



Cardiovascular Disease Dataset Analysis

By Jaspreet Kang



Introduction

Cardiovascular dataset contains 70,000 data points and 13 variables:

1. **Id**
2. **Age** (in days)
3. **Gender** (Female, Male)
4. **Height** (in cm)
5. **Weight** (in kg)
6. **Systolic Blood Pressure**
7. **Diastolic Blood Pressure**
8. **Cholesterol**: 3 levels (Normal, Above Normal, Well Above Normal)
9. **Glucose**: 3 levels (Normal, Above Normal, Well Above Normal)
10. **Smoke**: 2 levels (Non-Smoker, Smoker)
11. **Alcohol Intake**: 2 levels (Non-Drinker, Drinker)
12. **Active**: 2 levels (Not Active, Active)
13. **CVD**: 2 levels (No CVD, CVD)



Data Cleaning and Feature Engineering

Two New Variables Created

1. BMI
2. Age1 (in years)



Removing Potential Misentered Information

- Removed data points where diastolic blood pressure was greater than or equal to systolic blood pressure
- Removed data points where systolic blood pressure was greater than 300 or less than 70
- Removed data points where diastolic blood pressure was less than or equal to 20
- Removed data points where bmi was greater than 50 or less than 15



Final Data Set

- 1,234 of the 70,000 data points removed
- Categorical variables stored as integers converted to factor variables



Research Questions and Goals

Research Questions/Topics

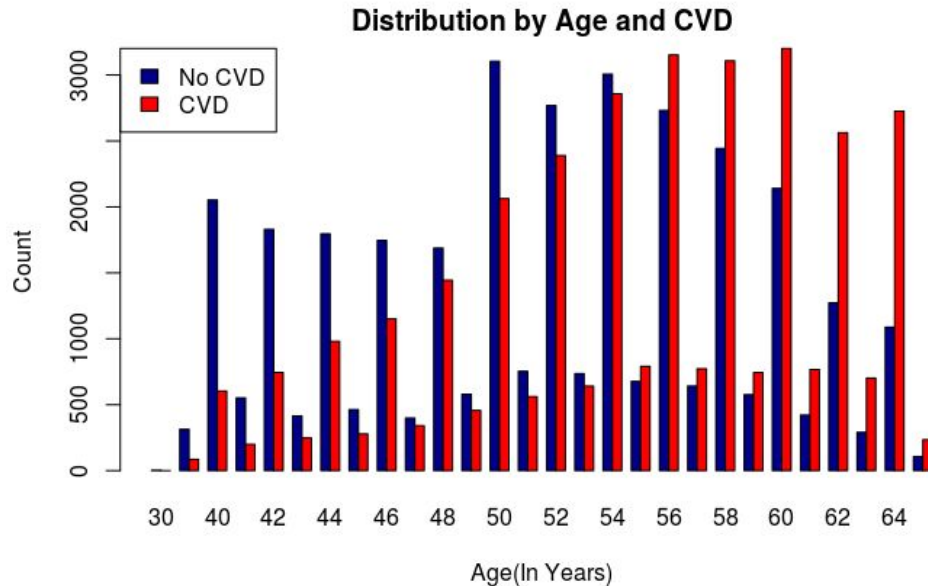
1. At what age does the presence of CVD surpass the absence of CVD?
2. What variables are most correlated with cardiovascular disease?
3. Take a closer look at the relationship between gender and bmi with CVD.

Predict Cardiovascular Disease

Use and compare the performance of several classification modeling techniques in order to predict cardiovascular disease.

1. Logistic Regression
2. XG Boosting Classifier
3. Random Foresting Classifier

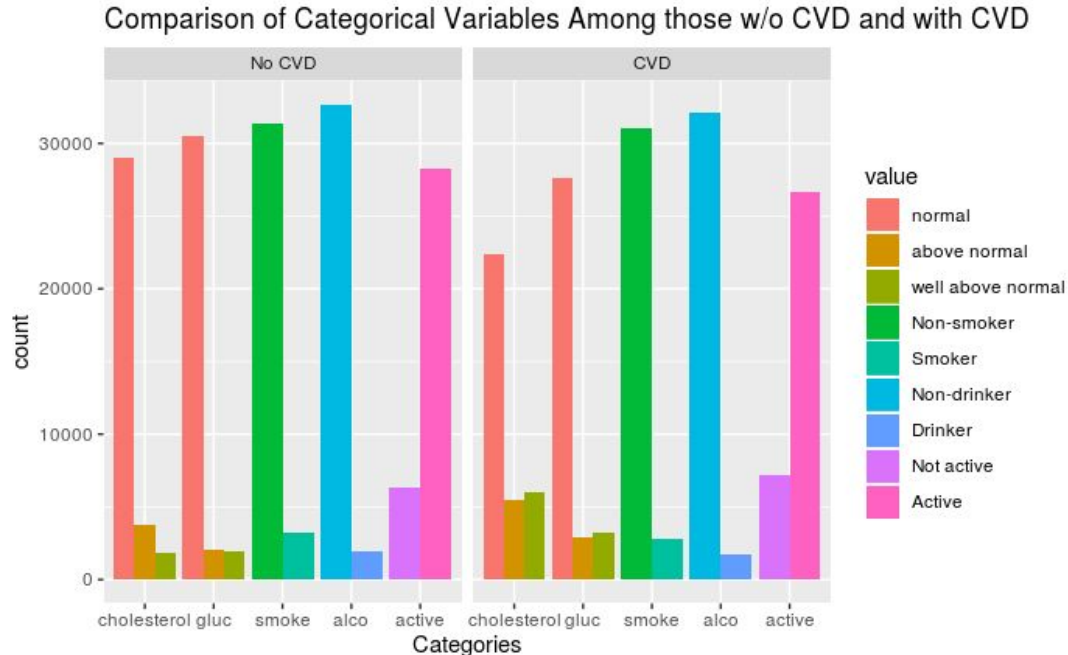
Age vs CVD



The presence of CVD surpasses the absence of CVD after the age of 54.

Furthermore, the ratio of CVD:NoCVD quickly heads towards a 1:1 ratio in the mid 40s.

Variable Relationship with CVD

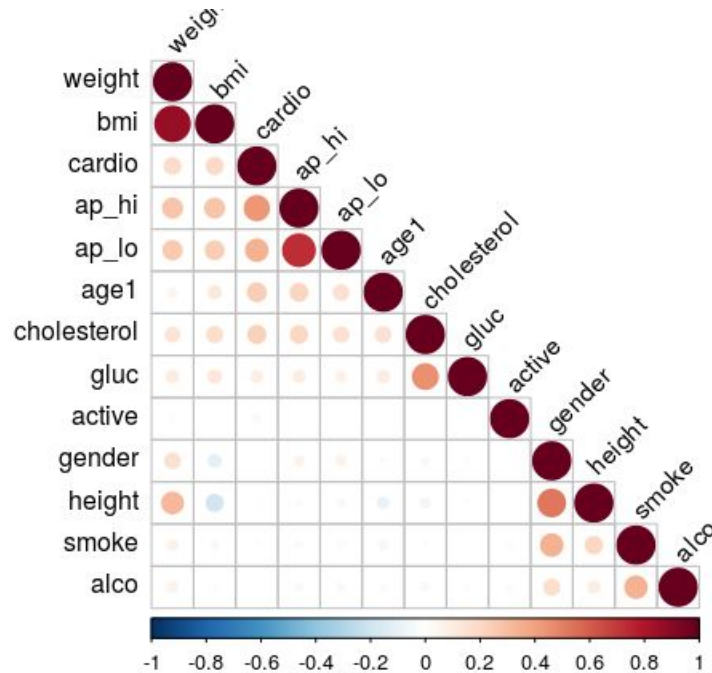


The ratio of (normal:above normal+well above normal) with regards to cholesterol levels drastically reduces from No CVD group to CVD group. The same can be said for glucose levels to a smaller extent.

Smoking and alcohol consumptions show no significant changes between groups.

Those with CVD seem to be less active

Variable Relationship with CVD contd.



Note: ap_hi = systolic blood pressure
ap_lo = diastolic blood pressure
cardio = CVD

$\text{correlation}(\text{ap_hi}, \text{cardio}) = 0.43$

$\text{correlation}(\text{ap_lo}, \text{cardio}) = 0.34$

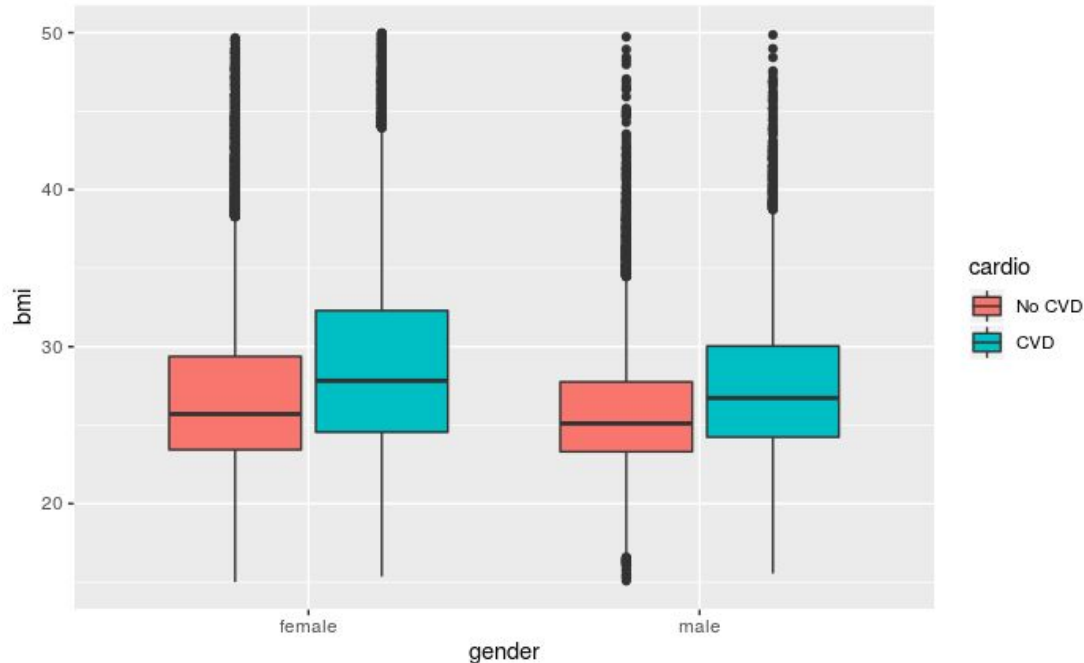
$\text{correlation}(\text{age1}, \text{cardio}) = 0.24$

$\text{correlation}(\text{cholesterol}, \text{cardio}) = 0.22$

$\text{correlation}(\text{bmi}, \text{cardio}) = 0.19$

$\text{correlation}(\text{weight}, \text{cardio}) = 0.18$

Gender and BMI vs CVD



Women tend to have on average a higher BMI value than men.

The median of BMI for both genders is higher among the CVD group.

Cardiovascular Disease Prediction

72.9%

Logistic Regression

Actual → Predicted ↓	No CVD	CVD
No CVD	5487	2222
CVD	1490	4488

- False Positive = 10.9%
- False Negative = 16.2%
- Used One-Hot Encoding with Stepwise Feature Selection

73.3%

XG Boosting Classifier

Actual → Predicted ↓	No CVD	CVD
No CVD	5456	2132
CVD	1521	4578

- False Positive = 11.1%
- False Negative = 15.6%
- Performed 500 iterations with eta = 0.01 and max depth = 6

72.7%

Random Forest

Actual → Predicted ↓	No CVD	CVD
No CVD	5304	2059
CVD	1673	4651

- False Positive = 12.2%
- False Negative = 15.1%
- Used 300 trees with 3 variables tried at each split

80% of dataset was split up into the training set and the remaining 20% was used as the testing set.


All classification models yielded a successful classification rate of around 73% on the testing set.

XG Boosting Model performed the best with a classification rate of 73.3%.



Conclusion

- The prevalence of CVD rapidly increases in the mid 40s, with more people in the dataset having CVD than not having CVD after the age of 54.
- Systolic Blood Pressure, Diastolic Blood Pressure, Age, Cholesterol Levels, and BMI are most significantly associated with CVD. However, none of these variables show a strong correlation.
- The BMI for both gender groups are higher among the CVD group.
- Logistic Regression model, XG Boosting Classifier model, and Random Foresting model all yield a successful classification rate of around 73%. The XG Boosting model performs the best with a successful classification rate of 73.3%.
- **Drawback:** Due to my limited knowledge of abnormal ranges of blood pressure and BMI, I relied on online research to determine blood pressure and BMI values that are impossible or highly improbable to achieve in order to identify and remove potential faulty data. Getting an expert's feedback would have allowed me to address outliers more effectively.

- 
- Cardiovascular dataset can be found below:
 - <https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>
 - My extensive CVD analysis write-up can be found below:
 - http://jaspreetkang.com/CVD_analysis.pdf
 - My R code for this analysis can be found below:
 - <https://github.com/kangiaspreet/Cardiovascular-Disease-Analysis/tree/master/CVD>