AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

FINAL-TERM PROJECT

Course: INTRODUCTION TO DATA SCIENCE

Sec: A

SUBMITTED TO

Faculty: Tohedul Islam

SUBMITTED BY

Group: 15

NAME	ID
1. Abdullah Muhammad Hamja	20 -43465-1
2. Sumaiya Ahmed Susmita	21-45266-2

DATA SCIENCE FINAL PROJECT
Student Mental Health

Dataset Description:

The 'Student Mental Health' dataset is a comprehensive collection aimed at exploring factors related to students' mental health. It includes key variables such as age, gender, CGPA (Cumulative Grade Point Average), course enrolment, and indicators for depression, anxiety, panic attacks, and seeking help. This dataset provides insights into demographic distributions, academic performance correlations with mental health, the prevalence of mental health conditions, and patterns of help-seeking behaviour. Potential use cases involve identifying risk factors, developing predictive models, and informing targeted interventions. Ethical considerations emphasize responsible data handling due to the sensitivity of mental health information. Overall, the dataset is a valuable resource for researchers, educators, and policymakers interested in addressing mental health challenges in the student population.

CODES – CONSOLE – DETAILS:

1. Import CSV file.

   Code:

   ```
   dataset <- read.csv("D:/FALL2023/IntroToDataScience/Student_Mental_health.csv", header = TRUE, sep = ",")
   print(dataset)
   ```

```
2:15   (Top Level) ÷                                                                                R Script ÷
Console   Terminal ×   Background Jobs ×
R 4.3.1 · ~/
   Choose.your.gender Age    What.is.your.course. Your.current.year.of.Study What.is.your.CGPA. Marital.status Do.you.have.Depression.
1              Female  18             Engineering                     year 1        3.00 - 3.49             No                     Yes
2                Male  21       Islamic education                     year 2        3.00 - 3.49             No                      No
3                Male  19                     BIT                     Year 1        3.00 - 3.49             No                     Yes
4              Female  22                    Laws                     year 3        3.00 - 3.49            Yes                     Yes
5                Male  23             Mathemathics                    year 4        3.00 - 3.49             No                      No
6                Male  19             Engineering                     Year 2        3.50 - 4.00             No                      No
7              Female  23        Pendidikan islam                     year 2       3.50 - 4.00             Yes                     Yes
8              Female  18                     BCS                     year 1        3.50 - 4.00             No                      No
9              Female  19         Human Resources                     Year 2        2.50 - 2.99             No                      No
10               Male  18                   Irkhs                     year 1        3.50 - 4.00             No                      No
11             Female  20              Psychology                     year 1        3.50 - 4.00             No                      No
12             Female  24             Engineering                     Year 3        3.50 - 4.00            Yes                     Yes
13             Female  18                     BCS                     year 1        3.00 - 3.49             No                     Yes
14               Male  19             Engineering                     year 1        3.00 - 3.49             No                      No
15             Female  18                   KENMS                     Year 2        3.50 - 4.00             No                      No
16               Male  24                     BCS                     Year 3        3.50 - 4.00             No                      No
17             Female  24             Accounting                      year 3        3.00 - 3.49             No                      No
18             Female  24                     ENM                     year 4        3.00 - 3.49            Yes                     Yes
19             Female  20                     BIT                     Year 2        3.50 - 4.00             No                      No
20             Female  18          Marine science                     year 2        3.50 - 4.00            Yes                     Yes
21             Female  19             Engineering                     year 1        3.00 - 3.49             No                      No
22             Female  18                     KOE                     Year 2        3.00 - 3.49             No                      No
23             Female  24                     BCS                     year 1        3.50 - 4.00             No                      No
24             Female  24             Engineering                     year 1        3.00 - 3.49             No                      No
25             Female  23                     BCS                     Year 3        3.50 - 4.00             No                     Yes
26             Female  18         Banking Studies                     year 1        3.50 - 4.00             No                      No
27             Female  19             Engineering                     year 1        3.50 - 4.00             No                      No
```

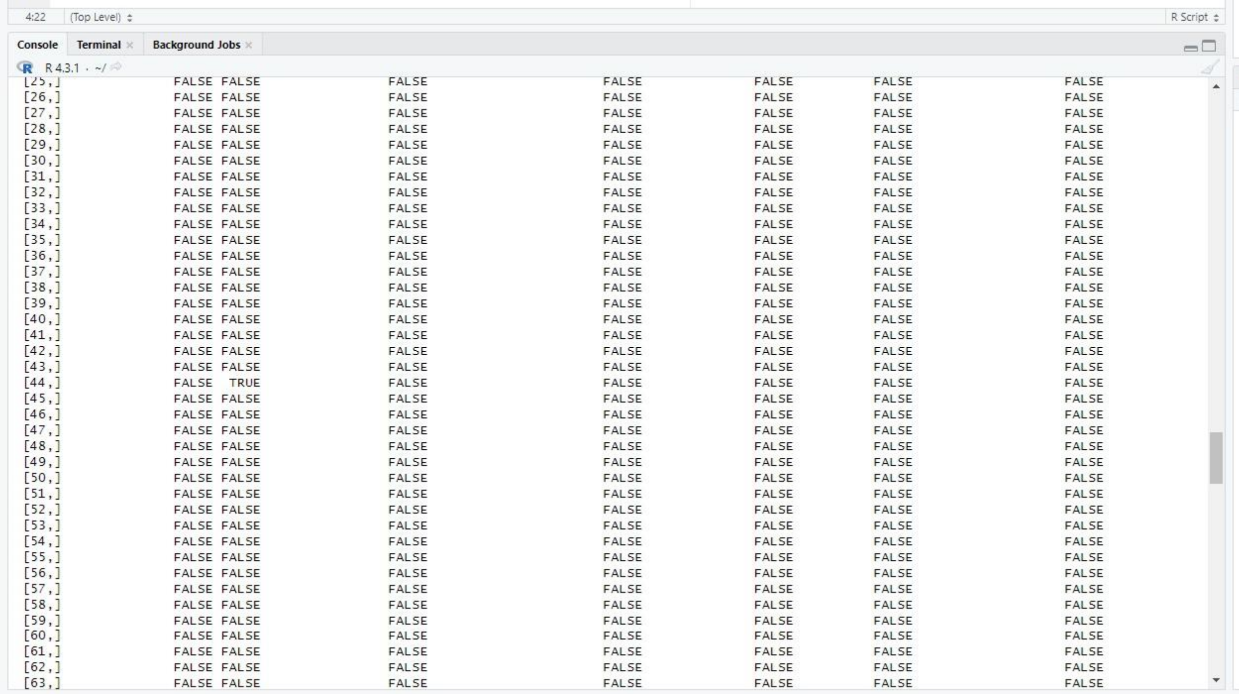
This code is used to import external Excel files (in CSV format) into R.

2. <u>Find missing values.</u>

   Code:

   missing_values <- is.na(dataset)

   print(missing_values)



```
4:22   (Top Level)                                                R Script
Console   Terminal   Background Jobs
R 4.3.1 · ~/
[25,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[26,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[27,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[28,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[29,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[30,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[31,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[32,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[33,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[34,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[35,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[36,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[37,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[38,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[39,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[40,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[41,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[42,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[43,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[44,]   FALSE  TRUE   FALSE   FALSE   FALSE   FALSE   FALSE
[45,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[46,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[47,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[48,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[49,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[50,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[51,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[52,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[53,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[54,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[55,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[56,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[57,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[58,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[59,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[60,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[61,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[62,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
[63,]   FALSE FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
```

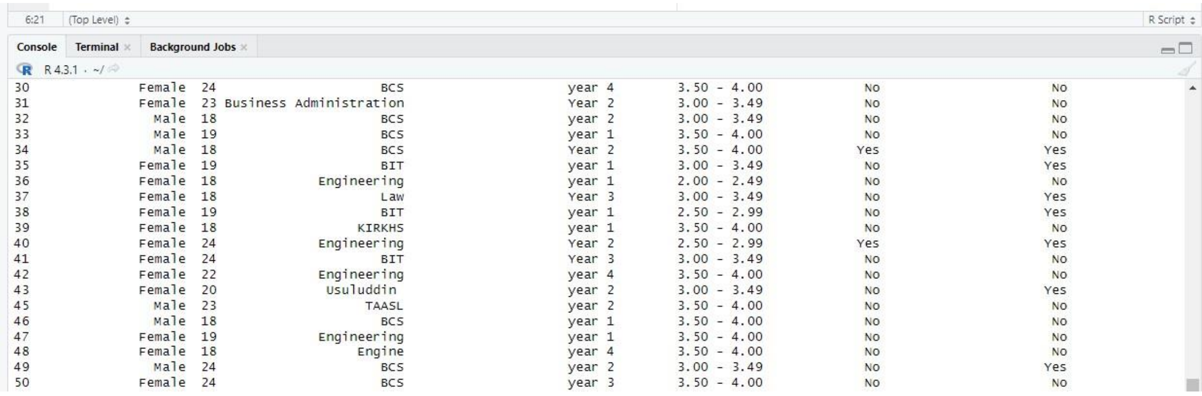
This code is used to identify the missing values in a dataset. The one that has TRUE written in it is a missing value.

3. <u>Discard missing values.</u>

   Code:

   dataset <- na.omit(dataset)

   print(dataset)



```
6:21   (Top Level)                                                R Script
Console   Terminal   Background Jobs
R 4.3.1 · ~/
30   Female  24                     BCS   year 4   3.50 - 4.00   No    No
31   Female  23 Business Administration   Year 2   3.00 - 3.49   No    No
32     Male  18                     BCS   year 2   3.00 - 3.49   No    No
33     Male  19                     BCS   year 1   3.50 - 4.00   No    No
34     Male  18                     BCS   Year 2   3.50 - 4.00   Yes   Yes
35   Female  19                     BIT   year 1   3.00 - 3.49   No    Yes
36   Female  18             Engineering   year 1   2.00 - 2.49   No    No
37   Female  18                     Law   Year 3   3.00 - 3.49   No    Yes
38   Female  19                     BIT   year 1   2.50 - 2.99   No    Yes
39   Female  18                  KIRKHS   year 1   3.50 - 4.00   No    No
40   Female  24             Engineering   Year 2   2.50 - 2.99   Yes   Yes
41   Female  24                     BIT   Year 3   3.00 - 3.49   No    No
42   Female  22             Engineering   year 4   3.50 - 4.00   No    No
43   Female  20              Usuluddin    year 2   3.00 - 3.49   No    Yes
45     Male  23                   TAASL   year 2   3.50 - 4.00   No    No
46     Male  18                     BCS   year 1   3.50 - 4.00   No    No
47   Female  19             Engineering   year 1   3.50 - 4.00   No    No
48   Female  18                  Engine   year 4   3.50 - 4.00   No    No
49     Male  24                     BCS   year 2   3.00 - 3.49   No    Yes
50   Female  24                     BCS   year 3   3.50 - 4.00   No    No
```

This code is used to remove the instance that had a missing value.

4. <u>Converting Numeric Age values to Categorical Age values.</u>

Code:

```
catage <- "Age"
breaks <- c(0, 19, 25, Inf)
labels <- c("Teenager", "Young Adult", "Adult")
dataset$CatAge <- cut(dataset[[catage]], breaks = breaks, labels = labels,
include.lowest = TRUE)
print(dataset)
```

```
14:1   (Top Level) ÷
Console   Terminal ×   Background Jobs ×
R 4.3.1 · ~/
86            Female  18            psychology                     year 1      3.50 - 4.00         No
87            Female  19           Fiqh fatwa                      Year 3      3.00 - 3.49         No
88            Female  18            psychology                     year 1      3.50 - 4.00         No
89              Male  24                   BIT                     year 1      3.00 - 3.49         No
90              Male  24           Engineering                     Year 2      2.00 - 2.49         No
   Do.you.have.Anxiety. Do.you.have.Panic.attack. Did.you.seek.any.specialist.for.a.treatment.      CatAge
1                    No                       Yes                                           No    Teenager
2                   Yes                        No                                           No Young Adult
3                   Yes                       Yes                                           No    Teenager
4                    No                        No                                           No Young Adult
5                    No                        No                                           No Young Adult
6                    No                       Yes                                           No    Teenager
7                    No                       Yes                                           No Young Adult
8                   Yes                        No                                           No    Teenager
9                    No                        No                                           No    Teenager
10                  Yes                       Yes                                           No    Teenager
11                   No                        No                                           No Young Adult
12                   No                        No                                           No Young Adult
```

This code is used to convert the categorize the age range to teenager, young adult, and adult. It is saved as a new column.

5. <u>Pearson's Chi-squared Test</u>

Code:

```
dataset <- data.frame(
  "what.is.your.cgpa" = sample(c("0-1.99", "2-2.49", "2.5-2.99", "3-3.49", "3.5-4.00"),
100, replace = TRUE),
  "Do.you.have.anxiety" = sample(c("Yes", "No"), 100, replace = TRUE)
)
contingency_table <- table(dataset$What.is.your.cgpa.,
dataset$Do.you.have.anxiety.)
chi_squared_test <- chisq.test(contingency_table)
print(chi_squared_test)
```

```
> dataset <- data.frame(
+ "what.is.your.cgpa" = sample(c("0-1.99", "2-2.49", "2.5-2.99", "3-3.49", "3.5-4.00"), 100, replace = TRUE),
+ "Do.you.have.anxiety" = sample(c("Yes", "No"), 100, replace = TRUE)
+ )
> contingency_table <- table(dataset$what.is.your.cgpa, dataset$Do.you.have.anxiety)
> chi_squared_test <- chisq.test(contingency_table)
> print(chi_squared_test)

        Pearson's Chi-squared test

data:  contingency_table
X-squared = 2.9557, df = 4, p-value = 0.5653
```

This code is used to find the significant attributes using Pearson's Chi-Squared Test.

6. Install Package

a) e1071

Code:

install.packages("e1071")

library(e1071)

```
> library(e1071)
Warning message:
package 'e1071' was built under R version 4.3.2
> |
```

This code is used to install the necessary package for the 'Naïve Bayes' function.

b) caret

Code:

install.packages("caret")

library(caret)

```
> library(caret)
Loading required package: ggplot2
Need help? Try Stackoverflow: https://stackoverflow.com/tags/ggplot2
Loading required package: lattice
Warning messages:
1: package 'caret' was built under R version 4.3.2
2: package 'ggplot2' was built under R version 4.3.2
```

This code is used to install the necessary package for the 'Naïve Bayes' classification.

7. <u>Naïve Bayes</u>
   Code:
   nb_model <- naiveBayes(What.is.your.course. ~ ., data = dataset)
   print(nb_model)

```
Console   Terminal   Background Jobs
R 4.3.1 · ~/
> nb_model <- naiveBayes(What.is.your.course. ~ ., data = dataset)
> print(nb_model)

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
          Accounting                  ALA       Banking Studies                  BCS                      Benl
                0.01                 0.01                  0.01                 0.18                      0.01
                BENL    Biomedical science        Biotechnology                  BIT   Business Administration
                0.02                 0.04                  0.01                 0.09                      0.01
       Communication                  CTS      Diploma Nursing         DIPLOMA TESL                     Econs
                0.01                 0.01                  0.01                 0.01                      0.01
               engin               Engine          Engineering                  ENM                      Fiqh
                0.01                 0.02                  0.17                 0.01                      0.01
          Fiqh fatwa      Human Resources       Human Sciences                Irkhs         Islamic education
                0.01                 0.01                  0.01                 0.01                      0.01
   Islamic Education                   IT                KENMS               Kirkhs                    KIRKHS
                0.01                 0.01                  0.01                 0.01                      0.01
                 koe                  Koe                  KOE                  Kop                       Law
                0.01                 0.01                  0.04                 0.01                      0.01
                Laws               Malcom       Marine science         Mathemathics                      MHSC
                0.02                 0.01                  0.01                 0.01                      0.01
             Nursing     Pendidikan islam     Pendidikan Islam     Pendidikan Islam                psychology
                0.01                 0.01                  0.01                 0.01                      0.02
          Psychology          Radiography                TAASL            Usuluddin
                0.01                 0.01                  0.01                 0.01

Conditional probabilities:
                         Choose.your.gender
Y                            Female      Male
  Accounting              1.0000000 0.0000000
  ALA                     1.0000000 0.0000000
  Banking Studies         1.0000000 0.0000000
  BCS                     0.6111111 0.3888889
  Benl                    1.0000000 0.0000000
  BENL                    1.0000000 0.0000000
  Biomedical science      0.5000000 0.5000000
  Biotechnology           1.0000000 0.0000000
  BIT                     0.6666667 0.3333333
  Business Administration 1.0000000 0.0000000
  Communication           1.0000000 0.0000000
  CTS                     1.0000000 0.0000000
  Diploma Nursing         1.0000000 0.0000000
  DIPLOMA TESL            1.0000000 0.0000000
  Econs                   1.0000000 0.0000000
  engin                   1.0000000 0.0000000
  Engine                  1.0000000 0.0000000
  Engineering             0.7058824 0.2941176
  ENM                     1.0000000 0.0000000
  Fiqh                    1.0000000 0.0000000
  Fiqh fatwa              1.0000000 0.0000000
```

This code is used to find the Naïve Bayes function to predict, based on the course of the student.

8. <u>Confusion Matrix</u>
   Code:
   pred <- predict(nb_model, dataset)
   confusion_matrix <- table(dataset$What.is.your.course., pred)
   print(confusion_matrix)

```
Console   Terminal   Background Jobs
R 4.3.1 · ~/

> pred <- predict(nb_model, dataset)
> confusion_matrix <- table(dataset$What.is.your.course., pred)
> print(confusion_matrix)
                            pred
                             Accounting  ALA Banking Studies BCS Benl BENL Biomedical science Biotechnology BIT
  Accounting                          0    0               0   0    0    0                  1             0   0
  ALA                                 0    0               0   0    0    0                  0             0   0
  Banking Studies                     0    0               0   1    0    0                  0             0   0
  BCS                                 0    0               0  11    0    1                  1             0   1
  Benl                                0    0               0   0    0    0                  1             0   0
  BENL                                0    0               0   0    0    2                  0             0   0
  Biomedical science                  0    0               0   0    0    0                  4             0   0
  Biotechnology                       0    0               0   0    0    0                  1             0   0
  BIT                                 0    0               0   0    0    0                  0             0   7
  Business Administration             0    0               0   0    0    0                  1             0   0
  Communication                       0    0               0   0    0    0                  0             0   1
  CTS                                 0    0               0   0    0    0                  0             0   1
  Diploma Nursing                     0    0               0   1    0    0                  0             0   0
  DIPLOMA TESL                        0    0               0   1    0    0                  0             0   0
  Econs                               0    0               0   1    0    0                  0             0   0
  engin                               0    0               0   1    0    0                  0             0   0
  Engine                              0    0               0   1    0    0                  0             0   0
  Engineering                         0    0               0   2    0    1                  2             0   0
```

This code has been used to find the confusion matrix on the student's courses.

9. <u>Accuracy</u>
   Code:
   accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
   print(paste("Accuracy:", accuracy))

```
> accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
> print(paste("Accuracy:", accuracy))
[1] "Accuracy: 0.39"
```

This code is used to find the predictive accuracy of the Naïve Bayes classification.

10. <u>10-Fold Cross Validation</u>

Code:

```
set.seed(123)
folds <- createFolds(dataset$Do.you.have.Depression., k = 10, list = TRUE, returnTrain
= TRUE)

for (i in 1:10) {
  train_fold <- dataset[unlist(folds[i]), ]
  test_fold <- dataset[-unlist(folds[i]), ]

  nb_model_fold <- naiveBayes(Do.you.have.Depression. ~ ., data = train_fold)

  predictions <- predict(nb_model_fold, test_fold)

  confusion_matrix <- table(test_fold$Do.you.have.Depression., predictions)
  accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)

  cat("Fold", i, "Accuracy:", accuracy, "\n")
}
```

```
+
+   cat("Fold", i, "Accuracy:", accuracy, "\n")
+ }
Fold 1 Accuracy: 0.8
Fold 2 Accuracy: 0.9
Fold 3 Accuracy: 0.8181818
Fold 4 Accuracy: 0.7777778
Fold 5 Accuracy: 0.7
Fold 6 Accuracy: 0.8181818
Fold 7 Accuracy: 0.9
Fold 8 Accuracy: 0.8
Fold 9 Accuracy: 0.7777778
Fold 10 Accuracy: 0.8
> 
```

The code is used for 10 Fold Cross Validation and Accuracy.

11. <u>Train Set and Test Set</u>

Code:

```
train_indices <- sample(1:nrow(dataset), 0.7 * nrow(dataset))
train_data <- dataset[train_indices, ]
test_data <- dataset[-train_indices, ]
```

```
> train_indices <- sample(1:nrow(dataset), 0.7 * nrow(dataset))
> train_data <- dataset[train_indices, ]
> test_data <- dataset[-train_indices, ]
```

This code is used to generate the train set and test set according to this dataset.

12. Recall, Precision and F-measure value

Code:

```
metrics <- data.frame(Recall = numeric(10), Precision = numeric(10), F_measure =
numeric(10))

for (i in 1:10) {
  train_fold <- dataset[unlist(folds[i]), ]
  test_fold <- dataset[-unlist(folds[i]), ]

  nb_model_fold <- naiveBayes(Do.you.have.Depression. ~ ., data = train_fold)

  predictions <- predict(nb_model_fold, test_fold)

  confusion_matrix <- table(test_fold$Do.you.have.Depression., predictions)


  recall <- confusion_matrix[2, 2] / sum(confusion_matrix[2, ])
  precision <- confusion_matrix[2, 2] / sum(confusion_matrix[, 2])
  f_measure <- 2 * (precision * recall) / (precision + recall)

  metrics[i, ] <- c(Recall = recall, Precision = precision, F_measure = f_measure)

  cat("Fold", i, "Recall:", recall, "Precision:", precision, "F-measure:", f_measure, "\n")
}


average_metrics <- colMeans(metrics)
cat("Average Recall:", average_metrics["Recall"], "Average Precision:",
average_metrics["Precision"], "Average F-measure:", average_metrics["F_measure"],
"\n")
```

```
Console  Terminal  Background Jobs
R 4.3.1 · C:/Users/mdham/OneDrive/Desktop/
+   cat("Fold", i, "Recall:", recall, "Precision:", precision, "F-measure:", f_measure, "\n")
+ }
Fold 1 Recall: 0.75 Precision: 0.75 F-measure: 0.75
Fold 2 Recall: 0.6666667 Precision: 1 F-measure: 0.8
Fold 3 Recall: 0.75 Precision: 0.75 F-measure: 0.75
Fold 4 Recall: 0.6666667 Precision: 0.6666667 F-measure: 0.6666667
Fold 5 Recall: 0.25 Precision: 1 F-measure: 0.4
Fold 6 Recall: 0.5 Precision: 1 F-measure: 0.6666667
Fold 7 Recall: 0.75 Precision: 1 F-measure: 0.8571429
Fold 8 Recall: 1 Precision: 0.6 F-measure: 0.75
Fold 9 Recall: 0.6666667 Precision: 0.6666667 F-measure: 0.6666667
Fold 10 Recall: 1 Precision: 0.6 F-measure: 0.75
>
>
> average_metrics <- colMeans(metrics)
> cat("Average Recall:", average_metrics["Recall"], "Average Precision:", average_metrics["Precision"], "Average F-measure:", average_metrics["F_measur
e"], "\n")
Average Recall: 0.7 Average Precision: 0.8033333 Average F-measure: 0.7057143
>
```

The code shows 10 Fold (Recall, Precision, F-measure) value and it also shows the Average (Recall, Precision, F-measure) value.