WISE
厦|门|大|学|王|亚|南|经|济|研|究|院
The Wang Yanan Institute for Studies in Economics, Xiamen University

关于肥胖的研究

——基于Random Forest和Adaboost

汇报人：王湘晴、邱芸茜

汇报时间：2024年5月25日

小组分工

WISE
厦|门|大|学|王|亚|南|经|济|研|究|院
The Wang Yanan Institute for Studies in Economics, Xiamen University

姓名	分工
王湘晴	数据描述性统计，模型拟合、检验
邱芸茜	研究背景，数据、方法介绍

目录
CATALOGUE

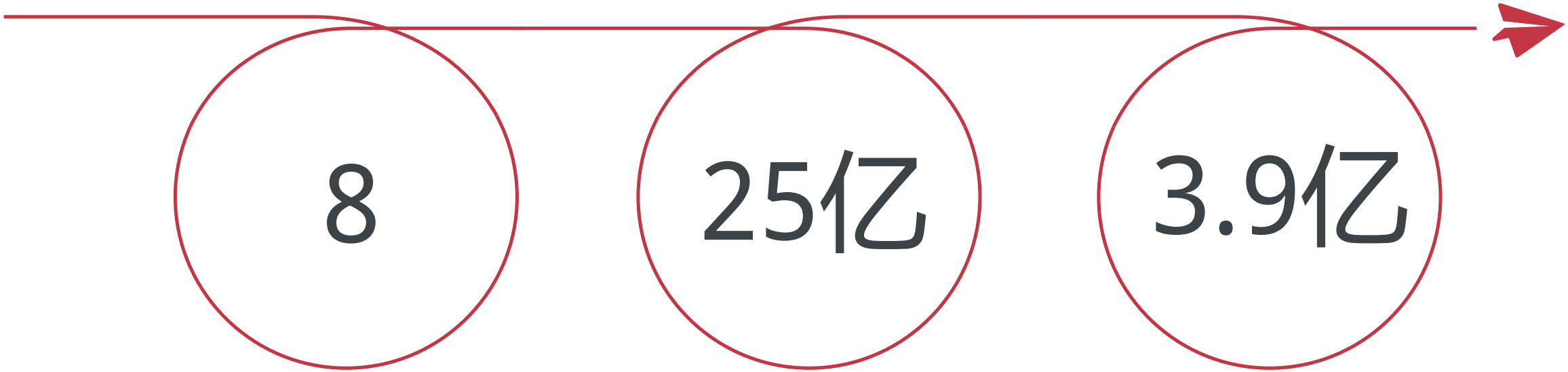
WISE
厦|门|大|学|王|亚|南|经|济|研|究|院
The Wang Yanan Institute for Studies in Economics, Xiamen University

01.

研 究 背 景

研究背景

研究背景



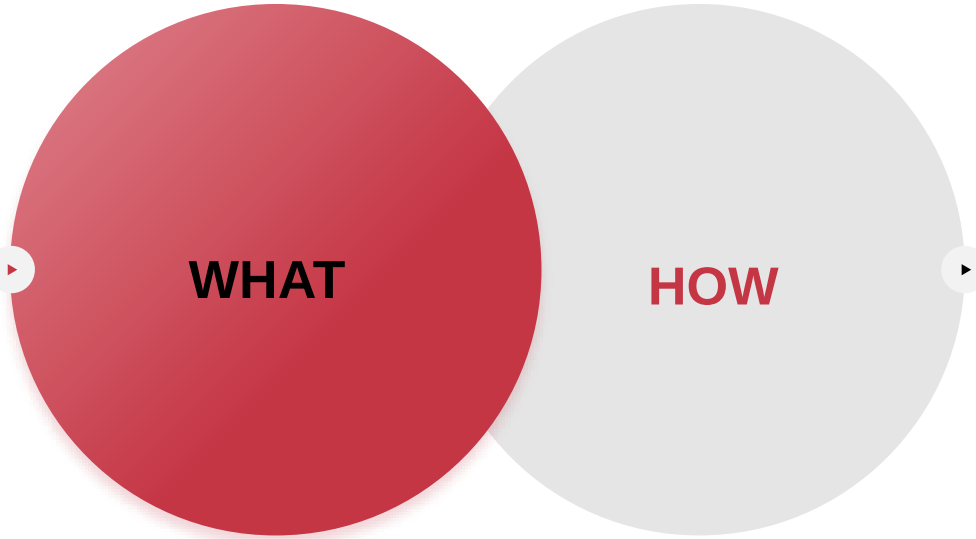
- 2022年，世界上每8人中就有1人患有肥胖症。
- 2022年，有25亿成人（18 岁及以上）超重。其中，8.9亿人患有肥胖症。
- 2022年，超过3.9亿5-19岁儿童和青少年超重，其中1.6亿患有肥胖症。

肥胖会增加患糖尿病和心脏病的风险，会影响到骨骼健康和生殖系统，并增加罹患某些癌症的风险。肥胖对生活质量造成影响，如睡眠或活动。在生理和心理层面都会对人带来巨大伤害。

研究背景

WISE
厦|门|大|学|王|亚|南|经|济|研|究|院
The Wang Yanan Institute for Studies in Economics, Xiamen University

什么影响肥胖？
哪些因素？



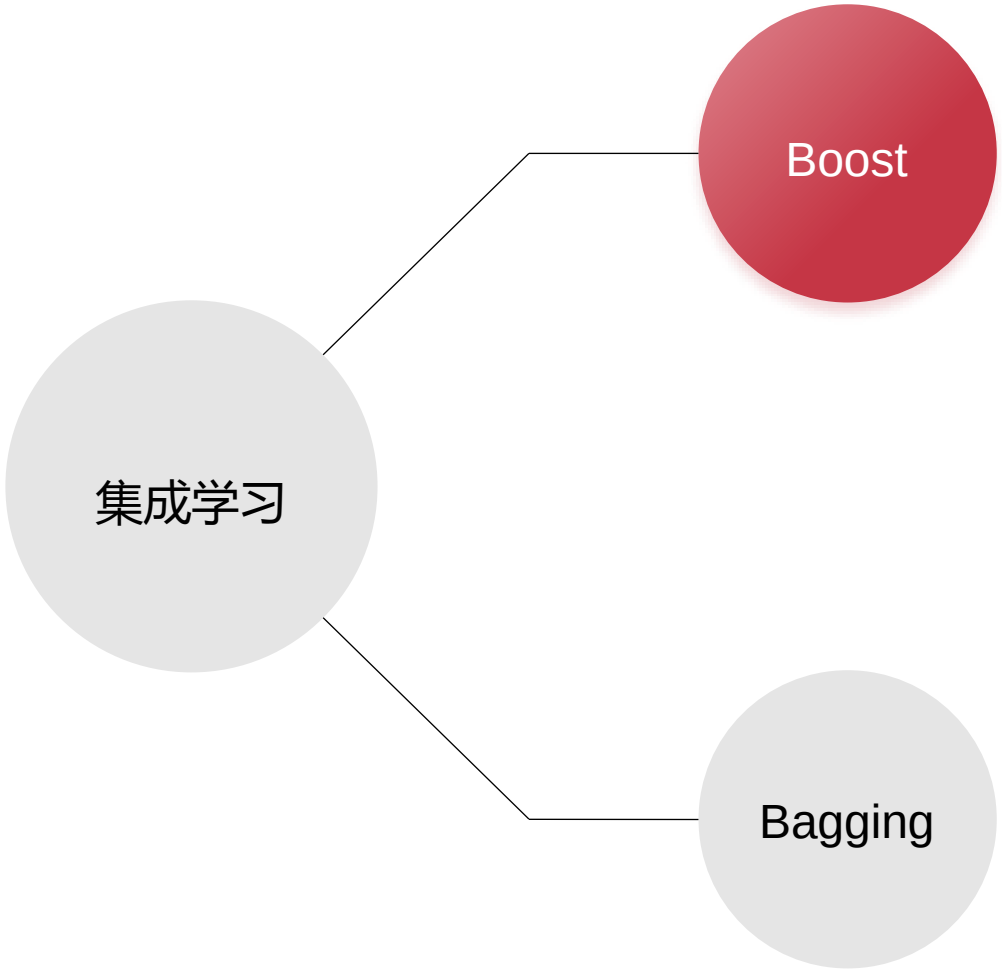
用什么方法去分类，预测？
效果的好坏？

WISE
厦|门|大|学|王|亚|南|经|济|研|究|院
The Wang Yanan Institute for Studies in Economics, Xiamen University

02.

研究方法介绍

研究方法介绍



Adaboost

随机森林 Random Forest

研究方法介绍——随机森林

第1步：T中共有N个样本，有放回的随机选择N个样本来训练一个决策树，作为决策树根节点处的样本。

第2步：当每个样本有M个属性时，在决策树的每个节点需要分裂时，随机从这M个属性中选取出m个属性，满足条件m << M。然后从这m个属性中采用信息增益等选择来选择1个属性作为该节点的分裂属性。

第3步：每个节点都要按照步骤2来分裂，一直到不能够再分裂为止。

第4步：按照步骤1~3建立大量的决策树，这样就构成了随机森林。

众多决策树构成了随机森林，每棵决策树都会有一个投票结果，最终投票结果最多的类别，就是最终的模型预测结果。

优点

可以处理大量的输入变量

在决定类别时评估变量的重要性

对于不平衡的分类资料集来说，可以平衡误差

缺点

会出现过拟合问题

在多个分类变量的问题种，随机森林无法提高基学习器的准确性

研究方法介绍——adaboost

第1步：对所有观测数据初始化权重。

第2步：在数据子集上建立一个模型。利用该模型对整个数据集进行预测。通过比较预测值和实际值来计算误差。

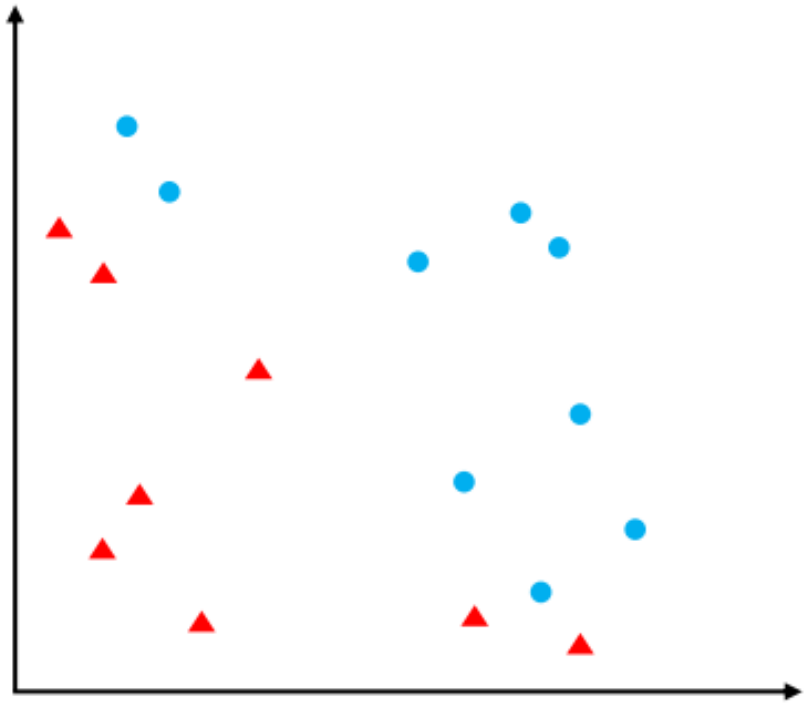
第3步：根据误差调整数据集的权重，误差越大，分配给观测点的权重就越大。

第4步：重复2-3步，直到达到某个预定的足够小的错误率或预先指定的最大迭代次数。

第5步：将所有的模型结果进行加权组合。

优点：可以迭代纠正弱分类器的错误，并通过组合弱学习器来提高准确率；AdaBoost 不容易过度拟合。

缺点：AdaBoost 对噪声数据很敏感；受异常值的影响很大。



研究方法介绍——adaboost

第1步：对所有观测数据初始化权重。

第2步：在数据子集上建立一个模型。利用该模型对整个数据集进行预测。通过比较预测值和实际值来计算误差。

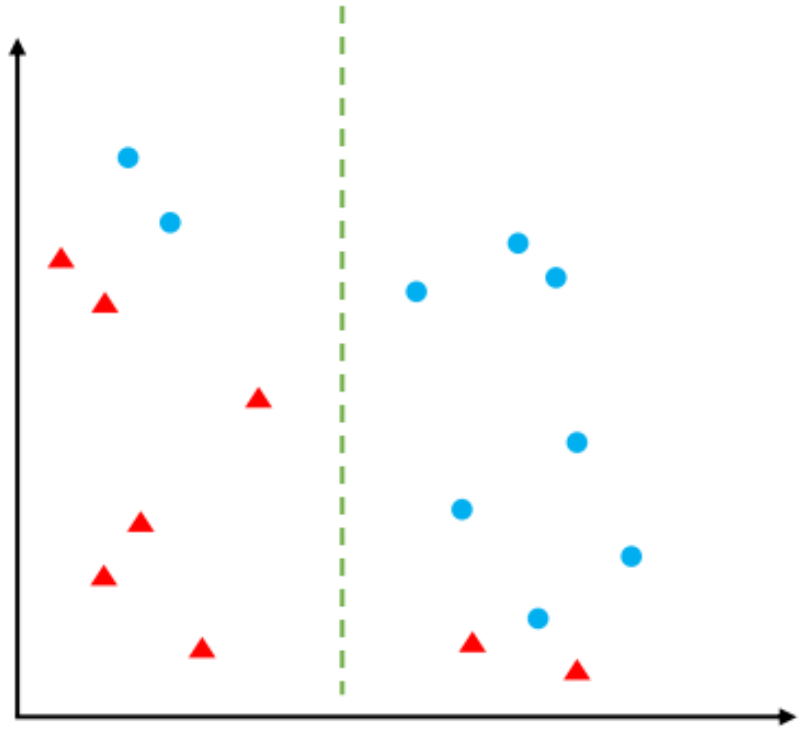
第3步：根据误差调整数据集的权重，误差越大，分配给观测点的权重就越大。

第4步：重复2-3步，直到达到某个预定的足够小的错误率或预先指定的最大迭代次数。

第5步：将所有的模型结果进行加权组合。

优点：可以迭代纠正弱分类器的错误，并通过组合弱学习器来提高准确率；AdaBoost 不容易过度拟合。

缺点：AdaBoost 对噪声数据很敏感；受异常值的影响很大。



研究方法介绍——adaboost

第1步：对所有观测数据初始化权重。

第2步：在数据子集上建立一个模型。利用该模型对整个数据集进行预测。通过比较预测值和实际值来计算误差。

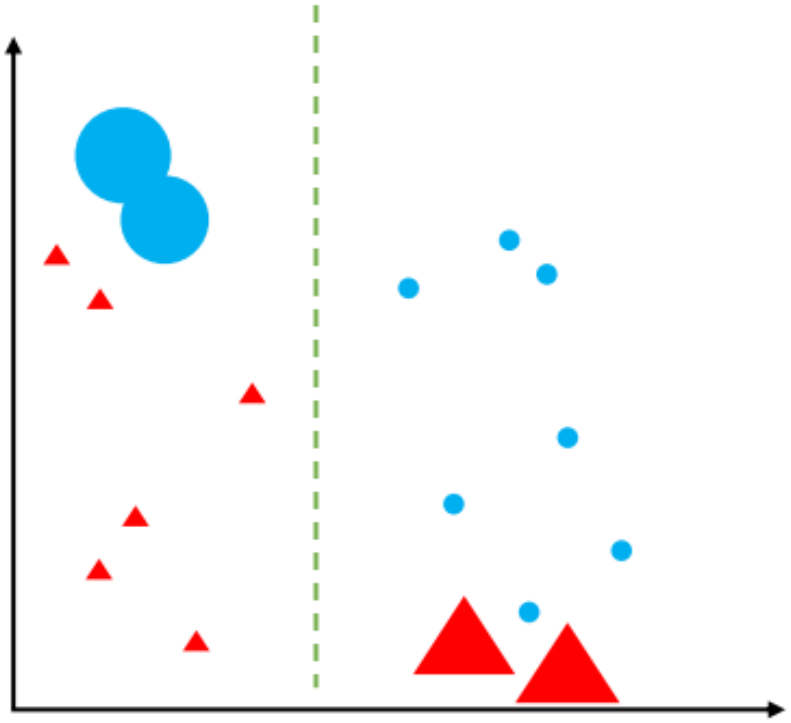
第3步：根据误差调整数据集的权重，误差越大，分配给观测点的权重就越大。

第4步：重复2-3步，直到达到某个预定的足够小的错误率或预先指定的最大迭代次数。

第5步：将所有的模型结果进行加权组合。

优点：可以迭代纠正弱分类器的错误，并通过组合弱学习器来提高准确率；AdaBoost 不容易过度拟合。

缺点：AdaBoost 对噪声数据很敏感；受异常值的影响很大。



研究方法介绍——adaboost

第1步：对所有观测数据初始化权重。

第2步：在数据子集上建立一个模型。利用该模型对整个数据集进行预测。通过比较预测值和实际值来计算误差。

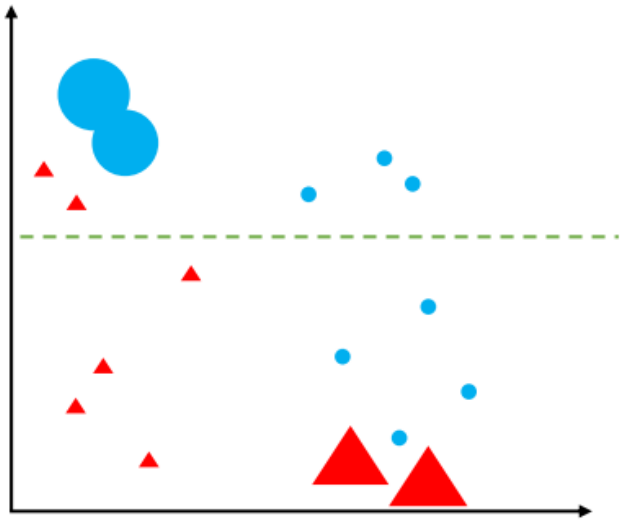
第3步：根据误差调整数据集的权重，误差越大，分配给观测点的权重就越大。

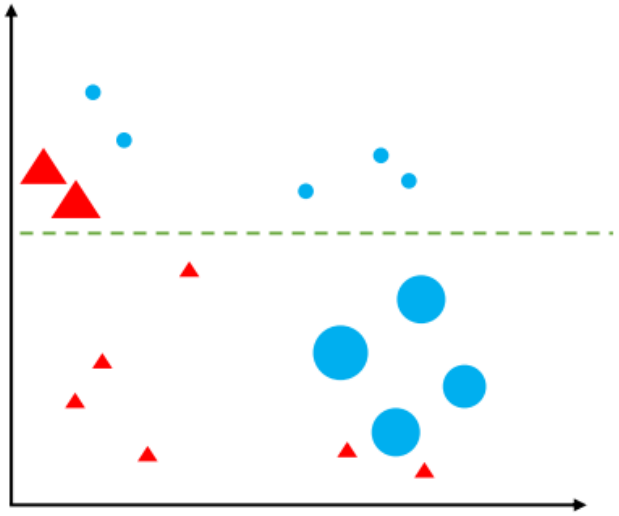
第4步：重复2-3步，直到达到某个预定的足够小的错误率或预先指定的最大迭代次数。

第5步：将所有的模型结果进行加权组合。

优点：可以迭代纠正弱分类器的错误，并通过组合弱学习器来提高准确率；AdaBoost 不容易过度拟合。

缺点：AdaBoost 对噪声数据很敏感；受异常值的影响很大。





研究方法介绍——adaboost

第1步：对所有观测数据初始化权重。

第2步：在数据子集上建立一个模型。利用该模型对整个数据集进行预测。通过比较预测值和实际值来计算误差。

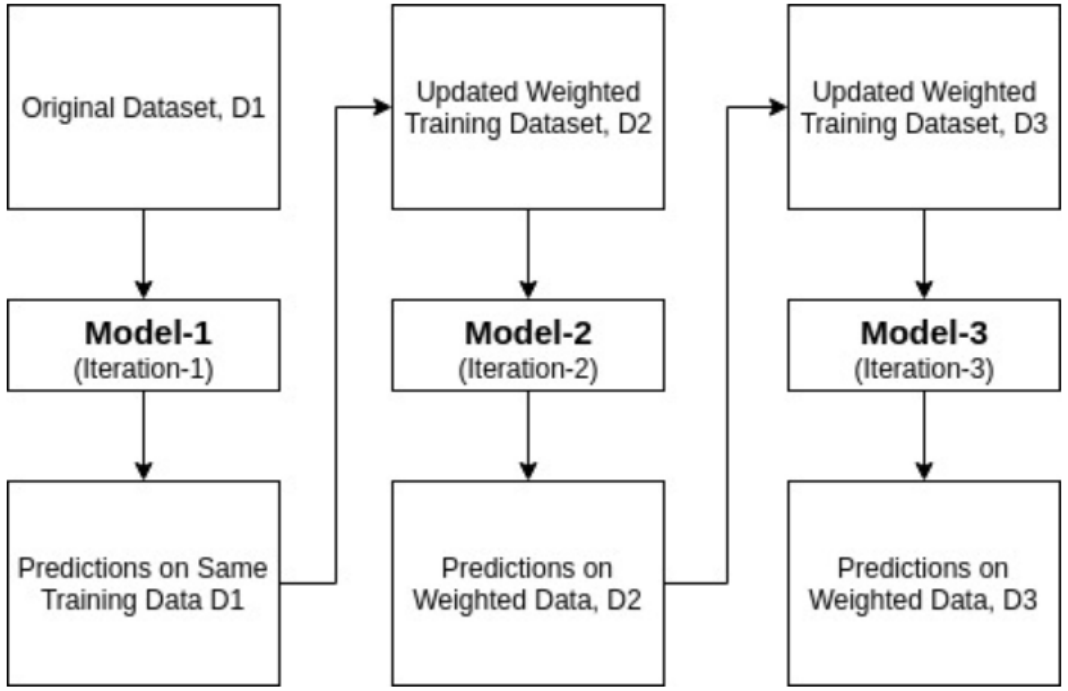
第3步：根据误差调整数据集的权重，误差越大，分配给观测点的权重就越大。

第4步：重复2-3步，直到达到某个预定的足够小的错误率或预先指定的最大迭代次数。

第5步：将所有的模型结果进行加权组合。

优点：可以迭代纠正弱分类器的错误，并通过组合弱学习器来提高准确率；AdaBoost 不容易过度拟合。

缺点：AdaBoost 对噪声数据很敏感；受异常值的影响很大。



图片来源：AdaBoost Classifier Tutorial (kaggle.com)

WISE
厦|门|大|学|王|亚|南|经|济|研|究|院
The Wang Yanan Institute for Studies in Economics, Xiamen University

03.

数 据 介 绍

数据介绍

WISE
厦门大学王亚南经济研究院
The Wang Yanan Institute for Studies in Economics, Xiamen University

VARIABLES	DEFINITIONS	Charactors
Gender	Gender	Categorical
Age	Age	Continuous
Height	Height	Continuous
Weight	Weight	Continuous
Family_history	Has a family member suffered or suffers from overweight?	Binary
FAVC	Do you eat high caloric food frequently?	Binary
FCVC	Do you usually eat vegetables in your meals?	Integer
NCP	How many main meals do you have daily?	Continuous
CAEC	Do you eat any food between meals?	Categorical
SMOKE	Do you smoke?	Binary
CH2O	How much water do you drink daily?	Continuous
SCC	Do you monitor the calories you eat daily?	Binary
FAF	How often do you have physical activity?	Continuous
TUE	How much time do you use technological devices such as cell phone, videogames, television, computer and others?	Integer
CALC	How often do you drink alcohol?	Categorical
MTRANS	Which transportation do you usually use?	Categorical
NObeyesdad	Obesity level	Categorical

Obesity level classification

Insufficient Weight
Normal Weight
Overweight Level I
Overweight Level II
Obesity Type I
Obesity Type II
Obesity Type III

数据来源：Obesity Levels (kaggle.com)

数据介绍

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Age	2111	24.31	6.35	14.00	19.95	26.00	61.00
Height	2111	1.70	0.09	1.45	1.63	1.77	1.98
Weight	2111	86.59	26.19	39.00	65.47	107.43	173.00
FCVC	2111	2.42	0.53	1.00	2.00	3.00	3.00
NCP	2111	2.69	0.78	1.00	2.66	3.00	4.00
CH2O	2111	2.01	0.61	1.00	1.58	2.48	3.00
FAF	2111	1.01	0.85	0.00	0.12	1.67	3.00
TUE	2111	0.66	0.61	0.00	0.00	1.00	2.00

```
'data.frame':   2111 obs. of  17 variables:
 $ Age                           : num  21 21 23 27 22 29 23 22 24 22 ...
 $ Gender                        : chr  "Female" "Female" "Male" "Male" ...
 $ Height                        : num  1.62 1.52 1.8 1.8 1.78 1.62 1.5 1.64 1.78 1.72 ...
 $ Weight                        : num  64 56 77 87 89.8 53 55 53 64 68 ...
 $ CALC                          : chr  "no" "Sometimes" "Frequently" "Frequently" ...
 $ FAVC                          : chr  "no" "no" "no" "no" ...
 $ FCVC                          : num  2 3 2 3 2 2 3 2 3 2 ...
 $ NCP                           : num  3 3 3 3 1 3 3 3 3 3 ...
 $ SCC                           : chr  "no" "yes" "no" "no" ...
 $ SMOKE                         : chr  "no" "yes" "no" "no" ...
 $ CH2O                          : num  2 3 2 2 2 2 2 2 2 2 ...
 $ family_history_with_overweight: chr  "yes" "yes" "yes" "no" ...
 $ FAF                           : num  0 3 2 2 0 0 1 3 1 1 ...
 $ TUE                           : num  1 0 1 0 0 0 0 0 1 1 ...
 $ CAEC                          : chr  "Sometimes" "Sometimes" "Sometimes" "Sometimes" ...
 $ MTRANS                        : chr  "Public_Transportation" "Public_Transportation" "Public_Transportation" "Walking" ...
 $ NObeyesdad                    : chr  "Normal_Weight" "Normal_Weight" "Normal_Weight" "Overweight_Level_I" ...
```

数据预处理

WISE
厦|门|大|学|王|亚|南|经|济|研|究|院
The Wang Yanan Institute for Studies in Economics, Xiamen University

缺失值&重复值处理

```
# test if the dataset has missing values
sum(is.na(data))
```

```
## [1] 0
```

```
# test if the dataset has duplicates
sum(duplicated(data))
```

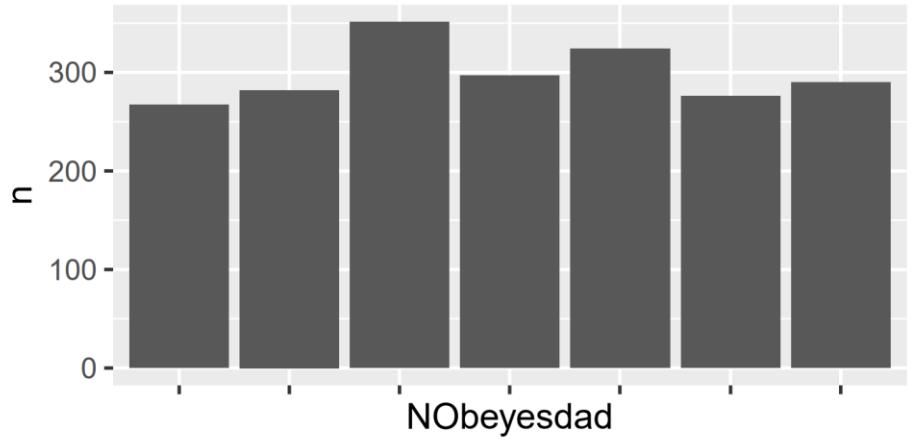
```
## [1] 24
```

```
data <- data[!duplicated(data),] # drop duplicates
sum(duplicated(data)) # test again if there are duplicates
```

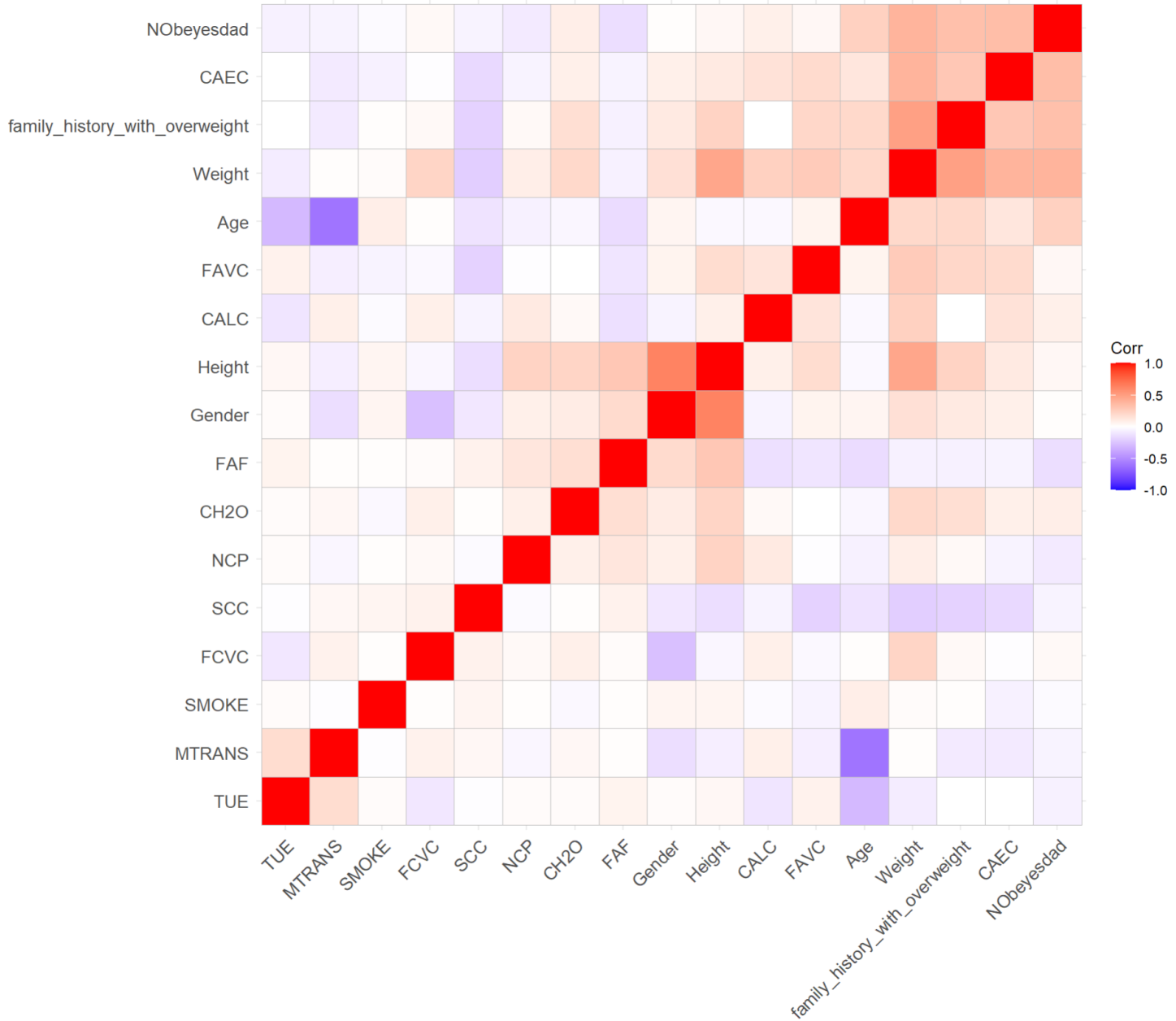
```
## [1] 0
```

平衡性检验

```
data_summary <- data %>% count(NObeyesdad)
require(ggplot2)
ggplot(data_summary,aes(x=NObeyesdad, y=n))+geom_col()+
  theme(axis.text.x = element_blank())
```



相关性



Step1: 将字符型数据转换为数值

Step2: ggcorrplot

WISE
厦|门|大|学|王|亚|南|经|济|研|究|院
The Wang Yanan Institute for Studies in Economics, Xiamen University

04.

模 型 拟 合

Random Forest


```r
require(randomForest)
set.seed(123)
# split the data (70% training, 30% testing)
train_index <- sample(1:nrow(data), 0.7 * nrow(data))
train_data <- data[train_index, ]
test_data <- data[-train_index, ]
rf_model <- randomForest(NObeyesdad ~ ., data = train_data,
      ntree = 100, importance = TRUE)
```

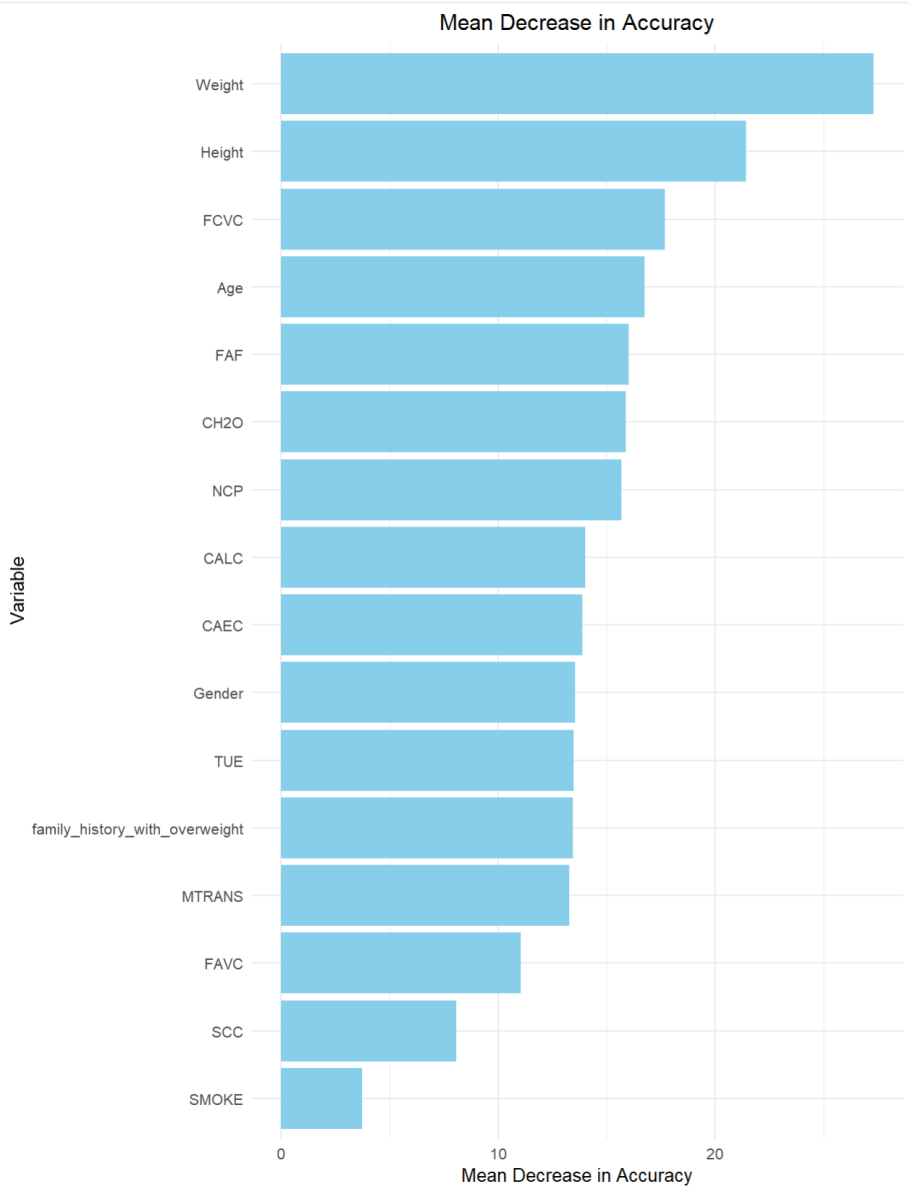
```
Call:
 randomForest(formula = NObeyesdad ~ ., data = train_data, ntree = 100,      importance = TRUE)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 4

        OOB estimate of  error rate: 5.48%
Confusion matrix:
                     Insufficient_Weight Normal_Weight Obesity_Type_I
Insufficient_Weight                 183            13              0
Normal_Weight                         5           182              0
Obesity_Type_I                        0             2            238
Obesity_Type_II                       0             1              1
Obesity_Type_III                      0             0              0
Overweight_Level_I                    0            19              1
Overweight_Level_II                   0             5              2
                     Obesity_Type_II Obesity_Type_III Overweight_Level_I
Insufficient_Weight               0                0                  0
Normal_Weight                     0                0                  9
Obesity_Type_I                    2                0                  0
Obesity_Type_II                 194                0                  0
Obesity_Type_III                  0              224                  0
Overweight_Level_I                0                0                168
Overweight_Level_II               1                0                  4
                     Overweight_Level_II class.error
Insufficient_Weight                    0  0.06632653
Normal_Weight                          1  0.07614213
Obesity_Type_I                         6  0.04032258
Obesity_Type_II                        0  0.01020408
Obesity_Type_III                       0  0.00000000
Overweight_Level_I                     8  0.14285714
Overweight_Level_II                  191  0.05911330
```

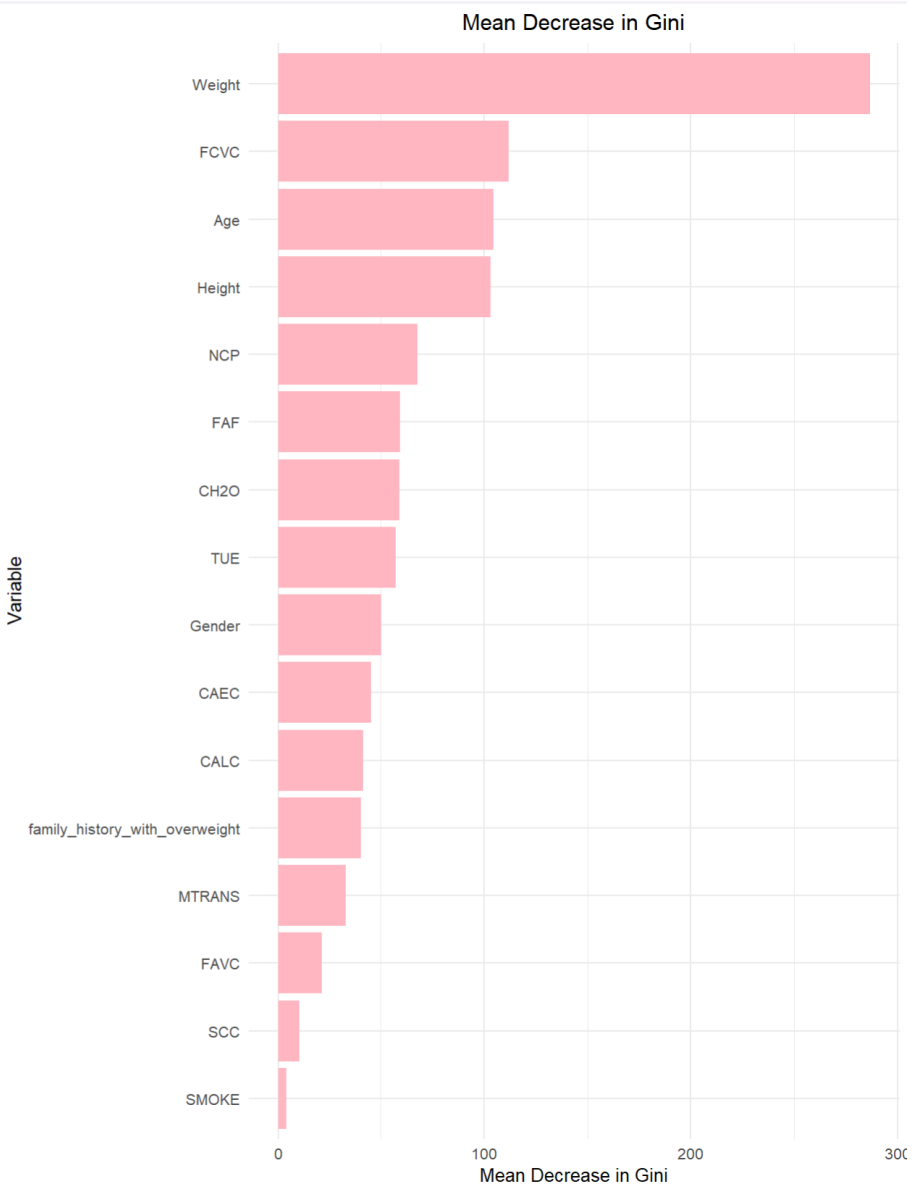
predictions	Insufficient_Weight	Normal_Weight	Obesity_Type_I	Obesity_Type_II
Insufficient_Weight	60	2	0	0
Normal_Weight	6	81	3	0
Obesity_Type_I	0	0	107	0
Obesity_Type_II	0	0	1	95
Obesity_Type_III	0	0	0	0
Overweight_Level_I	0	4	0	0
Overweight_Level_II	0	2	1	0

predictions	Obesity_Type_III	Overweight_Level_I	Overweight_Level_II
Insufficient_Weight	0	0	0
Normal_Weight	0	6	5
Obesity_Type_I	0	0	1
Obesity_Type_II	0	0	0
Obesity_Type_III	103	0	0
Overweight_Level_I	0	72	2
Overweight_Level_II	0	1	75

Random Forest: Number of Trees

Random Forest: Variable of Importance





Adaboost


```
require(adabag)
require(rpart)
```

```
set.seed(123)
# split the data (70% training, 30% testing)
train_index <- sample(1:nrow(data), 0.7 * nrow(data))
train_data <- data[train_index, ]
test_data <- data[-train_index, ]
# Train AdaBoost model
adaboost_model <- boosting(NObeyesdad ~ ., data = train_data,
      control = rpart.control(maxdepth = 6))
```

```
## [1] "Training Misclassification Error: 0"

## [1] "Testing Misclassification Error: 0.03"
```

逐步提升maxdepth，发现maxdepth>=5后，testing misclassification error没有较大改变

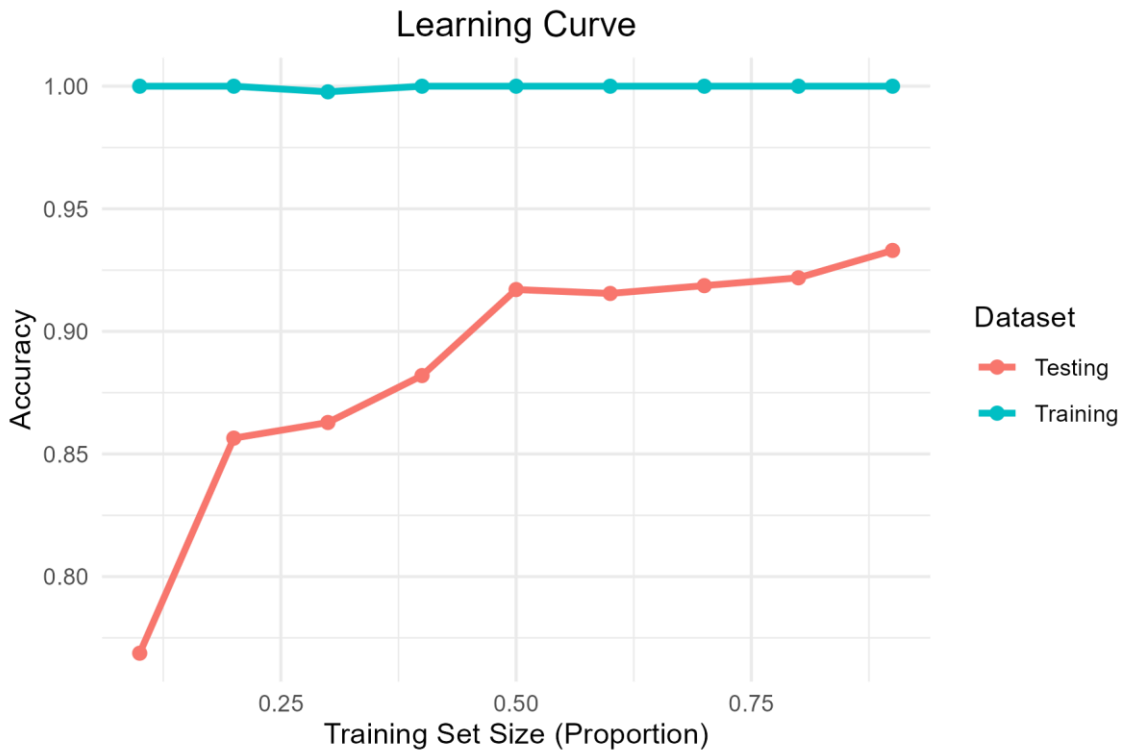
Adaboost


Training and Testing Errors vs Max Depth
Misclassification Error
0.6
0.4
0.2
0.0
2.5
5.0
7.5
10.0
Max Depth
colour
Testing Error
Training Error

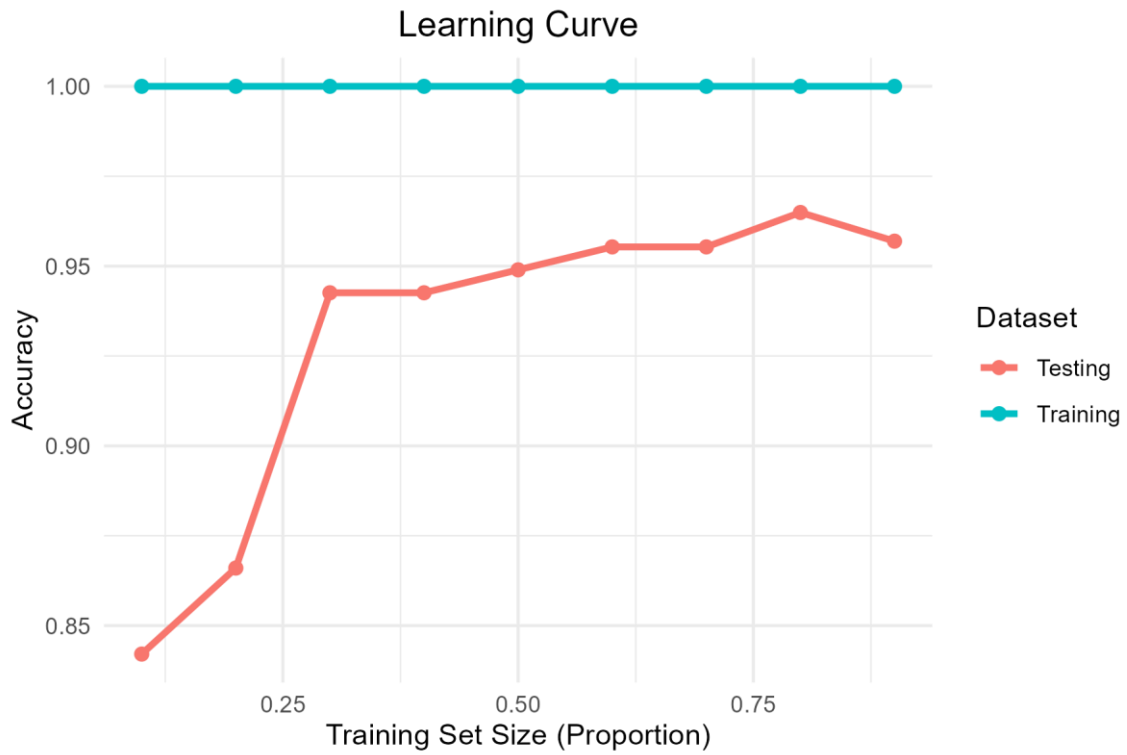
Overfitting Diagnosis: Learning Curve

WISE
厦|门|大|学|王|亚|南|经|济|研|究|院
The Wang Yanan Institute for Studies in Economics, Xiamen University

Random Forest



Adaboost



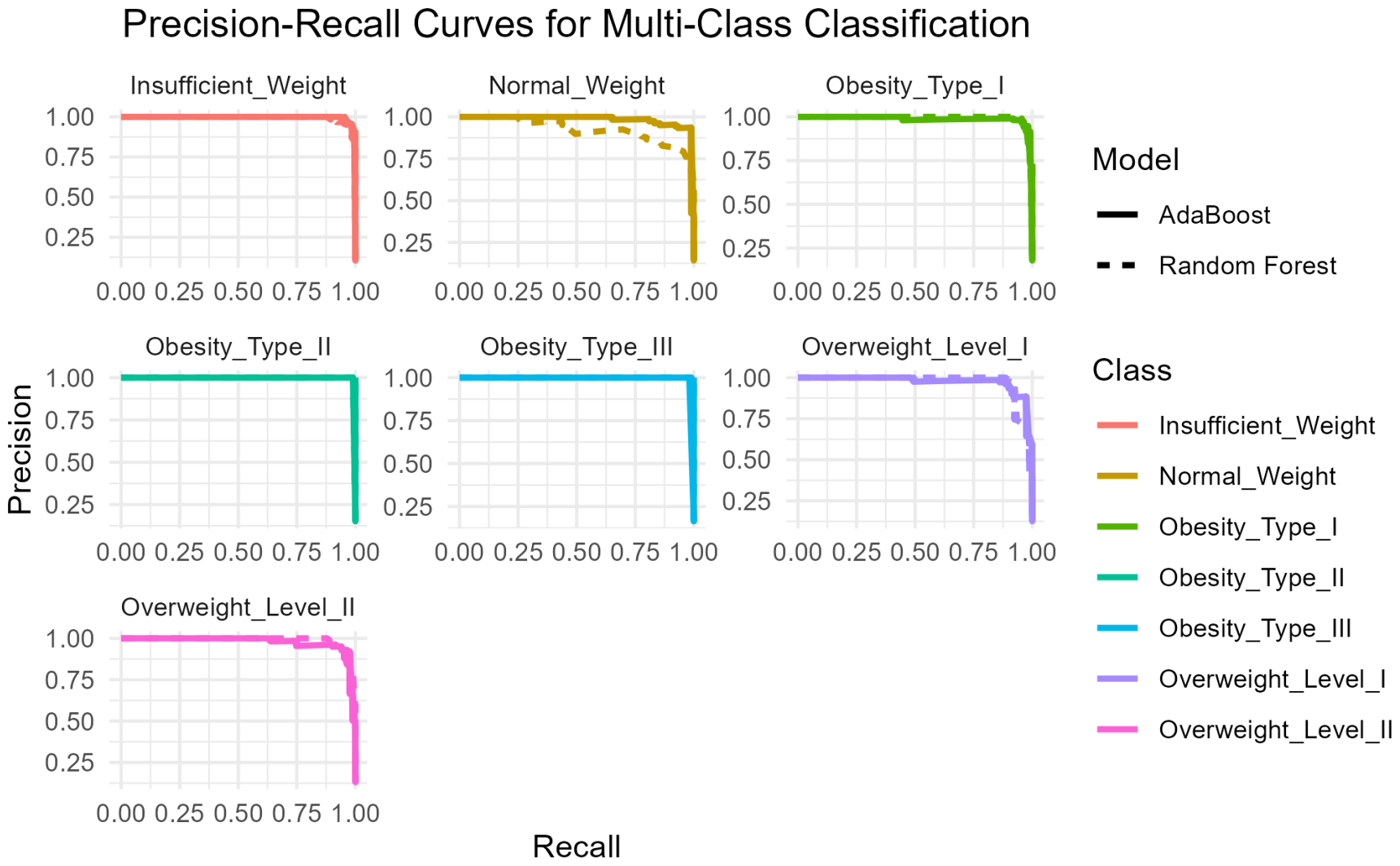
Note: different y scales

WISE
厦|门|大|学|王|亚|南|经|济|研|究|院
The Wang Yanan Institute for Studies in Economics, Xiamen University

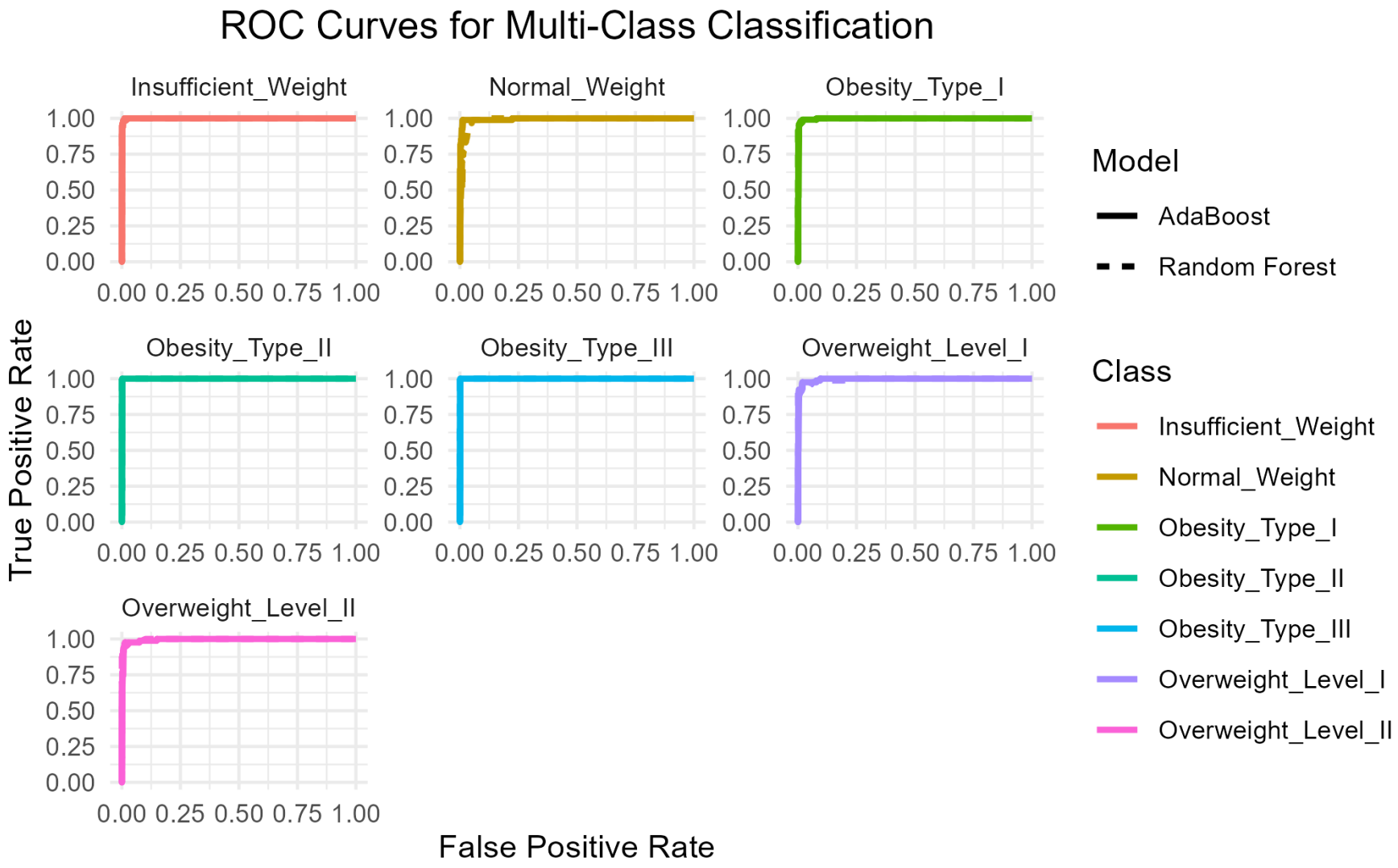
05.

模 型 评 估

PR curve


Precision-Recall Curves for Multi-Class Classification
Insufficient_Weight
Normal_Weight
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Overweight_Level_I
Overweight_Level_II
Precision
Recall
1.00
0.75
0.50
0.25
0.00 0.25 0.50 0.75 1.00
Model
AdaBoost
Random Forest
Class
Insufficient_Weight
Normal_Weight
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Overweight_Level_I
Overweight_Level_II

ROC curve


ROC Curves for Multi-Class Classification
Insufficient_Weight
Normal_Weight
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Overweight_Level_I
Overweight_Level_II
True Positive Rate
False Positive Rate
0.00
0.25
0.50
0.75
1.00
Model
AdaBoost
Random Forest
Class
Insufficient_Weight
Normal_Weight
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Overweight_Level_I
Overweight_Level_II

AUC comparison

WISE
厦|门|大|学|王|亚|南|经|济|研|究|院
The Wang Yanan Institute for Studies in Economics, Xiamen University

Description: df [7 × 5]

Class <chr>	AUC_RF <dbl>	AUC_AdaBoost <dbl>	diff <dbl>	Better_Model <chr>
Insufficient_Weight	0.9992708	0.9997029	-0.0004321288	Adaboost
Normal_Weight	0.9885030	0.9957604	-0.0072574245	Adaboost
Obesity_Type_I	0.9986217	0.9977115	0.0009101942	Random Forest
Obesity_Type_II	1.0000000	1.0000000	0.0000000000	Adaboost
Obesity_Type_III	1.0000000	1.0000000	0.0000000000	Adaboost
Overweight_Level_I	0.9937171	0.9957498	-0.0020327081	Adaboost
Overweight_Level_II	0.9973202	0.9955705	0.0017496456	Random Forest

7 rows

Running Time


```r
start_time <- Sys.time()
rf_model <- randomForest(NObeyesdad ~ ., data = train_data,
                         ntree = 100, importance = TRUE)
end_time <- Sys.time()
end_time - start_time
```

```
Time difference of 0.2858849 secs
```

```r
start_time <- Sys.time()
adaboost_model <- boosting(NObeyesdad ~ ., data = train_data,
                           control = rpart.control(maxdepth = 5))
end_time <- Sys.time()
end_time - start_time
```

```
Time difference of 3.927259 secs
```

WISE
厦|门|大|学|王|亚|南|经|济|研|究|院
The Wang Yanan Institute for Studies in Economics, Xiamen University

06.

研 究 结 果

研究结果

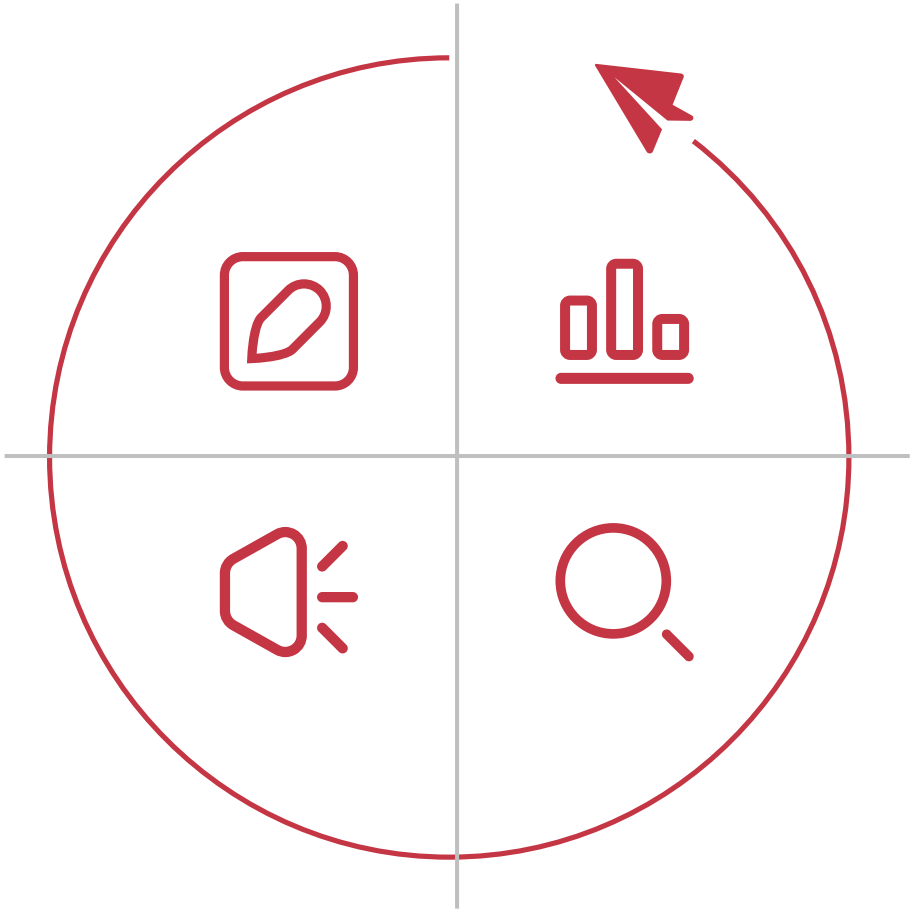
WISE
厦门大学王亚南经济研究院
The Wang Yanan Institute for Studies in Economics, Xiamen University

重要性前四

身高
体重
年龄
吃蔬菜的频率

重要性后五

吸烟频率
监控卡路里次数
是否经常吃高卡路里食物
交通方式
家族肥胖史



效果比较—准确性

对大部分的类别，Adaboost优于Random forest

效果比较—处理时间

Random forest优于Adaboost

WISE
厦|门|大|学|王|亚|南|经|济|研|究|院
The Wang Yanan Institute for Studies in Economics, Xiamen University

THANKS

Thanks for watching