Dimension Reduction on the Heart Dataset Report

I downloaded the heart dataset from Kaggle. This dataset contains detailed information on the risk factors for cardiovascular disease. It includes information on age, gender, height, weight, blood pressure values, cholesterol levels, glucose levels, smoking habits, and alcohol consumption of over 70 thousand individuals. Most of the variables were binary.

First, I started with preprocessing and cleaning data. I checked for duplicates and null/missing data. There were no duplicates and missing data. I removed the unnecessary columns. I found the summary statistics for the data. I converted the age from days to years. Then I used the height variable and found that 1 represents women and 2 represents men. Then I divided the dataset into 70% train data and 30% test data. I did some visualization. I plotted histogram, correlation matrix, and heat map. Then I checked for multicollinearity.

Then, I used PCA method for this dataset. PCA is one of the methods for dimension reduction. I found its accuracy. Its accuracy was 0.824. My other model was Random Forest Feature importance. The accuracy for this model was 1 for train data and 0.722 for test data. Then I used Decision Tree. The accuracy for this model was 1 for train data and 0.99 for test data. In conclusion, Decision Tree was the most accurate model for this dataset. (I realized that I forgot to clean data and scale it for Decision Tree. That is why I think the accuracy was that high.)

For future, I will use better models for my dataset. I will use the same notebook for my dataset so that I don’t have to repeat the preprocessing and cleaning data.