Nazmus Sakib
Id: 20-43156-1

Body Mass Index Dataset

Dataset Reference: https://www.kaggle.com/datasets/sjagkoo7/bmi-body-mass-index

This body mass index (BMI) dataset encompassing gender, height, weight, and index values. The goal is to uncover relationships between these attributes and the index, potentially reflecting health or body composition. Through data preparation, including normalization and feature scaling, we apply the K-Nearest Neighbors (KNN) algorithm to predict index values based on attribute information. Our focus on attribute-index correlations guides the selection of significant predictors, enhancing model performance. Ultimately, this analysis strives to reveal connections between gender, height, weight, and index, contributing to the classification of individuals according to the provided index and offering insights into health-related trends.

Attribute Information:

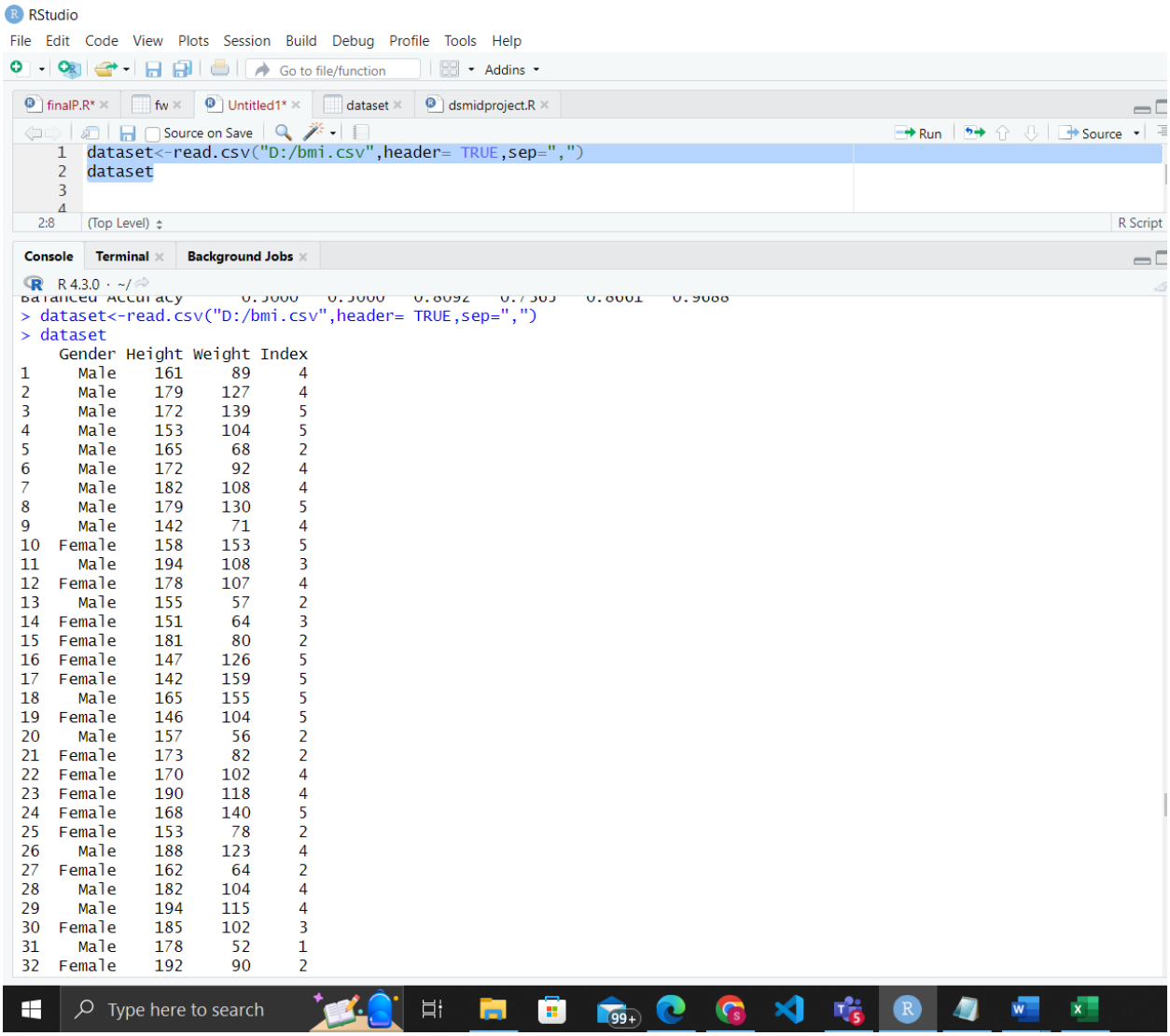
Gender: Represents the gender of individuals, categorized as 'Male' or 'Female'.

Height: Indicates the height of individuals in centimeters.

Weight: Signifies the weight of individuals in kilograms.

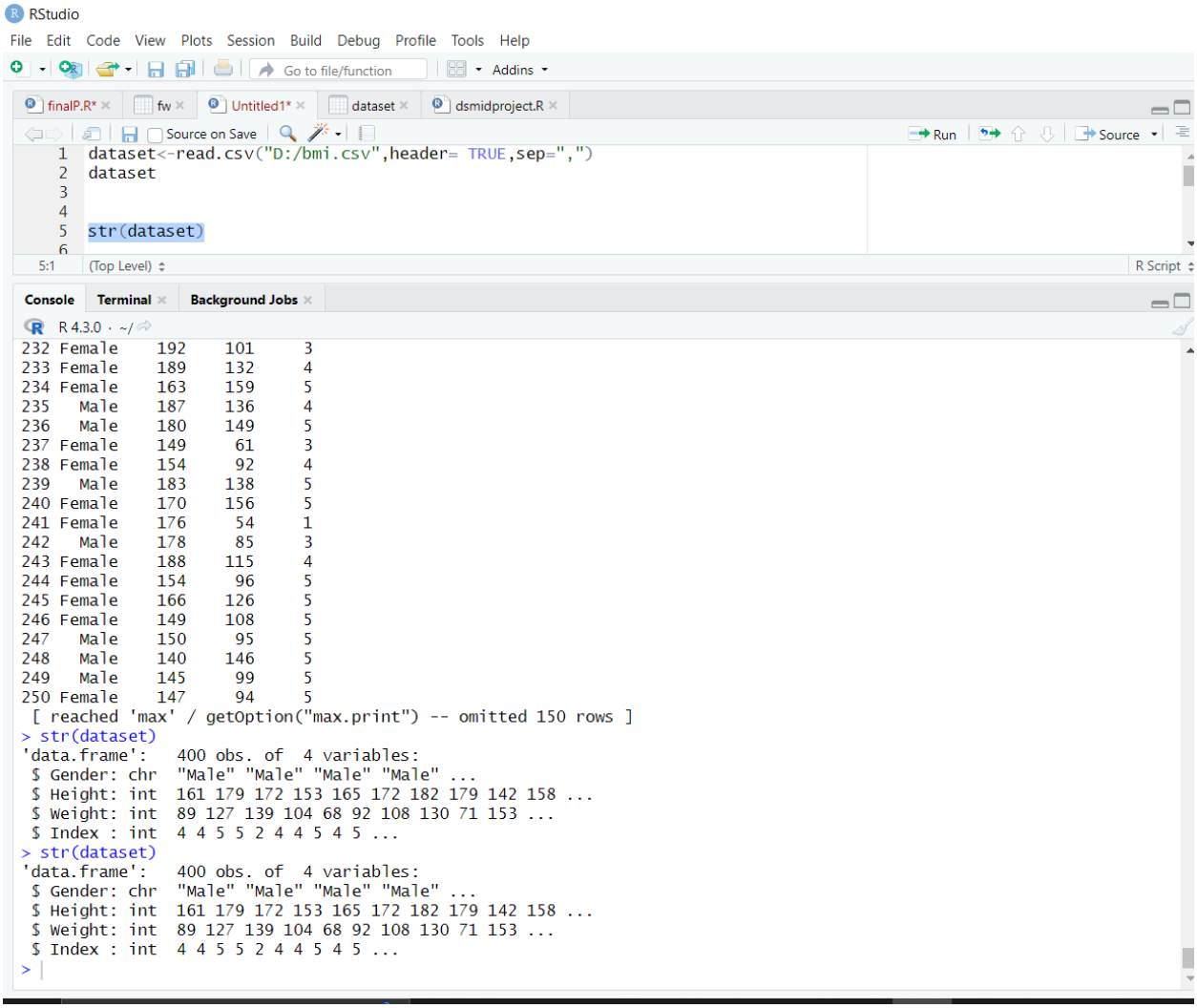
Index: Denotes an index value associated with health or body composition, which the analysis aims to predict.

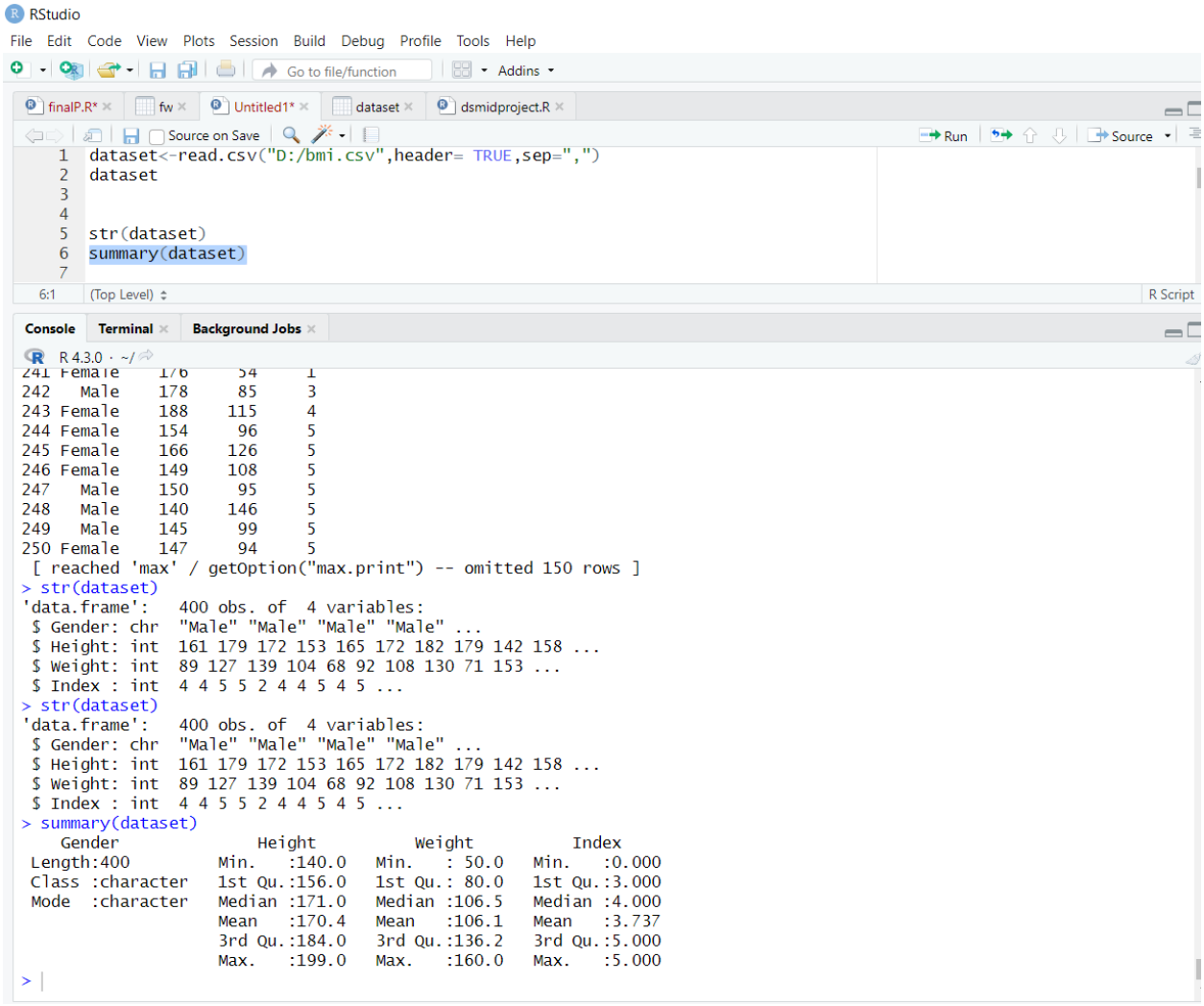
Dataset import:



Data Preparation:

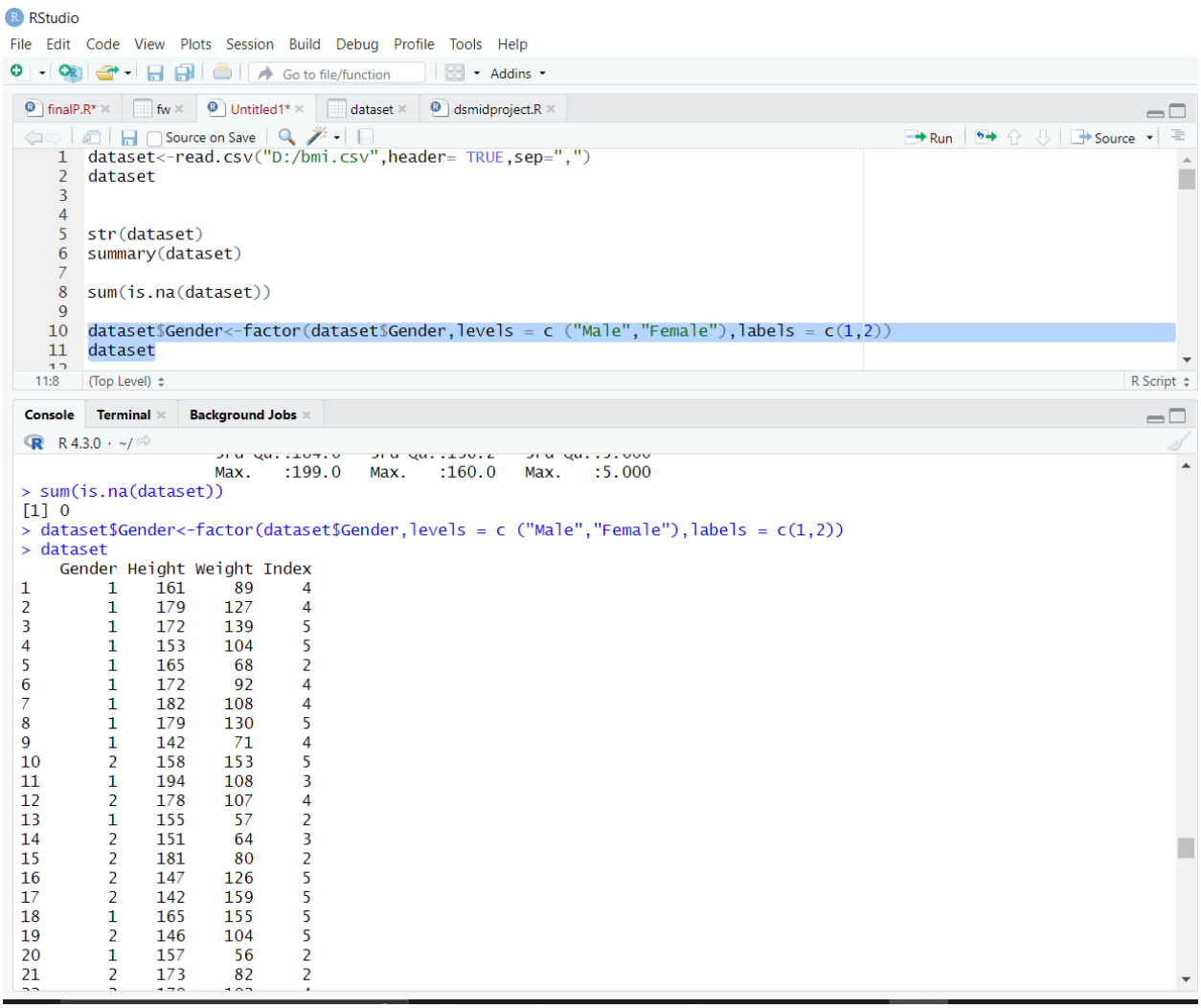
Finding the datatypes of this dataset,

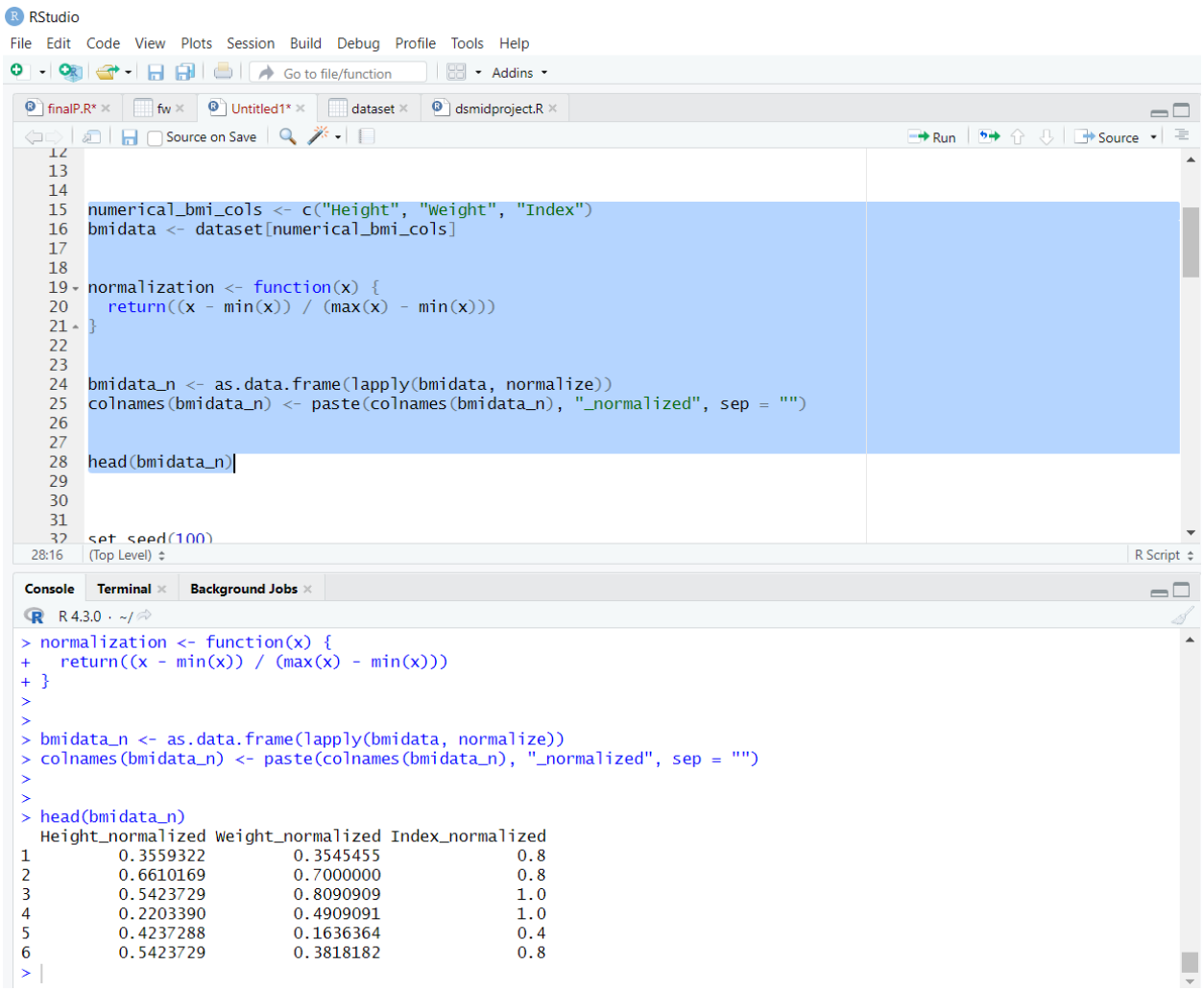

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
finalP.R*
fw
Untitled1*
dataset
dsmidproject.R
Source on Save
Run
Source
1 dataset<-read.csv("D:/bmi.csv",header= TRUE,sep=",")
2 dataset
3
4
5 str(dataset)
6
5:1
(Top Level)
R Script
Console
Terminal
Background Jobs
R 4.3.0 · ~/
232 Female 192 101 3
233 Female 189 132 4
234 Female 163 159 5
235 Male 187 136 4
236 Male 180 149 5
237 Female 149 61 3
238 Female 154 92 4
239 Male 183 138 5
240 Female 170 156 5
241 Female 176 54 1
242 Male 178 85 3
243 Female 188 115 4
244 Female 154 96 5
245 Female 166 126 5
246 Female 149 108 5
247 Male 150 95 5
248 Male 140 146 5
249 Male 145 99 5
250 Female 147 94 5
[reached 'max' / getOption("max.print") -- omitted 150 rows]
> str(dataset)
'data.frame': 400 obs. of 4 variables:
$ Gender: chr "Male" "Male" "Male" "Male" ...
$ Height: int 161 179 172 153 165 172 182 179 142 158 ...
$ Weight: int 89 127 139 104 68 92 108 130 71 153 ...
$ Index : int 4 4 5 5 2 4 4 5 4 5 ...
> str(dataset)
'data.frame': 400 obs. of 4 variables:
$ Gender: chr "Male" "Male" "Male" "Male" ...
$ Height: int 161 179 172 153 165 172 182 179 142 158 ...
$ Weight: int 89 127 139 104 68 92 108 130 71 153 ...
$ Index : int 4 4 5 5 2 4 4 5 4 5 ...
>



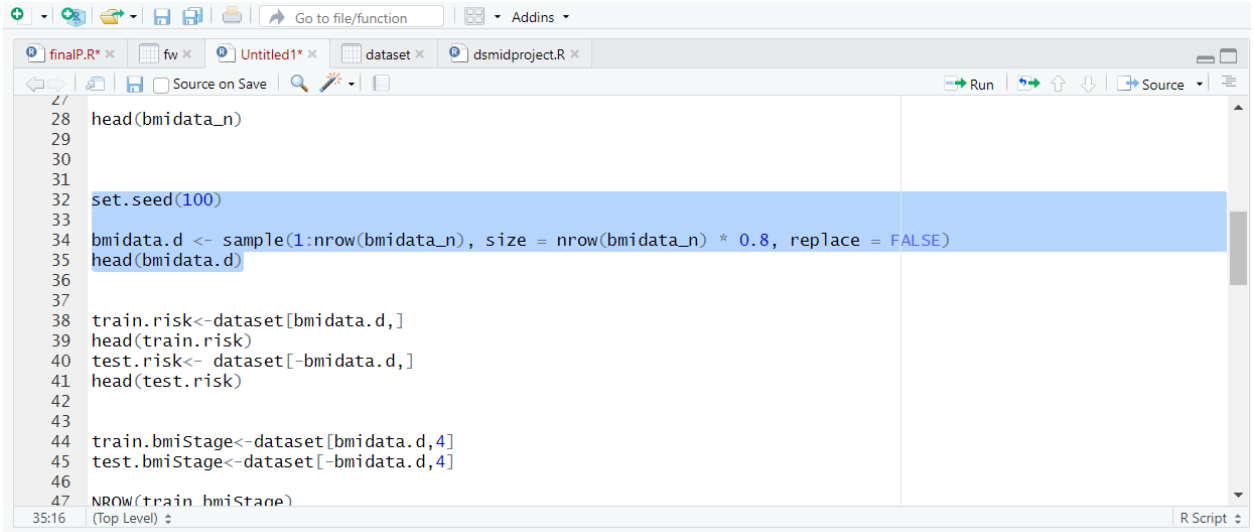
Applying Data preparation steps:

Checking missing value, make numeric column, normalize dataset:


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
finalP.R*
fw
Untitled1*
dataset
dsmidproject.R
Source on Save
Run
Source
1 dataset<-read.csv("D:/bmi.csv",header= TRUE,sep=",")
2 dataset
3
4
5 str(dataset)
6 summary(dataset)
7
8 sum(is.na(dataset))
9
10 dataset$Gender<-factor(dataset$Gender,levels = c ("Male","Female"),labels = c(1,2))
11 dataset
11:8
(Top Level)
R Script
Console
Terminal
Background Jobs
R 4.3.0 · ~/
Max. :199.0 Max. :160.0 Max. :5.000
> sum(is.na(dataset))
[1] 0
> dataset$Gender<-factor(dataset$Gender,levels = c ("Male","Female"),labels = c(1,2))
> dataset
Gender Height Weight Index
1 1 161 89 4
2 1 179 127 4
3 1 172 139 5
4 1 153 104 5
5 1 165 68 2
6 1 172 92 4
7 1 182 108 4
8 1 179 130 5
9 1 142 71 4
10 2 158 153 5
11 1 194 108 3
12 2 178 107 4
13 1 155 57 2
14 2 151 64 3
15 2 181 80 2
16 2 147 126 5
17 2 142 159 5
18 1 165 155 5
19 2 146 104 5
20 1 157 56 2
21 2 173 82 2

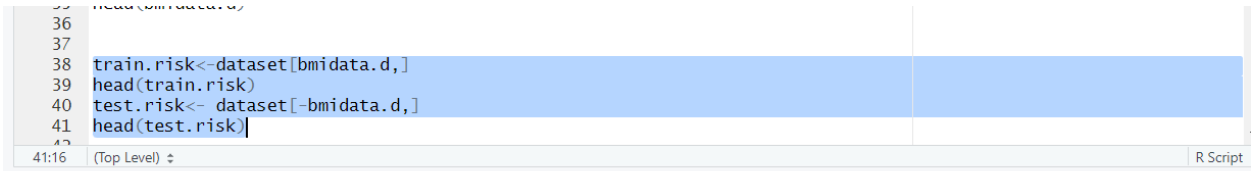


Splitting the data to prepare them for train and testing data:



```
head(bmidata_n)



set.seed(100)

bmidata.d <- sample(1:nrow(bmidata_n), size = nrow(bmidata_n) * 0.8, replace = FALSE)
head(bmidata.d)


train.risk<-dataset[bmidata.d,]
head(train.risk)
test.risk<- dataset[-bmidata.d,]
head(test.risk)


train.bmiStage<-dataset[bmidata.d,4]
test.bmiStage<-dataset[-bmidata.d,4]

NROW(train.bmiStage)
```

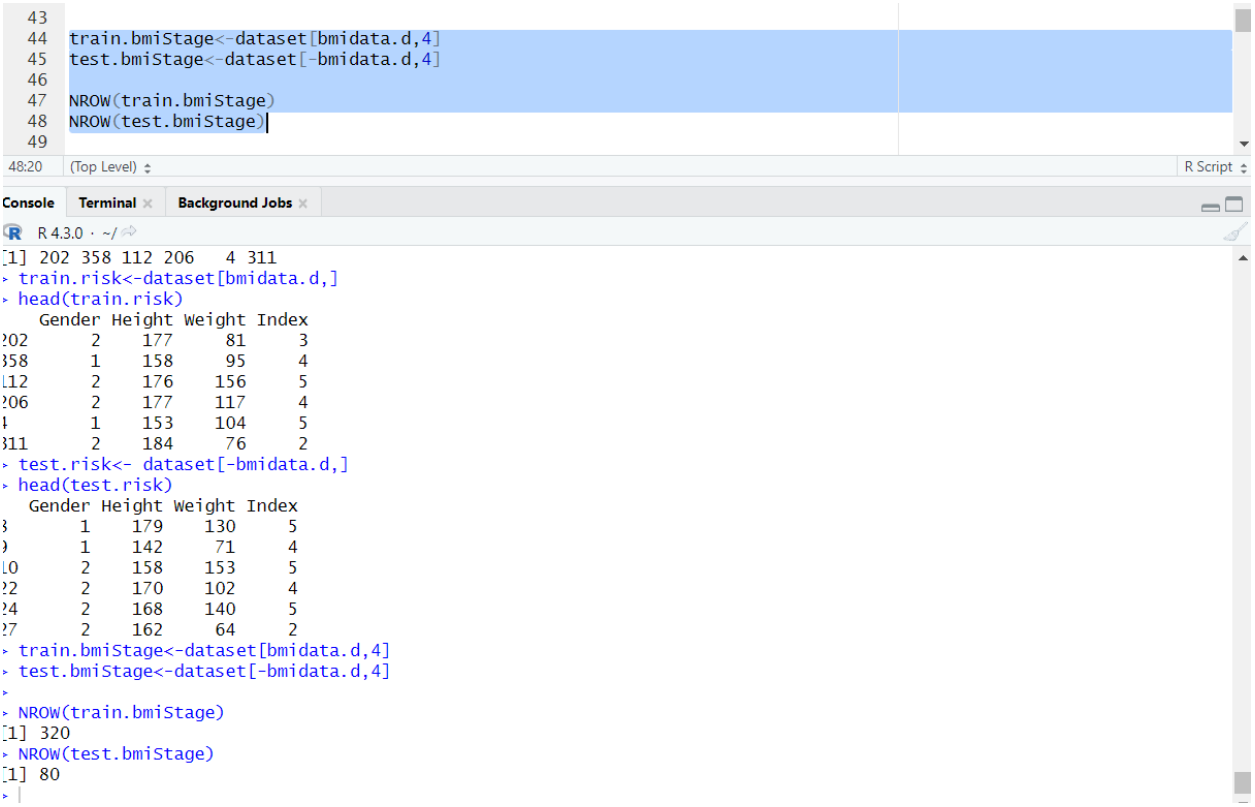
```
> bmidata_n <- as.data.frame(lapply(bmidata, normalize))
> colnames(bmidata_n) <- paste(colnames(bmidata_n), "_normalized", sep = "")
>
>
> head(bmidata_n)
  Height_normalized Weight_normalized Index_normalized
1         0.3559322         0.3545455              0.8
2         0.6610169         0.7000000              0.8
3         0.5423729         0.8090909              1.0
4         0.2203390         0.4909091              1.0
5         0.4237288         0.1636364              0.4
6         0.5423729         0.3818182              0.8
> set.seed(100)
>
> bmidata.d <- sample(1:nrow(bmidata_n), size = nrow(bmidata_n) * 0.8, replace = FALSE)
> head(bmidata.d)
[1] 202 358 112 206   4 311
>
```



```
train.risk<-dataset[bmidata.d,]
head(train.risk)
test.risk<- dataset[-bmidata.d,]
head(test.risk)
```

```
> bmidata.d <- sample(1:nrow(bmidata_n), size = nrow(bmidata_n) * 0.8, replace = FALSE)
> head(bmidata.d)
[1] 202 358 112 206   4 311
> train.risk<-dataset[bmidata.d,]
> head(train.risk)
    Gender Height Weight Index
202      2    177     81     3
358      1    158     95     4
112      2    176    156     5
206      2    177    117     4
4        1    153    104     5
311      2    184     76     2
> test.risk<- dataset[-bmidata.d,]
> head(test.risk)
   Gender Height Weight Index
8       1    179    130     5
9       1    142     71     4
10      2    158    153     5
22      2    170    102     4
24      2    168    140     5
27      2    162     64     2
>
```

Targeting an attribute and counting the train and test values:



```
43
44 train.bmiStage<-dataset[bmidata.d,4]
45 test.bmiStage<-dataset[-bmidata.d,4]
46
47 NROW(train.bmiStage)
48 NROW(test.bmiStage)
49
```

48:20 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.0 · ~/

```
[1] 202 358 112 206   4 311
> train.risk<-dataset[bmidata.d,]
> head(train.risk)
    Gender Height Weight Index
202      2    177     81     3
358      1    158     95     4
112      2    176    156     5
206      2    177    117     4
4        1    153    104     5
311      2    184     76     2
> test.risk<- dataset[-bmidata.d,]
> head(test.risk)
   Gender Height Weight Index
3       1    179    130     5
9       1    142     71     4
10      2    158    153     5
22      2    170    102     4
24      2    168    140     5
27      2    162     64     2
> train.bmiStage<-dataset[bmidata.d,4]
> test.bmiStage<-dataset[-bmidata.d,4]
>
> NROW(train.bmiStage)
[1] 320
> NROW(test.bmiStage)
[1] 80
>
```

Building the KNN model and finding accuracy. We found accuracy is 78.75.

For find the accuracy we need to install a package which is class and need to call library function.

```
55
56 knn.50<- knn(train=train.risk, test=test.risk, cl=train.bmiStage, k=50)
57
58
59 ACC.50 <- 100 * sum(test.bmiStage == knn.50)/NROW(test.bmiStage)
60 ACC.50
61
62
63 table(knn.50,test.bmiStage)
64 knn.50
```

60:7 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.0 · ~/

```
22      2    170    102     4
24      2    168    140     5
27      2    162     64     2
> train.bmiStage<-dataset[bmidata.d,4]
> test.bmiStage<-dataset[-bmidata.d,4]
>
> NROW(train.bmiStage)
[1] 320
> NROW(test.bmiStage)
[1] 80
> install.packages("class")
Error in install.packages : Updating loaded packages
> library(class)
> install.packages("class")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropr
iate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Warning in install.packages :
  package 'class' is in use and will not be installed
> knn.50<- knn(train=train.risk, test=test.risk, cl=train.bmiStage, k=50)
>
>
> ACC.50 <- 100 * sum(test.bmiStage == knn.50)/NROW(test.bmiStage)
> ACC.50
[1] 78.75
>
```

The accuracy table for K=50 is,

```
62
63  table(knn.50,test.bmiStage)
64  knn.50
65
66  install.packages("lattice")
67  library(lattice)
68  install.packages("caret")
```

64:7 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.0 · ~/

```
> install.packages("class")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropr
iate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Warning in install.packages :
  package 'class' is in use and will not be installed
> knn.50<- knn(train=train.risk, test=test.risk, cl=train.bmiStage, k=50)
>
>
> ACC.50 <- 100 * sum(test.bmiStage == knn.50)/NROW(test.bmiStage)
> ACC.50
[1] 78.75
> table(knn.50,test.bmiStage)
      test.bmiStage
knn.50  0  1  2  3  4  5
     0  0  0  0  0  0  0
     1  0  0  0  0  0  0
     2  3  1  5  3  0  0
     3  0  0  2  6  3  0
     4  0  0  0  2 20  0
     5  0  0  0  0  3 32
> knn.50
 [1] 5 4 5 4 5 2 4 3 5 2 5 5 5 4 5 3 4 3 3 4 2 5 4 3 2 5 2 4 4 5 3 4 2 5 3 4 4 5 5 5 5 4 3 5 4 4 4 4 2 4 5 3 5 5 2
[56] 5 5 5 3 5 2 5 5 4 3 5 5 4 5 4 5 2 2 5 5 4 2 5 5 5
Levels: 0 1 2 3 4 5
>
```

The confusion matrix is applied to understand the prediction:

```
76  finalResult<- confusionMatrix(table(knn.50,test.bmiStage))
77  print(finalResult)
78
79
80
81
82
```

77:19 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.0 · ~/

```
  package 'ggplot2' is in use and will not be installed
> finalResult<- confusionMatrix(table(knn.50,test.bmiStage))
> print(finalResult)
Confusion Matrix and Statistics

      test.bmiStage
knn.50  0  1  2  3  4  5
     0  0  0  0  0  0  0
     1  0  0  0  0  0  0
     2  3  1  5  3  0  0
     3  0  0  2  6  3  0
     4  0  0  0  2 20  0
     5  0  0  0  0  3 32

Overall Statistics

               Accuracy : 0.7875
                 95% CI : (0.6817, 0.8711)
    No Information Rate : 0.4
    P-Value [Acc > NIR] : 1.77e-12

                  Kappa : 0.698

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
```

```
finalResult<- confusionMatrix(table(knn.50,test.bmiStage))
print(finalResult)
```

```
Console  Terminal  Background Jobs
R 4.3.0 · ~/
     3  0  0  2  6  3  0
     4  0  0  0  2 20  0
     5  0  0  0  0  3 32

Overall Statistics

               Accuracy : 0.7875
                 95% CI : (0.6817, 0.8711)
    No Information Rate : 0.4
    P-Value [Acc > NIR] : 1.77e-12

                  Kappa : 0.698

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            0.0000   0.0000   0.7143   0.5455   0.7692   1.0000
Specificity            1.0000   1.0000   0.9041   0.9275   0.9630   0.9375
Pos Pred Value            NaN      NaN   0.4167   0.5455   0.9091   0.9143
Neg Pred Value         0.9625   0.9875   0.9706   0.9275   0.8966   1.0000
Prevalence             0.0375   0.0125   0.0875   0.1375   0.3250   0.4000
Detection Rate         0.0000   0.0000   0.0625   0.0750   0.2500   0.4000
Detection Prevalence   0.0000   0.0000   0.1500   0.1375   0.2750   0.4375
Balanced Accuracy      0.5000   0.5000   0.8092   0.7365   0.8661   0.9688
>
```

Kappa value of 0.698 indicates a substantial level of agreement beyond what would be expected by chance. This suggests that the KNN classifier is performing quite well in terms of classification accuracy and that the observed agreement between predicted and actual values.

For getting the final result we need to install 4 package and use the library function to use these packages after installation.

install.packages("lattice")

library(lattice)

install.packages("caret")

library(caret)

install.packages("stringi")

library(stringi)

install.packages("ggplot2")

library(ggplot2)