



# Predictive Classification

## Risk Factors for Heart Disease

### Group 2:

Jessie Browne

Sindhu Sai Sangani

Rama Sai Arun Varma Penmasta

Murali Krishna Vattikunta



# Presentation Outline

## Introduction

- Heart Disease Predictive Classification Problem

## Models and Results

- Exploratory Data Analysis in Python
- Linear Discriminant Analysis (LDA)
- Decision Tree
- Support Vector Machine (SVM)

## Discussion and Conclusions

---

# Heart Disease: Predictive Classification

## Personal Key Indicators of Heart Disease

2020 annual CDC survey data of >400k adults related to their health status

Dataset was retrieved from Kaggle

- 584,844 observations
- 18 features (9 booleans, 5 strings, and 4 decimals)



HeartDisease, BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, DiffWalking, Sex, AgeCategory, Race, Diabetic, PhysicalActivity, GenHealth, SleepTime, Asthma, KidneyDisease, SkinCancer



# Predictive Classification Model

## The goal of this project:


- Predict the likelihood of being diagnosed with heart disease using 17 covariates


## Motivation:

- Some risk factors are preventable, such as smoking amount of sleep each night
- Predicting high heart disease risk allows for intervention

---

# Exploratory Data Analysis in Python

- 
- Identifying the null values and datatypes of each data attribute.
  - Identifying the numerical and categorical variables.
  - Using visualization from matplotlib and seaborn to understand the relation between the numerical attributes.
  - None of numerical attributes alone is a indicator of heart disease.
  - Using Kernel Density Plot observed there is any strong correlation in categorical variables.

- 
- Converting the categorical variables to numerical by using label encoder.
  - Standardization of numerical data using StandardScaler.
  - Splitting the train and test data.(80/20 ratio)
  - Data set is imbalanced used oversampling technique to make it balanced.
  - Converted the dataframes into csv files train and test as an input to statistical methods.



---

# Linear Discriminant Analysis

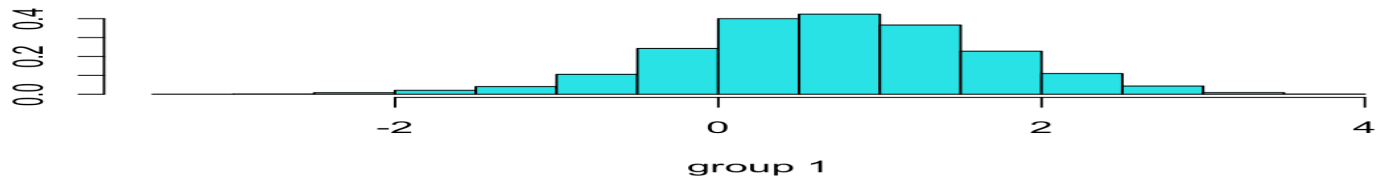
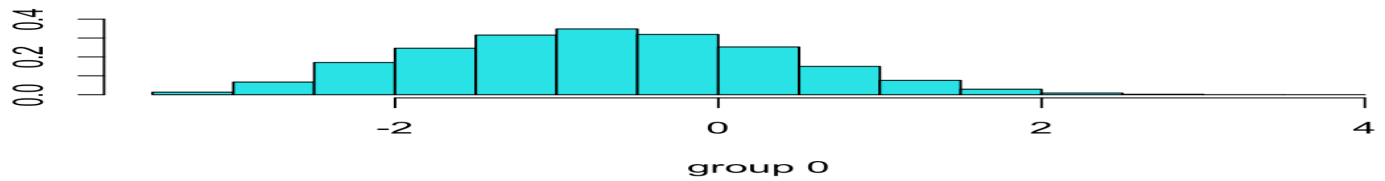
Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0  
<https://cran.r-project.org/web/packages/MASS/citation.html>




## Basic Concepts

- LDA is one of the popular machine learning techniques, which is mainly used to tackle classification problems rather than the supervised classification problems. It's fundamentally a dimensionality decrease procedure. Utilizing the Linear blends of indicators, LDA attempts to foresee the class of the offered perspectives.
- LDA can be computed in R using the `lda()` function which is present in the package MASS. LDA is used to determine group means and also for each individual.

# Histogram:



- 
- Initially we have trained 10000 samples and tested 2000 samples of data and achieved the accuracy as below:
    - Training accuracy: 0.76 %
    - Testing accuracy: 0.75 %
  - When we ran this model for whole trained dataset and test dataset, we achieved the below accuracy:
    - Training accuracy: 0.76 %
    - Testing accuracy: 0.76 %
  - From the obtained accuracy in both the cases we can conclude that, even if we increase the training and testing dataset there would not be much change in the accuracy for this model.

---

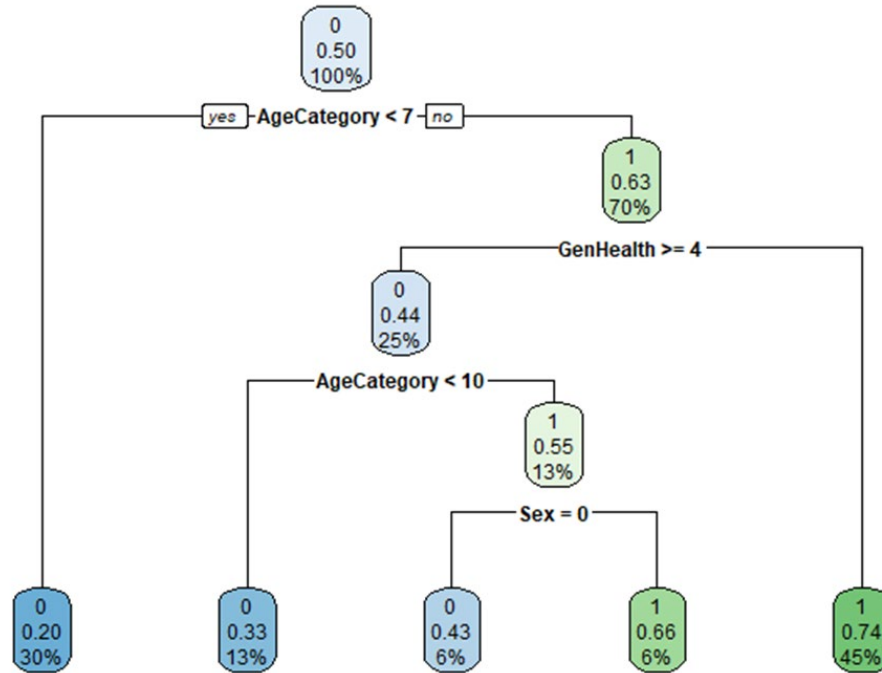
# Decision Trees

*Terry Therneau and Beth Atkinson (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. <https://CRAN.R-project.org/package=rpart>*



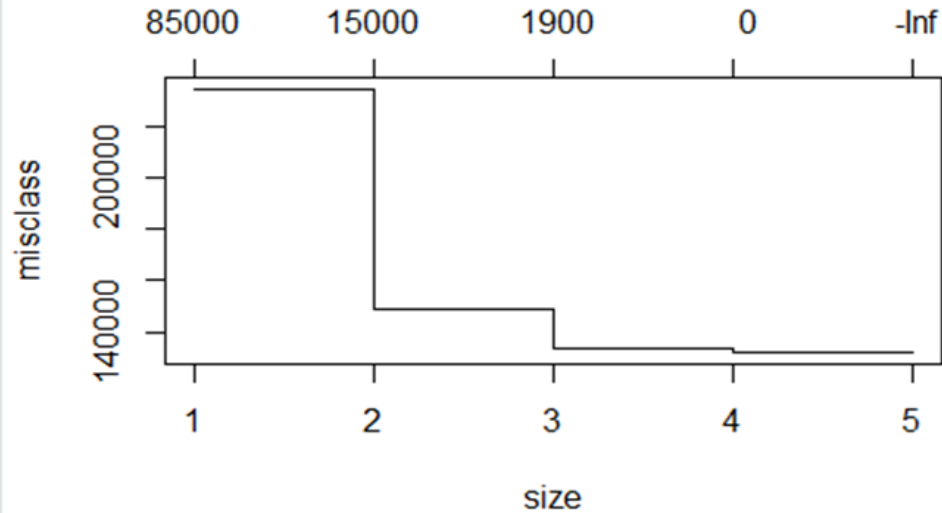
## Basic Concepts

- Decision tree model works on both classification and regression methods. The only difference between classification and regression tree is that in the former we use the categorical variable as target variable whereas in the latter we use continuous variable as target variable.
- It is basically a tree like structure which has target node/root node, decision node and terminal node.
- The main components of a decision tree model are nodes and branches and the most important steps in building a model are splitting, stopping, and pruning.

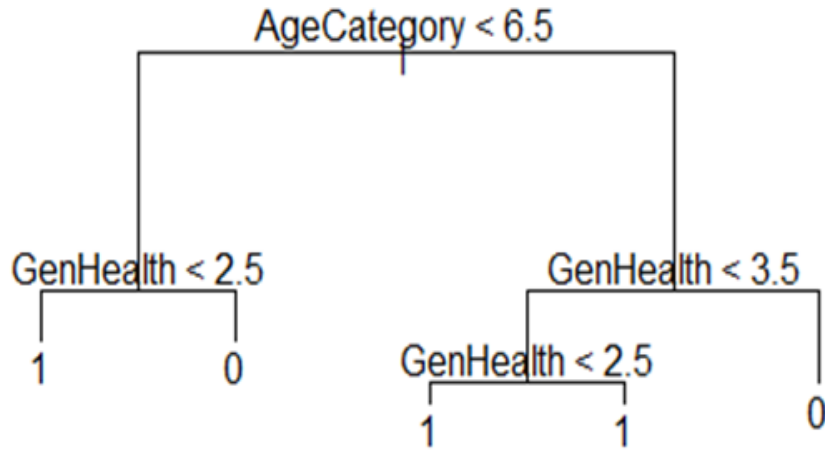


- The first importance covariate is AgeCategory(55-59) in this model followed by the GenHealth(Very good) and the gender
- 30% probability of people in AgeCategory(55-59) reportedly have heart disease
- And the remaining people who are not in this category have 63% probability of having heart disease.
- The accuracy reported on test data is 72.99%

## Low deviation at 5th sized tree







- Performed Post Pruning on trained data to eliminate the imbalance between the train set and validation set.
- Used Cross Validation on the data using `cv.tree` function
- The accuracy reported after pruning is 71.6 percent which is validly classified against the test observations.
- The Pruning process provided better understanding of tree at the cost of little classification accuracy rate.

---

# Support Vector Machine

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2021). *e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), TU wien*. Retrieved from <https://CRAN.R-project.org/package=e1071>



# SVM Parameters

## Kernel Function

- projects data from a two-dimensional space to a higher dimension

$$K(x_1, x_2) = \Phi(x_1) \cdot \Phi(x_2)$$

## Cost Parameter

- Bounds the sum of  $\epsilon_i$
- When  $C = 0$ , no violations are allowed
- Higher values of  $C$  allows more violation of hyperplane

## Exploratory SVM Models

Kernel	Cost	Accuracy	Sensitivity	Precision
Linear	0.001	0.745	0.814433	0.7053571
Radial	0.1	0.747	0.8061856	0.7109091
Quadratic	0.1	0.75	0.7690722	0.7299413
Polynomial	0.1	0.745	0.7587629	0.7272727
Sigmoid	0.01	0.49	0.0804124	0.3786408

### Method:

- Random sample from the training data,  $n = 1000$ 
  - Tune cost parameters
- Random sample from the the training data,  $n = 10,000$ 
  - Fit exploratory models
- Select kernel: linear
  - Highest sensitivity



## SVM Model Ensemble

### Method:

- Linear Kernel
- Cost parameter = 0.001
- Training partitioned,  $n = 15,000$
- Each member predicted test set,  $n = 117,142$
- Weighted average of model predictions
  - Accuracy range 0.73 to 0.75
- If prediction  $> 0.5$ , 1, else 0

SVM Ensemble Evaluation	
Accuracy	0.744
Sensitivity	0.821
Precision	0.702

---

# Comparison of Models



**The LDA model predictions have the highest accuracy, precision, and recall.**

Model	Accuracy	Precision	Recall
LDA	0.760	0.774	0.753
Decision Tree	0.716	0.715	0.717
SVM ensemble	0.744	0.702	0.821

---

# Conclusion





## Key Points

- In our work, we applied LDA, Decision trees and SVM techniques on our dataset to check if the person has a heart disease and how all the covariates are in relation with the target variable.
- As the data is highly imbalanced, we have performed oversampling techniques in python to overcome the underfitting/overfitting of the data.
- We observed that LDA is the most effective technique with respect to our dataset as LDA performs well on binary classification data.



## Future Direction

### Our analysis could be improved by:

- Applying discoveries concerning feature importance from LDA and decision tree to the SVM model (feature reduction)
- stratifying subsamples for the SVM exploratory and ensemble models
- applying a stochastic gradient descent (SGD) SVM, which has lower complexity than SVM and could process the full dataset