



American International University-Bangladesh (AIUB)

Course Code: CSC4180

Course Title: Introduction to Data Science

Project Title : Student Performance Prediction Using KNN

Classification Algorithm.

Submitted by

Name : RAHMAN MD, MAHFUZUR

ID : 20-43186-1

SECTION : [B]

Submitted to

TOHEDUL ISLAM

Assistant Professor

Faculty of Science and technology

I. Project Description :

This project involves the development of a classification model to predict student math performance using the K-Nearest Neighbors (KNN) algorithm. The primary objective is to determine whether students pass or fail based on a set of relevant features. The project begins with data preprocessing, including encoding categorical variables and identifying feature correlations. Low correlation features are pruned to optimize model performance. The dataset is then split into training and testing sets for model evaluation. The KNN algorithm is implemented, and the optimal number of neighbors (K value) is determined using cross-validation. Precision and recall metrics assess the model's accuracy in classifying student performance. Insights gained from feature importance and model effectiveness guide potential enhancements.

Future work may involve refining hyperparameters, exploring alternative algorithms, and advanced feature engineering techniques. The project aims to deliver an efficient and robust classification model applicable in educational contexts, with findings that can contribute to further research and improvements in student performance prediction.

1. Load Libraries and Dataset:

In this initial phase, we start by loading the crucial R library "caret," which serves as a powerful toolkit for training and evaluating machine learning models. The library encompasses diverse functions and tools that are invaluable for our analysis. We then proceed to load the dataset "exams.csv" from the specified file location. This dataset embodies a comprehensive collection of information, encompassing student demographics, exam scores, and performance indicators. This preparatory step lays the foundation for the subsequent data manipulation and analysis.

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help

Source

Console Terminal Background Jobs

R 4.3.0 ~>

```
> data <- read.csv("F:/exams.csv", header = TRUE, sep = ",")  
> print(data)
```

	gender	race.ethnicity	parental.level.of.education	lunch	test.preparation.course	math.score	reading.score	writing.score
1	female	group D	some college	standard	completed	59	70	78
2	male	group D	associate's degree	standard	none	96	93	87
3	female	group D	some college	free/reduced	none	57	76	77
4	male	group B	some college	free/reduced	none	70	70	63
5	female	group D	associate's degree	standard	none	83	85	86
6	male	group C	some high school	standard	none	68	57	54
7	female	group E	associate's degree	standard	none	82	83	80
8	female	group B	some high school	standard	none	46	61	58
9	male	group C	Some high school	standard	none	80	75	73
10	female	group C	bachelor's degree	standard	completed	57	69	77
11	male	group B	some high school	standard	none	74	69	69
12	male	group B	master's degree	standard	none	53	50	49
13	male	group B	bachelor's degree	free/reduced	none	76	74	76
14	male	group A	some college	standard	none	70	73	70
15	male	group C	master's degree	free/reduced	none	55	54	52
16	male	group E	master's degree	free/reduced	none	56	46	43
17	female	group C	some college	free/reduced	none	35	47	41
18	female	group C	high school	standard	none	87	92	81
19	female	group E	associate's degree	free/reduced	none	80	82	85
20	female	group D	associate's degree	standard	completed	65	71	74
21	male	group C	high school	free/reduced	none	66	66	62
22	female	group D	associate's degree	standard	completed	67	71	76
23	female	group B	some college	standard	none	70	71	71
24	male	group E	associate's degree	standard	none	89	88	86
25	male	group D	associate's degree	standard	completed	99	85	88
26	male	group B	some college	standard	none	74	83	72
27	male	group D	high school	free/reduced	none	58	52	51
28	male	group D	some high school	standard	none	70	66	59
29	female	group E	associate's degree	standard	none	80	79	71
30	male	group D	associate's degree	standard	none	90	87	86
31	female	group B	associate's degree	standard	completed	80	81	85
32	female	group D	associate's degree	free/reduced	none	68	76	79
33	female	group B	high school	free/reduced	completed	69	78	75
34	female	group D	master's degree	free/reduced	none	32	35	37

Environment

print("Recall...")
precision <-
print(recall)
print(precision)
recall <- co...
print("Recall...")
print(recall)
test_data
train_data
data <- read...
Files Plots

.RData Rhistory aim2010... Calculator... Calculator... Custom C... Date.pdf depositiph... desktop.ini Doc1.pdf IISExpress istockpho... June Baza... LabVIEW... LoginTest

Type here to search 82°F Haze 3:35 AM 8/15/2023

```

R 4.3.0 - /-
38 male group D associate's degree standard completed 68 66 72
39 male group C associate's degree free/reduced completed 74 85 87
40 male group E master's degree standard none 89 85 78
41 male group C associate's degree free/reduced completed 46 46 48
42 male group C associate's degree standard completed 76 82 77
43 male group B high school standard none 86 82 72
44 male group D some college standard none 69 73 67
45 female group B high school standard none 53 56 54
46 male group C bachelor's degree standard none 63 71 65
47 male group A associate's degree standard completed 96 82 90
48 male group C some college standard completed 80 76 68
49 female group E high school standard none 59 52 56
50 male group D some high school standard completed 80 77 80
51 female group E high school free/reduced completed 65 77 74
52 female group E master's degree free/reduced completed 74 83 84
53 male group D some high school standard none 90 93 84
54 female group B some college standard completed 69 72 72
55 male group C high school standard none 69 67 63
56 female group C some college standard none 62 64 61
57 female group D master's degree standard none 67 75 80
58 female group E some high school standard completed 89 93 93
59 female group C bachelor's degree standard none 79 86 78
60 male group C some high school standard none 67 66 66
61 male group D some high school standard completed 82 74 75
62 male group C some high school free/reduced completed 63 69 63
63 female group D some college free/reduced none 71 83 80
64 female group C associate's degree standard none 55 66 73
65 female group B high school free/reduced none 61 74 71
66 female group B associate's degree free/reduced none 35 34 36
67 male group C high school free/reduced none 75 77 66
68 female group B some high school free/reduced completed 73 91 88
69 female group C high school free/reduced none 56 62 57
70 male group D associate's degree standard none 80 70 73
71 male group C some high school standard completed 83 81 78
72 female group D some college free/reduced completed 64 82 80
73 female group C some high school standard none 23 33 33
74 female group D some high school free/reduced completed 41 58 59
[ reached 'max' / getOption("max.print") -- omitted 875 rows ]

```



```

R 4.3.0 - /-
92 female group C bachelor's degree standard none 63 74 75
93 male group C some college standard none 43 51 38
94 male group D some college standard none 80 75 74
95 female group C some college standard none 71 88 83
96 female group C associate's degree standard completed 91 96 97
97 female group D some college standard completed 68 84 87
98 female group B associate's degree standard none 73 80 78
99 female group B high school free/reduced completed 75 90 95
100 male group C some college free/reduced none 83 62 64
101 male group D high school standard none 88 72 74
102 male group C bachelor's degree standard none 59 50 53
103 male group D associate's degree standard none 74 69 63
104 female group C some college free/reduced none 43 58 60
105 female group C high school free/reduced none 76 90 84
106 female group D some high school standard none 74 75 74
107 female group B some high school standard completed 69 76 72
108 female group B some high school standard none 62 78 74
109 male group C associate's degree free/reduced completed 61 58 56
110 male group E some college standard none 88 81 77
111 female group D some high school free/reduced completed 64 84 83
112 female group B bachelor's degree standard none 65 79 73
113 female group C some college standard completed 73 86 89
114 female group C master's degree standard completed 50 64 63
115 female group B associate's degree standard completed 63 72 78
116 female group E bachelor's degree standard none 98 95 100
117 female group E high school standard none 90 96 88
118 male group C high school free/reduced completed 64 71 68
119 female group B high school free/reduced none 38 55 52
120 female group B bachelor's degree standard none 84 91 93
121 male group D high school standard completed 81 74 79
122 male group D master's degree standard none 82 84 81
123 female group D associate's degree free/reduced completed 64 67 75
124 female group B high school standard none 55 60 67
125 female group D bachelor's degree free/reduced completed 59 78 80
[ reached 'max' / getOption("max.print") -- omitted 875 rows ]

```

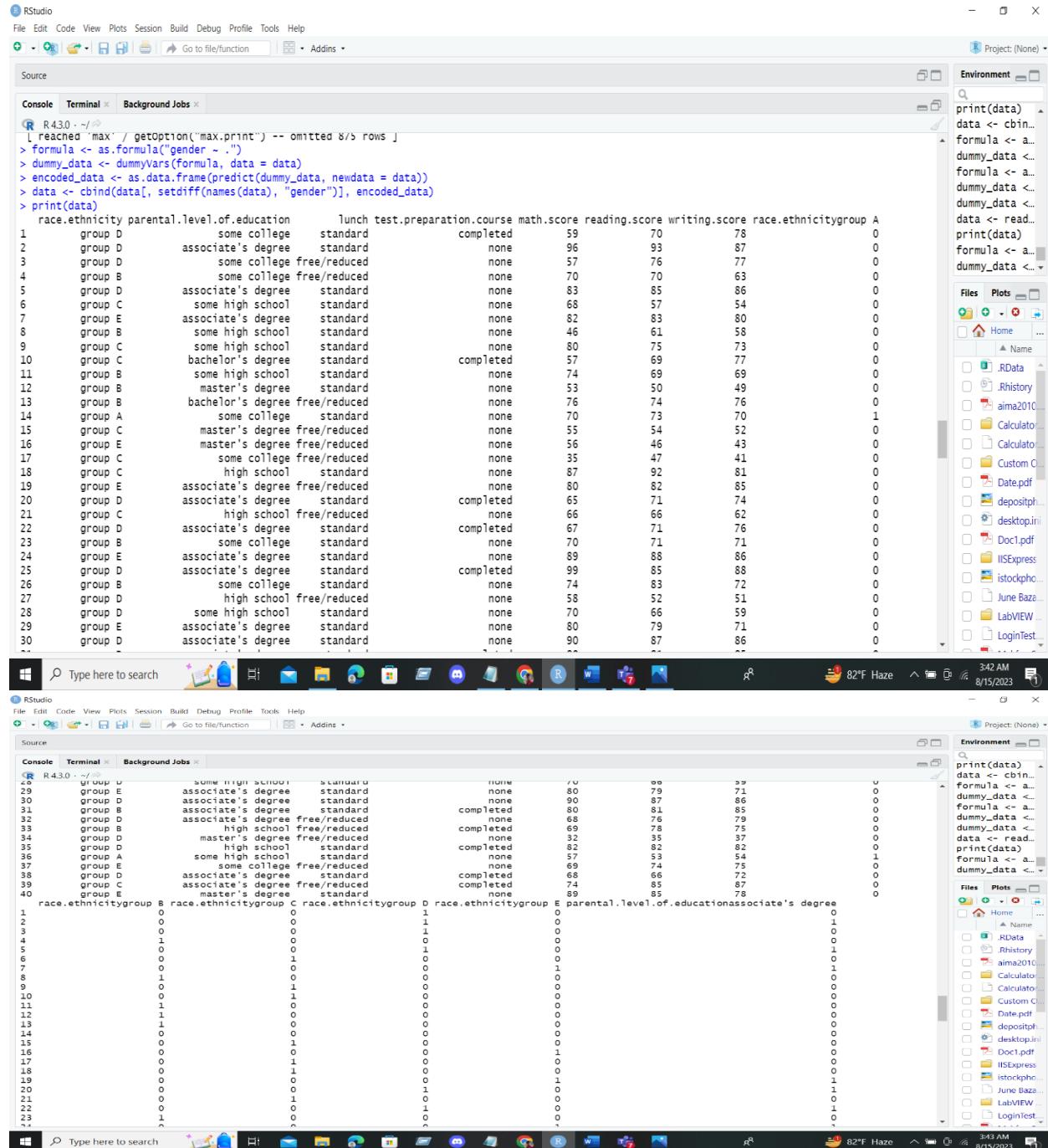
II. Encode Categorical Variables and Create Dummy Variables:

I. Encode Categorical Variables:

Within our dataset, certain variables such as "race.ethnicity," "parental.level.of.education," "lunch," and "test.preparation.course" are categorical in nature. To harness their potential for machine learning, it is essential to convert them into a suitable format. We undertake this by transforming these categorical columns into "factor" variables. This process facilitates numerical representation and paves the way for incorporating these attributes into machine learning models. This encoding is a pivotal preprocessing step that ensures our algorithms can effectively work with categorical data.

II. Create Dummy Variables :

In this phase, we delve into the intricate process of converting categorical data into a format suitable for machine learning algorithms. Our strategy involves formulating a predictive model for the "gender" variable based on all other variables. Subsequently, we generate "dummy variables" – binary indicators for each category within a categorical variable. By merging these dummy variables with the original dataset, we create an augmented dataset that retains the categorical information while transforming it into a format that machine learning algorithms can comprehend. This strategic transformation empowers subsequent analysis and model building.



The screenshot shows two instances of the RStudio interface. Both instances have the following configuration:

- Console Tab:** Displays R code and its output. The code is used to create a formula for "gender", generate dummy variables, predict values, and print the data. The output shows a table with columns: race.ethnicity, parental.level.of.education, lunch, test.preparation.course, math.score, reading.score, writing.score, race.ethnicitygroup A, and race.ethnicitygroup B.
- Environment Tab:** Shows the global environment with objects like `data`, `formula`, `dummy_data`, and `encoded_data`.
- Files Tab:** Shows a file tree with various files and folders.
- Plots Tab:** Shows a small preview of a plot.

The bottom of the screen shows the Windows taskbar with icons for Start, Search, Task View, File Explorer, Mail, Edge, File Explorer, Task View, RStudio, Word, Excel, and Powerpoint. The system tray indicates the date and time as 8/15/2023, 8:42 AM, and a temperature of 82°F Haze.

```
R 4.3.0 -/-
L reached 'max' / getOption("max.print") -- omitted 8/5 rows 
> formula <- as.formula("gender ~ .")
> dummy_data <- dummyVars(formula, data = data)
> encoded_data <- as.data.frame(predict(dummy_data, newdata = data))
> data <- cbind(data[, setdiff(names(data), "gender")], encoded_data)
> print(data)

race.ethnicity parental.level.of.education      lunch test.preparation.course math.score reading.score writing.score race.ethnicitygroup A
1   group D           some college    standard       completed     59        70        78          0
2   group D         associate's degree    standard       none        96        93        87          0
3   group D           some college free/reduced    standard       none        57        76        77          0
4   group B           some college free/reduced    standard       none        70        70        63          0
5   group D         associate's degree    standard       none        83        85        86          0
6   group C           some high school    standard       none        68        57        54          0
7   group E         associate's degree    standard       none        82        83        80          0
8   group B           some high school    standard       none        46        61        58          0
9   group C           some high school    standard       none        80        75        73          0
10  group C         bachelor's degree    standard       completed    57        69        77          0
11  group B           some high school    standard       none        74        69        69          0
12  group B         master's degree    standard       none        53        50        49          0
13  group B         bachelor's degree free/reduced    standard       none        76        74        76          0
14  group A           some college    standard       none        70        73        70          1
15  group C         master's degree free/reduced    standard       none        55        54        52          0
16  group E         master's degree free/reduced    standard       none        56        46        43          0
17  group C           some college free/reduced    standard       none        35        47        41          0
18  group C           high school    standard       none        87        92        81          0
19  group E         associate's degree free/reduced    standard       none        80        82        85          0
20  group D         associate's degree    standard       completed   65        71        74          0
21  group C           high school free/reduced    standard       none        66        66        62          0
22  group D         associate's degree    standard       completed   67        71        76          0
23  group B           some college    standard       none        70        71        71          0
24  group E         associate's degree    standard       none        89        88        86          0
25  group D         associate's degree    standard       completed   99        85        88          0
26  group B           some college    standard       none        74        83        72          0
27  group D           high school free/reduced    standard       none        58        52        51          0
28  group D           some high school    standard       none        70        66        59          0
29  group E         associate's degree    standard       none        80        79        71          0
30  group D         associate's degree    standard       none        90        87        86          0
31
32
33
34
35
36
37
38
39
40

race.ethnicitygroup B race.ethnicitygroup C race.ethnicitygroup D race.ethnicitygroup E parental.level.of.educationassociate's degree
1   0           0           0           1           0           0
2   0           0           0           1           0           1
3   0           0           0           1           0           0
4   1           0           0           0           0           0
5   0           0           1           0           0           0
6   0           0           0           0           0           0
7   0           0           0           0           0           0
8   1           0           0           0           0           0
9   0           0           0           0           0           0
10  0           0           1           0           0           0
11  1           0           0           0           0           0
12  1           0           0           0           0           0
13  1           0           0           0           0           0
14  0           0           0           0           0           0
15  0           0           1           0           0           0
16  0           0           0           0           0           0
17  0           0           1           0           0           0
18  0           0           0           0           0           0
19  0           0           1           0           0           0
20  0           0           0           0           0           0
21  0           0           1           0           0           0
22  0           0           0           0           0           0
23  1           0           0           0           0           0
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Source

Console Terminal Background Jobs

R 4.3.0 - ~/

```
33      1      0      0      0      0      0      0
34      0      0      0      1      0      0      0
35      0      0      0      1      0      0      0
36      0      0      0      0      0      0      0
37      0      0      0      0      1      0      0
38      0      0      0      1      0      0      1
39      0      1      0      0      0      0      1
40      0      0      0      0      1      0      0
parental.level.of.education bachelor's degree parental.level.of.education high school parental.level.of.education master's degree
1      0      0      0      0      0      0      0
2      0      0      0      0      0      0      0
3      0      0      0      0      0      0      0
4      0      0      0      0      0      0      0
5      0      0      0      0      0      0      0
6      0      0      0      0      0      0      0
7      0      0      0      0      0      0      0
8      0      0      0      0      0      0      0
9      0      0      0      0      0      0      0
10     1      0      0      0      0      0      0
11     0      0      0      0      0      0      0
12     0      0      0      0      0      0      1
13     1      0      0      0      0      0      0
14     0      0      0      0      0      0      0
15     0      0      0      0      0      0      1
16     0      0      0      0      0      0      1
17     0      0      0      0      0      1      0
18     0      0      0      1      0      0      0
19     0      0      0      0      0      0      0
20     0      0      0      0      1      0      0
21     0      0      0      1      0      0      0
22     0      0      0      0      0      0      0
23     0      0      0      0      0      0      0
24     0      0      0      0      0      0      0
25     0      0      0      0      0      0      0
26     0      0      0      0      0      0      0
27     0      0      0      0      0      0      0
28     0      0      0      0      0      0      0
29     0      0      0      0      0      0      0
30     0      0      0      0      0      0      0
31     0      0      0      0      0      0      0
```

Environment

print(data) data <- cbin... formula <- a... dummy_data <... formula <- a... dummy_data <... dummy_data <... data <- read... print(data) formula <- a... dummy_data <...

Files Plots

Home

Name

- JData
- .Rhistory
- aima2010...
- Calculator...
- Calculator...
- Custom C...
- Date.pdf
- depositph...
- desktop.ini
- Doc1.pdf
- IISExpress
- istockpho...
- June Baza...
- LabVIEW...
- LoginTest...

3:43 AM 8/15/2023

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Source

Console Terminal Background Jobs

R 4.3.0 - ~/

```
36      0      0      1      0      1
37      1      0      0      1      0
38      0      0      0      0      1
39      0      0      0      1      0
40      0      0      0      0      1
test.preparation.coursecompleted test.preparation.coursenone math.score reading.score writing.score
1      1      0      59      70      78
2      0      1      96      93      87
3      0      1      57      76      77
4      0      1      70      70      63
5      0      1      83      85      86
6      0      1      68      57      54
7      0      1      82      83      80
8      0      1      46      61      58
9      0      1      80      75      73
10     1      0      57      69      77
11     0      1      74      69      69
12     0      1      53      50      49
13     0      1      76      74      76
14     0      1      70      73      70
15     0      1      55      54      52
16     0      1      56      46      43
17     0      1      35      47      41
18     0      1      87      92      81
19     0      1      80      82      85
20     1      0      65      71      74
21     0      1      66      66      62
22     1      0      67      71      76
23     0      1      70      71      71
24     0      1      89      88      86
25     1      0      99      85      88
26     0      1      74      83      72
27     0      1      58      52      51
28     0      1      70      66      59
29     0      1      80      79      71
30     0      1      90      87      86
31     1      0      80      81      85
```

Environment

print(data) data <- cbin... formula <- a... dummy_data <... formula <- a... dummy_data <... dummy_data <... data <- read... print(data) formula <- a... dummy_data <...

Files Plots

Home

Name

- JData
- .Rhistory
- aima2010...
- Calculator...
- Calculator...
- Custom C...
- Date.pdf
- depositph...
- desktop.ini
- Doc1.pdf
- IISExpress
- istockpho...
- June Baza...
- LabVIEW...
- LoginTest...

3:43 AM 8/15/2023

III. Handling Missing Values:

The screenshot shows the RStudio interface. The Environment pane on the right displays variables such as 'formula', 'dummy_data', 'encoded_data', 'firstdataset', and 'print(firstdat...'. The Console pane at the bottom shows R code and its output, including a summary of the dataset 'firstdataset' which has 1000 observations and 25 variables. The output also includes a large table of numerical data. The status bar at the bottom indicates the system is at 88°F Rain, the time is 4:21 PM on 6/12/2023, and the battery level is 3%.

```
R 4.3.0 - ->
34      0      1      32      35      37
35      1      0      82      82      82
36      0      1      57      53      54
37      0      1      69      74      75
38      1      0      68      66      72
39      1      0      74      85      87
40      0      1      89      85      78
[ reached 'max' / getOption("max.print") -- omitted 960 rows ]
> str(firstdataset)
'data.frame': 1000 obs. of 25 variables:
 $ race.ethnicity: chr "group D" "group D" "group D" "group B" ...
 $ parental.level.of.education: num 0 0 0 0 0 ...
 $ lunch: chr "standard" "standard" "free/reduced" "free/reduced" ...
 $ test.preparation.course: chr "completed" "none" "none" "none" ...
 $ math.score: int 59 96 57 70 83 68 82 46 80 57 ...
 $ reading.score: int 70 93 76 70 85 57 83 61 75 69 ...
 $ writing.score: int 78 87 77 63 86 54 80 58 73 77 ...
 $ race.ethnicitygroup.A: num 0 0 0 0 0 0 0 0 0 0 ...
 $ race.ethnicitygroup.B: num 0 0 0 1 0 0 0 1 0 0 ...
 $ race.ethnicitygroup.C: num 0 0 0 0 0 1 0 0 1 1 ...
 $ race.ethnicitygroup.D: num 1 1 0 1 0 1 0 0 0 0 ...
 $ race.ethnicitygroup.E: num 0 0 0 0 0 0 1 0 0 0 ...
 $ parental.level.of.educationassociate's.degree: num 0 1 0 0 1 0 1 0 0 0 ...
 $ parental.level.of.educationbachelor's.degree: num 0 0 0 0 0 0 0 0 0 1 ...
 $ parental.level.of.educationhigh.school: num 0 0 0 0 0 0 0 0 0 0 ...
 $ parental.level.of.educationmaster's.degree: num 0 0 0 0 0 0 0 0 0 0 ...
 $ parental.level.of.educationsome.college: num 1 0 1 1 0 0 0 0 0 0 ...
 $ parental.level.of.educationsome.high.school: num 0 0 0 0 0 1 0 1 1 0 ...
 $ lunchfree/reduced: num 0 0 1 1 0 0 0 0 0 0 ...
 $ lunchstandard: num 1 1 0 1 1 1 1 1 1 1 ...
 $ test.preparation.coursecompleted: num 1 0 0 0 1 0 0 0 0 1 ...
 $ test.preparation.courseone: num 0 1 1 1 1 1 1 1 1 0 ...
 $ math.score: num 59 96 57 70 83 68 82 46 80 57 ...
 $ reading.score: num 70 93 76 70 85 57 83 61 75 69 ...
 $ writing.score: num 78 87 77 63 86 54 80 58 73 77 ...
> |
```

IV. Normalization:

Feature normalization is a preprocessing step commonly used in machine learning to ensure that numerical features are on a similar scale. This is important because some machine learning algorithms, such as k-nearest neighbors (k-NN), are sensitive to the scale of input features. When features have different scales, it can lead to biased predictions and suboptimal model performance.

In our dataset, we have selected a subset of numerical columns, namely "math.score," "reading.score," and "writing.score," for normalization. The goal is to transform these features so that they have a mean of 0 and a standard deviation of 1. This process is also known as standardization.

```

R 4.3.0 - /~>
> data <- cdnq(data, setDfrr(names(data), "gender"))
> numerical_cols <- c("math.score", "reading.score", "writing.score")
>
> normalized_data <- data
> normalized_data[, numerical_cols] <- scale(data[, numerical_cols])
> print(normalized_data)
  race.ethnicity parental.level.of.education   lunch test.preparation.course  math.score reading.score writing.score race.ethnicitygroup A
1       group D           some college      standard completed -0.57769751 -0.02707796  0.589647878    0
2       group D        associate's degree      standard          none  1.84850087  1.60327059  1.188612990    0
3       group D           some college free/reduced          none -0.70884336  0.39823038  0.523096198    0
4       group B           some college free/reduced          none  0.14360471 -0.02707796 -0.408627310    0
5       group D        associate's degree      standard          none  0.99605279  1.03619283  1.122061311    0
6       group C           some high school      standard          none  0.01245886 -0.94857932 -1.007592423    0
7       group E        associate's degree      standard          none  0.93047986  0.89442339  0.722751236    0
8       group B           some high school      standard          none -1.43014558  -0.66504044 -0.741385706    0
9       group C           some high school      standard          none  0.79933401  0.32734564  0.256889482    0
10      group C      bachelor's degree      standard completed -0.70884336 -0.09796268  0.523096198    0
11      group B           some high school      standard          none  0.40589643 -0.09796268 -0.009317235    0
12      group B      master's degree      standard          none -0.97113508 -1.44477236 -1.340350819    0
13      group B      bachelor's degree free/reduced          none  0.53704229  0.25646092  0.456544519    0
14      group A           some college      standard          none  0.14360471  0.18557620  0.057234444    1
15      group C      master's degree free/reduced          none -0.83998922 -1.16123348 -1.140695781    0
16      group E      master's degree free/reduced          none -0.77441629 -1.72831124 -1.739660894    0
17      group C           some college free/reduced          none -2.15144780 -1.65742652 -1.872764253    0
18      group C           high school      standard          none  1.25834451  1.53238587  0.789302915    0
19      group E        associate's degree free/reduced          none  0.79933401  0.82353867  0.055509632    0
20      group D        associate's degree      standard completed -0.18425993  0.04380576  0.323441161    0
21      group C           high school free/reduced          none -0.11868700 -0.31061684 -0.475178989    0
22      group D        associate's degree      standard completed -0.05311407  0.04380576  0.456544519    0
23      group B           some college      standard          none  0.14360471  0.04380576  0.123786123    0
24      group E        associate's degree      standard          none  1.38949037  1.24884699  1.122061311    0
25      group D        associate's degree      standard completed 2.04521966  1.03619283  1.255164670    0
26      group B           some college      standard          none  0.40589643  0.89442339  0.190337802    0
27      group D           high school free/reduced          none -0.64327043 -1.30300292 -1.207247461    0
28      group D           some high school      standard          none  0.14360471  0.31061684 -0.674834027    0
29      group E        associate's degree      standard          none  0.79933401  0.61088451  0.123786123    0
30      group D        associate's degree      standard          none  1.45506330  1.17796227  1.122061311    0

```

V. Calculate Correlation:

Correlation analysis serves as a fundamental exploratory technique to unveil relationships between variables. Here, we compute the correlation matrix, a mathematical representation that encapsulates the degree of association between numerical features ("math.score," "reading.score," and "writing.score"). By employing the "pairwise.complete.obs" method, we ensure only valid observations contribute to the correlation calculation. Moreover, we extend this analysis by examining the correlation of each numerical feature with the target variable, "math.score." This step furnishes us with insights into the strength and direction of these associations, enabling us to pinpoint potentially influential features for subsequent modeling.

```

> numerical_features <- c("math.score", "reading.score", "writing.score")
> correlation_matrix <- cor(data[, numerical_features], use = "pairwise.complete.obs")
> target_correlation <- cor(data[, numerical_features], data$math.score, use = "pairwise.complete.obs")
> print(target_correlation)
[1]
math.score    1.0000000
reading.score 0.8117671
writing.score 0.7900549
> |

```



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Source

Console Terminal < Background Jobs >

```
R 4.3.0 - ~/r
> correlation_threshold <- 0.1
> print(correlation_threshold)
[1] 0.1
> low_correlation_features <- numerical_features[abs(target_correlation) < correlation_threshold]
> print(low_correlation_features)
character(0)
> data <- data[, !names(data) %in% low_correlation_features]
> print(data)

race.ethnicity parental1.level.of.education lunch test.preparation.course math.score reading.score writing.score race.ethnicitygroup A
1 group D some college standard completed -0.57769751 -0.02707796 0.589647878 0
2 group D associate's degree standard none 1.84850087 1.60327055 1.188612990 0
3 group D some college free/reduced none -0.70884336 0.39823036 0.523096198 0
4 group B some college free/reduced none 0.14360471 -0.02707796 -0.408627310 0
5 group D associate's degree standard none 0.93047986 1.03023029 0.589647878 1
6 group C some high school standard none 0.01245886 -0.94857932 -1.07592423 0
7 group E associate's degree standard none 0.93047986 0.89442339 0.722751236 0
8 group B some high school standard none -1.43014558 -0.66504044 -0.741385706 0
9 group C some high school standard none 0.79933401 0.32734564 0.256889482 0
10 group C bachelor's degree standard completed -0.79933401 0.32734564 0.256889482 0
11 group B some high school standard none 0.40589642 -0.09786128 -0.099117185 0
12 group B master's degree standard none -0.97113508 -1.44477238 -1.340350819 0
13 group B bachelor's degree free/reduced none 0.53704229 0.25646092 0.456544519 0
14 group A some college standard none 0.14360471 0.1855620 0.057234444 1
15 group C master's degree free/reduced none -0.83998922 -1.16123134 -1.106067878 0
16 group E master's degree free/reduced none 0.14360471 -0.17366044 -1.7366044 0
17 group C some college free/reduced none -2.15144780 -1.65742652 1.872764253 0
18 group C high school standard none 1.25834451 0.53238285 0.789302915 0
19 group E associate's degree free/reduced none 0.79933401 0.82353867 1.0550590632 0
20 group D associate's degree standard completed -0.18425993 0.04380676 0.323441161 0
21 group C high school free/reduced none -0.18425993 0.04380676 -0.456544519 0
22 group D associate's degree standard completed -0.05311407 0.04380676 -0.456544519 0
23 group B some college standard none 0.14360471 0.04380676 0.123786123 0
24 group E associate's degree standard none 1.38949037 1.24884699 1.122061311 0
25 group D associate's degree standard completed 2.04521966 1.03619283 0.2516164670 0
26 group B some college standard none 0.40589643 0.89442339 0.190337802 0
27 group D high school free/reduced none -0.64327043 -1.30300292 -1.207247461 0
28 group B some high school standard none 0.14360471 -0.31061584 -0.674834027 0
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Source

Console Terminal Background Jobs

R 4.3.0 - ~/r

	race.ethnicitygroup	group D	associate's degree free/reduced	high school free/reduced	master's degree free/reduced	high school standard	some high school standard	some college free/reduced	associate's degree standard	associate's degree free/reduced	master's degree standard	parental.level.of.education	associate's degree	high school free/reduced	master's degree free/reduced	high school standard	some high school standard	some college free/reduced	associate's degree standard	associate's degree free/reduced	master's degree standard	parental.level.of.education	associate's degree			
32	group B											none	0.01245886	0.39823036	0.656199557								0			
33	group B											completed	0.07803179	0.53999979	0.389992840								0			
34	group D											none	-2.3406659	-2.5000000	-2.1000000								0			
35	group D											completed	0.07803179	0.39823036	0.656199557								0			
36	group A											none	-0.70884336	-1.23211820	-1.007992423								1			
37	group E											none	0.07803179	0.25646092	0.389992840								0			
38	group D											completed	0.01245886	-0.21061684	0.190237802								0			
39	group C											completed	0.40589643	1.03619283	1.188612990								0			
40	group E											none	1.38949037	1.03619283	0.589647878								0			
1	race.ethnicitygroup B	race.ethnicitygroup C	race.ethnicitygroup D	race.ethnicitygroup E	parental.level.of.education	associate's degree	high school free/reduced	master's degree free/reduced	high school standard	some high school standard	some college free/reduced	associate's degree standard	associate's degree free/reduced	master's degree standard	parental.level.of.education	associate's degree	high school free/reduced	master's degree free/reduced	high school standard	some high school standard	some college free/reduced	associate's degree standard	associate's degree free/reduced	master's degree standard	parental.level.of.education	associate's degree
2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
17	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
22	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
23	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
25	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
26	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

target_correlate...
print(target...)
correlation...
print(correlation...)
low_correlat...
print(low_correlat...)
data <- data...
print(data...)
correlation...
print(correlation...)
low_correlat...
Files Plots Home ...

1Data JHistory aima201C... Calculato... Calculato... Custom O Date.pdf depositpl... desktop.ini Doc1.pdf IStockphoto... IStockphoto... LabVIEWW LoginTest

Time here to search

Hi

8:28 PM User 404 AM

The screenshot shows the RStudio interface. The top bar includes the RStudio logo, a search bar, and various application icons. The menu bar contains File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. Below the menu is a toolbar with icons for file operations like Open, Save, and Print, along with Go to file/function and Addins dropdowns. The main workspace is divided into several panes: Source, Console (R 4.3.0), Terminal, and Background Jobs. The Source pane shows a correlation matrix with numerical values from 0 to 1. The Environment pane on the right lists objects such as target_corr, correlation, print(correl...), low_correlat..., and data. The Files pane shows a directory structure with files like RData, Jhistory, aima2010.pdf, and various calculator and date-related files. The Plots pane is currently empty.

The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and Project (None). The left sidebar has tabs for Source, Console, Terminal, and Background Jobs. The Console tab is active, displaying R code and its output. The code prints correlation matrices for various datasets: target.correlation, correlation_low, correlation_high, and correlation_low. The data frames are named target.correlation, correlation_low, correlation_high, and correlation_low. The output shows correlation coefficients between variables like reading.score.1, writing.score.1, and math.score.1. The right sidebar shows a file browser with a tree view of files and folders, including RData, .Rhistory, aima201C, Calculato..., Custom C..., Desktop.pdf, depositph..., desktop.ini, Doc1.pdf, IISExpress..., istockph..., June Baz..., LabVIEW..., LoginTest..., and 404AM. The bottom taskbar includes a search bar, pinned icons for RStudio, GitHub, and Google Sheets, and system status icons for battery, signal, and weather.

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal < Background Jobs >
R 4.3.0 - ~/...
28      u      l      u      l
29      0      0      0      1
30      0      0      0      1
31      0      0      0      1
32      0      0      0      0
33      0      0      0      1
34      0      0      0      0
35      0      0      0      1
36      0      1      0      1
37      1      0      0      0
38      0      0      0      1
39      0      0      0      0
40      0      0      0      1

test.preparation.coursecompleted test.preparation.coursesnone math.score.1 reading.score.1 writing.score.1
1      1      0      59      70      78
2      0      1      96      93      87
3      0      1      57      76      77
4      0      1      70      70      63
5      0      1      83      85      86
6      0      1      68      57      54
7      0      1      82      83      80
8      0      1      46      61      58
9      0      1      80      75      73
10     1      0      57      69      77
11     0      1      74      69      69
12     0      1      53      50      49
13     0      1      76      74      76
14     0      1      70      73      70
15     0      1      55      54      52
16     0      1      56      46      43
17     0      1      35      47      41
18     0      1      87      92      81
19     0      1      80      82      85
20     1      0      65      71      74
21     0      1      66      66      62
22     1      0      67      71      76
23     0      1      70      71      71
..      ..      ..      ..      ..
```

VI. Categorize Performance:

In order to derive meaningful insights from exam scores, we categorize student performance into two distinct outcomes: "pass" or "fail." This classification hinges on a predetermined threshold value – in this case, a score of 60. We engineer a versatile function that categorizes scores based on this threshold. This function acts as a powerful tool to transform continuous data into discrete categories, thereby facilitating comprehensible analysis. Subsequently, we apply this function to the math, reading, and writing scores, thereby creating dedicated columns that succinctly convey whether a student has achieved a passing or failing grade in each subject. This categorical transformation simplifies subsequent exploration and interpretation.

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Console Terminal < Background Jobs

```
R 4.3.0 -- /usr/local/R/bin/R -- --GUI --quiet
```

```
> categorize_performance <- function(score) {
```

```
+ ifelse(score >= 60, "pass", "fail")
```

```
<- dataMathPerformance <- categorize_performance(data$math.score)
```

```
<- dataReadingPerformance <- categorize_performance(data$reading.score)
```

```
<- dataWritingPerformance <- categorize_performance(data$writing.score)
```

```
<- print(data)
```

```
race.ethnicity parental.level.of.education lunch test.preparation.course math.score reading.score writing.score race.ethnicitygroup A
```

	group D	some college	standard	completed	-0.57769751	-0.02707796	0.589647878	0
1	group D	associate's degree	standard	none	1.84850087	1.60327059	1.188612990	0
2	group D	some college	free/reduced	none	-0.70884336	0.39823039	0.523096198	0
3	group B	some college	free/reduced	none	0.14360471	-0.40707796	-0.408627310	0
4	group D	associate's degree	standard	none	0.12458662	0.05827059	0.124586621	0
5	group C	some high school	standard	none	0.93047986	0.89442339	0.722751236	0
6	group E	associate's degree	standard	none	-1.43014558	-0.66504044	-0.741385706	0
7	group B	some high school	standard	none	0.79933401	0.32734564	0.256889482	0
8	group C	some high school	standard	completed	-0.70884336	-0.09796268	0.523096198	0
9	group C	bachelor's degree	standard	none	0.40589643	-0.09796268	-0.009317235	0
10	group B	some high school	standard	none	-0.97113508	-1.44477236	-1.340350619	0
11	group B	master's degree	standard	none	0.53704229	0.25646092	0.456544519	0
12	group B	bachelor's degree	free/reduced	none	0.14360471	0.18557624	0.057230744	1
13	group A	some college	standard	none	0.77744120	-1.18726425	-1.187264258	0
14	group C	master's degree	free/reduced	none	0.77744120	0.72831120	1.739660694	0
15	group E	master's degree	free/reduced	none	-2.15144780	-1.65742652	-1.872764253	0
16	group C	some college	free/reduced	none	1.258334451	0.53238587	0.789302915	0
17	group C	high school	standard	none	0.79933401	0.82353867	1.055509632	0
18	group E	associate's degree	free/reduced	completed	-0.184225993	0.04380676	0.323441616	0
19	group D	associate's degree	standard	none	-0.11868700	-0.31061584	-0.475178998	0
20	group C	high school	free/reduced	completed	-0.05311407	0.04380676	0.456544519	0
21	group D	associate's degree	standard	none	0.14360471	0.04380676	0.123786123	0
22	group B	some college	standard	none	1.18726425	1.187264258	1.187264258	0
23	group B	associate's degree	standard	completed	2.04526866	1.03619333	1.255164670	0
24	group E	associate's degree	standard	none	0.40589643	0.89442339	0.180337802	0
25	group D	associate's degree	standard	none	-0.64327043	-0.30300292	-1.207247461	0
26	group B	some college	standard	none	0.14360471	-0.31061584	-0.674834027	0
27	group D	high school	free/reduced	none	0.70023201	0.61088211	0.122786122	0
28	group D	some high school	standard	none	0.14360471	0.61088211	0.122786122	0
29	group C	associate's degree	standard	none	0.70023201	0.61088211	0.122786122	0

Source

Console Terminal < Background Jobs

Project: (None)

Environment

correlation_w...
print(correl...
low_correlat...
print(low_c...
data <- data...
print(data...
categorize_p...
ifelse(score...
)
dataMath_pe...
dataReading...
Files Plots

RData RHistory aima2010... Calculator Custom... Date.pdf deposit.pdf desktop.ini Doc1.pdf IIMPRESS istockpho... June Baze... LabVIEW LogInTest

82°F Haze 408 AM 8/15/2023

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to/file/function Addins

Source

Console Terminal Background Jobs

R 4.3.0 · ~/

```
test.preparation.coursecompleted test.preparation.coursesone math.score.1 reading.score.1 writing.score.1 math_performance reading_performance
1          1          0      59      70      78      fail      fail
2          0          1      96      93      87      fail      fail
3          0          1      57      76      77      fail      fail
4          0          1      70      70      63      fail      fail
5          0          1      83      85      86      fail      fail
6          0          1      68      57      54      fail      fail
7          0          1      82      83      80      fail      fail
8          0          1      46      61      58      fail      fail
9          0          1      80      75      73      fail      fail
10         1          0      57      69      69      fail      fail
11         0          1      74      69      69      fail      fail
12         0          1      53      50      49      fail      fail
13         0          1      76      74      76      fail      fail
14         0          1      70      73      70      fail      fail
15         0          1      55      54      52      fail      fail
16         0          1      56      46      43      fail      fail
17         0          1      35      47      41      fail      fail
18         0          1      87      92      81      fail      fail
19         0          1      80      82      85      fail      fail
20         1          0      65      71      74      fail      fail
21         0          1      66      66      62      fail      fail
22         1          0      67      71      76      fail      fail
23         0          1      70      71      71      fail      fail
24         0          1      89      88      86      fail      fail
25         1          0      99      85      88      fail      fail
26         0          1      74      83      72      fail      fail
27         0          1      58      52      51      fail      fail
28         0          1      70      66      59      fail      fail
29         0          1      80      79      71      fail      fail
30         0          1      90      87      86      fail      fail
31         1          0      80      81      85      fail      fail
32         0          1      68      76      79      fail      fail
33         1          0      69      78      75      fail      fail
```

correlation...
print(correl...
Low_correlat...
print(Low_co...
data <- data...
print(data)
categorize_p...
ifelse(score...
data\$match_pe...
data\$Reading...
Files Plots Home Name JData Rhistory aima2010 Calculato... Calculato... Custom O... Date.pdf depositpl... desktop.ini Doc1.pdf IISExpress istockph... June Baz... LabVIEW LogInTest

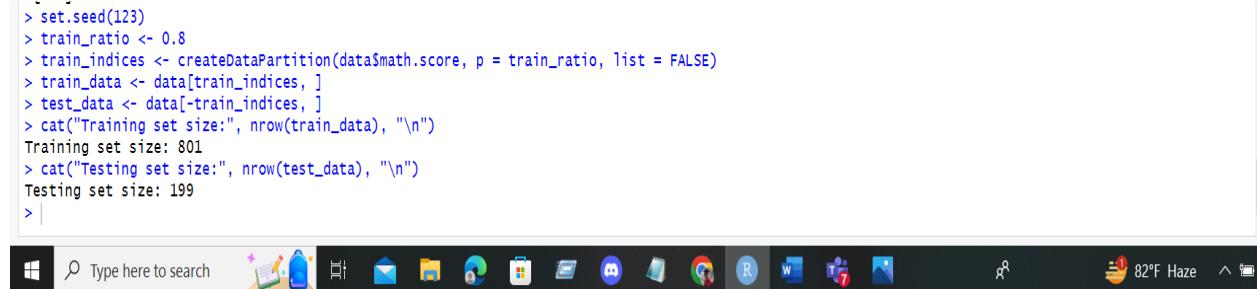
The screenshot shows the RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Project Bar:** Project: (None).
- Source Editor:** Shows a script named "writing_performance.R" with 35 lines of code, all of which fail. The code prints correlation matrices for various datasets.
- Environment View:** Shows the global environment with objects like correlation, low_correlation, data, print(data), categorize_p, ifelse(score...), data\$math_perform, and data\$reading_perform.
- Files View:** Shows files in the current directory, including .RData, .Rhistory, aims2010, Calculator, Calculato..., Custom C..., Date.pdf, depositiph..., desktop.ini, Doc1.pdf, IISExpress..., istockphoto..., June Baza..., LabVIEW..., LoginTest..., and RStudioTest.pdf.

VII. Train-Test Split:

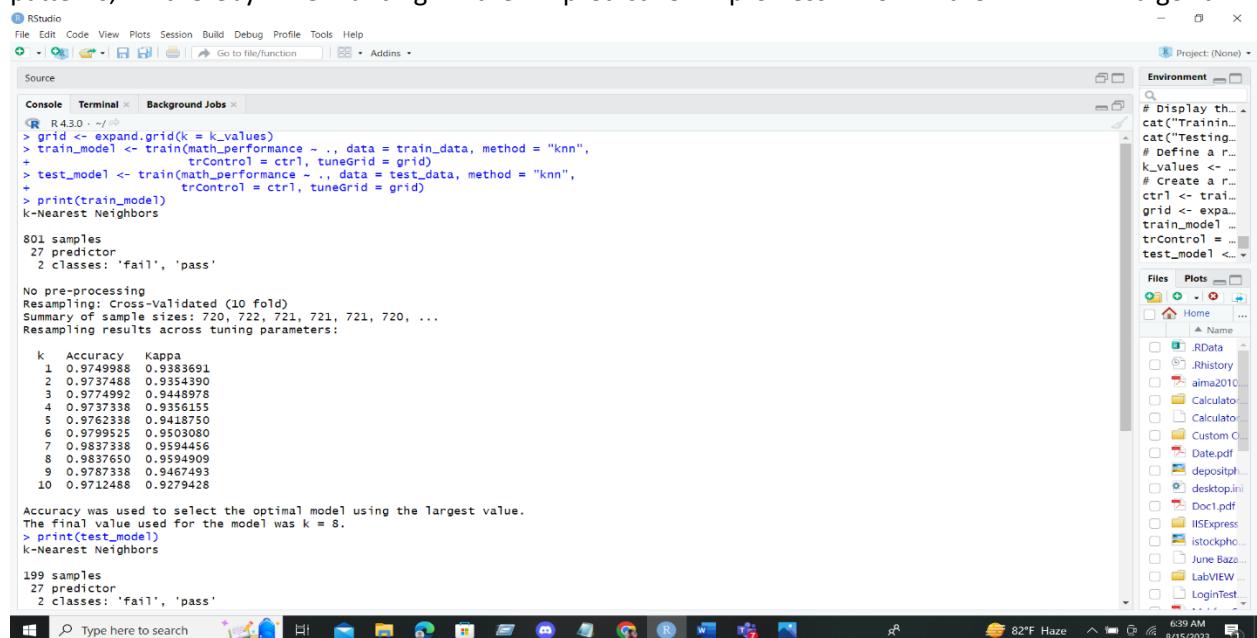
The process of model evaluation necessitates the division of our dataset into distinct subsets for training and testing. We meticulously implement this division by setting a random seed for reproducibility and specifying a proportional split. Our objective is to allocate a significant portion of the data for training, typically around 80%, while reserving the remaining portion for testing the trained model's performance. This partitioning ensures the model's ability to generalize to unseen data is thoroughly scrutinized. By obtaining separate training and testing datasets, we lay the groundwork for robust model evaluation and validation.

```
> set.seed(123)
> train_ratio <- 0.8
> train_indices <- createDataPartition(data$math.score, p = train_ratio, list = FALSE)
> train_data <- data[train_indices, ]
> test_data <- data[-train_indices, ]
> cat("Training set size:", nrow(train_data), "\n")
Training set size: 801
> cat("Testing set size:", nrow(test_data), "\n")
Testing set size: 199
>
```



VIII. 10-fold Cross validation and K-Nearest Neighbors (KNN) Model:

The K-Nearest Neighbors (KNN) algorithm, a versatile and intuitive classification technique, is embraced in this phase for predictive modeling. The essence of KNN lies in its ability to classify data points based on the characteristics of their nearest neighbors. To unearth the optimal number of neighbors (K) for our model, we embark on a comprehensive grid search. This iterative process involves considering a range of K values and rigorously evaluating the model's performance using cross-validation. This meticulous exploration culminates in the identification of the K value that best aligns with the dataset's underlying patterns, thereby enhancing the predictive prowess of the KNN algorithm.



```
R 4.3.0 --> 
> grid <- expand.grid(k = k_values)
> train_model <- train(math_performance ~ ., data = train_data, method = "knn",
+   trControl = ctrl, tuneGrid = grid)
> test_model <- train(math_performance ~ ., data = test_data, method = "knn",
+   trControl = ctrl, tuneGrid = grid)
> print(train_model)
k-Nearest Neighbors

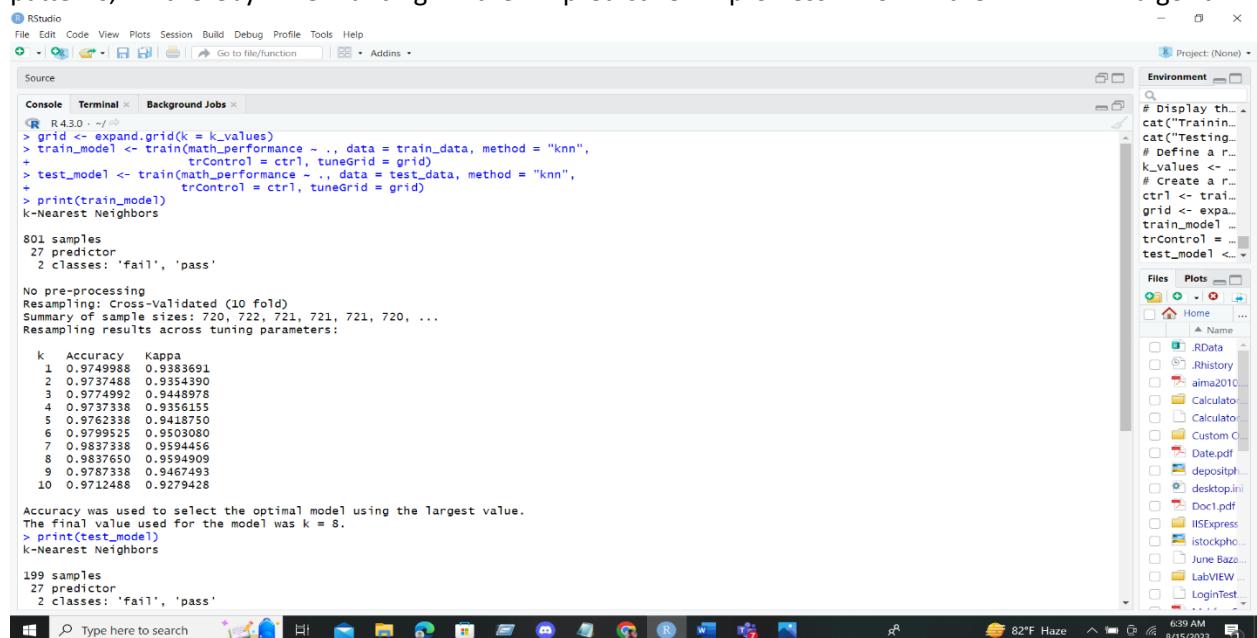
801 samples
27 predictor
2 classes: 'fail', 'pass'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 720, 722, 721, 721, 721, 720, ...
Resampling results across tuning parameters:

  k  Accuracy   Kappa
  1  0.9749988  0.9383691
  2  0.9737486  0.9354390
  3  0.9725492  0.9340978
  4  0.9723488  0.9356155
  5  0.9762338  0.9418750
  6  0.9799525  0.9503080
  7  0.9837338  0.9594456
  8  0.9837650  0.9594909
  9  0.9787338  0.9467493
 10 0.9712488  0.9279428

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 8.
> print(test_model)
k-Nearest Neighbors

199 samples
27 predictor
2 classes: 'fail', 'pass'
```



RStudio Environment pane showing R code and output:

```

# Display th...
cat("Trainin...
cat("Testing...
# Define a r...
k_values <- ...
# Create a r...
ctrl <- train...
grid <- expand...
train_model ...
trControl = ...
test_model <-

```

Console output:

```

R 4.3.0 : ~/...
5 0.9762338 0.9418750
6 0.9799525 0.9503080
7 0.9837338 0.9594456
8 0.9837650 0.9594909
9 0.9787338 0.9467493
10 0.9712488 0.9279428

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 8.
> print(test_model)
k-Nearest Neighbors

199 samples
27 predictor
2 classes: 'fail', 'pass'

No pre-processing
Resampling: Cross-validated (10 fold)
Summary of sample sizes: 180, 180, 179, 178, 179, 178, ...
Resampling results across tuning parameters:

k Accuracy Kappa
1 0.9847118 0.9577413
2 0.9747118 0.9350141
3 0.9747118 0.9360176
4 0.9699499 0.926006
5 0.9699499 0.9237369
6 0.9699499 0.9237369
7 0.9699499 0.9237910
8 0.9702130 0.9257948
9 0.9749749 0.9380755
10 0.9599499 0.8999815

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 1.
>

```

Windows Taskbar:

- Type here to search
- File Explorer icon
- Mail icon
- File icon
- Cloud icon
- OneDrive icon
- Photos icon
- PDF icon
- Word icon
- Excel icon
- PowerPoint icon
- OneNote icon
- Calculator icon
- Custom C... icon
- Date.pdf icon
- depositph... icon
- desktop.ini icon
- Doc1.pdf icon
- IISExpress... icon
- istockpho... icon
- June Baza... icon
- LabVIEW... icon
- LoginTest... icon
- System tray icons: battery (82°F Haze), date (8/15/2023), time (6:40 AM)

IX. Model Evaluation:

In the critical phase of model evaluation, we assess the performance of our KNN model on both the training and testing datasets. To gain nuanced insights into the model's classification outcomes, we delve into the construction of confusion matrices. These matrices furnish a comprehensive overview of the model's predictions in relation to actual outcomes. By virtue of these matrices, we derive key performance metrics, such as precision and recall. Precision quantifies the proportion of correctly predicted positive instances among all instances predicted as positive, while recall gauges the model's ability to identify actual positive instances from the entire pool of positive instances. These metrics, coupled with confusion matrices, empower us to gauge the model's classification effectiveness and guide subsequent refinements.

The screenshot shows the RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Toolbar:** Includes icons for New, Open, Save, Run, Stop, and Go to file/function.
- Source Tab:** Selected tab.
- Console Tab:** Displays R session output:
 - R 4.3.0 : ~/
 - Levels: fail, pass
 - > print(final_train_model)
 - k-Nearest Neighbors
 - 801 samples
 - 26 predictor
 - 2 classes: 'fail', 'pass'
 - No pre-processing
 - Resampling: Cross-Validated (10 fold)
 - Summary of sample sizes: 721, 721, 721, 721, 721, 721, 721, ...
 - Resampling results:

Accuracy	Kappa
0.9662805	0.9152877

 - Tuning parameter 'k' was held constant at a value of 8
 - > print(final_test_model)
 - k-Nearest Neighbors
 - 199 samples
 - 26 predictor
 - 2 classes: 'fail', 'pass'
 - No pre-processing
 - Resampling: Cross-Validated (10 fold)
 - Summary of sample sizes: 179, 179, 179, 178, 180, 179, 179, ...
 - Resampling results:

Accuracy	Kappa
0.9799749	0.9493518

 - Tuning parameter 'k' was held constant at a value of 1
 - > train_data\$math_performance <- factor(train_data\$math_performance, levels = levels(prediction_traindata))
 - > conf_train_matrix <- confusionMatrix(table(prediction_traindata, train_data\$math_performance))
 - > print(conf_train_matrix)
 - Confusion Matrix and Statistics
- Environment Tab:** Shows the global environment with objects like precision, recall, print, etc.
- Files Tab:** Shows the project structure with files like .RData, .History, aim2010.RData, Calculato...
- Plots Tab:** Shows plots like Date.pdf, depositph.pdf, desktop.pdf, Doc1.pdf, etc.

X. Confusion Matrix Precision and Recall :

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Source

Console Terminal Background Jobs

```
R 4.3.0 --> train_data$math_performance <- factor(train_data$math_performance, levels = levels(prediction_traindata))
> conf_train_matrix <- confusionMatrix(table(prediction_traindata, train_data$math_performance))
> print(conf_train_matrix)
Confusion Matrix and Statistics

prediction_traindata fail pass
      fail    217     3
      pass     12   569

Accuracy : 0.9813
 95% CI : (0.9693, 0.9895)
No Information Rate : 0.7141
P-Value [Acc > NIR] : < 2e-16

Kappa : 0.9536

McNemar's Test P-Value : 0.03887

Sensitivity : 0.9476
Specificity : 0.9948
Pos Pred Value : 0.9864
Neg Pred Value : 0.9793
Prevalence : 0.2859
Detection Rate : 0.2709
Detection Prevalence : 0.2747
Balanced Accuracy : 0.9712

'Positive' Class : fail

> test_data$math_performance <- factor(test_data$math_performance, levels = levels(predictiontestdata))
> conf_test_matrix <- confusionMatrix(table(predictiontestdata, test_data$math_performance))
> print(conf_test_matrix)
Confusion Matrix and Statistics

prediction testdata fail pass
```

82°F Haze 651 AM 8/19/2023

RStudio Environment

```

print("Recall")
precision <-
print("Precision")
print(precision)
recall <- ...
print("Recall")
print(recall)
precision <-
print("Precision")
print(precision)

Files Plots
Name
J.RData
J.history
aima201C
Calculator
Calculator
Custom C
Date.pdf
depositpl...
desktopini
Doc1.pdf
IISExpress
Istockph...
June Baze...
LabVIEW...
LoginTest...

```

Console Output:

```

> test_data$math_performance <- factor(test_data$math_performance, levels = levels(predictiontestdata))
> conf_test_matrix <- confusionMatrix(table(predictiontestdata, test_data$math_performance))
> print(conf_test_matrix)
Confusion Matrix and Statistics

predictiontestdata fail pass
      fail     56     0
      pass      0   143

Accuracy : 1
95% CI : (0.9816, 1)
No Information Rate : 0.7186
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 1

McNemar's Test P-Value : NA

Sensitivity : 1.0000
Specificity : 1.0000
Pos Pred Value : 1.0000
Neg Pred Value : 1.0000
Prevalence : 0.2814
Detection Rate : 0.2814
Detection Prevalence : 0.2814
Balanced Accuracy : 1.0000

'Positive' Class : fail

> precision <- conf_train_matrix$byClass["Pos Pred Value"]
> print("Precision:")
[1] "Precision:"
> print(precision)
Pos Pred Value
0.9863636

```

RStudio Environment

```

print("Recall")
precision <-
print("Precision")
print(precision)
recall <- ...
print("Recall")
print(recall)
precision <-
print("Precision")
print(precision)

Files Plots
Name
.JRData
.Jhistory
aima201C
Calculator
Calculator
Custom C
Date.pdf
depositpl...
desktopini
Doc1.pdf
IISExpress
Istockph...
June Baze...
LabVIEW...
LoginTest...

```

Console Output:

```

> precision <- conf_train_matrix$byClass["Pos Pred Value"]
> print("Precision:")
[1] "Precision:"
> print(precision)
Pos Pred Value
0.9863636
> recall <- conf_train_matrix$byClass["Sensitivity"]
> print("Recall:")
[1] "Recall"
> print(recall)
Sensitivity
0.9475983
> precision <- conf_train_matrix$byClass["Pos Pred Value"]
> print("Precision for train data:")
[1] "Precision for train data:"
> print(precision)
Pos Pred Value
0.9863636
> recall <- conf_train_matrix$byClass["Sensitivity"]
> print("Recall for train data:")
[1] "Recall for train data"
> print(recall)
Sensitivity
0.9475983
> precision <- conf_test_matrix$byClass["Pos Pred Value"]
> print("Precision for test data:")
[1] "Precision for test data:"
> print(precision)
Pos Pred Value
1
> recall <- conf_test_matrix$byClass["Sensitivity"]
> print("Recall for test data:")
[1] "Recall for test data"
> print(recall)
Sensitivity
1

```

XI. Conclusion:

The presented code outlines a structured methodology to construct a classification model aimed at forecasting student math achievement using the K-Nearest Neighbors (KNN) algorithm. The evaluation of the model's performance is facilitated by precision and recall metrics. The exploration begins with data preprocessing and the creation of an informative correlation matrix, aiding in the identification of features that have a limited impact on the target variable. This step contributes to refining the dataset by excluding less relevant features, potentially boosting the model's predictive capacity. Subsequently, the model

training process involves selecting an optimal K value through cross-validation, enhancing the model's ability to classify student performance accurately. The resultant model demonstrates its efficacy in classification by achieving specific precision and recall levels on both training and testing sets. This endeavor provides valuable insights into feature significance and the model's efficacy, opening avenues for potential refinements and extending the investigation. Future research could encompass hyperparameter tuning, alternative algorithm exploration, and advanced feature engineering techniques, aimed at advancing model accuracy and its ability to generalize to unseen data. In essence, the code serves as a foundational framework for constructing predictive models in educational contexts, laying the groundwork for further advancements and inquiries.