BISST0663_Final_Project

Jueshen Hou Zimeng Ren

2024-10-09

```
##This chunk is only needed when running on Jason's laptop.If it is on Jason's device change eval=TRUE
options("install.lock"=FALSE)
```

```
library(xfun)
```

```
## Warning: package 'xfun' was built under R version 4.4.2
```

```
##
## Attaching package: 'xfun'
```

```
## The following objects are masked from 'package:base':
##
##     attr, isFALSE
```

```
#This is a test2222
```

Load Datas

```
ALZH<-read.csv("https://raw.githubusercontent.com/jasonh0509/StatsLearningFinal/refs/heads/main/alzheim
```

Take a Look

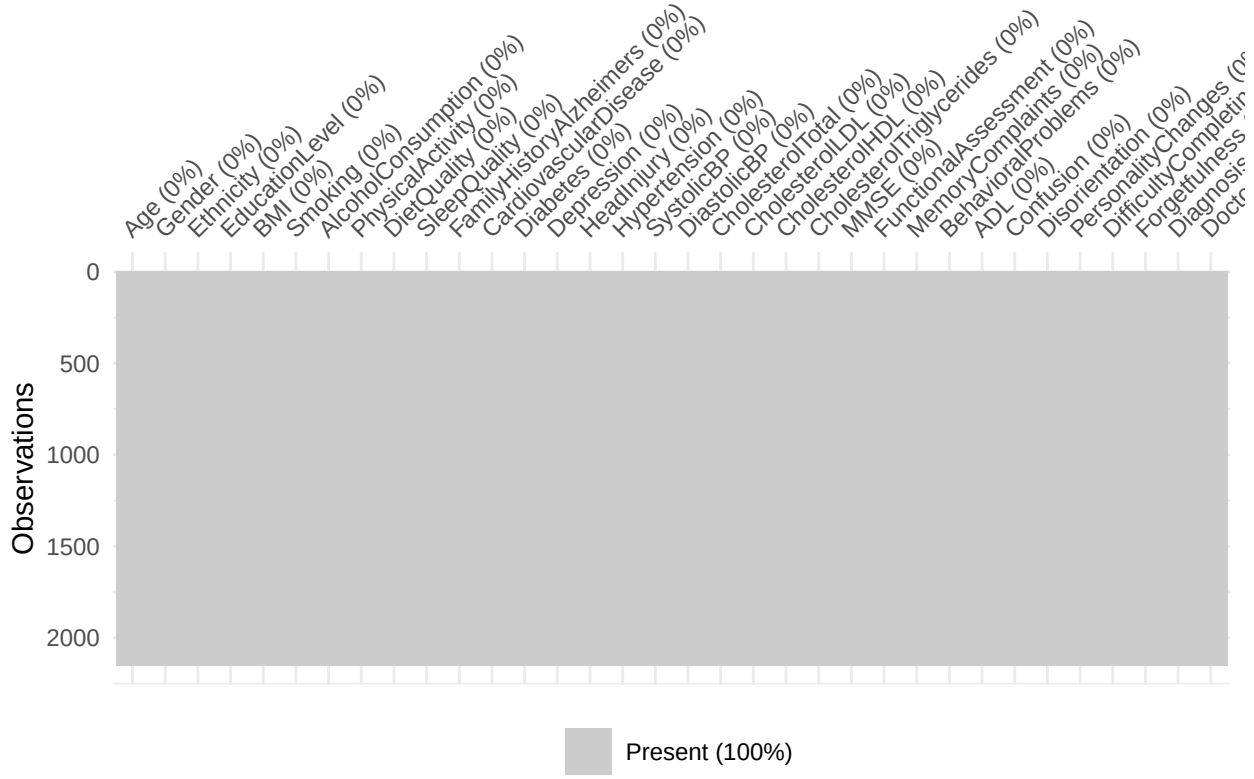
```
glimpse(ALZH)
```

```
## Rows: 2,149
## Columns: 35
## $ PatientID                <int> 4751, 4752, 4753, 4754, 4755, 4756, 4757, 47~
## $ Age                      <int> 73, 89, 73, 74, 89, 86, 68, 75, 72, 87, 89, ~
## $ Gender                   <int> 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1,~
## $ Ethnicity                <int> 0, 0, 3, 0, 0, 1, 3, 0, 1, 0, 3, 0, 0, 0, 0,~
## $ EducationLevel           <int> 2, 0, 1, 1, 0, 1, 2, 1, 0, 0, 1, 2, 1, 1, 2,~
## $ BMI                      <dbl> 22.92775, 26.82768, 17.79588, 33.80082, 20.7~
## $ Smoking                  <int> 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0,~
## $ AlcoholConsumption       <dbl> 13.2972177, 4.5425238, 19.5550845, 12.209265~
## $ PhysicalActivity         <dbl> 6.3271125, 7.6198845, 7.8449878, 8.4280014, ~
## $ DietQuality              <dbl> 1.34721431, 0.51876714, 1.82633466, 7.435604~
```

```
## $ SleepQuality              <dbl> 9.025679, 7.151293, 9.673574, 8.392554, 5.59~
## $ FamilyHistoryAlzheimers   <int> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0,~
## $ CardiovascularDisease     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,~
## $ Diabetes                  <int> 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1,~
## $ Depression                <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1,~
## $ HeadInjury                <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Hypertension              <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0,~
## $ SystolicBP                <int> 142, 115, 99, 118, 94, 168, 143, 117, 117, 1~
## $ DiastolicBP               <int> 72, 64, 116, 115, 117, 62, 88, 63, 119, 78, ~
## $ CholesterolTotal          <dbl> 242.3668, 231.1626, 284.1819, 159.5822, 237.~
## $ CholesterolLDL            <dbl> 56.15090, 193.40800, 153.32276, 65.36664, 92~
## $ CholesterolHDL            <dbl> 33.68256, 79.02848, 69.77229, 68.45749, 56.8~
## $ CholesterolTriglycerides  <dbl> 162.18914, 294.63091, 83.63832, 277.57736, 2~
## $ MMSE                      <dbl> 21.4635324, 20.6132673, 7.3562486, 13.991127~
## $ FunctionalAssessment      <dbl> 6.5188770, 7.1186955, 5.8950773, 8.9651063, ~
## $ MemoryComplaints          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,~
## $ BehavioralProblems        <int> 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0,~
## $ ADL                       <dbl> 1.72588346, 2.59242413, 7.11954774, 6.481225~
## $ Confusion                 <int> 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1,~
## $ Disorientation            <int> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1,~
## $ PersonalityChanges        <int> 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1,~
## $ DifficultyCompletingTasks <int> 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,~
## $ Forgetfulness             <int> 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0,~
## $ Diagnosis                 <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0,~
## $ DoctorInCharge            <chr> "XXXConfid", "XXXConfid", "XXXConfid", "XXXC~
```

```
ALZH_noID<-ALZH[,-1]
```

```
na_plot_ALZH<-vis_miss(ALZH_noID);na_plot_ALZH
```




```
colSums(is.na(ALZH_noID))
```

```
##                      Age                   Gender                Ethnicity
##                        0                        0                        0
##           EducationLevel                      BMI                  Smoking
##                        0                        0                        0
##        AlcoholConsumption         PhysicalActivity              DietQuality
##                        0                        0                        0
##              SleepQuality   FamilyHistoryAlzheimers     CardiovascularDisease
##                        0                        0                        0
##                 Diabetes               Depression               HeadInjury
##                        0                        0                        0
##             Hypertension              SystolicBP             DiastolicBP
##                        0                        0                        0
##          CholesterolTotal          CholesterolLDL          CholesterolHDL
##                        0                        0                        0
##  CholesterolTriglycerides                     MMSE      FunctionalAssessment
##                        0                        0                        0
##          MemoryComplaints        BehavioralProblems                      ADL
##                        0                        0                        0
##                 Confusion          Disorientation       PersonalityChanges
##                        0                        0                        0
## DifficultyCompletingTasks             Forgetfulness                Diagnosis
##                        0                        0                        0
##            DoctorInCharge
##                        0
```

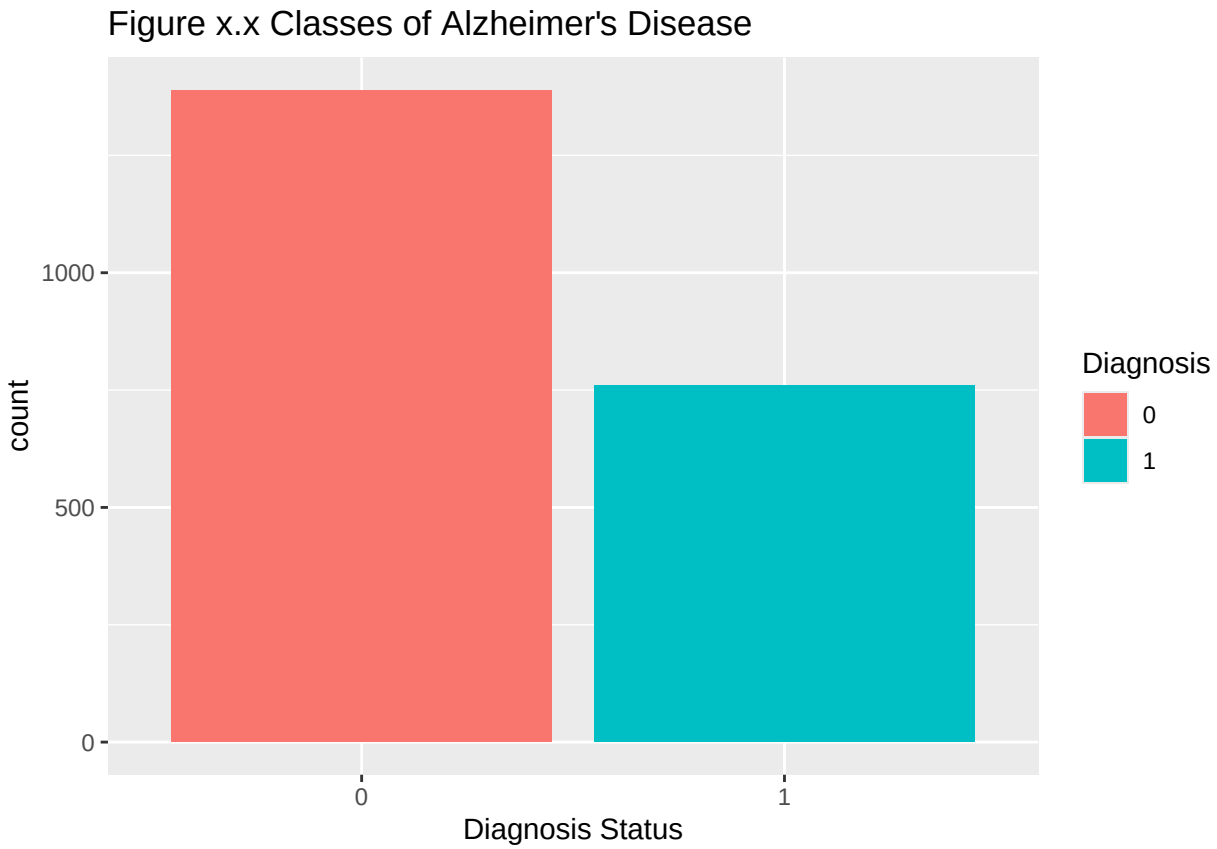
```
ALZH_noID$Diagnosis<-as.factor(ALZH_noID$Diagnosis)
ALZH_noID <- ALZH_noID %>%
  mutate(across(c(Gender, Ethnicity,EducationLevel,Smoking,FamilyHistoryAlzheimers,CardiovascularDiseas
```

Set up Data Set(Keep Same Across All Stats Leraning Models)

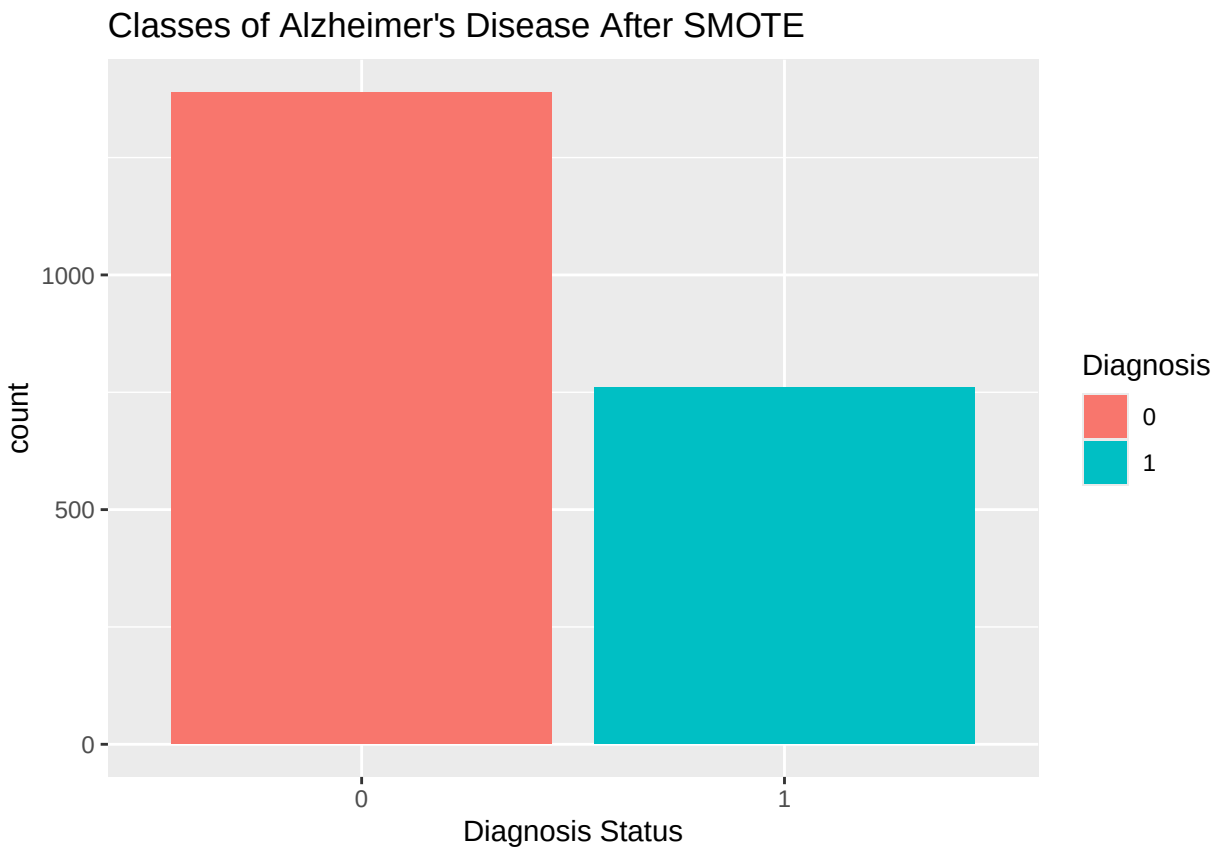
```
ALZH.raw <- ALZH_noID %>% select(-DoctorInCharge) %>% mutate(Diagnosis = as.numeric(as.character(Diagnos
ALZH.gbm <- ALZH.raw
##Dispite the name, this ALZH.gbm is the data set will be used for all models, the set in other models
```

```
ALZH_for_explore<-ALZH.gbm
```

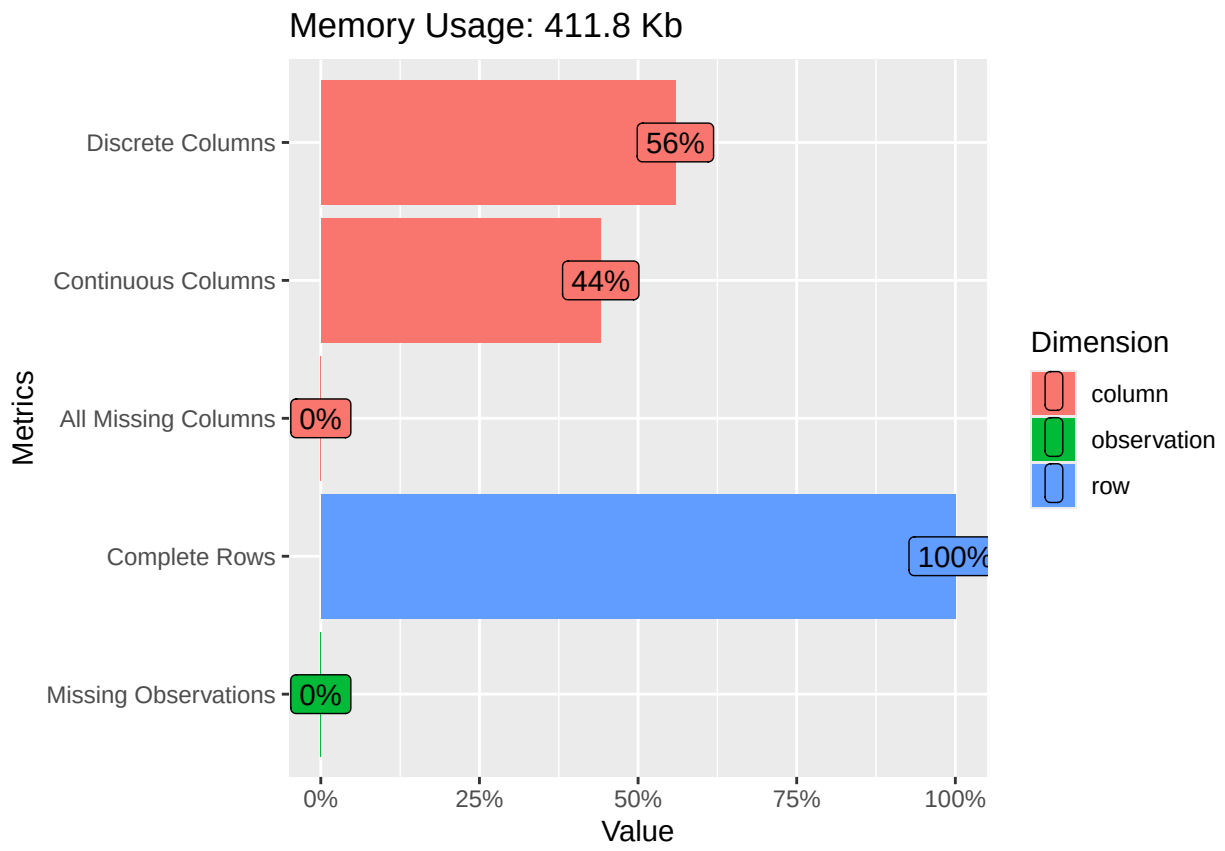
```
alzh_classes<-ggplot(data = ALZH_noID, mapping = aes(x=Diagnosis,fill=Diagnosis))+
  geom_bar()+
  xlab("Diagnosis Status")+
  ggtitle("Figure x.x Classes of Alzheimer's Disease")

alzh_classes
```




```
ggplot(data = ALZH_noID, mapping = aes(x=Diagnosis,fill=Diagnosis))+
  geom_bar()+
  xlab("Diagnosis Status")+
  ggtitle("Classes of Alzheimer's Disease After SMOTE")
```



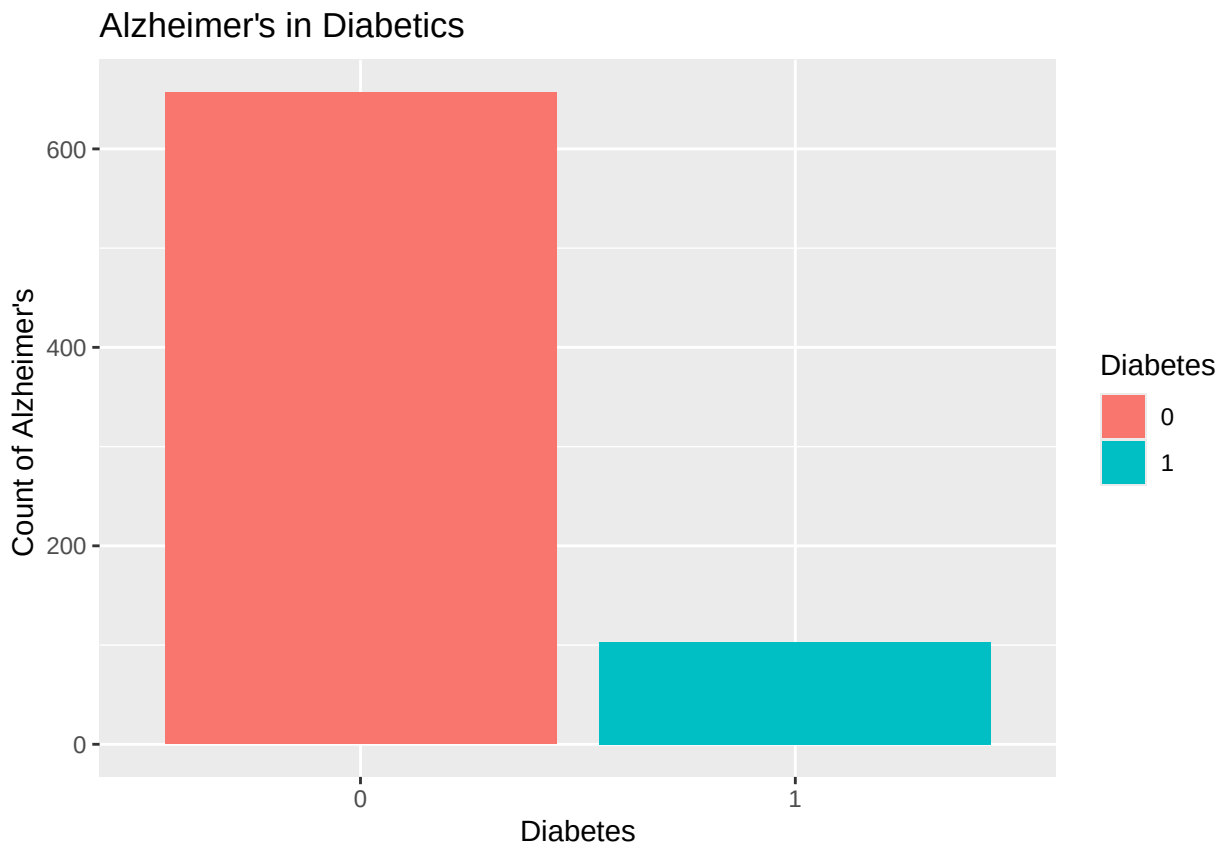

```
plot_intro(ALZH_noID)
```



More EDA focused on positive cases

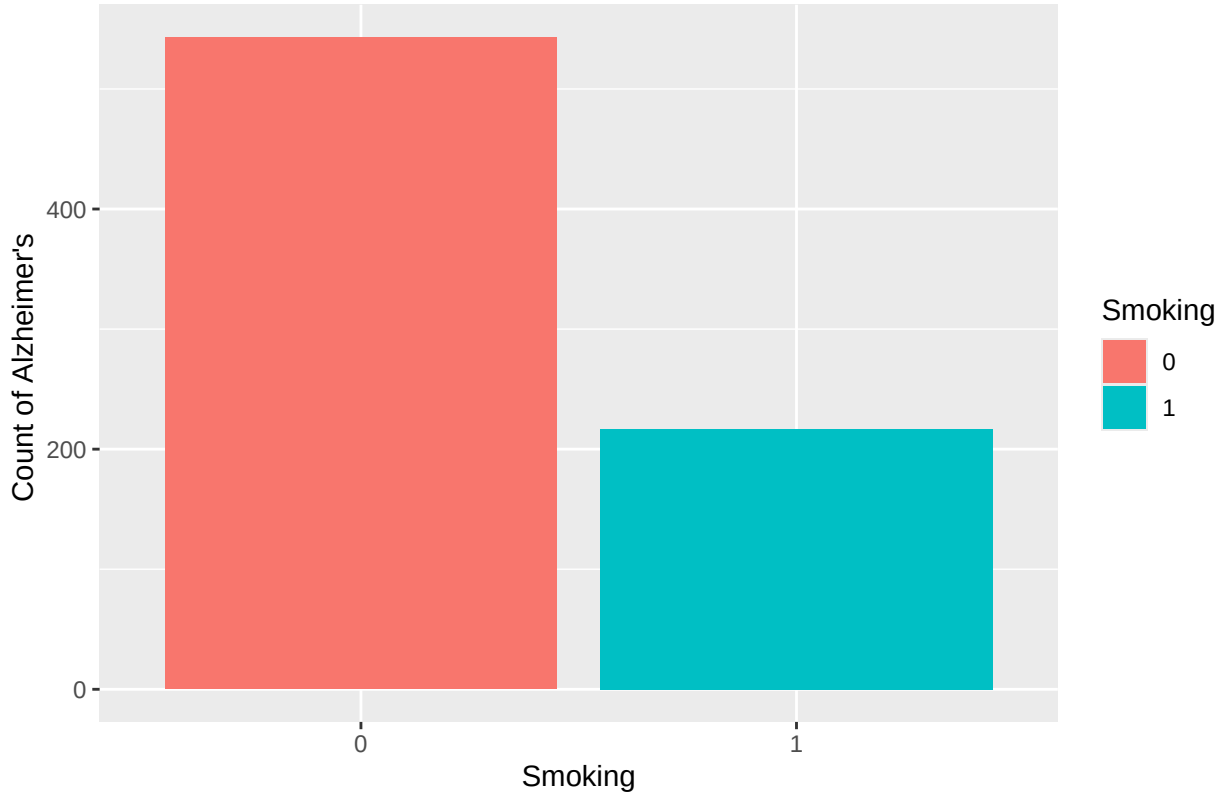
```
alzh_pos<-subset(ALZH_noID,Diagnosis==1)

alzh_gender<-alzh_pos%>%
  group_by(Gender)%>%
  summarise(n = n()) %>%
  ggplot(aes(x = Gender, y = n,fill=Gender))+
  geom_col()+
  labs(y="Count of Alzheimer's ")
alzh_gender+ggtitle("Alzheimer's  Across Gender")
```


```
alzh_diab<-alzh_pos%>%
  group_by(Diabetes)%>%
  summarise(n = n()) %>%
  ggplot(aes(x = Diabetes, y = n,fill=Diabetes))+
  geom_col()+
  labs(y="Count of Alzheimer's ")
alzh_diab+ggtitle("Alzheimer's in Diabetics")
```

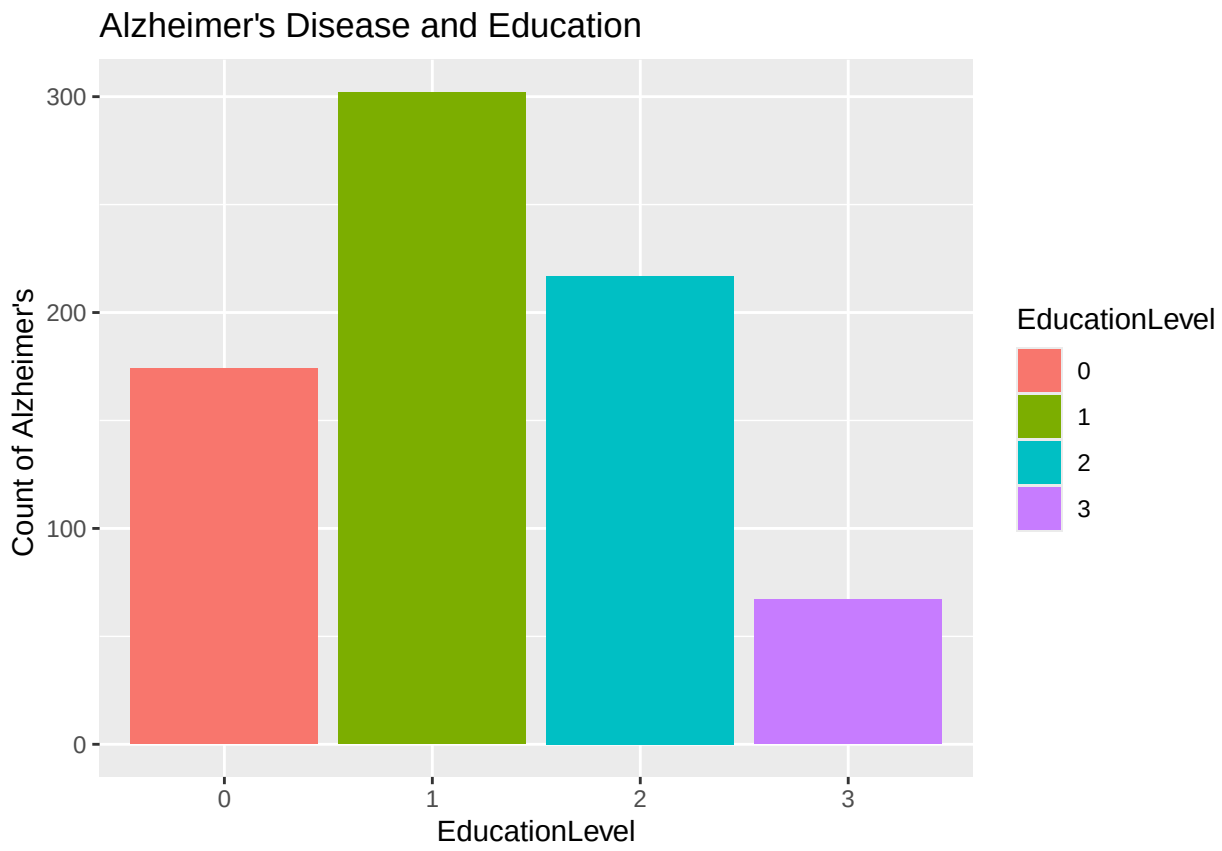



```
alzh_smoke<-alzh_pos%>%
  group_by(Smoking)%>%
  summarise(n = n()) %>%
  ggplot(aes(x = Smoking, y = n,fill=Smoking))+
  geom_col()+
  labs(y="Count of Alzheimer's ")
alzh_smoke+ggtitle("Figure x.x Alzheimer's in Cigarette Users")
```

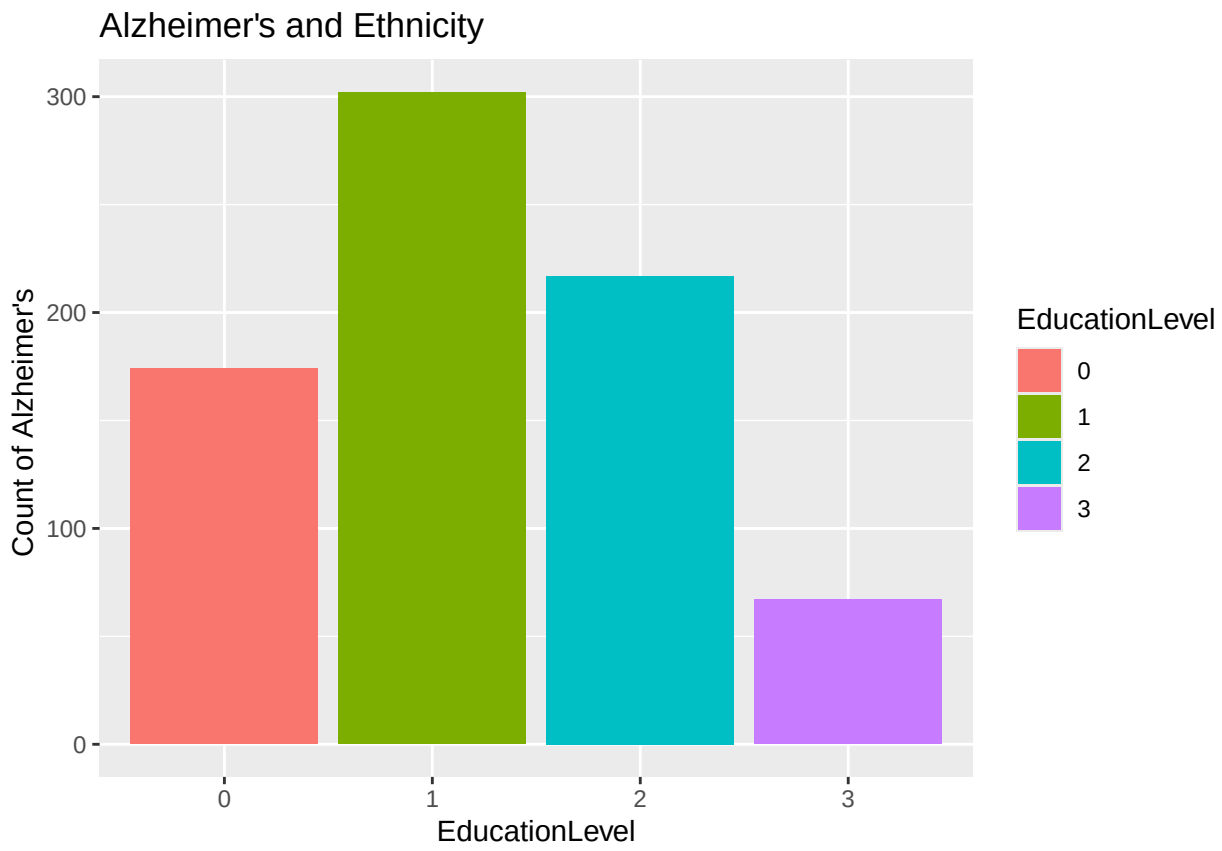
Figure x.x Alzheimer's in Cigarette Users




```
alzh_edu<-alzh_pos%>%
  group_by(EducationLevel)%>%
  summarise(n = n()) %>%
  ggplot(aes(x = EducationLevel, y = n,fill=EducationLevel))+
  geom_col()+
  labs(y="Count of Alzheimer's ")
alzh_edu+ggtitle("Alzheimer's Disease and Education")
```




```
alzh_ethnicity<-alzh_pos%>%
  group_by(Ethnicity)%>%
  summarise(n = n()) %>%
  ggplot(aes(x = Ethnicity, y = n,fill=Ethnicity))+
  geom_col()+
  labs(y="Count of Alzheimer's ")
alzh_edu+ggtitle("Alzheimer's and Ethnicity")
```



Exploration of Corerlation of 4 Variables related to

```
correlation.mtx.4var<-cor(ALZH_for_explore[,c("CholesterolHDL", "CholesterolLDL", "CholesterolTotal", "C
correlation.mtx.4var
```

```
##                          CholesterolHDL CholesterolLDL CholesterolTotal
## CholesterolHDL              1.00000000   -0.037148129      0.010116206
## CholesterolLDL             -0.03714813    1.000000000      0.010335686
## CholesterolTotal            0.01011621    0.010335686      1.000000000
## CholesterolTriglycerides    0.01523465   -0.005582058     -0.001959256
##                          CholesterolTriglycerides
## CholesterolHDL                        0.015234649
## CholesterolLDL                       -0.005582058
## CholesterolTotal                     -0.001959256
## CholesterolTriglycerides              1.000000000
```

```
correlation.mtx.4var<-as.data.frame(correlation.mtx.4var)
knitr::kable(correlation.mtx.4var)
```

	CholesterolHDL	CholesterolLDL	CholesterolTotal	CholesterolTriglycerides
CholesterolHDL	1.0000000	-0.0371481	0.0101162	0.0152346
CholesterolLDL	-0.0371481	1.0000000	0.0103357	-0.0055821

	CholesterolHDL	CholesterolLDL	CholesterolTotal	CholesterolTriglycerides
CholesterolTotal	0.0101162	0.0103357	1.0000000	-0.0019593
CholesterolTriglycerides	0.0152346	-0.0055821	-0.0019593	1.0000000

```
##4 logistic regression models with only response and one of the 4 cholesterol variables
hdl.logistic<-glm(Diagnosis~CholesterolHDL,data=ALZH_for_explore,family = binomial);summary(hdl.logisti
```

```
##
## Call:
## glm(formula = Diagnosis ~ CholesterolHDL, family = binomial,
##     data = ALZH_for_explore)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.833278   0.125691  -6.630 3.37e-11 ***
## CholesterolHDL  0.003853   0.001953   1.973   0.0485 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2792.3  on 2148  degrees of freedom
## Residual deviance: 2788.4  on 2147  degrees of freedom
## AIC: 2792.4
##
## Number of Fisher Scoring iterations: 4
```

```
ldl.logistic<-glm(Diagnosis~CholesterolLDL,data=ALZH_for_explore,family = binomial);summary(ldl.logisti
```

```
##
## Call:
## glm(formula = Diagnosis ~ CholesterolLDL, family = binomial,
##     data = ALZH_for_explore)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.411680   0.136372  -3.019  0.00254 **
## CholesterolLDL -0.001544   0.001042  -1.482  0.13839
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2792.3  on 2148  degrees of freedom
## Residual deviance: 2790.1  on 2147  degrees of freedom
## AIC: 2794.1
##
## Number of Fisher Scoring iterations: 4
```

```
total.logistic<-glm(Diagnosis~CholesterolTotal,data=ALZH_for_explore,family = binomial);summary(total.l
```

```
##
## Call:
## glm(formula = Diagnosis ~ CholesterolTotal, family = binomial,
##     data = ALZH_for_explore)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.6738669  0.2433093  -2.770  0.00561 **
## CholesterolTotal  0.0003145  0.0010609   0.296  0.76690
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2792.3  on 2148  degrees of freedom
## Residual deviance: 2792.2  on 2147  degrees of freedom
## AIC: 2796.2
##
## Number of Fisher Scoring iterations: 4
```

```
tryglycerides.logistic<-glm(Diagnosis~CholesterolTriglycerides,data=ALZH_for_explore,family = binomial)
```

```
##
## Call:
## glm(formula = Diagnosis ~ CholesterolTriglycerides, family = binomial,
##     data = ALZH_for_explore)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.7095795  0.1112808  -6.376 1.81e-10 ***
## CholesterolTriglycerides  0.0004653  0.0004428   1.051    0.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2792.3  on 2148  degrees of freedom
## Residual deviance: 2791.2  on 2147  degrees of freedom
## AIC: 2795.2
##
## Number of Fisher Scoring iterations: 4
```

```
##HDL is most correlated with the response variable
```

logistic

```
ALZH.logistic<-ALZH.gbm%>%select(-c(CholesterolTotal,CholesterolLDL,CholesterolTriglycerides))###set fo
n <-nrow(ALZH.logistic);n
```

```
## [1] 2149
```

```
set.seed(114514)
draw<-sample(1:n,size = 1934)##1934 is 90% of the data,here is the rows we use fall all trainig and val
##This is the ultimate sample data indces!
train <-ALZH.logistic[draw,]
train_x<-train%>%dplyr::select(-Diagnosis)
train_y<-train%>%dplyr::select(Diagnosis)

test <- ALZH.logistic[-draw,]
test_x<-test%>%dplyr::select(-Diagnosis)
test_y <-test$Diagnosis


x <-model.matrix(Diagnosis~.,data=ALZH.logistic)
y <- ALZH.logistic$Diagnosis
```

```
ALZH_logistic <-glm(Diagnosis~.,data=train,family = binomial)
summary(ALZH_logistic)
```

```
##
## Call:
## glm(formula = Diagnosis ~ ., family = binomial, data = train)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 4.868e+00  9.281e-01   5.246 1.56e-07 ***
## Age                        -1.075e-02  7.399e-03  -1.452   0.1464
## Gender1                    -2.646e-02  1.339e-01  -0.198   0.8433
## Ethnicity1                 -2.031e-01  1.741e-01  -1.167   0.2434
## Ethnicity2                  2.548e-01  2.318e-01   1.099   0.2716
## Ethnicity3                 -2.037e-01  2.361e-01  -0.863   0.3882
## EducationLevel1            -2.070e-01  1.815e-01  -1.141   0.2540
## EducationLevel2            -1.131e-01  1.920e-01  -0.589   0.5557
## EducationLevel3            -2.875e-01  2.583e-01  -1.113   0.2658
## BMI                        -4.568e-03  9.238e-03  -0.494   0.6210
## Smoking1                   -1.409e-01  1.500e-01  -0.939   0.3476
## AlcoholConsumption         -7.706e-03  1.157e-02  -0.666   0.5052
## PhysicalActivity           -1.073e-02  2.313e-02  -0.464   0.6427
## DietQuality                 2.592e-02  2.333e-02   1.111   0.2666
## SleepQuality               -5.889e-02  3.802e-02  -1.549   0.1214
## FamilyHistoryAlzheimers1   -6.129e-02  1.555e-01  -0.394   0.6936
## CardiovascularDisease1      1.089e-01  1.825e-01   0.596   0.5509
## Diabetes1                   1.081e-02  1.916e-01   0.056   0.9550
## Depression1                 1.066e-01  1.638e-01   0.651   0.5149
## HeadInjury1                -2.628e-01  2.287e-01  -1.149   0.2506
## Hypertension1               1.636e-01  1.846e-01   0.886   0.3755
## SystolicBP                 -6.674e-05  2.575e-03  -0.026   0.9793
## DiastolicBP                 2.271e-03  3.775e-03   0.602   0.5475
## CholesterolHDL              6.086e-03  2.921e-03   2.084   0.0372 *
## MMSE                       -1.079e-01  8.573e-03 -12.582  < 2e-16 ***
## FunctionalAssessment       -4.412e-01  2.780e-02 -15.869  < 2e-16 ***
## MemoryComplaints1           2.574e+00  1.748e-01  14.727  < 2e-16 ***
```

```
## BehavioralProblems1         2.461e+00  1.923e-01  12.800  < 2e-16 ***
## ADL                        -4.013e-01  2.686e-02 -14.939  < 2e-16 ***
## Confusion1                 -2.260e-01  1.685e-01  -1.341   0.1798
## Disorientation1            -1.468e-01  1.833e-01  -0.801   0.4232
## PersonalityChanges1        -1.557e-01  1.925e-01  -0.809   0.4187
## DifficultyCompletingTasks1  1.337e-01  1.818e-01   0.735   0.4621
## Forgetfulness1             -3.051e-02  1.462e-01  -0.209   0.8347
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2502.0  on 1933  degrees of freedom
## Residual deviance: 1437.3  on 1900  degrees of freedom
## AIC: 1505.3
##
## Number of Fisher Scoring iterations: 6
```

```
pred_test<-predict(ALZH_logistic,type='response',newdata = test)
glm.pred <- ifelse(pred_test > 0.5, 1, 0)
table(glm.pred, test_y)
```

```
##         test_y
## glm.pred   0   1
##        0 121  19
##        1   9  66
```

```
Recall.regular.glm<-sum(glm.pred == 1 & test_y == 1)/sum(test_y == 1);Recall.regular.glm
```

```
## [1] 0.7764706
```

```
Precision.regular.glm<-sum(glm.pred == 1 & test_y == 1)/sum(glm.pred == 1);Precision.regular.glm
```

```
## [1] 0.88
```

```
F1Score.regular.glm<-2*Precision.regular.glm*Recall.regular.glm/(Precision.regular.glm+Recall.regular.gl
```

```
## [1] 0.825
```

```
Accuracy.regular.glm<-sum(glm.pred == test_y)/length(test_y);Accuracy.regular.glm
```

```
## [1] 0.8697674
```

MLR with best subset

```
ALZH_leanning<-ALZH_noID%>%dplyr::select(-DoctorInCharge)
bestsubset <- regsubsets(Diagnosis~., data = ALZH_leanning)
bestsubsum<-summary(bestsubset)
bestsubsum
```

```
## Subset selection object
## Call: regsubsets.formula(Diagnosis ~ ., data = ALZH_leanning)
## 36 Variables  (and intercept)
##                              Forced in Forced out
## Age                              FALSE      FALSE
## Gender1                          FALSE      FALSE
## Ethnicity1                       FALSE      FALSE
## Ethnicity2                       FALSE      FALSE
## Ethnicity3                       FALSE      FALSE
## EducationLevel1                  FALSE      FALSE
## EducationLevel2                  FALSE      FALSE
## EducationLevel3                  FALSE      FALSE
## BMI                              FALSE      FALSE
## Smoking1                         FALSE      FALSE
## AlcoholConsumption               FALSE      FALSE
## PhysicalActivity                 FALSE      FALSE
## DietQuality                      FALSE      FALSE
## SleepQuality                     FALSE      FALSE
## FamilyHistoryAlzheimers1         FALSE      FALSE
## CardiovascularDisease1           FALSE      FALSE
## Diabetes1                        FALSE      FALSE
## Depression1                      FALSE      FALSE
## HeadInjury1                      FALSE      FALSE
## Hypertension1                    FALSE      FALSE
## SystolicBP                       FALSE      FALSE
## DiastolicBP                      FALSE      FALSE
## CholesterolTotal                 FALSE      FALSE
## CholesterolLDL                   FALSE      FALSE
## CholesterolHDL                   FALSE      FALSE
## CholesterolTriglycerides         FALSE      FALSE
## MMSE                             FALSE      FALSE
## FunctionalAssessment             FALSE      FALSE
## MemoryComplaints1                FALSE      FALSE
## BehavioralProblems1              FALSE      FALSE
## ADL                              FALSE      FALSE
## Confusion1                       FALSE      FALSE
## Disorientation1                  FALSE      FALSE
## PersonalityChanges1              FALSE      FALSE
## DifficultyCompletingTasks1       FALSE      FALSE
## Forgetfulness1                   FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          Age Gender1 Ethnicity1 Ethnicity2 Ethnicity3 EducationLevel1
## 1  ( 1 ) " " " "     " "         " "         " "         " "
## 2  ( 1 ) " " " "     " "         " "         " "         " "
## 3  ( 1 ) " " " "     " "         " "         " "         " "
## 4  ( 1 ) " " " "     " "         " "         " "         " "
## 5  ( 1 ) " " " "     " "         " "         " "         " "
## 6  ( 1 ) " " " "     " "         " "         " "         " "
## 7  ( 1 ) "*" " "     " "         " "         " "         " "
## 8  ( 1 ) "*" " "     " "         " "         " "         " "
##          EducationLevel2 EducationLevel3 BMI Smoking1 AlcoholConsumption
## 1  ( 1 ) " "             " "             " " " "      " "
## 2  ( 1 ) " "             " "             " " " "      " "
```

```
## 3  ( 1 ) " "             " "             " " " "       " "
## 4  ( 1 ) " "             " "             " " " "       " "
## 5  ( 1 ) " "             " "             " " " "       " "
## 6  ( 1 ) " "             " "             " " " "       " "
## 7  ( 1 ) " "             " "             " " " "       " "
## 8  ( 1 ) " "             " "             " " " "       " "
##          PhysicalActivity DietQuality SleepQuality FamilyHistoryAlzheimers1
## 1  ( 1 ) " "               " "         " "          " "
## 2  ( 1 ) " "               " "         " "          " "
## 3  ( 1 ) " "               " "         " "          " "
## 4  ( 1 ) " "               " "         " "          " "
## 5  ( 1 ) " "               " "         " "          " "
## 6  ( 1 ) " "               " "         " "          " "
## 7  ( 1 ) " "               " "         " "          " "
## 8  ( 1 ) " "               " "         "*"          " "
##          CardiovascularDisease1 Diabetes1 Depression1 HeadInjury1 Hypertension1
## 1  ( 1 ) " "                    " "       " "         " "          " "
## 2  ( 1 ) " "                    " "       " "         " "          " "
## 3  ( 1 ) " "                    " "       " "         " "          " "
## 4  ( 1 ) " "                    " "       " "         " "          " "
## 5  ( 1 ) " "                    " "       " "         " "          " "
## 6  ( 1 ) " "                    " "       " "         " "          " "
## 7  ( 1 ) " "                    " "       " "         " "          " "
## 8  ( 1 ) " "                    " "       " "         " "          " "
##          SystolicBP DiastolicBP CholesterolTotal CholesterolLDL CholesterolHDL
## 1  ( 1 ) " "        " "         " "              " "            " "
## 2  ( 1 ) " "        " "         " "              " "            " "
## 3  ( 1 ) " "        " "         " "              " "            " "
## 4  ( 1 ) " "        " "         " "              " "            " "
## 5  ( 1 ) " "        " "         " "              " "            " "
## 6  ( 1 ) " "        " "         " "              "*"            " "
## 7  ( 1 ) " "        " "         " "              "*"            " "
## 8  ( 1 ) " "        " "         " "              "*"            " "
##          CholesterolTriglycerides MMSE FunctionalAssessment MemoryComplaints1
## 1  ( 1 ) " "                      " "  "*"                   " "
## 2  ( 1 ) " "                      " "  "*"                   " "
## 3  ( 1 ) " "                      " "  "*"                   "*"
## 4  ( 1 ) " "                      " "  "*"                   "*"
## 5  ( 1 ) " "                      "*"  "*"                   "*"
## 6  ( 1 ) " "                      "*"  "*"                   "*"
## 7  ( 1 ) " "                      "*"  "*"                   "*"
## 8  ( 1 ) " "                      "*"  "*"                   "*"
##          BehavioralProblems1 ADL Confusion1 Disorientation1 PersonalityChanges1
## 1  ( 1 ) " "                 " " " "        " "             " "
## 2  ( 1 ) " "                 "*" " "        " "             " "
## 3  ( 1 ) " "                 "*" " "        " "             " "
## 4  ( 1 ) "*"                 "*" " "        " "             " "
## 5  ( 1 ) "*"                 "*" " "        " "             " "
## 6  ( 1 ) "*"                 "*" " "        " "             " "
## 7  ( 1 ) "*"                 "*" " "        " "             " "
## 8  ( 1 ) "*"                 "*" " "        " "             " "
##          DifficultyCompletingTasks1 Forgetfulness1
## 1  ( 1 ) " "                        " "
## 2  ( 1 ) " "                        " "
```

```
## 3  ( 1 ) " "                          " "
## 4  ( 1 ) " "                          " "
## 5  ( 1 ) " "                          " "
## 6  ( 1 ) " "                          " "
## 7  ( 1 ) " "                          " "
## 8  ( 1 ) " "                          " "
```

```
which.min(bestsubsum$cp)
```

```
## [1] 8
```

```
which.min(bestsubsum$bic)
```

```
## [1] 5
```

```
which.min(bestsubsum$adjr2)
```

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

```
knitr::kable(coef(bestsubset,8))
```

	x
(Intercept)	2.1734842
Age	-0.0015068
SleepQuality	-0.0074936
CholesterolLDL	-0.0003843
MMSE	-0.0130444
FunctionalAssessment	-0.0557346
MemoryComplaints1	0.3528825
BehavioralProblems1	0.3154456
ADL	-0.0508926

```
bestSubset_vars <- names(coef(bestsubset, 8))[-1]
bestSubset_vars
```

```
## [1] "Age"                   "SleepQuality"          "CholesterolLDL"
## [4] "MMSE"                  "FunctionalAssessment" "MemoryComplaints1"
## [7] "BehavioralProblems1"   "ADL"
```

```
bestSubset_STR<-paste(bestSubset_vars,collapse = ",")
```

A MLR with best subset

```
set.with.BestsubsetVar<-train%>%dplyr::select(Diagnosis,Age,SleepQuality,CholesterolHDL,MMSE,FunctionalA
MLR_bestsubset<-glm(Diagnosis~.,data=set.with.BestsubsetVar,family="binomial")
pred.bestsubset<-predict(MLR_bestsubset,newdata = test,type = "response",family="binomial")
class.bestsubset<-ifelse(pred.bestsubset>0.5,1,0)
table(class.bestsubset,test_y)
```

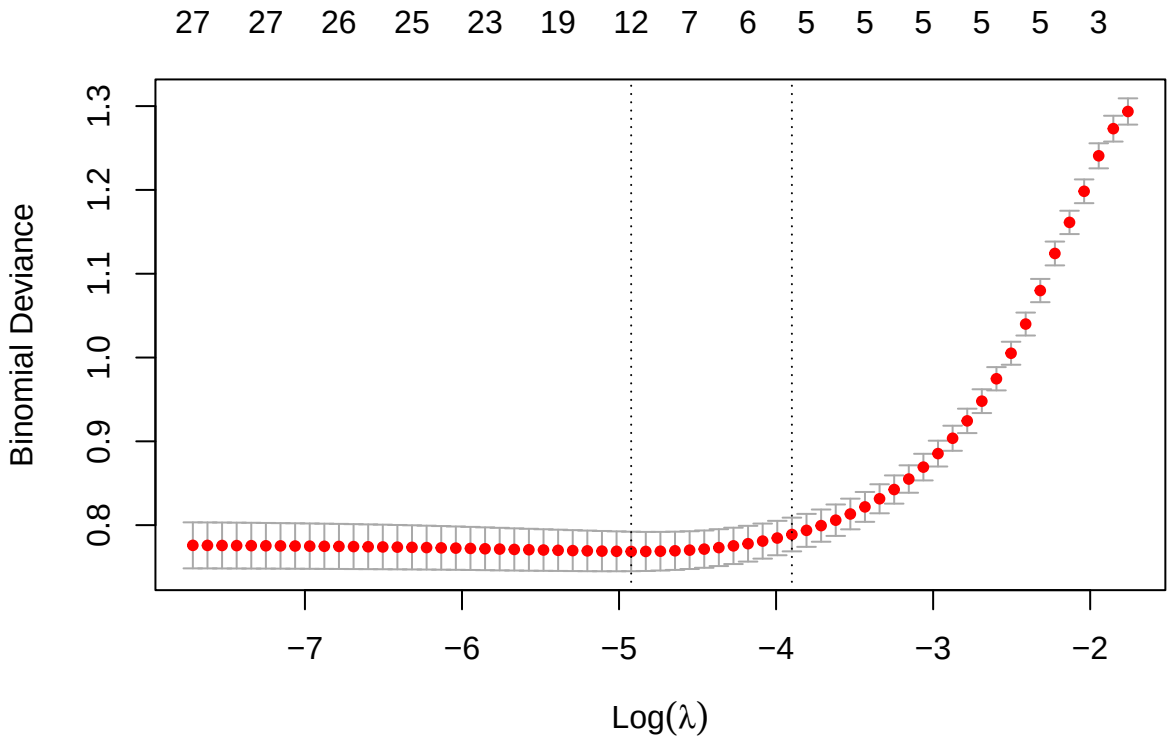
```
##                  test_y
## class.bestsubset   0   1
##                0 119  20
##                1  11  65
```

lasso

```
##subject to fix
library(glmnet)
grid <- 10^seq(10,-2, length = 100)
train_y.lasso<-train_y%>%mutate(Diagnosis=as.factor(Diagnosis))
train_x_lasso<-as.matrix(train_x)
lasso.mod<-glmnet(train_x_lasso,train_y.lasso$Diagnosis,alpha = 1,lambda = grid,family = "binomial")

summary(lasso.mod)
```

```
##            Length Class     Mode
## a0          100   -none-    numeric
## beta       2900   dgCMatrix S4
## df          100   -none-    numeric
## dim           2   -none-    numeric
## lambda      100   -none-    numeric
## dev.ratio   100   -none-    numeric
## nulldev       1   -none-    numeric
## npasses       1   -none-    numeric
## jerr          1   -none-    numeric
## offset        1   -none-    logical
## classnames    2   -none-    character
## call          6   -none-    call
## nobs          1   -none-    numeric
```

```
cv.out <-cv.glmnet(train_x_lasso, train_y.lasso$Diagnosis, alpha = 1,family="binomial",nfolds = 10)
plot(cv.out)
```




```
best_lambda <- cv.out$lambda.min;best_lambda
```

```
## [1] 0.007279412
```

```
train_y<-as.matrix(train_y)
lasso.final<-glmnet(train_x_lasso,train_y.lasso$Diagnosis,alpha = 1,lambda = best_lambda,family = "bino
lasso.pred <- predict(lasso.final, s = best_lambda, newx = as.matrix(test_x))
lasso.pred.class<-ifelse(lasso.pred > 0.5,1,0)
table(prediction=lasso.pred.class,actual=test_y)
```

```
##           actual
## prediction   0   1
##          0 126  34
##          1   4  51
```

```
Recall.lasso<-sum(lasso.pred.class == 1 & test_y == 1)/sum(test_y == 1);Recall.lasso
```

```
## [1] 0.6
```

```
Precision.lasso<-sum(lasso.pred.class == 1 & test_y == 1)/sum(lasso.pred.class == 1);Precision.lasso
```

```
## [1] 0.9272727
```

```
F1Score.lasso<-2*Precision.lasso*Recall.lasso/(Precision.lasso+Recall.lasso);F1Score.lasso
```

```
## [1] 0.7285714
```

```
Accuracy.lasso<-sum(lasso.pred.class == test_y)/length(test_y);Accuracy.lasso
```

```
## [1] 0.8232558
```

```
ALZH_noID_noCholest<-
  ALZH_noID%>%
  dplyr::select(-CholesterolTotal,-CholesterolHDL,-CholesterolLDL,-CholesterolTriglycerides)
```

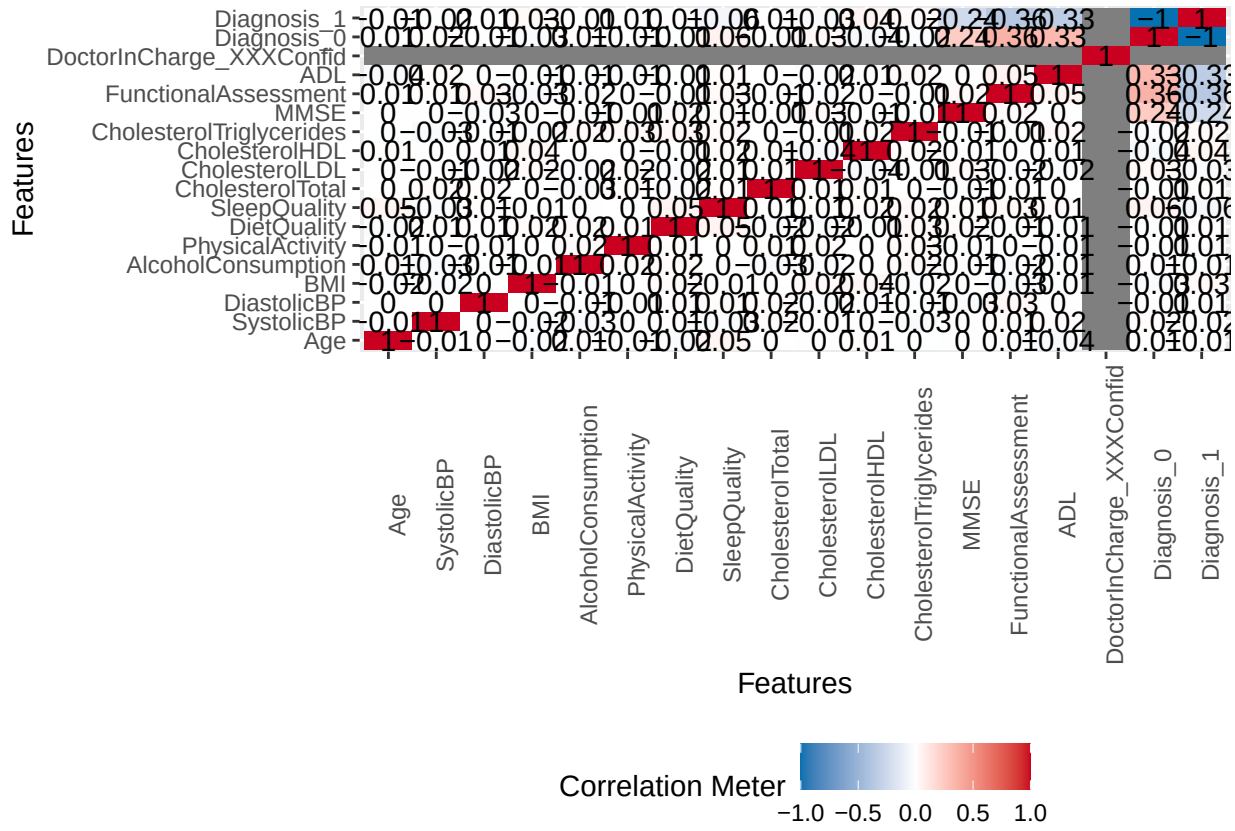
KNN

```
ALZH_IntOnly<-ALZH_noID[,sapply(ALZH.gbm,is.integer)]
ALZH_double<-ALZH_noID[,sapply(ALZH.gbm,is.double)]
ALZH_NumOnly<-cbind(ALZH_IntOnly,ALZH_double)
ALZH_fct<-ALZH_noID[,sapply(ALZH.gbm,is.factor)]
```

```
plot_correlation(ALZH_NumOnly)
```

```
## Warning in cor(x = structure(list(Age = c(73L, 89L, 73L, 74L, 89L, 86L, : the
## standard deviation is zero
```

```
## Warning: Removed 34 rows containing missing values or values outside the scale range
## (`geom_text()`).
```




```
#RFE.featureSet<-ALZH_NumOnly[draw,]%>%dplyr::select(-Diagnosis,DoctorInCharge)
#RFE.featureSet<-as.data.frame(RFE.featureSet)
#RFE.response<-ALZH_NumOnly[draw,]%>%dplyr::select(Diagnosis)

RFE.featureSet <- ALZH_NumOnly[draw,-which(names(ALZH_NumOnly) == "Diagnosis")]
RFE.featureSet<-RFE.featureSet[,-which(names(RFE.featureSet) == "DoctorInCharge")]
RFE.featureSet<-RFE.featureSet%>%select(-c(CholesterolTotal,CholesterolLDL,CholesterolTriglycerides))
RFE.response <- ALZH_NumOnly[draw, "Diagnosis"]

set.seed(12345)
control<-rfeControl(functions = rfFuncs, method = "cv", number = 10)
RFE.result<-rfe(RFE.featureSet,RFE.response, sizes = c(1:13), rfeControl = control)
print(RFE.result)
```

Feature Selection for kNN

```
##
## Recursive feature selection
##
## Outer resampling method: Cross-Validated (10 fold)
##
## Resampling performance over subset size:
##
```

```
##  Variables Accuracy  Kappa AccuracySD KappaSD Selected
##          1   0.6257 0.1754    0.03259 0.07033
##          2   0.7384 0.4200    0.02731 0.05662
##          3   0.8082 0.5522    0.02179 0.05236
##          4   0.8180 0.5737    0.02604 0.06161
##          5   0.8221 0.5819    0.02253 0.05103
##          6   0.8206 0.5757    0.02336 0.05581
##          7   0.8216 0.5787    0.02289 0.05282
##          8   0.8263 0.5876    0.02226 0.05202
##          9   0.8268 0.5886    0.01838 0.04473
##         10   0.8293 0.5934    0.02156 0.05172
##         11   0.8273 0.5885    0.01896 0.04517
##         12   0.8304 0.5936    0.01932 0.04743        *
##
## The top 5 variables (out of 12):
##    FunctionalAssessment, MMSE, ADL, SleepQuality, DietQuality
```

```
#alzh_secondKNN and the later alzh.gbm.forTuning are the same
alzh_secondKNN<-ALZH_NumOnly[draw,]%>%
dplyr::select(FunctionalAssessment, MMSE, ADL, DietQuality, SleepQuality,Diagnosis)%>%
mutate(Diagnosis=as.factor(Diagnosis))
alzh_secondKNN.test<-ALZH.gbm[-draw,]%>%dplyr::select(FunctionalAssessment, MMSE, ADL, DietQuality, Sle
mutate(Diagnosis=as.factor(Diagnosis))
```

```
set.seed(12345)
k_list<-seq(1,20,by=1)
nk<-length(k_list);nk
```

```
## [1] 20
```

```
Perf.Metric.knn<-data.frame(k=rep(0,nk),Recall=rep(0,length(k_list)),Precision=rep(0,length(k_list)),F1
Accuracy=rep(0,length(k_list)))

set.seed(12345)
n<-nrow(alzh_secondKNN)
pool<-rep(1:10,ceiling(n/10))
fold<-sample(pool,n,replace = FALSE)

for(k in 1:nk){
  Perf.Metric.knn$k[k]<-k

  recall.sum<-0
  precision.sum<-0
  f1_score.sum<-0
  accuracy.sum<-0


  for(i in 1:10){
    #Find data in each fold
    infold<-which(fold == i)

    #Create training and testing sets
    Train<-alzh_secondKNN[-infold,]
```

```
    Test<-alzh_secondKNN[infold,]
    #Run kNN
    k_preds<-knn(Train%>%select(-Diagnosis),Test%>%select(-Diagnosis),k=k,cl=Train$Diagnosis)

    Recall<-sum(k_preds == 1 & Test$Diagnosis == 1)/sum(Test$Diagnosis == 1);recall.sum<-recall.sum+Reca
    Precision<-sum(k_preds == 1 & Test$Diagnosis == 1)/sum(k_preds == 1);precision.sum<-precision.sum+Pr
    F1_Score<-2*Precision*Recall/(Precision+Recall);f1_score.sum<-f1_score.sum+F1_Score
    Accuracy<-sum(k_preds == Test$Diagnosis)/length(Test$Diagnosis);accuracy.sum<-accuracy.sum+Accuracy

  }

    Perf.Metric.knn$Recall[k]<-recall.sum/10
    Perf.Metric.knn$Precision[k]<-precision.sum/10
    Perf.Metric.knn$F1_Score[k]<-f1_score.sum/10
    Perf.Metric.knn$Accuracy[k]<-accuracy.sum/10

}

Perf.Metric.knn$k[which.max(Perf.Metric.knn$Recall)]
```

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

```
Perf.Metric.knn
```

```
##     k    Recall Precision  F1_Score  Accuracy
## 1   1 0.6403017 0.6298578 0.6338582 0.7440575
## 2   2 0.6329861 0.6525068 0.6414886 0.7539131
## 3   3 0.6295043 0.7049775 0.6642345 0.7787034
## 4   4 0.6311181 0.7129507 0.6680106 0.7827976
## 5   5 0.6083041 0.7309030 0.6624085 0.7843520
## 6   6 0.6062049 0.7484180 0.6678650 0.7900329
## 7   7 0.6059796 0.7604216 0.6726508 0.7946586
## 8   8 0.5963425 0.7485469 0.6621445 0.7889617
## 9   9 0.5943695 0.7772002 0.6719153 0.7993487
## 10 10 0.6100336 0.7785780 0.6820490 0.8029758
## 11 11 0.6004353 0.7807016 0.6772721 0.8019073
## 12 12 0.5978152 0.7877733 0.6782483 0.8024362
## 13 13 0.5912956 0.7864132 0.6735401 0.8008844
## 14 14 0.5978156 0.7955327 0.6807846 0.8055450
## 15 15 0.5918308 0.7970315 0.6776401 0.8045195
## 16 16 0.5927360 0.7979270 0.6783759 0.8055639
## 17 17 0.5902186 0.7992625 0.6774772 0.8050323
## 18 18 0.5964732 0.7960977 0.6803138 0.8055558
## 19 19 0.6023228 0.7998812 0.6856613 0.8091855
## 20 20 0.5931489 0.7914241 0.6756385 0.8035020
```

```
knn.final<-knn(train = alzh_secondKNN%>%select(-Diagnosis),test = alzh_secondKNN.test%>%select(-Diagnosi
table(knn.final,alzh_secondKNN.test$Diagnosis)
```

```
##
## knn.final   0   1
##         0 101  30
##         1  29  55
```

```
Recall.knn.final<-sum(knn.final == 1 & alzh_secondKNN.test$Diagnosis == 1)/sum(alzh_secondKNN.test$Diagn
```

```
## [1] 0.6470588
```

```
Precision.knn.final<-sum(knn.final == 1 & alzh_secondKNN.test$Diagnosis == 1)/sum(knn.final == 1);Precis
```

```
## [1] 0.6547619
```

```
F1Score.knn.final<-2*Precision.knn.final*Recall.knn.final/(Precision.knn.final+Recall.knn.final);F1Score
```

```
## [1] 0.6508876
```

```
Accuracy.knn.final<-sum(knn.final == alzh_secondKNN.test$Diagnosis)/length(alzh_secondKNN.test$Diagnosis
```

```
## [1] 0.7255814
```

gbm

```
set.seed(12345)
ALZH.boosting<-ALZH.gbm%>%select(-c(CholesterolTotal,CholesterolLDL,CholesterolTriglycerides))
boosting.try <- gbm(Diagnosis ~ ., data = ALZH.boosting[draw,], distribution = "bernoulli", n.trees = 5
```

```
yhat.gbm<-predict(boosting.try,newdata = ALZH.gbm[-draw,],n.trees = 5000,interaction.depth = 4,shrinkage
pred_gbm_class <- ifelse(yhat.gbm > 0.5, 1, 0)
table(pred_gbm_class,ALZH.gbm[-draw,]$Diagnosis)
```

```
##
## pred_gbm_class   0   1
##              0 125   5
##              1   5  80
```

```
set.seed(12345)
lambda_val <- seq(0.01, 0.05, by = 0.01)
result_container <- data.frame(Lambda = lambda_val, Recall = rep(0, length(lambda_val)), Precision = rep
ALZH.boosting.forTunine<-ALZH.gbm[draw,]
ALZH.boosting.realTest<-ALZH.gbm[-draw,]
```

Tune Together with 10 fold cv

This one is correct!!

```
ALZH.gbm.forTuning<-ALZH.gbm[draw,]
ALZH.gbm.realTest<-ALZH.gbm[-draw,]
```

```r
set.seed(12345)
lambda_val <- seq(0.01, 0.03, by = 0.01)
ntree_val <- c(1000, 2000, 3000)

ALZH.gbm.forGrid<-ALZH.gbm.forTuning%>%select(-c(CholesterolTotal,CholesterolLDL,CholesterolTriglyceride
ALZH.gbm.forGrid$Diagnosis <- factor(ALZH.gbm.forGrid$Diagnosis, levels = c(0, 1), labels = c("No", "Yes
ALZH.gbm.realTest<-ALZH.gbm[-draw,]%>%select(-c(CholesterolTotal,CholesterolLDL,CholesterolTriglycerides


### Grid Creation
train.control<-trainControl(method="cv",number=10,summaryFunction=twoClassSummary,classProbs=TRUE,savePr
grid<-expand.grid(shrinkage=lambda_val,
n.trees=ntree_val,
interaction.depth=4,n.minobsinnode=10)##default is 10


set.seed(12345)
Boosting_alzh_grid <- train(
  Diagnosis ~ .,
  data = ALZH.gbm.forGrid,
  method = "gbm",
  trControl = train.control,
  tuneGrid = grid,
  distribution = "bernoulli",
  metric = "Recall",
  verbose=TRUE,
  train.fraction = 0.9
)

Boosting_alzh_grid$results
Boosting_alzh_grid$bestTune

set.seed(12345)
for.final.gbm<-ALZH.gbm.forTuning%>%select(-c(CholesterolTotal,CholesterolLDL,CholesterolTriglycerides)
Boosting_alzh_grid.final<-gbm(Diagnosis~.,data=for.final.gbm,distribution="bernoulli",n.trees=1000,inter
yhat.boost.final<-predict(Boosting_alzh_grid.final,newdata = ALZH.gbm.realTest,n.trees = 1000,interacti
pred_gbm_class_final <- ifelse(yhat.boost.final > 0.5, 1, 0)
table(pred_gbm_class_final,ALZH.gbm.realTest$Diagnosis)


Recall.grid.gbm<-sum(pred_gbm_class_final == 1 & ALZH.gbm.realTest$Diagnosis == 1)/sum(ALZH.gbm.realTest
Precision.grid.gbm<-sum(pred_gbm_class_final == 1 & ALZH.gbm.realTest$Diagnosis == 1)/sum(pred_gbm_class
F1_Score.grid.gbm<-2*Precision.grid.gbm*Recall.grid.gbm/(Precision.grid.gbm+Recall.grid.gbm)
Accuracy.grid.gbm<-sum(pred_gbm_class_final == ALZH.gbm.realTest$Diagnosis)/length(ALZH.gbm.realTest$Dia

Recall.grid.gbm
Accuracy.grid.gbm
Precision.grid.gbm
F1_Score.grid.gbm
```