|  |
| --- |
| # Data(Train)  > train\_sal <- read.csv(file.choose())  > str(train\_sal)  'data.frame': 30161 obs. of 14 variables:  $ age : int 39 50 38 53 28 37 49 52 31 42 ...  $ workclass : chr " State-gov" " Self-emp-not-inc" " Private" " Private" ...  $ education : chr " Bachelors" " Bachelors" " HS-grad" " 11th" ...  $ educationno : int 13 13 9 7 13 14 5 9 14 13 ...  $ maritalstatus: chr " Never-married" " Married-civ-spouse" " Divorced" " Married-civ-spouse" ...  $ occupation : chr " Adm-clerical" " Exec-managerial" " Handlers-cleaners" " Handlers-cleaners" ...  $ relationship : chr " Not-in-family" " Husband" " Not-in-family" " Husband" ...  $ race : chr " White" " White" " White" " Black" ...  $ sex : chr " Male" " Male" " Male" " Male" ...  $ capitalgain : int 2174 0 0 0 0 0 0 0 14084 5178 ...  $ capitalloss : int 0 0 0 0 0 0 0 0 0 0 ...  $ hoursperweek : int 40 13 40 40 40 40 16 45 50 40 ...  $ native : chr " United-States" " United-States" " United-States" " United-States" ...  $ Salary : chr " <=50K" " <=50K" " <=50K" " <=50K" ...  > View(train\_sal)  > train\_sal$educationno <- as.factor(train\_sal$educationno)  > class(train\_sal)  [1] "data.frame"  > # Data(Test)  > test\_sal <- read.csv(file.choose())  > str(test\_sal)  'data.frame': 15060 obs. of 14 variables:  $ age : int 25 38 28 44 34 63 24 55 65 36 ...  $ workclass : chr " Private" " Private" " Local-gov" " Private" ...  $ education : chr " 11th" " HS-grad" " Assoc-acdm" " Some-college" ...  $ educationno : int 7 9 12 10 6 15 10 4 9 13 ...  $ maritalstatus: chr " Never-married" " Married-civ-spouse" " Married-civ-spouse" " Married-civ-spouse" ...  $ occupation : chr " Machine-op-inspct" " Farming-fishing" " Protective-serv" " Machine-op-inspct" ...  $ relationship : chr " Own-child" " Husband" " Husband" " Husband" ...  $ race : chr " Black" " White" " White" " Black" ...  $ sex : chr " Male" " Male" " Male" " Male" ...  $ capitalgain : int 0 0 0 7688 0 3103 0 0 6418 0 ...  $ capitalloss : int 0 0 0 0 0 0 0 0 0 0 ...  $ hoursperweek : int 40 50 40 40 30 32 40 10 40 40 ...  $ native : chr " United-States" " United-States" " United-States" " United-States" ...  $ Salary : chr " <=50K" " <=50K" " >50K" " >50K" ...  > View(test\_sal)  > test\_sal$educationno <- as.factor(test\_sal$educationno)  > class(test\_sal)  [1] "data.frame"  > #Visualization  > # Plot and ggplot  > ggplot(data=train\_sal,aes(x=train\_sal$Salary, y = train\_sal$age, fill = train\_sal$Salary)) +  + geom\_boxplot() +  + ggtitle("Box Plot") |
|  |
| |  | | --- | | >    train\_sal$workclass <- as.factor(train\_sal$workclass)  train\_sal$Salary <- as.factor(train\_sal$Salary)  plot(train\_sal$workclass,train\_sal$Salary)    plot(as.factor(train\_sal$education),as.factor(train\_sal$Salary))    ggplot(data=train\_sal,aes(x=train\_sal$Salary, y = train\_sal$capitalgain, fill = train\_sal$Salary)) +  geom\_boxplot() +  ggtitle("Box Plot")    ggplot(data=train\_sal,aes(x=train\_sal$Salary, y = train\_sal$hoursperweek, fill = train\_sal$Salary)) +  geom\_boxplot() +  ggtitle("Box Plot")    plot(as.factor(train\_sal$native),as.factor(train\_sal$Salary))    #Density Plot  ggplot(data=train\_sal,aes(x = train\_sal$age, fill = train\_sal$Salary)) +  geom\_density(alpha = 0.9, color = 'Violet')    > ggtitle("Age - Density Plot")  $title  [1] "Age - Density Plot"  attr(,"class")  [1] "labels"  ggplot(data=train\_sal,aes(x = train\_sal$occupation, fill = train\_sal$Salary)) +  geom\_density(alpha = 0.9, color = 'Violet') | |

> ggtitle("occupation Density Plot")

$title

[1] "occupation Density Plot"

attr(,"class")

[1] "labels"

# Naive Bayes Model

Model <- naiveBayes(train\_sal$Salary ~ ., data = train\_sal)

Model

|  |
| --- |
| Naive Bayes Classifier for Discrete Predictors  Call:  naiveBayes.default(x = X, y = Y, laplace = laplace)  A-priori probabilities:  Y  <=50K >50K  0.7510693 0.2489307  Conditional probabilities:  age  Y [,1] [,2]  <=50K 36.60826 13.46489  >50K 43.95911 10.26963  workclass  Y Federal-gov Local-gov Private Self-emp-inc  <=50K 0.0255153843 0.0643623361 0.7685074825 0.0209243809  >50K 0.0486148109 0.0811134790 0.6494405967 0.0799147576  workclass  Y Self-emp-not-inc State-gov Without-pay  <=50K 0.0787975103 0.0412748863 0.0006180197  >50K 0.0950985615 0.0458177944 0.0000000000  education  Y 10th 11th 12th 1st-4th 5th-6th  <=50K 0.0335937845 0.0436586766 0.0153622037 0.0064009182 0.0121838167  >50K 0.0078582845 0.0078582845 0.0038625466 0.0007991476 0.0015982952  education  Y 7th-8th 9th Assoc-acdm Assoc-voc Bachelors  <=50K 0.0230433055 0.0189820333 0.0331964861 0.0425109257 0.1288129608  >50K 0.0046616942 0.0033297816 0.0340969632 0.0458177944 0.2831646244  education  Y Doctorate HS-grad Masters Preschool Prof-school  <=50K 0.0041937050 0.3629982784 0.0312982828 0.0019864919 0.0060036198  >50K 0.0372935535 0.2153702717 0.1222695791 0.0000000000 0.0540756526  education  Y Some-college  <=50K 0.2357745111  >50K 0.1779435269  educationno  Y 1 2 3 4 5  <=50K 0.0019864919 0.0064009182 0.0121838167 0.0230433055 0.0189820333  >50K 0.0000000000 0.0007991476 0.0015982952 0.0046616942 0.0033297816  educationno  Y 6 7 8 9 10  <=50K 0.0335937845 0.0436586766 0.0153622037 0.3629982784 0.2357745111  >50K 0.0078582845 0.0078582845 0.0038625466 0.2153702717 0.1779435269  educationno  Y 11 12 13 14 15  <=50K 0.0425109257 0.0331964861 0.1288129608 0.0312982828 0.0060036198  >50K 0.0458177944 0.0340969632 0.2831646244 0.1222695791 0.0540756526  educationno  Y 16  <=50K 0.0041937050  >50K 0.0372935535  maritalstatus  Y Divorced Married-AF-spouse Married-civ-spouse  <=50K 0.1660707191 0.0004855869 0.3384099236  >50K 0.0602024507 0.0013319126 0.8522908897  maritalstatus  Y Married-spouse-absent Never-married Separated Widowed  <=50K 0.0149649053 0.4085551583 0.0385379420 0.0329757648  >50K 0.0041289291 0.0625998934 0.0087906233 0.0106553010  occupation  Y Adm-clerical Armed-Forces Craft-repair Exec-managerial  <=50K 0.1422769611 0.0003531541 0.1378183905 0.0907164614  >50K 0.0663292488 0.0001331913 0.1209376665 0.2579914758  occupation  Y Farming-fishing Handlers-cleaners Machine-op-inspct  <=50K 0.0385820863 0.0559307818 0.0759281331  >50K 0.0153169952 0.0110548748 0.0326318594  occupation  Y Other-service Priv-house-serv Prof-specialty Protective-serv  <=50K 0.1359643314 0.0062684854 0.0983092747 0.0191586103  >50K 0.0175812467 0.0001331913 0.2412093767 0.0279701652  occupation  Y Sales Tech-support Transport-moving  <=50K 0.1153931047 0.0279874630 0.0553127621  >50K 0.1291955248 0.0370271710 0.0424880128  relationship  Y Husband Not-in-family Other-relative Own-child  <=50K 0.299474683 0.304727851 0.037655057 0.194323048  >50K 0.756393181 0.109616409 0.004661694 0.008524241  relationship  Y Unmarried Wife  <=50K 0.132388646 0.031430716  >50K 0.028369739 0.092434736  race  Y Amer-Indian-Eskimo Asian-Pac-Islander Black Other  <=50K 0.011124354 0.028561338 0.108197590 0.009270295  >50K 0.004528503 0.033031433 0.048748002 0.002797017  race  Y White  <=50K 0.842846422  >50K 0.910895045  sex  Y Female Male  <=50K 0.3826866 0.6173134  >50K 0.1481087 0.8518913  capitalgain  Y [,1] [,2]  <=50K 148.9004 936.4124  >50K 3937.6798 14386.0600  capitalloss  Y [,1] [,2]  <=50K 53.35302 309.9476  >50K 193.75067 592.8256  hoursperweek  Y [,1] [,2]  <=50K 39.34856 11.95104  >50K 45.70658 10.73699  native  Y Cambodia Canada China Columbia Cuba  <=50K 0.0004855869 0.0031342427 0.0021189246 0.0023837902 0.0029576657  >50K 0.0009323388 0.0047948855 0.0026638253 0.0002663825 0.0033297816  native  Y Dominican-Republic Ecuador El-Salvador England  <=50K 0.0028693771 0.0010153181 0.0040171280 0.0024720788  >50K 0.0002663825 0.0005327651 0.0011987214 0.0039957379  native  Y France Germany Greece Guatemala Haiti  <=50K 0.0006621640 0.0037081181 0.0009270295 0.0026486558 0.0016774820  >50K 0.0015982952 0.0058604156 0.0010655301 0.0003995738 0.0005327651  native  Y Honduras Hong Hungary India Iran  <=50K 0.0004855869 0.0005738754 0.0004414426 0.0026486558 0.0010594623  >50K 0.0001331913 0.0007991476 0.0003995738 0.0053276505 0.0023974427  native  Y Ireland Italy Jamaica Japan Laos  <=50K 0.0008387410 0.0019423476 0.0030900984 0.0015891935 0.0006621640  >50K 0.0006659563 0.0031965903 0.0013319126 0.0030633990 0.0002663825  native  Y Mexico Nicaragua Outlying-US(Guam-USVI-etc)  <=50K 0.0254712400 0.0013684722 0.0006180197  >50K 0.0043953117 0.0002663825 0.0000000000  native  Y Peru Philippines Poland Portugal Puerto-Rico  <=50K 0.0012360394 0.0056504657 0.0019864919 0.0013243279 0.0042819936  >50K 0.0002663825 0.0079914758 0.0014651039 0.0005327651 0.0015982952  native  Y Scotland South Taiwan Thailand  <=50K 0.0003972984 0.0025162230 0.0010153181 0.0006180197  >50K 0.0002663825 0.0018646777 0.0025306340 0.0003995738  native  Y Trinadad&Tobago United-States Vietnam Yugoslavia  <=50K 0.0007063082 0.9053546992 0.0026045115 0.0004414426  >50K 0.0002663825 0.9316728823 0.0006659563 0.0007991476 |
|  |
| |  | | --- | | > | |

> Model\_pred <- predict(Model,test\_sal)

> mean(Model\_pred==test\_sal$Salary)

[1] 0.8187251

table(salary\_test$Salary,Model\_pred)

Model\_pred

<=50K >50K

<=50K 10549 811

>50K 1919 1781

> CrossTable(salary\_test$Salary,Model\_pred)

Cell Contents

|-------------------------|

| N |

| Chi-square contribution |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 15060

| Model\_pred

salary\_test$Salary | <=50K | >50K | Row Total |

-------------------|-----------|-----------|-----------|

<=50K | 10549 | 811 | 11360 |

| 139.202 | 669.585 | |

| 0.929 | 0.071 | 0.754 |

| 0.846 | 0.313 | |

| 0.700 | 0.054 | |

-------------------|-----------|-----------|-----------|

>50K | 1919 | 1781 | 3700 |

| 427.386 | 2055.808 | |

| 0.519 | 0.481 | 0.246 |

| 0.154 | 0.687 | |

| 0.127 | 0.118 | |

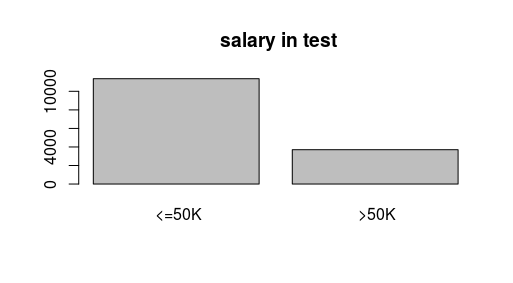
-------------------|-----------|-----------|-----------|

Column Total | 12468 | 2592 | 15060 |

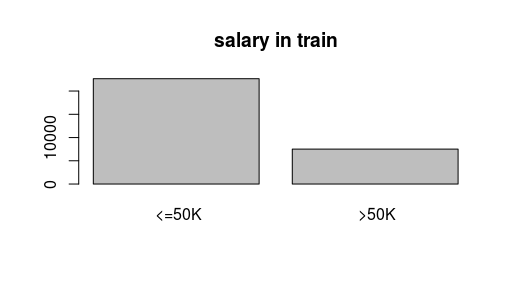
| 0.828 | 0.172 | |

-------------------|-----------|-----------|-----------|

barplot(table(salary\_test$Salary), main = " salary in test")



barplot(table(salary\_train$Salary) ,main = "salary in train")



confusionMatrix(Model\_pred,as.factor(test\_sal$Salary))

Confusion Matrix and Statistics

Reference

Prediction <=50K >50K

<=50K 10549 1919

>50K 811 1781

Accuracy : 0.8187

95% CI : (0.8125, 0.8248)

No Information Rate : 0.7543

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.456

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.9286

Specificity : 0.4814

Pos Pred Value : 0.8461

Neg Pred Value : 0.6871

Prevalence : 0.7543

Detection Rate : 0.7005

Detection Prevalence : 0.8279

Balanced Accuracy : 0.7050

'Positive' Class : <=50K