# Index

- Introduction
- Dataset
- Data Preprocessing
- Techniques Used
  - Logistic Regression
  - Random Forest
  - Naive Bayes
- Conclusion

# INTRODUCTION

- In 2008, heart disease and stroke were responsible for nearly 30.4% in united states. Coronary heart disease is the cause of more than 2/3 of these deaths.

- Coronary artery disease develops when the major blood vessels that supply your heart with blood, oxygen and nutrients (coronary arteries) become damaged or diseased.

- CSD is affected by many factors such as blood pressure, anxiety, cholesterol, diabetes, obesity, smoking , alcohol consumption , lack of physical activities etc.

# DATASET

- In our dataset we have considered the factors given below:

❑ Systolic blood pressure
❑ Tobacco cumulative (kg)
❑ LDL - low density lipoprotein cholesterol
❑ Adiposity - famihist - family history of heart disease (Present, Absent)
❑ Obesity
❑ Alcohol current alcohol consumption - age at onset
❑ CHD

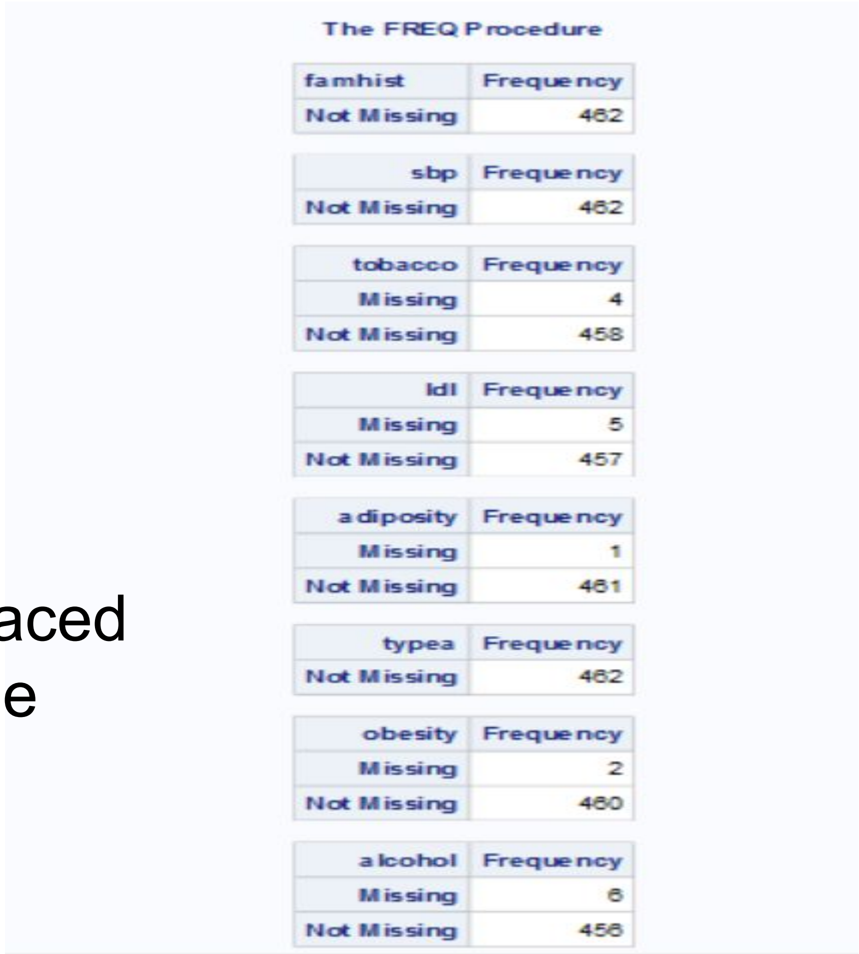
In this project, we are analyzing the most and least significant attributes that helps in predicting whether a person is having CHD or not.

# REPLACING MISSING VALUES

- Using proc *freq* , we find missing values in each of the attributes.

- There are missing values in features tobacco, ldl, obesity, alcohol.

- These missing values are replaced by the mean and median of the corresponding attributes.



The FREQ Procedure

| famhist | Frequency |
|---|---|
| Not Missing | 462 |

| sbp | Frequency |
|---|---|
| Not Missing | 462 |

| tobacco | Frequency |
|---|---|
| Missing | 4 |
| Not Missing | 458 |

| ldl | Frequency |
|---|---|
| Missing | 5 |
| Not Missing | 457 |

| adiposity | Frequency |
|---|---|
| Missing | 1 |
| Not Missing | 461 |

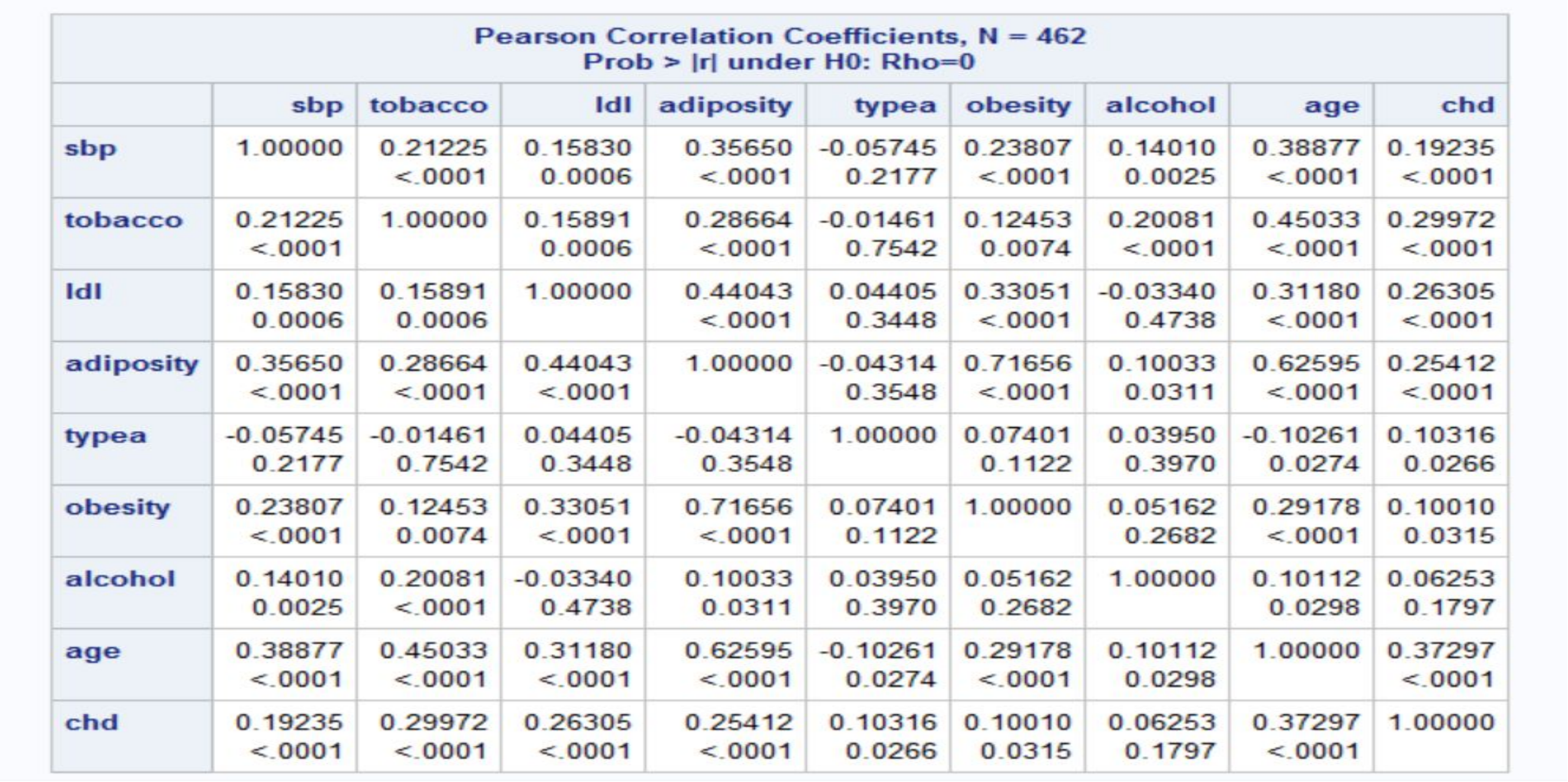
| typea | Frequency |
|---|---|
| Not Missing | 462 |

| obesity | Frequency |
|---|---|
| Missing | 2 |
| Not Missing | 460 |

| alcohol | Frequency |
|---|---|
| Missing | 6 |
| Not Missing | 456 |

# CORRELATION BETWEEN FEATURES

- Using proc *corr*, we find correlation between each of the attributes.

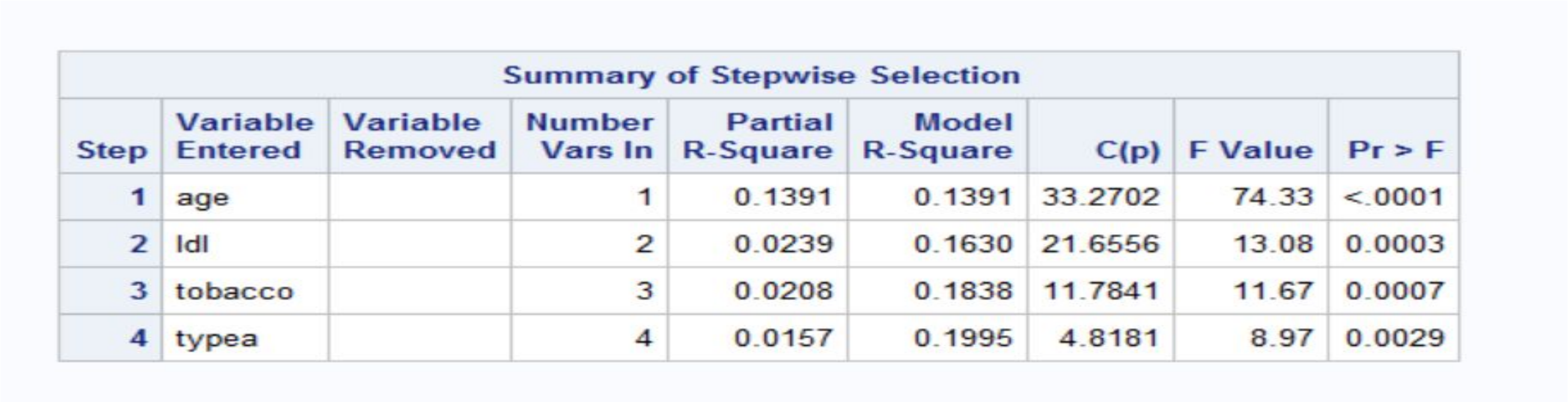


Pearson Correlation Coefficients, N = 462
Prob > |r| under H0: Rho=0

| | sbp | tobacco | ldl | adiposity | typea | obesity | alcohol | age | chd |
|---|---|---|---|---|---|---|---|---|---|
| sbp | 1.00000 | 0.21225<br><.0001 | 0.15830<br>0.0006 | 0.35650<br><.0001 | -0.05745<br>0.2177 | 0.23807<br><.0001 | 0.14010<br>0.0025 | 0.38877<br><.0001 | 0.19235<br><.0001 |
| tobacco | 0.21225<br><.0001 | 1.00000 | 0.15891<br>0.0006 | 0.28664<br><.0001 | -0.01461<br>0.7542 | 0.12453<br>0.0074 | 0.20081<br><.0001 | 0.45033<br><.0001 | 0.29972<br><.0001 |
| ldl | 0.15830<br>0.0006 | 0.15891<br>0.0006 | 1.00000 | 0.44043<br><.0001 | 0.04405<br>0.3448 | 0.33051<br><.0001 | -0.03340<br>0.4738 | 0.31180<br><.0001 | 0.26305<br><.0001 |
| adiposity | 0.35650<br><.0001 | 0.28664<br><.0001 | 0.44043<br><.0001 | 1.00000 | -0.04314<br>0.3548 | 0.71656<br><.0001 | 0.10033<br>0.0311 | 0.62595<br><.0001 | 0.25412<br><.0001 |
| typea | -0.05745<br>0.2177 | -0.01461<br>0.7542 | 0.04405<br>0.3448 | -0.04314<br>0.3548 | 1.00000 | 0.07401<br>0.1122 | 0.03950<br>0.3970 | -0.10261<br>0.0274 | 0.10316<br>0.0266 |
| obesity | 0.23807<br><.0001 | 0.12453<br>0.0074 | 0.33051<br><.0001 | 0.71656<br><.0001 | 0.07401<br>0.1122 | 1.00000 | 0.05162<br>0.2682 | 0.29178<br><.0001 | 0.10010<br>0.0315 |
| alcohol | 0.14010<br>0.0025 | 0.20081<br><.0001 | -0.03340<br>0.4738 | 0.10033<br>0.0311 | 0.03950<br>0.3970 | 0.05162<br>0.2682 | 1.00000 | 0.10112<br>0.0298 | 0.06253<br>0.1797 |
| age | 0.38877<br><.0001 | 0.45033<br><.0001 | 0.31180<br><.0001 | 0.62595<br><.0001 | -0.10261<br>0.0274 | 0.29178<br><.0001 | 0.10112<br>0.0298 | 1.00000 | 0.37297<br><.0001 |
| chd | 0.19235<br><.0001 | 0.29972<br><.0001 | 0.26305<br><.0001 | 0.25412<br><.0001 | 0.10316<br>0.0266 | 0.10010<br>0.0315 | 0.06253<br>0.1797 | 0.37297<br><.0001 | 1.00000 |

# REGRESSION ANALYSIS

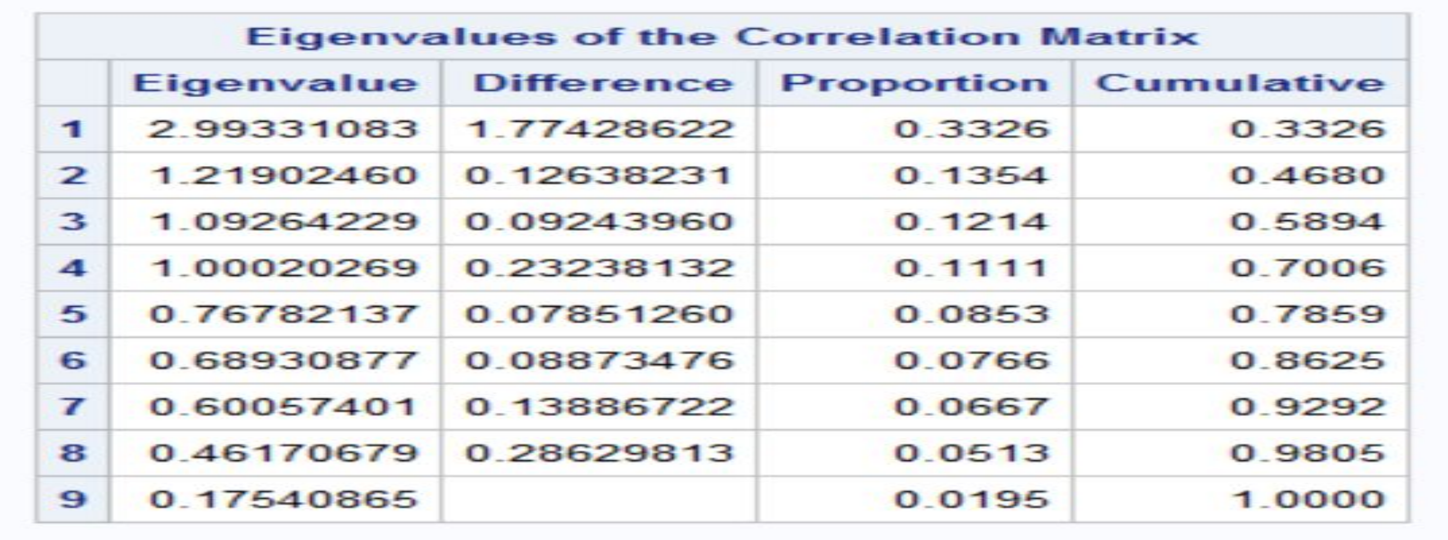
- Performed stepwise regression analysis using sing proc *reg,* with chd as dependent variable and sbp, tobacco, ldl, adiposity, typea, obesity, alcohol, age.



| Summary of Stepwise Selection | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable Entered | Variable Removed | Number Vars In | Partial R-Square | Model R-Square | C(p) | F Value | Pr > F |
| 1 | age | | 1 | 0.1391 | 0.1391 | 33.2702 | 74.33 | <.0001 |
| 2 | ldl | | 2 | 0.0239 | 0.1630 | 21.6556 | 13.08 | 0.0003 |
| 3 | tobacco | | 3 | 0.0208 | 0.1838 | 11.7841 | 11.67 | 0.0007 |
| 4 | typea | | 4 | 0.0157 | 0.1995 | 4.8181 | 8.97 | 0.0029 |

- We can see that even with age, ldl, tobacco and typea entering the model, the maximum variance is only 0.19995.

# PRINCIPAL COMPONENT ANALYSIS



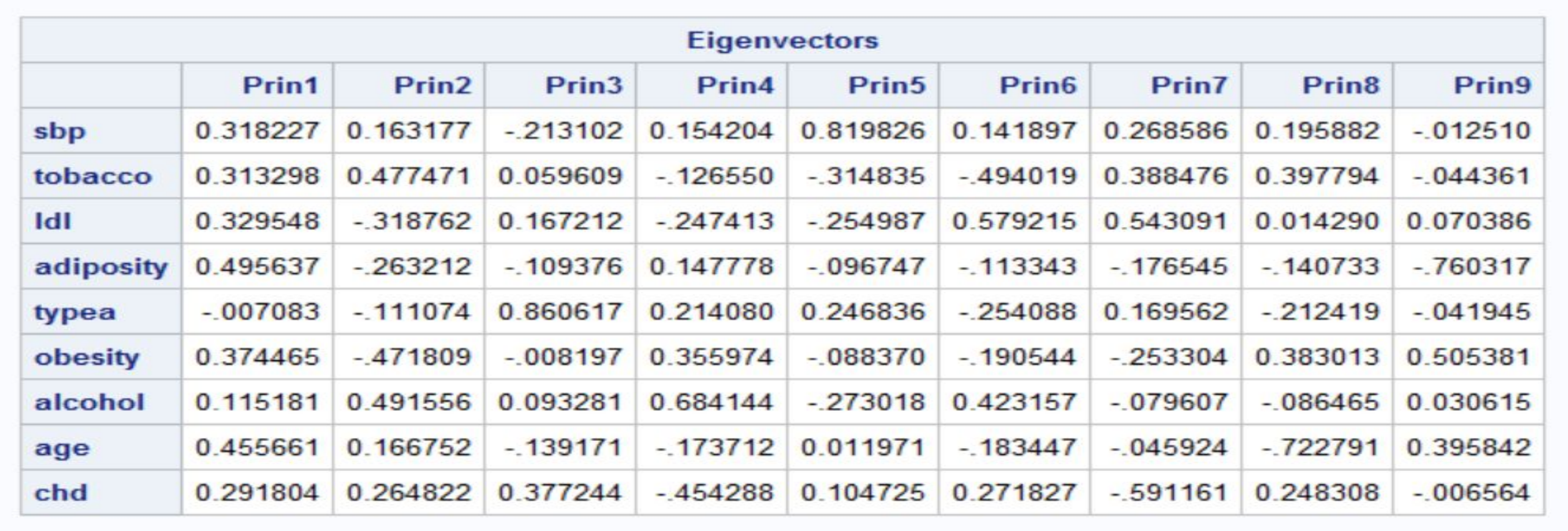
Eigenvalues of the Correlation Matrix

| | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| 1 | 2.99331083 | 1.77428622 | 0.3326 | 0.3326 |
| 2 | 1.21902460 | 0.12638231 | 0.1354 | 0.4680 |
| 3 | 1.09264229 | 0.09243960 | 0.1214 | 0.5894 |
| 4 | 1.00020269 | 0.23238132 | 0.1111 | 0.7006 |
| 5 | 0.76782137 | 0.07851260 | 0.0853 | 0.7859 |
| 6 | 0.68930877 | 0.08873476 | 0.0766 | 0.8625 |
| 7 | 0.60057401 | 0.13886722 | 0.0667 | 0.9292 |
| 8 | 0.46170679 | 0.28629813 | 0.0513 | 0.9805 |
| 9 | 0.17540865 | | 0.0195 | 1.0000 |

- Based on Eigenvalue correlation matrix, we select 4 Principal components as their cumulative Eigenvalue is greater than 70% for the first 4 principal components.

# PRINCIPAL COMPONENT ANALYSIS

Using PCA on the entire data, we get the below

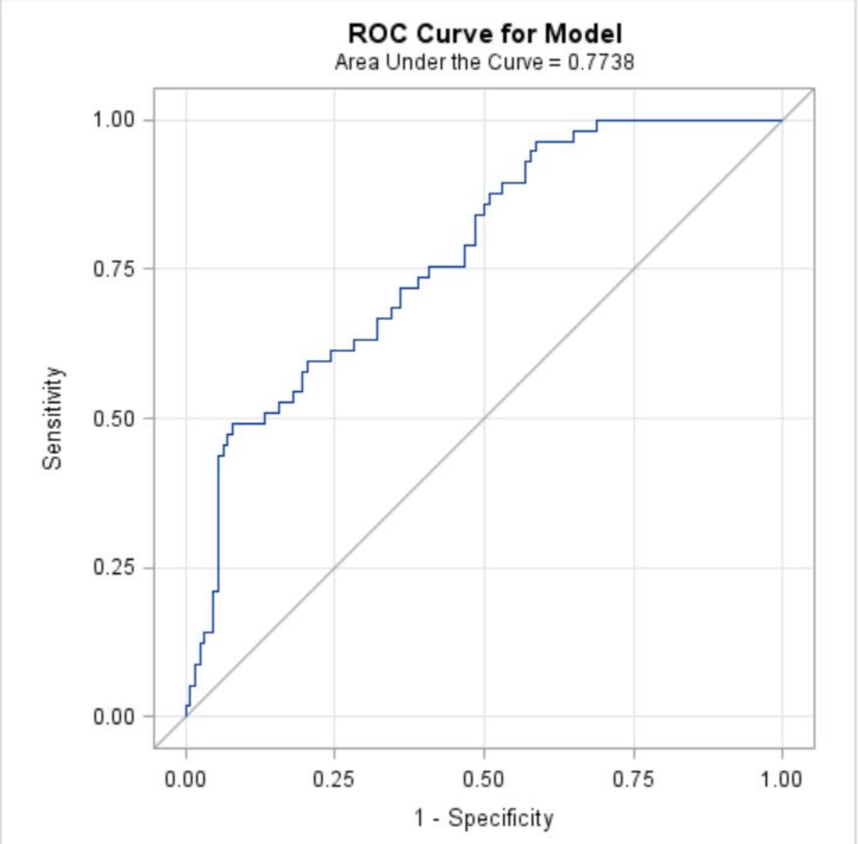


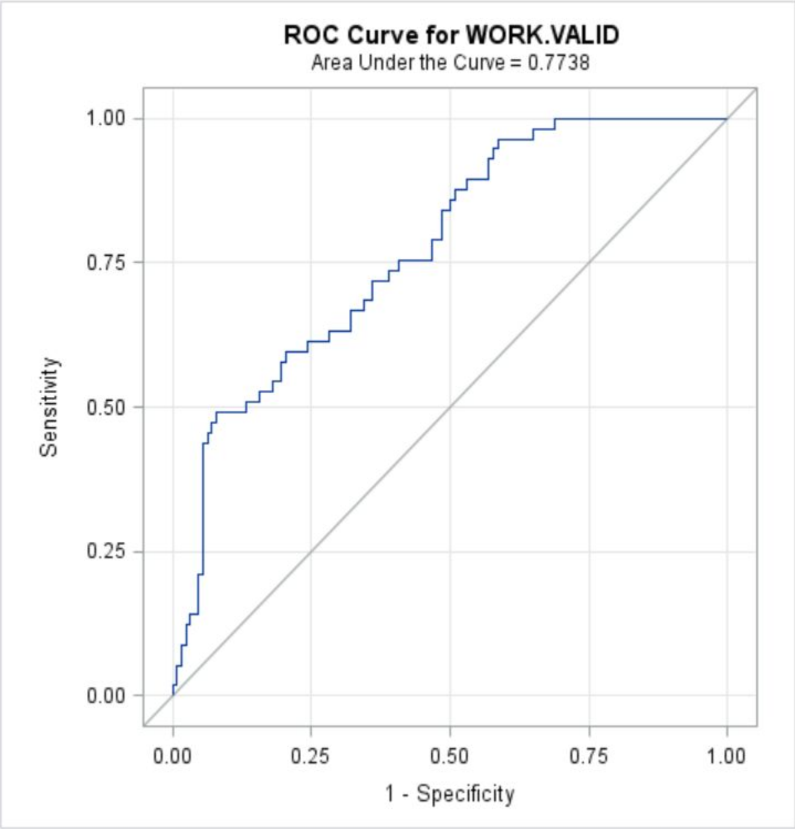
Eigenvectors

| | Prin1 | Prin2 | Prin3 | Prin4 | Prin5 | Prin6 | Prin7 | Prin8 | Prin9 |
|---|---|---|---|---|---|---|---|---|---|
| sbp | 0.318227 | 0.163177 | -.213102 | 0.154204 | 0.819826 | 0.141897 | 0.268586 | 0.195882 | -.012510 |
| tobacco | 0.313298 | 0.477471 | 0.059609 | -.126550 | -.314835 | -.494019 | 0.388476 | 0.397794 | -.044361 |
| ldl | 0.329548 | -.318762 | 0.167212 | -.247413 | -.254987 | 0.579215 | 0.543091 | 0.014290 | 0.070386 |
| adiposity | 0.495637 | -.263212 | -.109376 | 0.147778 | -.096747 | -.113343 | -.176545 | -.140733 | -.760317 |
| typea | -.007083 | -.111074 | 0.860617 | 0.214080 | 0.246836 | -.254088 | 0.169562 | -.212419 | -.041945 |
| obesity | 0.374465 | -.471809 | -.008197 | 0.355974 | -.088370 | -.190544 | -.253304 | 0.383013 | 0.505381 |
| alcohol | 0.115181 | 0.491556 | 0.093281 | 0.684144 | -.273018 | 0.423157 | -.079607 | -.086465 | 0.030615 |
| age | 0.455661 | 0.166752 | -.139171 | -.173712 | 0.011971 | -.183447 | -.045924 | -.722791 | 0.395842 |
| chd | 0.291804 | 0.264822 | 0.377244 | -.454288 | 0.104725 | 0.271827 | -.591161 | 0.248308 | -.006564 |

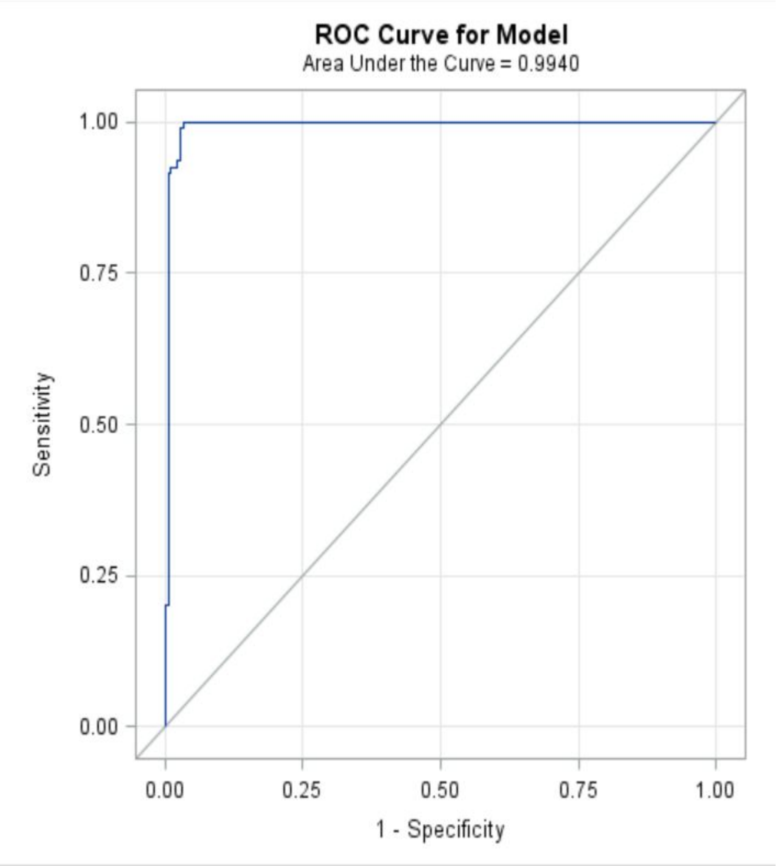
- PC1 has max loading for adiposity and age
- PC2 has max loading for tobacco, alcohol and obesity
- PC3 has max loading for typea and chd
- PC4 has max loading for alcohol and chd

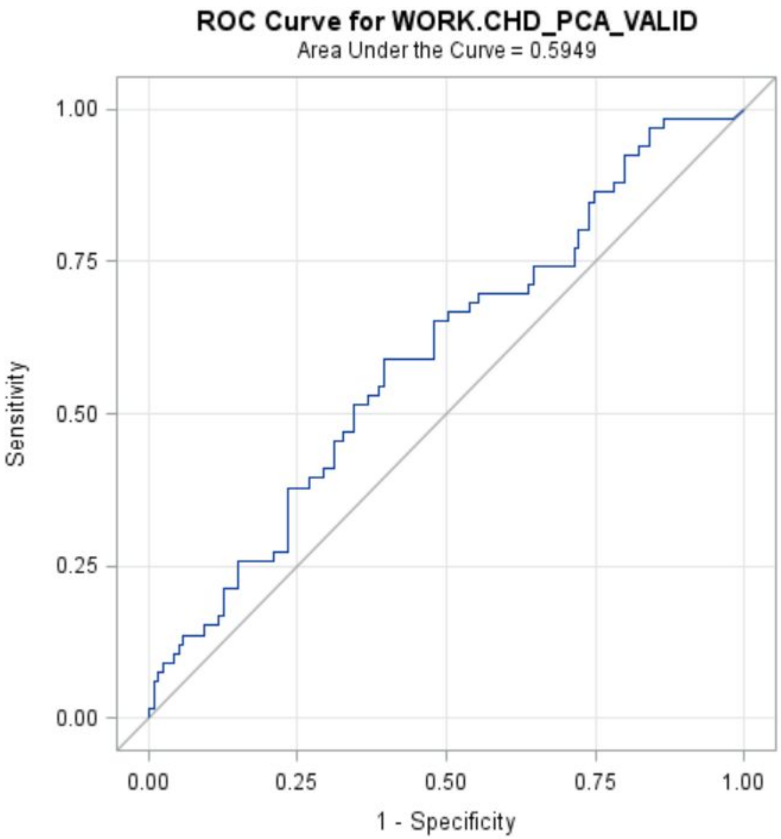
# ROC without PCA



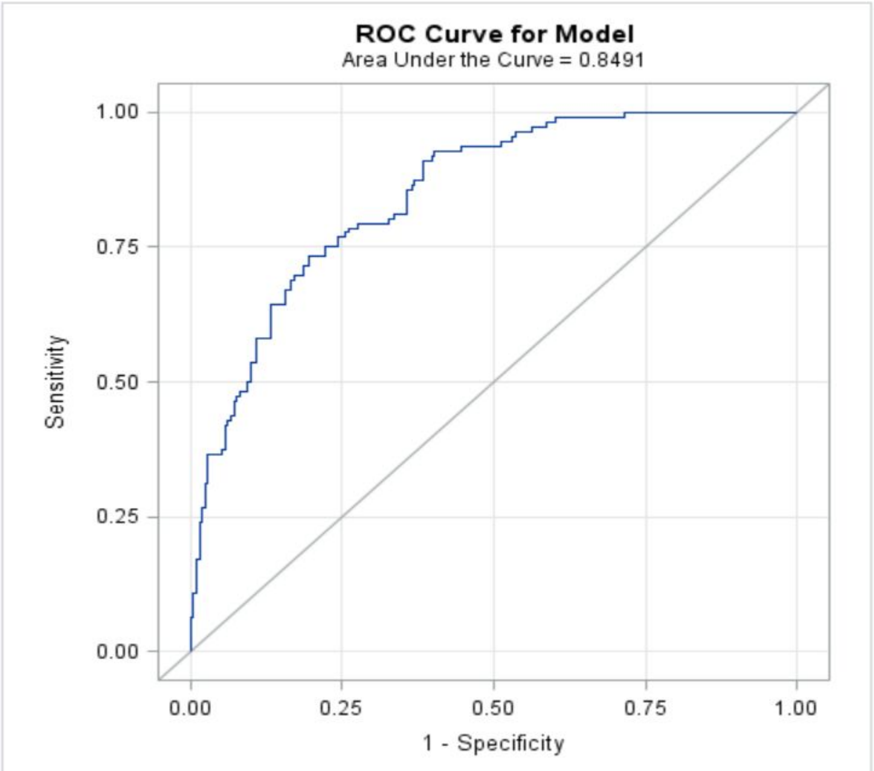


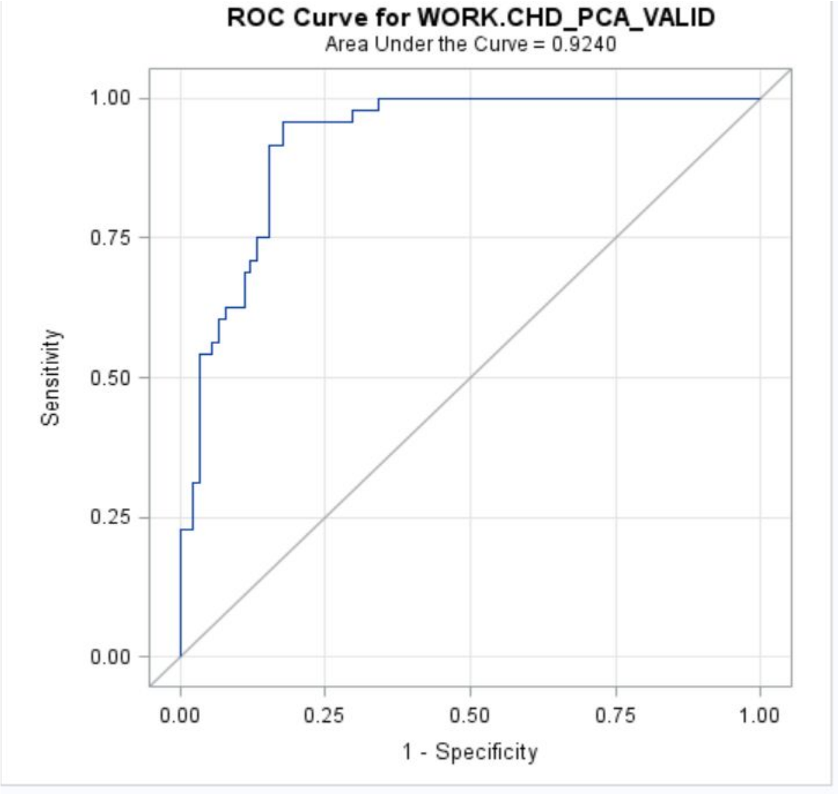
# ROC With PCA (Overfitting)





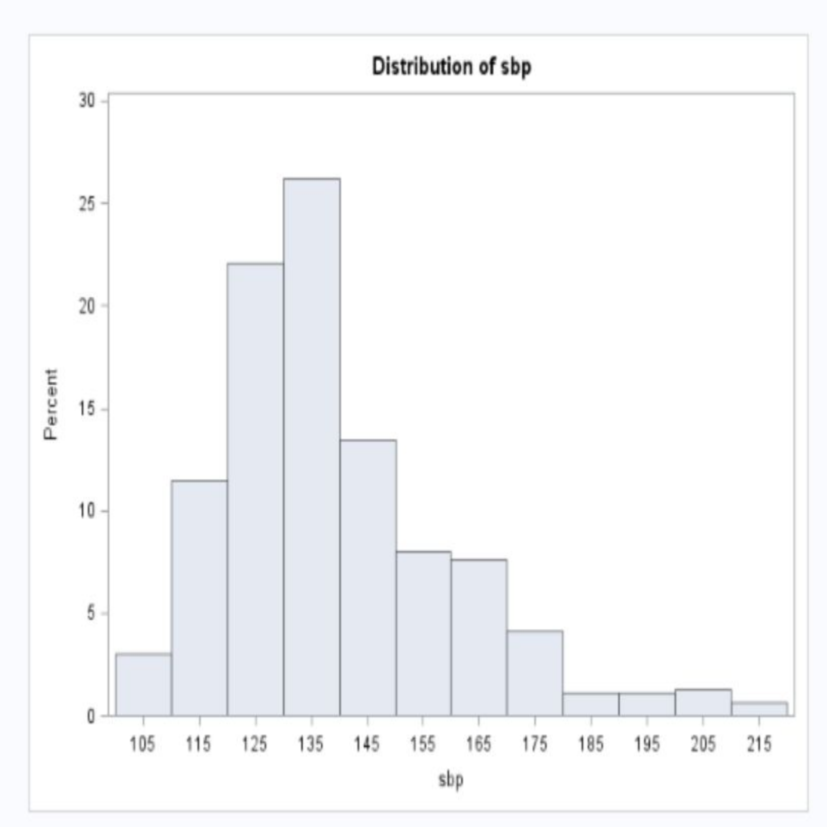
# Cross- Validated ROC

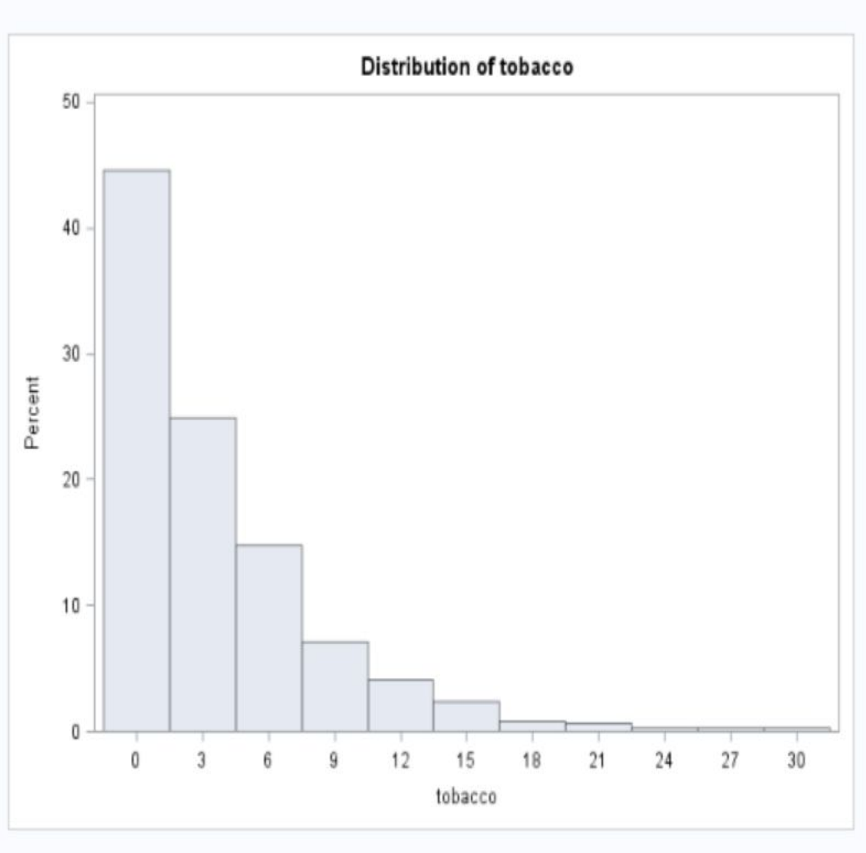




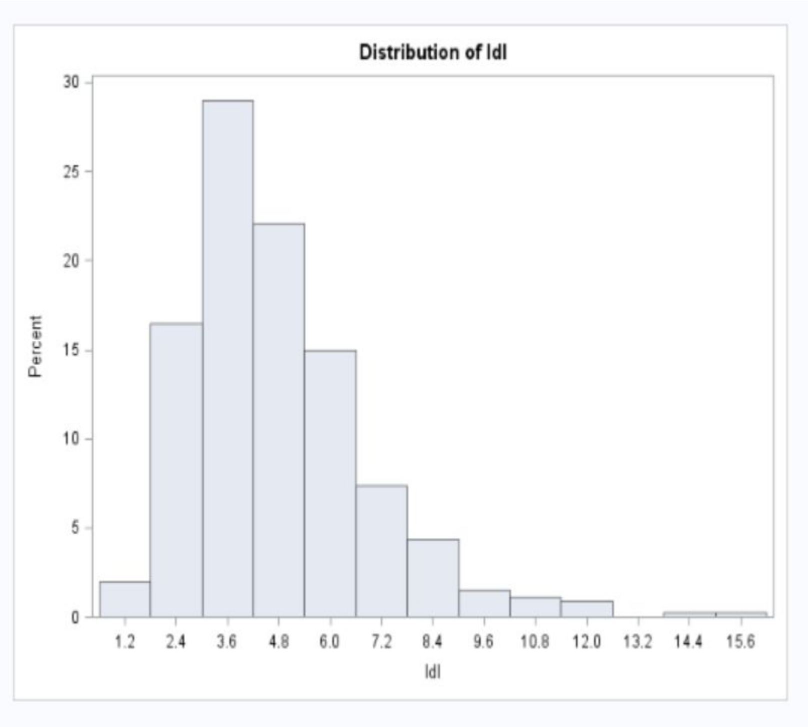
# DISTRIBUTION OF FEATURES

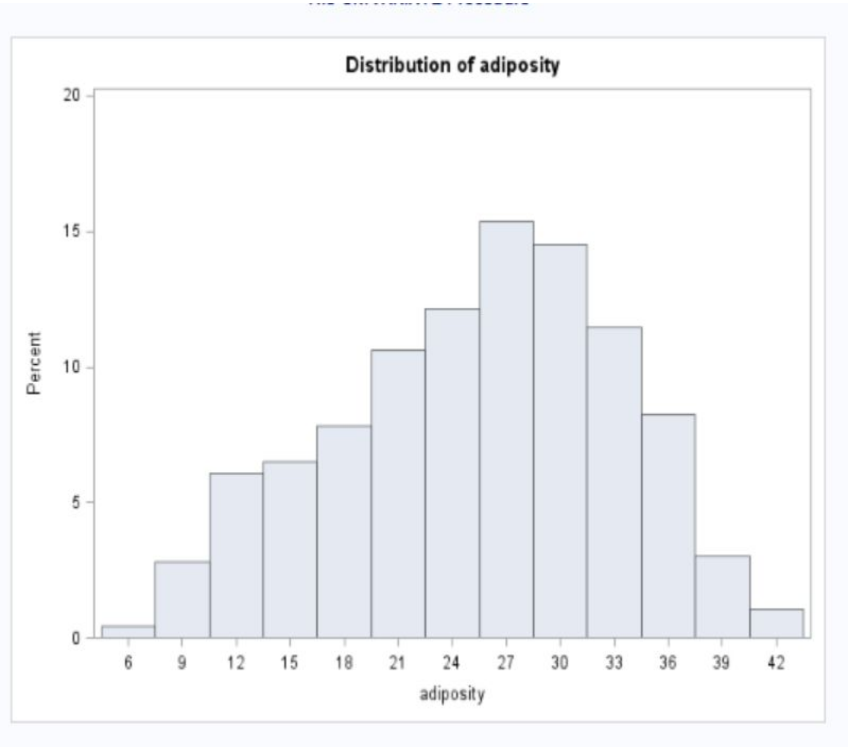
- Proc *univariate* is used to study whether each features are normally distributed



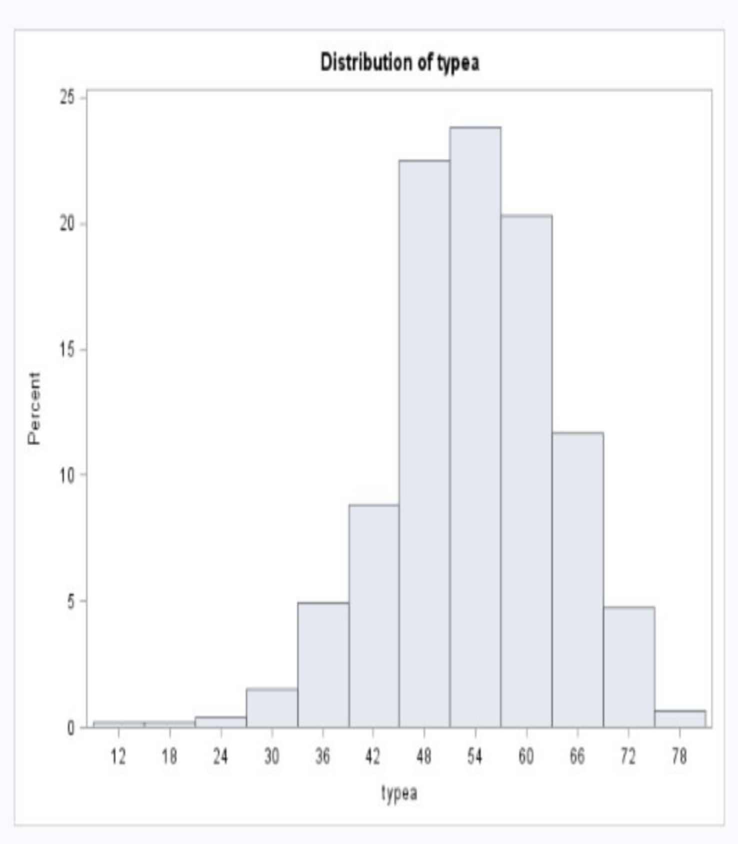


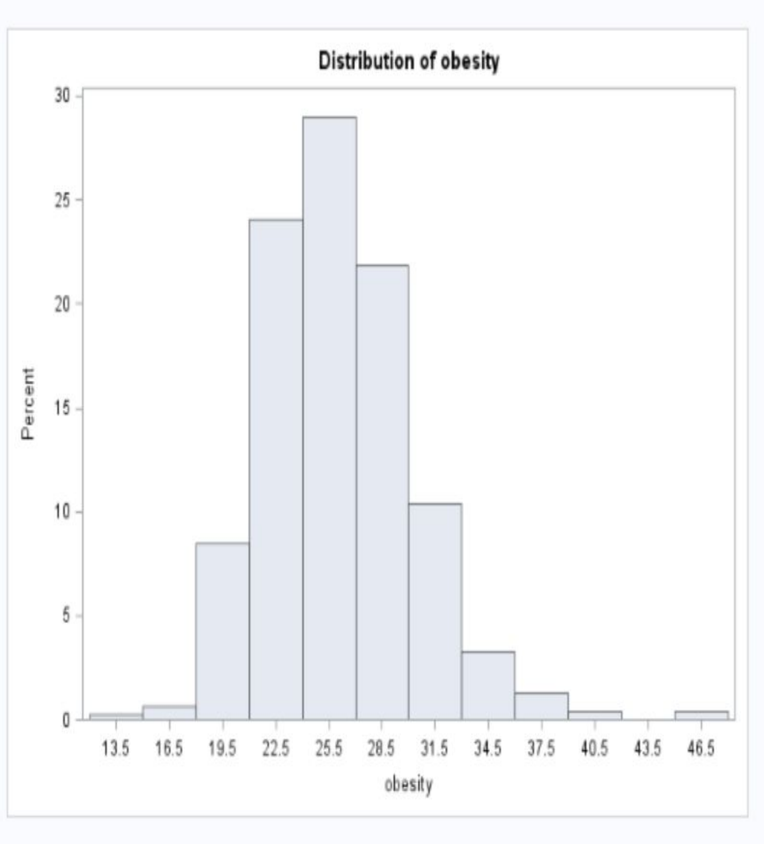
# DISTRIBUTION OF FEATURES



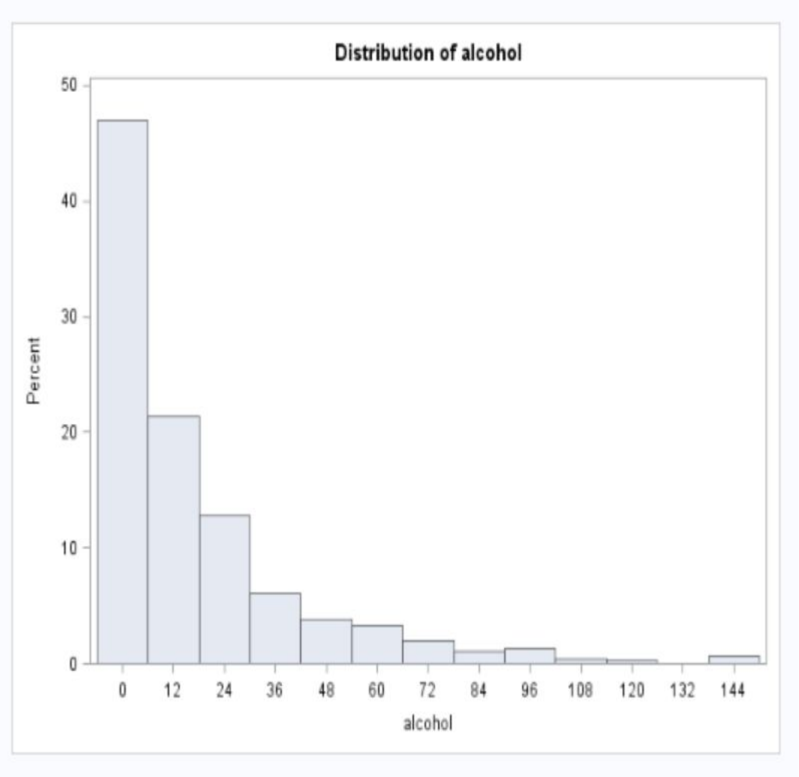


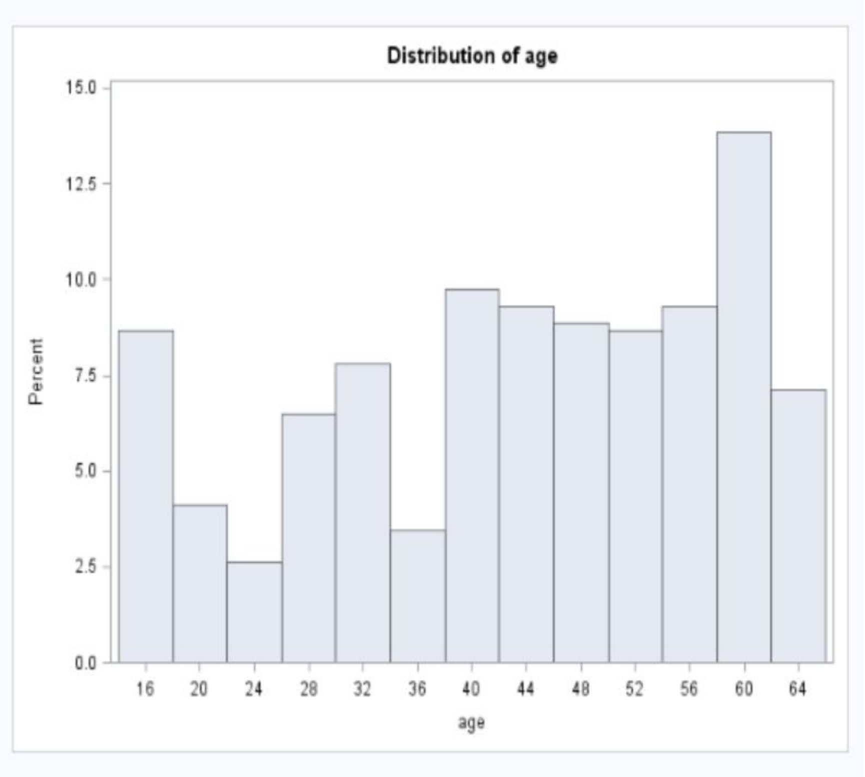
# DISTRIBUTION OF FEATURES





# DISTRIBUTION OF FEATURES



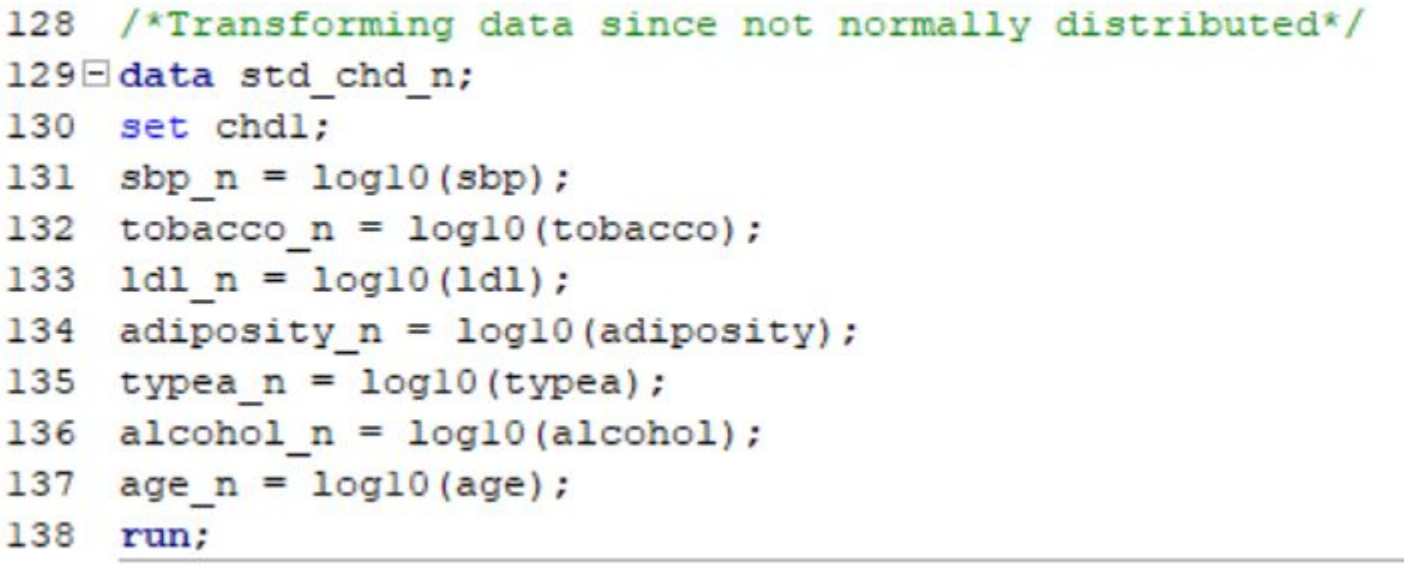


- We can see from the histograms that none of the features are normally distributed as they are either left or right skewed.

# RANDOM FOREST CLASSIFICATION

- All missing values in the datasets are initially replaced with the mean of the corresponding attributes.

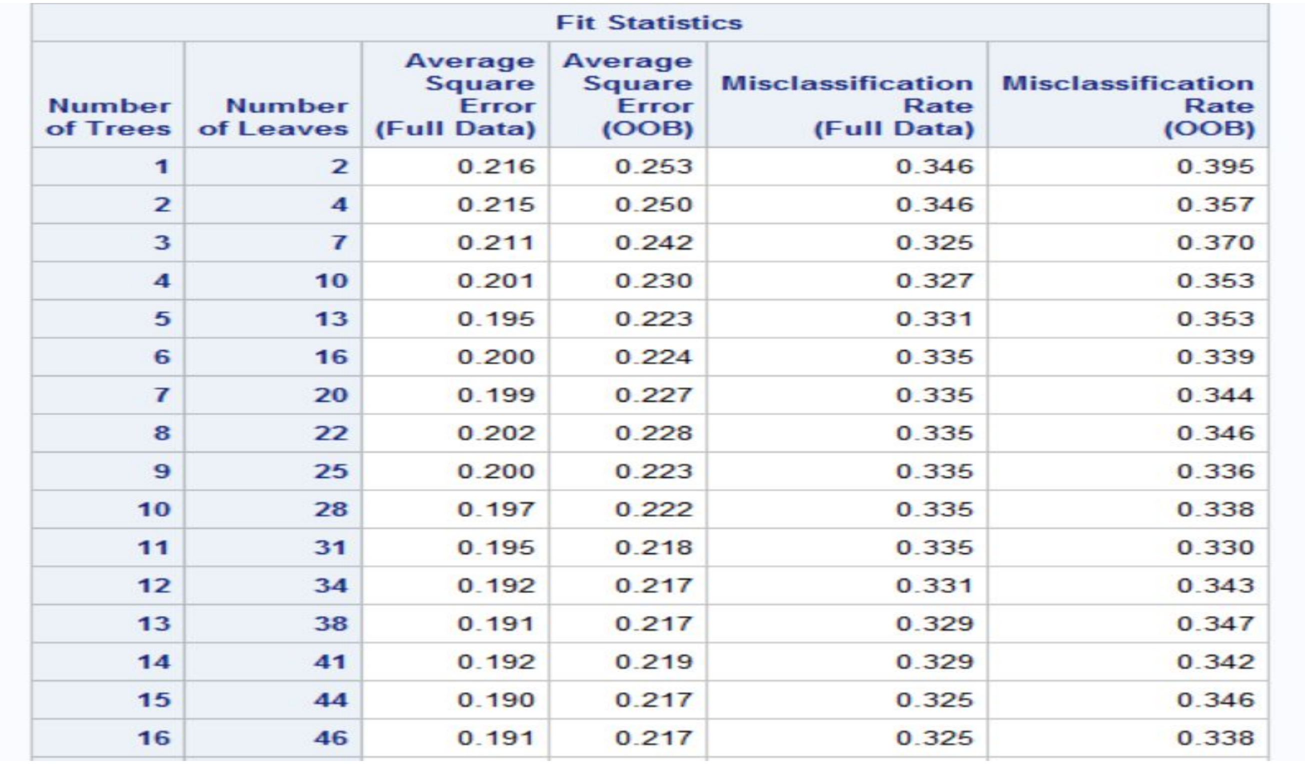
- As the features are not normally distributed, logarithmic transformation is performed on each of the 7 attributes.



```
/*Transforming data since not normally distributed*/
data std_chd_n;
set chdl;
sbp_n = log10(sbp);
tobacco_n = log10(tobacco);
ldl_n = log10(ldl);
adiposity_n = log10(adiposity);
typea_n = log10(typea);
alcohol_n = log10(alcohol);
age_n = log10(age);
run;
```

# RANDOM FOREST CLASSIFICATION

- Used proc *hpforest* on the dataset where missing values are replaced by their mean value.

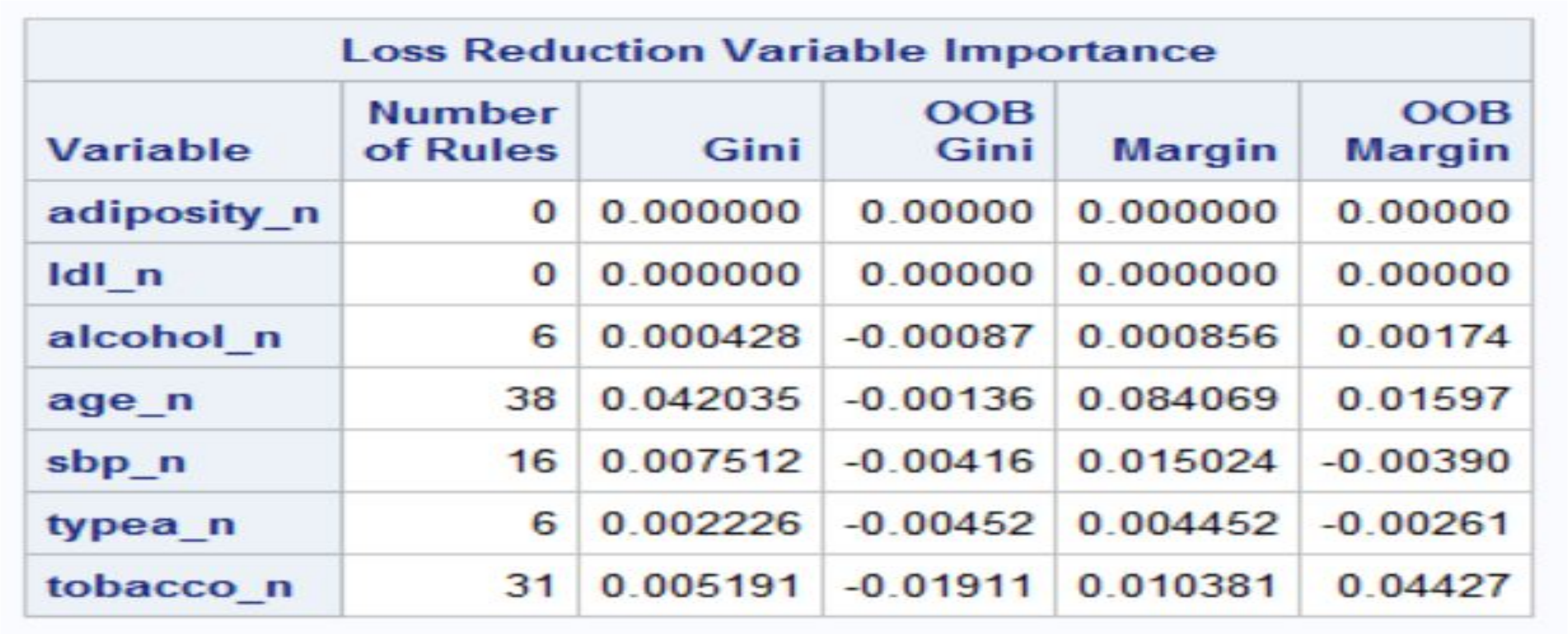


Fit Statistics

| Number of Trees | Number of Leaves | Average Square Error (Full Data) | Average Square Error (OOB) | Misclassification Rate (Full Data) | Misclassification Rate (OOB) |
|---|---|---|---|---|---|
| 1 | 2 | 0.216 | 0.253 | 0.346 | 0.395 |
| 2 | 4 | 0.215 | 0.250 | 0.346 | 0.357 |
| 3 | 7 | 0.211 | 0.242 | 0.325 | 0.370 |
| 4 | 10 | 0.201 | 0.230 | 0.327 | 0.353 |
| 5 | 13 | 0.195 | 0.223 | 0.331 | 0.353 |
| 6 | 16 | 0.200 | 0.224 | 0.335 | 0.339 |
| 7 | 20 | 0.199 | 0.227 | 0.335 | 0.344 |
| 8 | 22 | 0.202 | 0.228 | 0.335 | 0.346 |
| 9 | 25 | 0.200 | 0.223 | 0.335 | 0.336 |
| 10 | 28 | 0.197 | 0.222 | 0.335 | 0.338 |
| 11 | 31 | 0.195 | 0.218 | 0.335 | 0.330 |
| 12 | 34 | 0.192 | 0.217 | 0.331 | 0.343 |
| 13 | 38 | 0.191 | 0.217 | 0.329 | 0.347 |
| 14 | 41 | 0.192 | 0.219 | 0.329 | 0.342 |
| 15 | 44 | 0.190 | 0.217 | 0.325 | 0.346 |
| 16 | 46 | 0.191 | 0.217 | 0.325 | 0.338 |

The misclassification rate is lowest at 0.325 and the optimal number of trees in our model is 15

# RANDOM FOREST CLASSIFICATION



Loss Reduction Variable Importance

| Variable | Number of Rules | Gini | OOB Gini | Margin | OOB Margin |
|---|---|---|---|---|---|
| adiposity_n | 0 | 0.000000 | 0.00000 | 0.000000 | 0.00000 |
| ldl_n | 0 | 0.000000 | 0.00000 | 0.000000 | 0.00000 |
| alcohol_n | 6 | 0.000428 | -0.00087 | 0.000856 | 0.00174 |
| age_n | 38 | 0.042035 | -0.00136 | 0.084069 | 0.01597 |
| sbp_n | 16 | 0.007512 | -0.00416 | 0.015024 | -0.00390 |
| typea_n | 6 | 0.002226 | -0.00452 | 0.004452 | -0.00261 |
| tobacco_n | 31 | 0.005191 | -0.01911 | 0.010381 | 0.04427 |

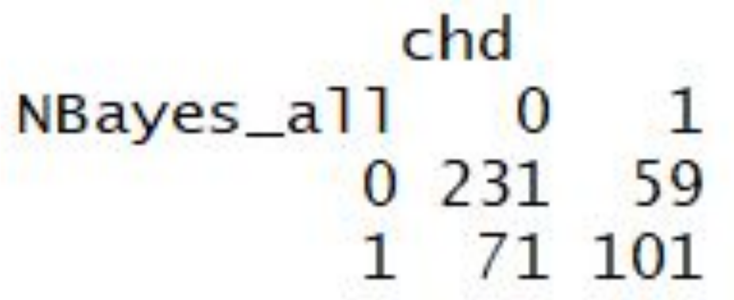
- The above variables in order contributes the most in prediction of chd in individuals.

# Naive Bayes

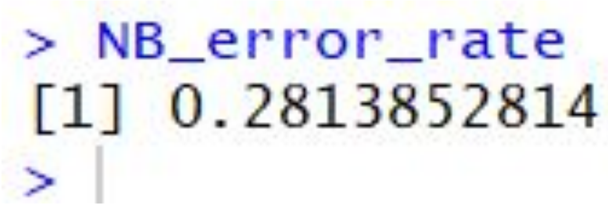
The sampling used is Simple random Sampling.

Test and train in the ratio 30- 70

Used Seed in R for Random Sampling.



```
           chd
NBayes_all   0   1
         0 231  59
         1  71 101
```



```
> NB_error_rate
[1] 0.2813852814
> 
```

# Conclusion

Analyzing various output from Different Classifiers.

- Logistic Regression - Error rate - 20%
- Random Forest - Error Rate- 33%
- Naive Bayes - Error rate- 28 % with high false positive

Hence logistic regression is the best classifier for our dataset. since the data is related to healthcare we need minimum false positive.