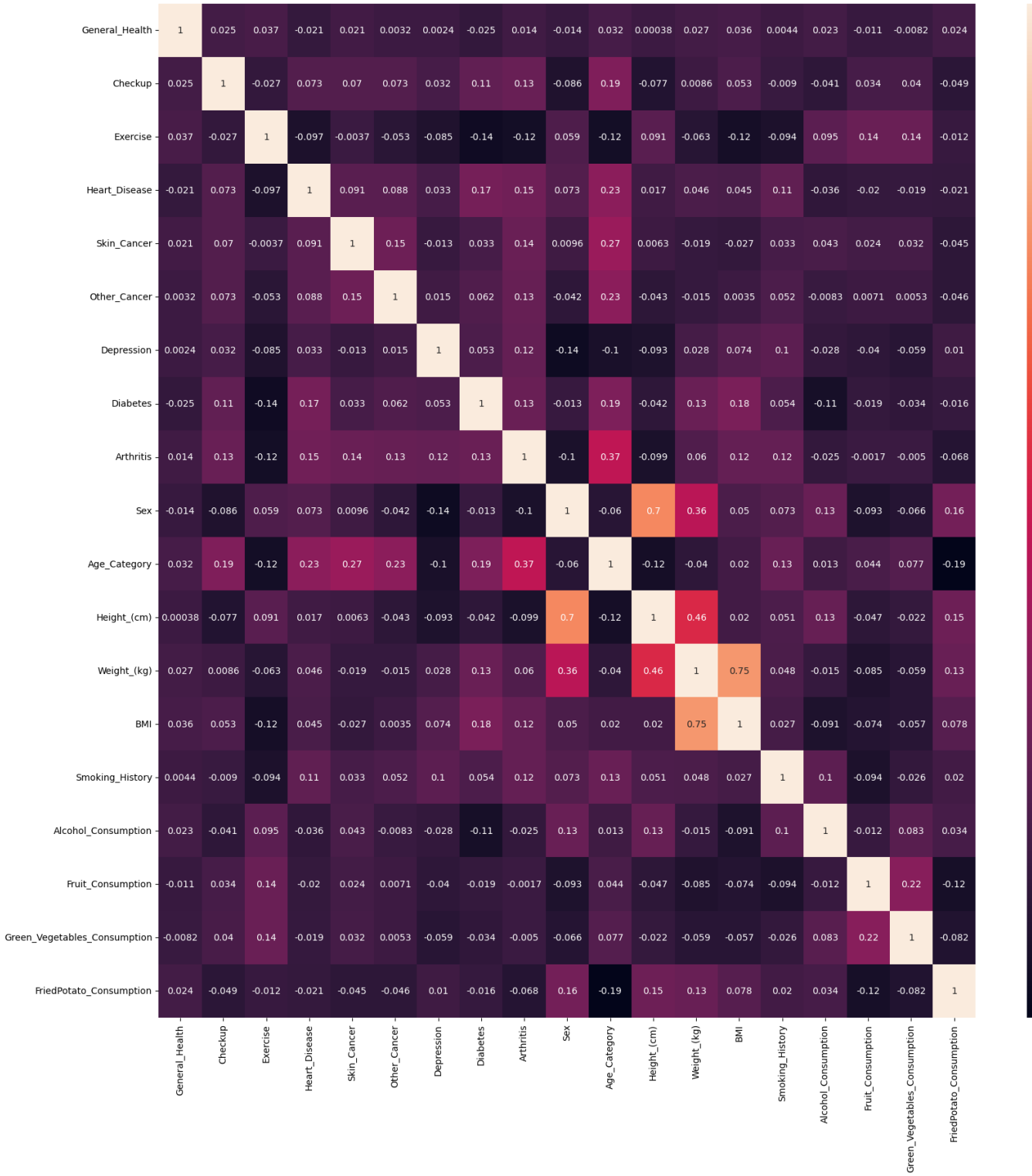


Figure C2 - If having a habit on exercise decreases the likelihood of getting heart disease? (T-test)

```
T-statistic: -53.47564507144645
P-value: 0.0
Reject the null hypothesis: having a habit on exercise decreases the likelihood of getting heart disease.
```

Figure C2 - If having a high BMI increases the likelihood of getting heart disease? (T-test)

```
T-statistic: 21.373434360753038
P-value: 2.806836359436406e-101
Reject the null hypothesis: High BMI increases the likelihood of getting heart disease.
```

Figure C3 - If having a high Age increases the likelihood of getting heart disease? (T-test)

```
T-statistic: 112.5937767311332
P-value: 0.0
Reject the null hypothesis: People older than 45 increases the likelihood of getting heart disease.
```

Figure C4 - If having Diabetes increases the likelihood of getting heart disease? (T-test)

```
T-statistic: 96.3260870751159
P-value: 0.0
Reject the null hypothesis: having diabetes increases the likelihood of getting heart disease.
```

Figure C5 - If having Arthritis increases the likelihood of getting heart disease? (T-test)

```
T-statistic: 85.72582918735579
P-value: 0.0
Reject the null hypothesis: having arthritis increases the likelihood of getting heart disease.
```

Figure C6

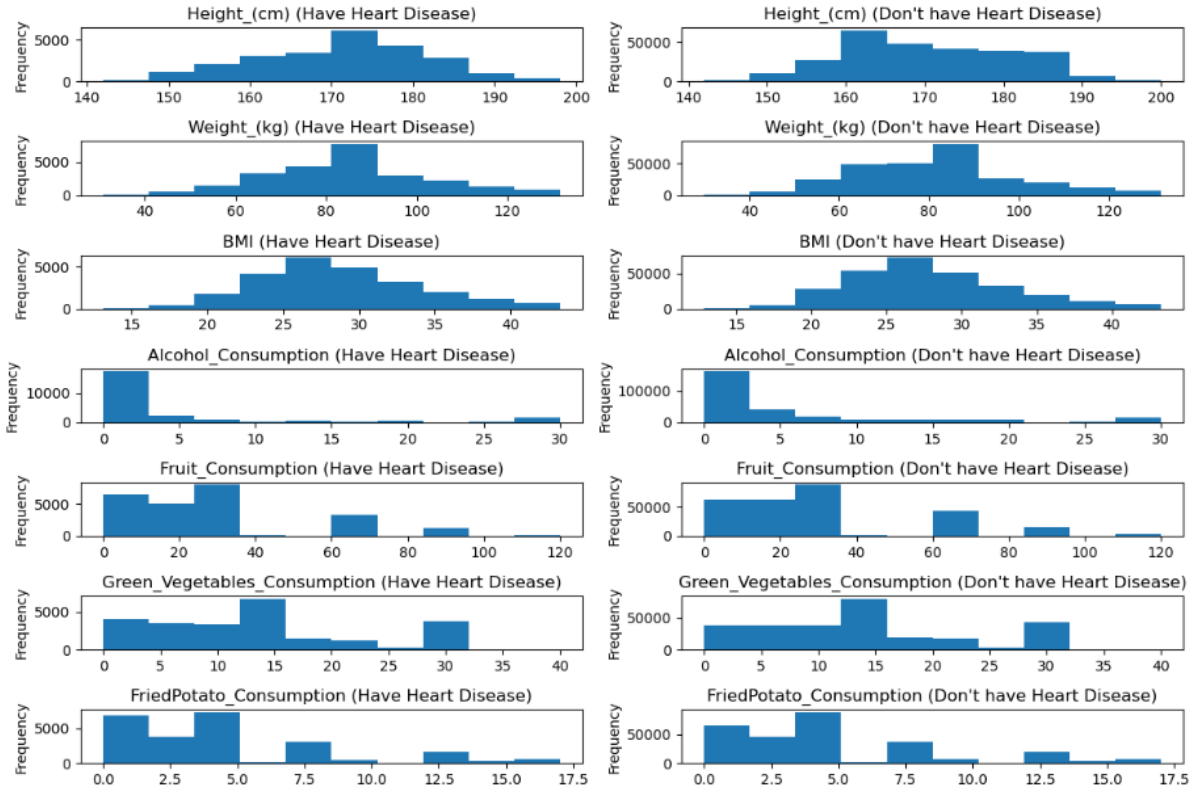


Figure C7

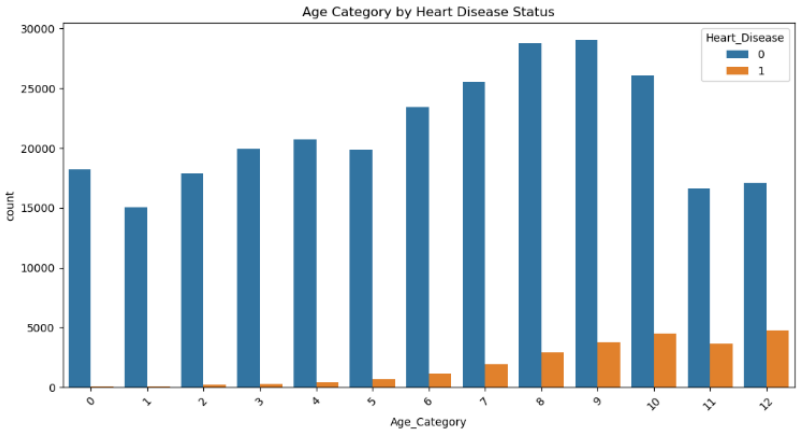


Figure C8

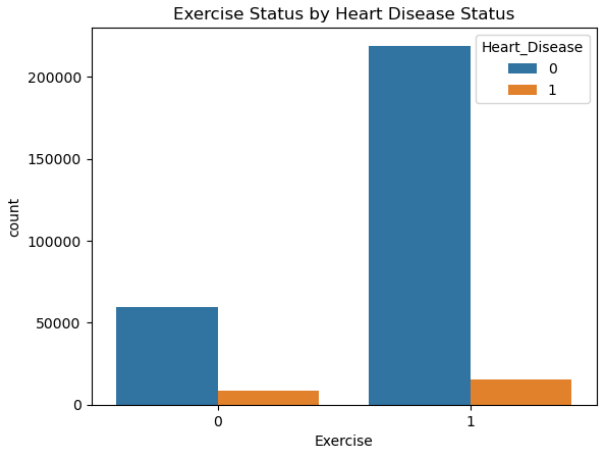
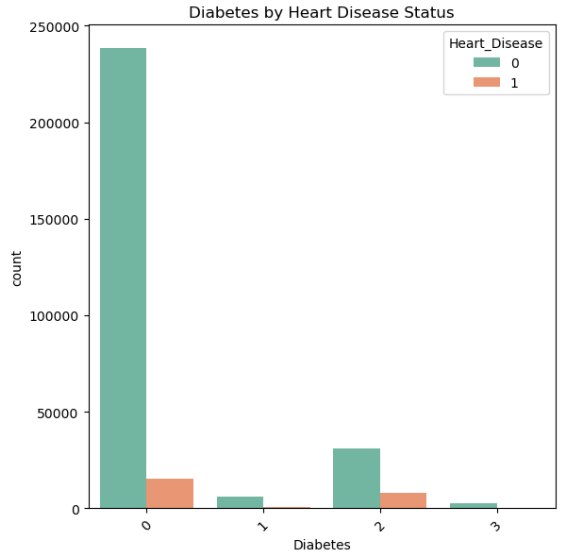


Figure C9



Appendix D

Evaluate Models

Table D1 - Evaluation Metrics

	Precision	Recall	F-1 score	Accuracy	False positive	False negative
Logistic Regression	0.20	0.78	0.32	0.73	14945	1064
Logistic Regression with PCA	0.68	0.42	0.52	0.93	940	2764
Decision Tree	0.19	0.24	0.21	0.85	4962	3681
Random Forest	0.36	0.11	0.17	0.91	976	4339

Figure D1 – Coefficients of Logistic Regression Model

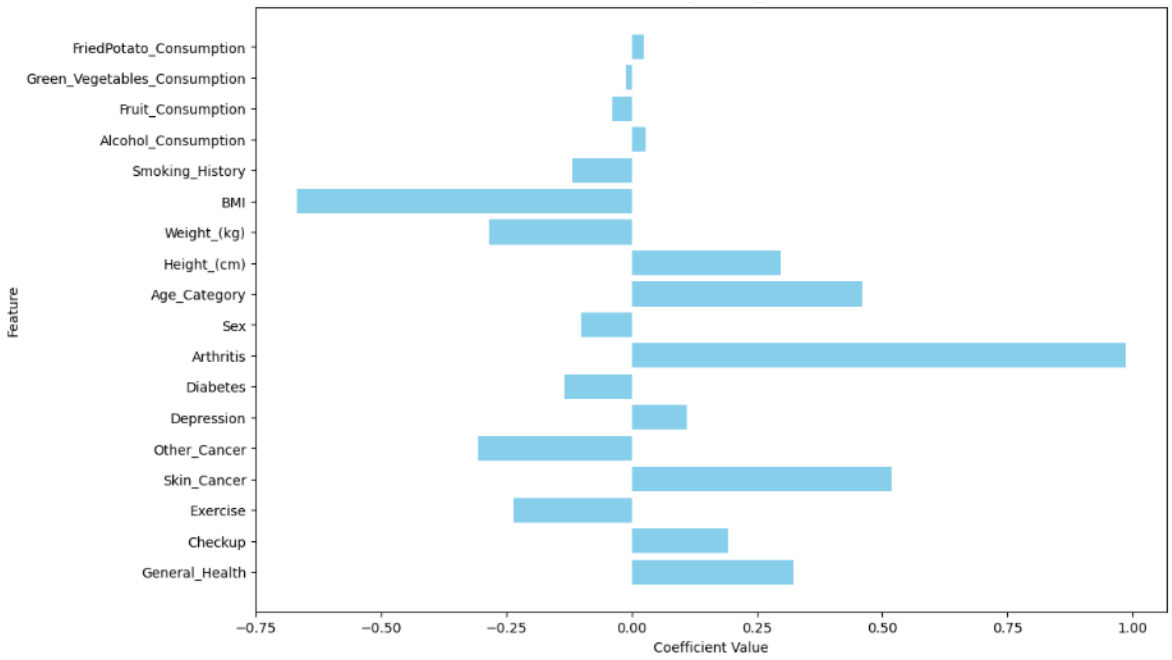


Figure D2 – Feature Importances of Decision Tree Model

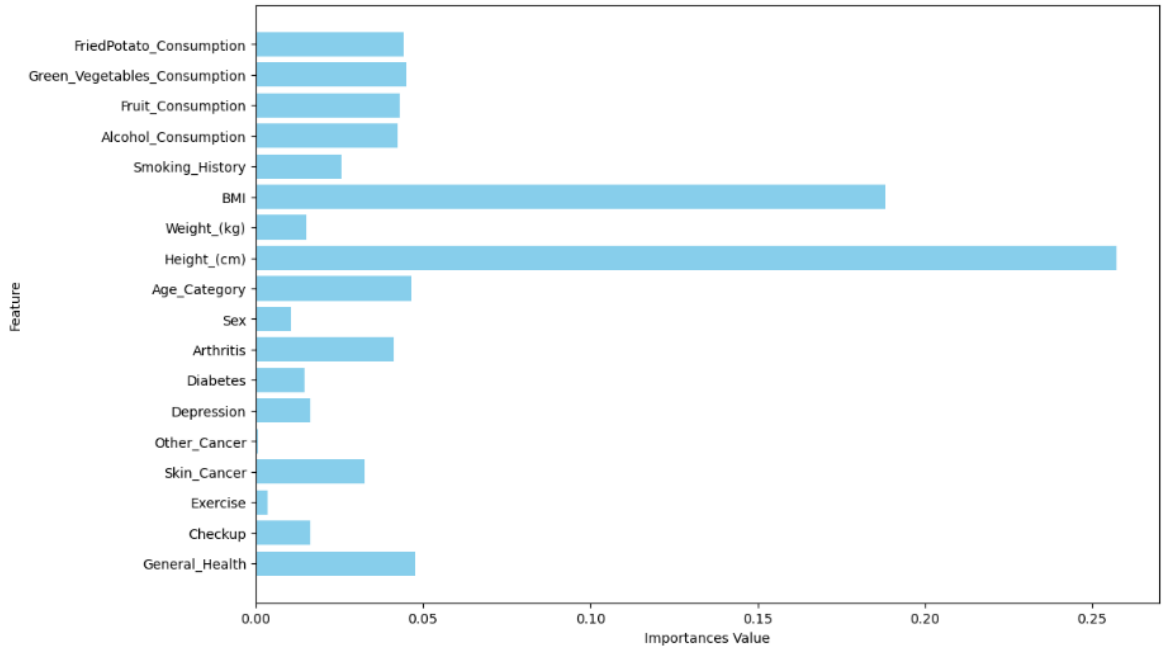


Figure D3 - Feature Importances of Random Forest Model