Data Science Project

Machine Learning

เสนอโดย

ดร. ขวัญกมล ดิฐกัญจน์

จัดทำโดย

นาย ธนวัฒน์ วิริยธรรมโสภณ 6530611033

ภาคเรียนที่ 2 ปีการศึกษา 2566

สาขาวิชา

การคอมพิวเตอร์

คณะวิทยาลัยการคอมพิวเตอร์ มหาวิทยาลัยสงขลานครินทร์

คำนำ

โปรเจคนี้เป็นส่วนหนึ่งของวิชา Data Science หรือ วิชา วิทยาศาสตร์ข้อมูล โดยได้รับมอบหมายให้ศึกษาหาข้อมูลจากสื่อออนไลน์เพื่อนำมาใช้ในรายงานนี้ โดยเลือกข้อมูลจากเว็บไซต์ออนไลน์ Kaggle.com โดยเลือกหาชุดข้อมูลสนใจที่มีข้อมูลเป็นลักษณะ Data set ที่เป็นแบบ Open Data โดยได้กำหนดข้อมูล และ สิ่งสำคัญไว้หลังจากนั้นจึงนำข้อมูลดังกล่าวมาอธิบายข้อมูลและสร้าง Model ที่จำเป็นต่อผลลัพธ์รายงานนี้และใช้ความรู้ในรายวิชาดังกล่าวในการสรุปข้อมูลหรือคาดการณ์ความเป็นไปได้ของข้อมูลนั้นๆ เช่น การใช้ Machine Learning Algorithm หรือ Deep Learning Model หรือ Software ทางการ เป็นต้น

ทั้งนี้นักศึกษาหวังว่าโปรเจคหรือรายงานชิ้นนี้จะเป็นประโยชน์ที่ใช้สำหรับการศึกษาชุดข้อมูลต่างๆที่สำคัญของรายวิชา Data Science และทางนักศึกษาขอขอบคุณ ดร. ขวัญกมล ดิฐกัญจน์ ที่ช่วยให้รายงานหรือโปรเจ็คชิ้นนี้เสร็จสมบูรณ์

10 เมษายน 2567

นาย ธนวัฒน์ วิริยธรรมโสภณ

6530611033

สารบัญ

Classification

a. ให้นักศึกษาเลือก Data Set พร้อมคำอธิบายตัวข้อมูล

ชุดข้อมูลที่นักศึกษาเลือกคือชุดข้อมูลของ **onlinefoods** จากเว็บไซต์ Kaggle เป็นชุดข้อมูล excel หรือ .csv ซึ่งชุดข้อมูลดังกล่าวเป็นคอลเลกชันที่ครอบคลุมที่ประกอบด้วยรายการที่สมจริง 389รายการซึ่งรวบรวมอย่างพิถีพิถันเพื่อรวบรวมแพลตฟอร์มการสั่งอาหารออนไลน์ในช่วงระยะเวลาหนึ่ง ประกอบด้วยคุณลักษณะต่างๆ ที่เกี่ยวข้องกับอาชีพ ขนาดครอบครัว ผลตอบรับ ฯลฯ ข้อมูลภายในคอลัมน์มีดังนี้ https://www.kaggle.com/datasets/sudarshan24byte/online-food-dataset

- Demographic Information:
- Age: Age of the customer.
- Gender: Gender of the customer.
- Marital Status: Marital status of the customer.
- Occupation: Occupation of the customer.
- Monthly Income: Monthly income of the customer.
- Educational Qualifications: Educational qualifications of the customer.
- Family Size: Number of individuals in the customer's family.
- Location Information:
- Latitude: Latitude of the customer's location.
- Longitude: Longitude of the customer's location.
- Pin Code: Pin code of the customer's location.
- Output: Current status of the order (yes, no).

B. Data Preparation รายละเอียดขั้นตอนของการเตรียมข้อมูลก่อนจะนำไปใช้ในการพัฒนา ด้วย Machine Learning

1. กำหนดเป้าหมายของ Data set นี้คือการจัดจำแนกชุดข้อมูลรวมถึงการ clean data, transformation และ combining data

2. ดาวน์โหลดข้อมูลจาก Kaggle.com โดยได้ดาวน์โหลดเป็นไฟล์ .csv ซึ่งข้อมูลมีทั้งหมด 389 ข้อมูล และมี Attribute 11 attribute โดยไฟล์มีชื่อว่า **onlinefood** ทำการ clean ข้อมูลที่มีเช่นข้อมูลที่ใช้งานไม่ได้ หรือข้อมูลที่ไม่สะอาดอ่านได้

3. ทำการแปลงข้อมูลเช่นการทำ Scaling, Normalization, หรือ Encoding ข้อมูลเพื่อให้ข้อมูลพร้อมนำเข้าสู่ Model ได้อย่างเหมาะสม

4. นำชุดข้อมูลที่ได้ Clean แล้วมารวมกันหรือรวมข้อมูลจากหลายแหล่งเพื่อเตรียมสำหรับการใช้งานใน Machine Learning Model

5. นำชุดข้อมูลไปใช้กับโปรแกรมหรือซอฟแวร์

C. บอกวัตถุสงค์ของการสร้าง Classification Model ว่าต้องการจำแนกอะไร เพื่ออะไร

1. การจำแนกลูกค้า ใช้ข้อมูลเพื่อแยกแยะลูกค้าในกลุ่มที่มีลักษณะ ที่แตกต่างกัน เช่น อายุ, เพศ, สถานะภาพการสมรส, อาชีพ, รายได้เฉลี่ยต่อเดือน, ระดับการศึกษา, ขนาดครอบครัว เพื่อให้สามารถดำเนินการตลาดและบริหารจัดการลูกค้าในแต่ละกลุ่มได้

2. การทำนายสถานะคำสั่งซื้อ ใช้ข้อมูลเพื่อทำนายสถานะคำสั่งซื้อของลูกค้า เช่น คำสั่งซื้ออยู่ในขั้นตอนการดำเนินการใดๆ จะช่วยในการวางแผนการจัดส่งสินค้าและบริการลูกค้า

3. การทำนายความน่าจะเป็น เพื่อทำนายความน่าจะเป็นของเหตุการณ์ต่างๆ เช่น ความน่าจะเป็นที่ลูกค้าจะทำรายการซื้อสินค้าในช่วงเวลาหนึ่งๆ หรือความน่าจะเป็นที่คำสั่งซื้อจะเสร็จสิ้นภายในเวลาที่กำหนด

4. การวิเคราะห์ข้อมูลเพื่อการตัดสินใจ ใช้ข้อมูลเพื่อวิเคราะห์และช่วยในการตัดสินใจ เช่น การอนุมัติสินเชื่อ, การจัดส่งสินค้าตามความต้องการของลูกค้า

5. การตรวจจับและป้องกันการฉ้อโกง ใช้ข้อมูลเพื่อตรวจจับและป้องกันการฉ้อโกง เช่น การตรวจจับการใช้บัตรเครดิตโดยไม่ได้รับอนุญาต, การปลอมแปลงข้อมูลการเงิน

D. สร้าง Classification Model

โมเดลที่ใช้สร้าง Classification Model: Rstudio

Decision tree:

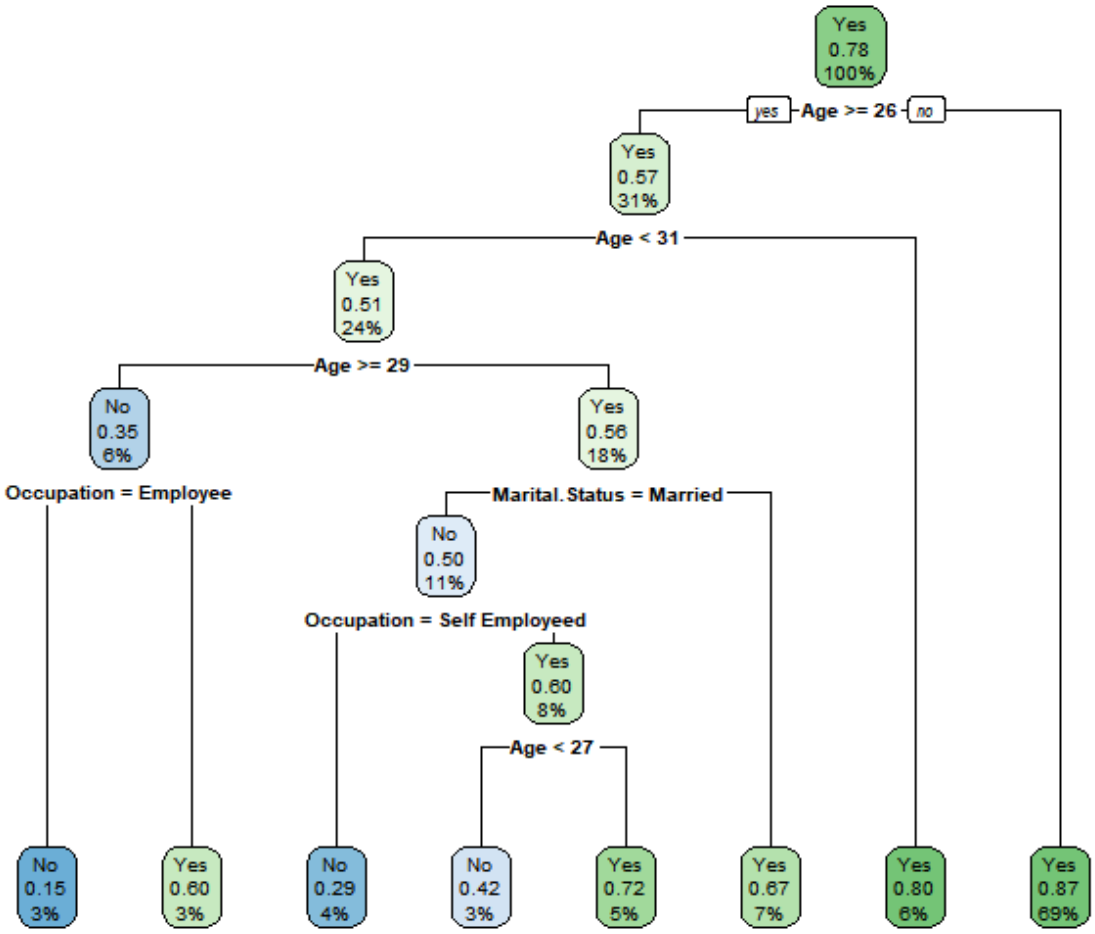
```
#แผนภูมิ Tree
install.packages("rpart.plot")  # Install necessary package
library("rpart")
library("rpart.plot")
#อ่านข้อมูล
onlinefood <- read.csv("C:/Users/Admin/Downloads/datasci+prj/onlinefoods.csv", header=TRUE, sep=",")

#check summary
summary(onlinefood)
onlinefood$Monthly.Income <- as.numeric(as.character(onlinefood$Monthly.Income))

# สร้างแผนภูมิ Tre
Tree <- rpart(Output ~ Age + Gender + Marital.Status + Occupation,
              data = onlinefood,
              method = "class",
              control = rpart.control(minsplit = 20))  # Adjust minsplit as needed
# จำลอง the tree
rpart.plot(Tree, type = 2, clip.right.labs = FALSE, varlen = 0, faclen = 0)
# เตรียมข้อมูลสำหรับการคาดเดา
newdata <- data.frame(
  Age = 26,
  Gender = factor("Male", levels = levels(onlinefood$Gender)),
  Marital.Status = factor("Married", levels = levels(onlinefood$Marital.Status)),
  Occupation = factor("Employed", levels = levels(onlinefood$Occupation)),
  Monthly.Income = 10001,
  Family.size = 1,
  latitude = 12.9579,
  longitude = 77.6309,
  Pin.code = 560007
)
# คาดเดา
prediction <- predict(Tree, newdata = newdata, type = "class")
prediction
```

```
> # เตรียมข้อมูลสำหรับการคาดเดา
> newdata <- data.frame(
+   Age = 26,
+   Gender = factor("Male", levels = levels(onlinefood$Gender)),
+   Marital.Status = factor("Married", levels = levels(onlinefood$Marital.Status)),
+   Occupation = factor("Employed", levels = levels(onlinefood$Occupation)),
+   Monthly.Income = 10001,
+   Family.size = 1,
+   latitude = 12.9579,
+   longitude = 77.6309,
+   Pin.code = 560007
+ )
> # คาดเดา
> prediction <- predict(Tree, newdata = newdata, type = "class")
> prediction
  1
Yes
Levels: No Yes
```

ใช้ library(rpart),(rpart.plot) สำหรับสร้างโมเดล decision treeและทำการอ่านไฟล์ด้วย read.csv เช็คไฟล์และดูว่าผลลัพธ์ออกไหมด้วย summary สรุปในส่วนของ onlinefood$monthly.income เป็นโค้ดเปลี่ยนให้ string เป็น num บรรทัดต่อไปคือส่วนชองการสร้าง แผนภูมิต้นไม้ และ ตัวคาดการณ์newdata เพื่อคาดการ

จากข้อมูล newdata ใหม่ที่ทำการทำนายเช่น เพศชาย สถานะแต่งงานแล้ว ฯลฯ ซึ่งทำนายว่า ในกลุ่มผู้ชาย อายุ 26 นั้นข้อมูลดังกล่าว เป็นไปตามที่คาดเดาโดยให้ตารางสุดท้ายเป็นตัวบอก คือ output ใน attribute ตัวสุดท้าย1 **โดยได้เลือกในส่วนนี้**

```
# สร้างแผนภูมิ Tre
Tree <- rpart(Output ~
```



Naïve Bayes:

โมเดลที่ใช้: Rstuido + Weka

```
# Naïve Bayes
install.packages("e1071")
library(e1071)

weatherBayes <- naiveBayes(Output ~ Age + Gender + Marital.Status + Occupation +
                             Monthly.Income + Family.size + latitude + longitude + Pin.code +Educational.Qualifications,
                           data = onlinefood)
weatherBayes

newdata1 <- data.frame(
  Age = 26,
  Gender = "Male",
  Marital.Status = "Married",
  Occupation = factor("Employed", levels = levels(onlinefood$Occupation)),
  Monthly.Income = 10001,
  Family.size = 1,
  latitude = 12.9579,
  longitude = 77.6309,
  Pin.code = 560007
)
newdata1$Monthly.Income <- factor(newdata1$Monthly.Income, levels = levels(onlinefood$Monthly.Income))

predictions <- predict(weatherBayes, newdata = newdata1, type = "class")
predictions
```

```
> newdata1$Monthly.Income <- factor(newdata1$Monthly.Income, levels = levels(onlinefood$Monthly.Income))
> predictions <- predict(weatherBayes, newdata = newdata1, type = "class")
> predictions
[1] Yes
Levels: No Yes
```

เช่นกันกับ algorithm นี้เพียงแต่เปลี่ยน algorithm เท่านั้น

ใช้ library(e1071) สำหรับ algorithm Naïve bayes

Naive Bayes Classifier

Attribute	Class Yes (0.77)	No (0.23)
Age		
mean	24.2326	26
std. dev.	2.8714	2.9046
weight sum	301	87
precision	1	1
Gender		
Female	127.0	41.0
Male	176.0	48.0
[total]	303.0	89.0
Marital Status		
Single	230.0	40.0
Married	67.0	43.0
Prefer not to say	7.0	7.0
[total]	304.0	90.0
Occupation		
Student	185.0	24.0
Employee	77.0	43.0
Self Employeed	35.0	21.0
House wife	8.0	3.0
[total]	305.0	91.0
Monthly Income		
No Income	165.0	24.0
Below Rs.10000	20.0	7.0
More than 50000	45.0	19.0
10001 to 25000	33.0	14.0
25001 to 50000	43.0	28.0
[total]	306.0	92.0
Educational Qualifications		
Post Graduate	148.0	28.0
Graduate	128.0	51.0
Ph.D	17.0	8.0
Uneducated	2.0	2.0
School	11.0	3.0
[total]	306.0	92.0
Family size		
mean	3.2492	3.3908
std. dev.	1.2944	1.5189
weight sum	301	87

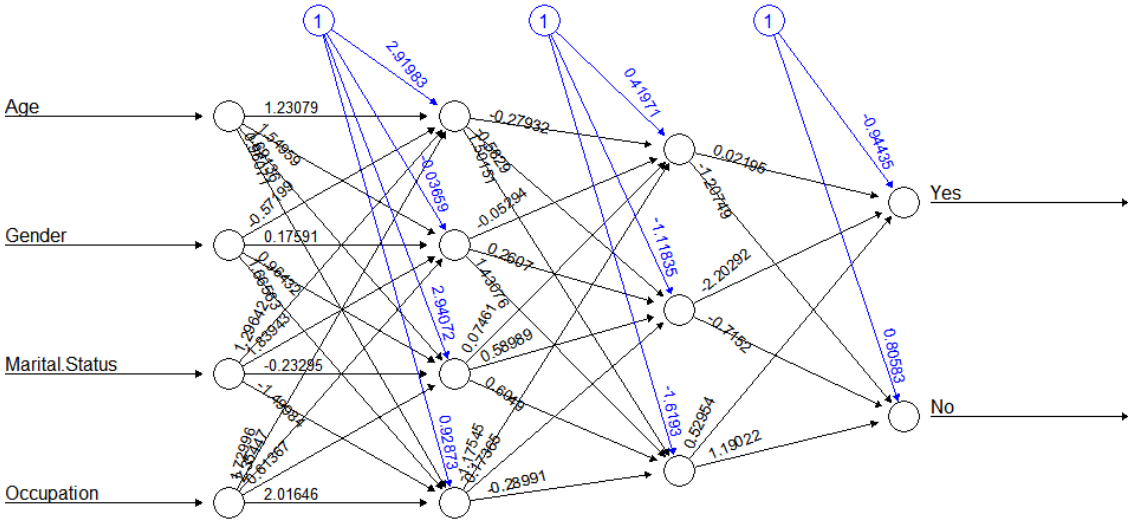
Neural Network:

```
#ลง neuralnet
install.packages("neuralnet")
library(neuralnet)

# อ่านไฟล์
onlinefood <- read.csv("C:/Users/Admin/Downloads/datasci+prj/onlinefoods.csv", header = TRUE, sep = ",")

# Convert ค่าเป็นตัว numeric
onlinefood$Gender <- as.numeric(as.factor(onlinefood$Gender))
onlinefood$Marital.Status <- as.numeric(as.factor(onlinefood$Marital.Status))
onlinefood$Occupation <- as.numeric(as.factor(onlinefood$Occupation))
onlinefood$Educational.Qualifications <- as.numeric(as.factor(onlinefood$Educational.Qualifications))

# แบ่งข้อมูล test กับ train
traindata <- onlinefood[1:387, ]
testdata <- onlinefood[388:nrow(onlinefood), ]

# สร้าง model network
onlinefoodNN <- neuralnet(Output ~ Age + Gender + Marital.Status + Occupation,
                          data = traindata,
                          hidden = c(4, 3),
                          linear.output = FALSE)
#รายละเอียดและโชว์ model
summary(onlinefood)
plot(onlinefoodNN)
```

โหลด library(neuralnet) อ่านไฟล์.csv หลังจากนั้น convert ค่า string ของไฟล์ให้เป็น numbric

แบ่งข้อมูล test and train 1-387 และ testdata เป็น 388 ตามด้วยโค้ดสร้างโมเดลโดยใช้ attribure 4 ตัวเพราะถ้าใช้ 10 ตัว ค่าจะไม่อ่านและเกิด error เพราะรับค่าไม่ได้

สุดท้ายรายละเอียด summary และ โชว์ plot



E. ผลการสร้างโมเดลว่ามีความถูกต้องเป็นเท่าใด โดยใช้การประเมินแบบใด n-fold cross validation/percent-split ให้เปรียบเทียบผลจากการสร้าง Model ไม่ต่ำากว่า 3 algorithm ใช้แอป Weka เปรียบเทียบ

Decision Tree:

Validation:

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         325               83.7629 %
Incorrectly Classified Instances        63               16.2371 %
Kappa statistic                          0.4338
Mean absolute error                      0.249
Root mean squared error                  0.3811
Relative absolute error                 71.3904 %
Root relative squared error             91.3611 %
Total Number of Instances              388

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.967    0.609    0.846      0.967    0.902      0.470    0.609     0.801     Yes
                 0.391    0.033    0.773      0.391    0.519      0.470    0.609     0.474     No
Weighted Avg.    0.838    0.480    0.830      0.838    0.816      0.470    0.609     0.728

=== Confusion Matrix ===

   a   b   <-- classified as
 291  10 |   a = Yes
  53  34 |   b = No
```

percent-split 66%

```
=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          92               69.697  %
Incorrectly Classified Instances        40               30.303  %
Kappa statistic                          0.1384
Mean absolute error                      0.3325
Root mean squared error                  0.5071
Relative absolute error                 93.8688 %
Root relative squared error            118.1337 %
Total Number of Instances              132

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.820    0.688    0.788      0.820    0.804      0.139    0.645     0.827     Yes
                 0.313    0.180    0.357      0.313    0.333      0.139    0.645     0.310     No
Weighted Avg.    0.697    0.564    0.684      0.697    0.690      0.139    0.645     0.701

=== Confusion Matrix ===

  a  b   <-- classified as
 82 18 |  a = Yes
 22 10 |  b = No
```

Naïve Bayes:

Validation:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         282               72.6804 %
Incorrectly Classified Instances       106               27.3196 %
Kappa statistic                          0.2953
Mean absolute error                      0.2974
Root mean squared error                  0.4469
Relative absolute error                 85.2703 %
Root relative squared error            107.1384 %
Total Number of Instances              388

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.777    0.448    0.857      0.777    0.815      0.301    0.697     0.867     Yes
                 0.552    0.223    0.417      0.552    0.475      0.301    0.697     0.381     No
Weighted Avg.    0.727    0.398    0.759      0.727    0.739      0.301    0.697     0.758

=== Confusion Matrix ===

   a   b   <-- classified as
 234  67 |   a = Yes
  39  48 |   b = No
```

percent-split 66%

```
=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          96               72.7273 %
Incorrectly Classified Instances        36               27.2727 %
Kappa statistic                          0.3157
Mean absolute error                      0.3002
Root mean squared error                  0.4373
Relative absolute error                 84.7486 %
Root relative squared error            101.8626 %
Total Number of Instances              132

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.780    0.438    0.848      0.780    0.813      0.319    0.712     0.864     Yes
                 0.563    0.220    0.450      0.563    0.500      0.319    0.712     0.543     No
Weighted Avg.    0.727    0.385    0.751      0.727    0.737      0.319    0.712     0.786

=== Confusion Matrix ===

  a  b   <-- classified as
 78 22 |  a = Yes
 14 18 |  b = No
```

Neural Network:

Validation:

```
Time taken to build model: 0.83 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         307               79.1237 %
Incorrectly Classified Instances        81               20.8763 %
Kappa statistic                          0.4166
Mean absolute error                      0.2183
Root mean squared error                  0.4285
Relative absolute error                 62.5759 %
Root relative squared error            102.7206 %
Total Number of Instances              388

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.854    0.425    0.874      0.854    0.864      0.417    0.726     0.860     Yes
               0.575    0.146    0.532      0.575    0.552      0.417    0.726     0.501     No
Weighted Avg.  0.791    0.363    0.797      0.791    0.794      0.417    0.726     0.780

=== Confusion Matrix ===

   a   b   <-- classified as
 257  44 |   a = Yes
  37  50 |   b = No
```

percent-split 66%

```
=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         108               81.8182 %
Incorrectly Classified Instances        24               18.1818 %
Kappa statistic                          0.483
Mean absolute error                      0.2086
Root mean squared error                  0.4074
Relative absolute error                 58.8927 %
Root relative squared error             94.8879 %
Total Number of Instances              132

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.900    0.438    0.865      0.900    0.882      0.485    0.722     0.825     Yes
               0.563    0.100    0.643      0.563    0.600      0.485    0.722     0.587     No
Weighted Avg.  0.818    0.356    0.811      0.818    0.814      0.485    0.722     0.767

=== Confusion Matrix ===

  a  b   <-- classified as
 90 10 |  a = Yes
 14 18 |  b = No
```

Classification ทั้ง 3 ตัวที่ได้ทำการลอง test แบบ cross validation และ Percentage split 66% มีการเปลี่ยนแปลงที่ค่อนข้างน้อยแม้จะใช้ algorithm

Clustering (20 คะแนน)

a. ให้นักศึกษาเลือก Data Set พร้อมคำอธิบายตัวข้อมูล

ชุดข้อมูลที่นักศึกษาเลือกคือชุดข้อมูลของ **onlinefoods** จากเว็บไซต์ Kaggle เป็นชุดข้อมูล excel หรือ .csv ซึ่งชุดข้อมูลดังกล่าวเป็นคอลเลกชันที่ครอบคลุมที่ประกอบด้วยรายการที่สมจริง 389รายการซึ่งรวบรวมอย่างพิถีพิถันเพื่อรวบรวมแพลตฟอร์มการสั่งอาหารออนไลน์ในช่วงระยะเวลาหนึ่ง ประกอบด้วยคุณลักษณะต่างๆ ที่เกี่ยวข้องกับอาชีพ ขนาดครอบครัว ผลตอบรับ ฯลฯ ข้อมูลภายในคอลัมน์มีดังนี้ https://www.kaggle.com/datasets/sudarshan24byte/online-food-dataset

- Demographic Information:
- Age: Age of the customer.
- Gender: Gender of the customer.
- Marital Status: Marital status of the customer.
- Occupation: Occupation of the customer.
- Monthly Income: Monthly income of the customer.
- Educational Qualifications: Educational qualifications of the customer.
- Family Size: Number of individuals in the customer's family.
- Location Information:
- Latitude: Latitude of the customer's location.
- Longitude: Longitude of the customer's location.
- Pin Code: Pin code of the customer's location.
- Output: Current status of the order (yes, no).

b. Data Preparation รายละเอียดขั้นตอนของการเตรียมข้อมูลก่อนจะนำไปใช้ในการพัฒนาด้วย Machine Learning

1. กำหนดเป้าหมายของ Data set นี้คือการจัดจำแนกชุดข้อมูลรวมถึงการ clean data, transformation และ combining data

2. ดาวน์โหลดข้อมูลจาก Kaggle.com โดยได้ดาวน์โหลดเป็นไฟล์ .csv ซึ่งข้อมูลมีทั้งหมด 389 ข้อมูล และมี Attribute 11 attribute โดยไฟล์มีชื่อว่า **onlinefood** ทำการ clean ข้อมูลที่มีเช่นข้อมูลที่ใช้งานไม่ได้ หรือข้อมูลที่ไม่สะอาดอ่านได้

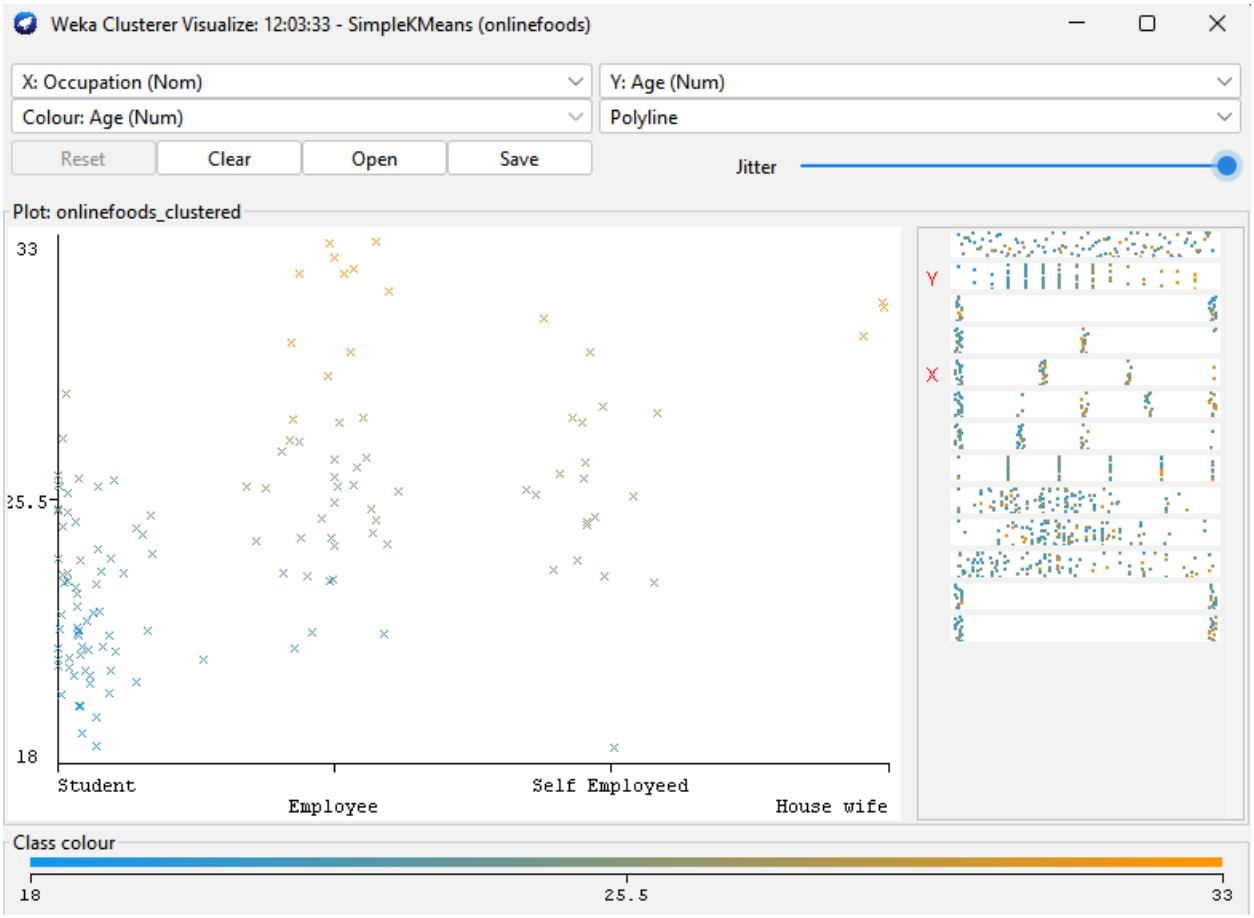
3. ทำการแปลงข้อมูลเช่นการทำ Scaling, Normalization, หรือ Encoding ข้อมูลเพื่อให้ข้อมูลพร้อมนำเข้าสู่ Model ได้อย่างเหมาะสม

4. นำชุดข้อมูลที่ได้ Clean แล้วมารวมกันหรือรวมข้อมูลจากหลายแหล่งเพื่อเตรียมสำหรับการใช้งานใน Machine Learning Model

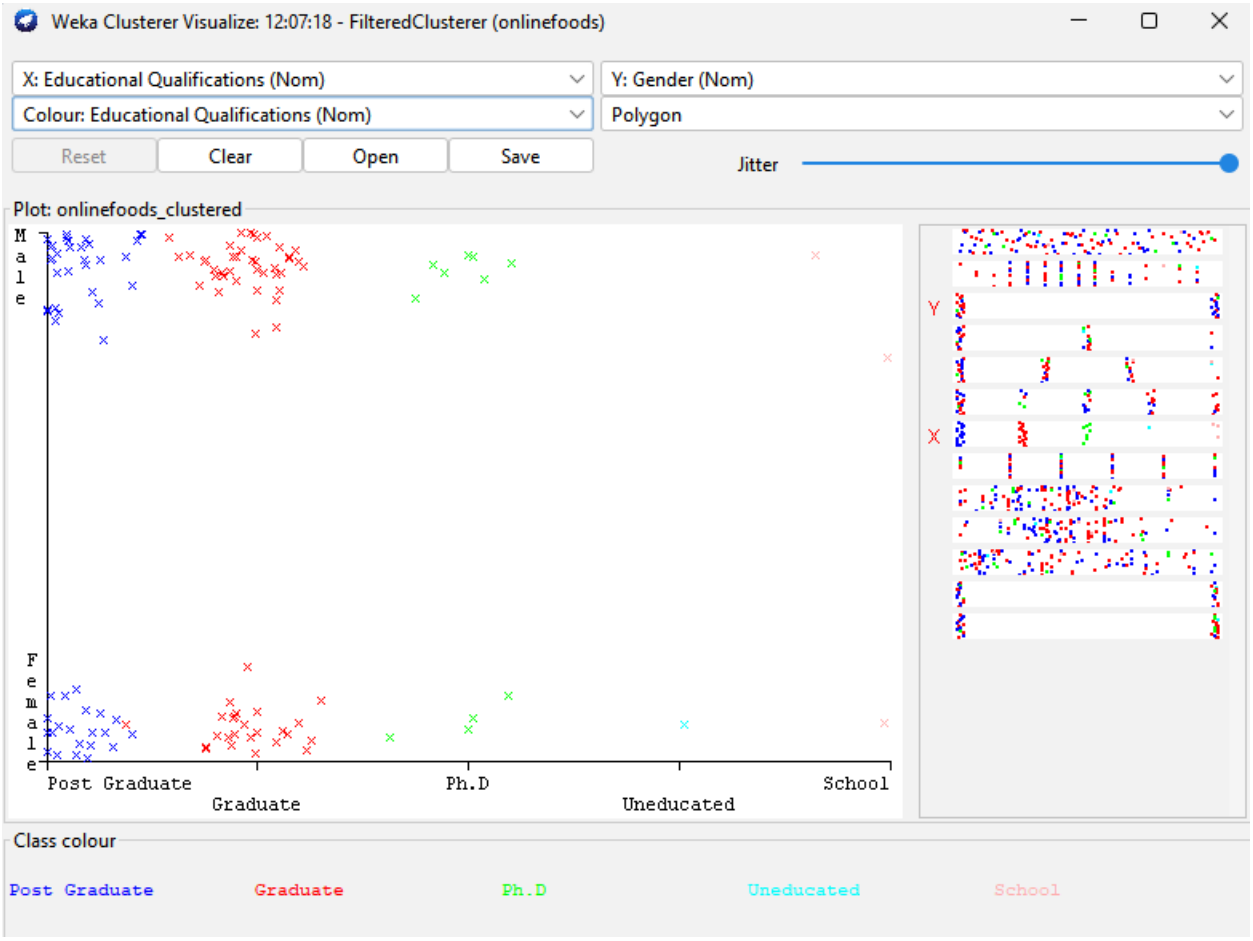
5. นำชุดข้อมูลไปใช้กับโปรแกรมหรือซอฟแวร์

c. สร้าง Cluster อย่างน้อย 3 cluster

จัดกลุ่มอายุ(Age)ของผู้ใช้งานและสถานะอาชีพ(Occupation) ด้วย Simple Kmean

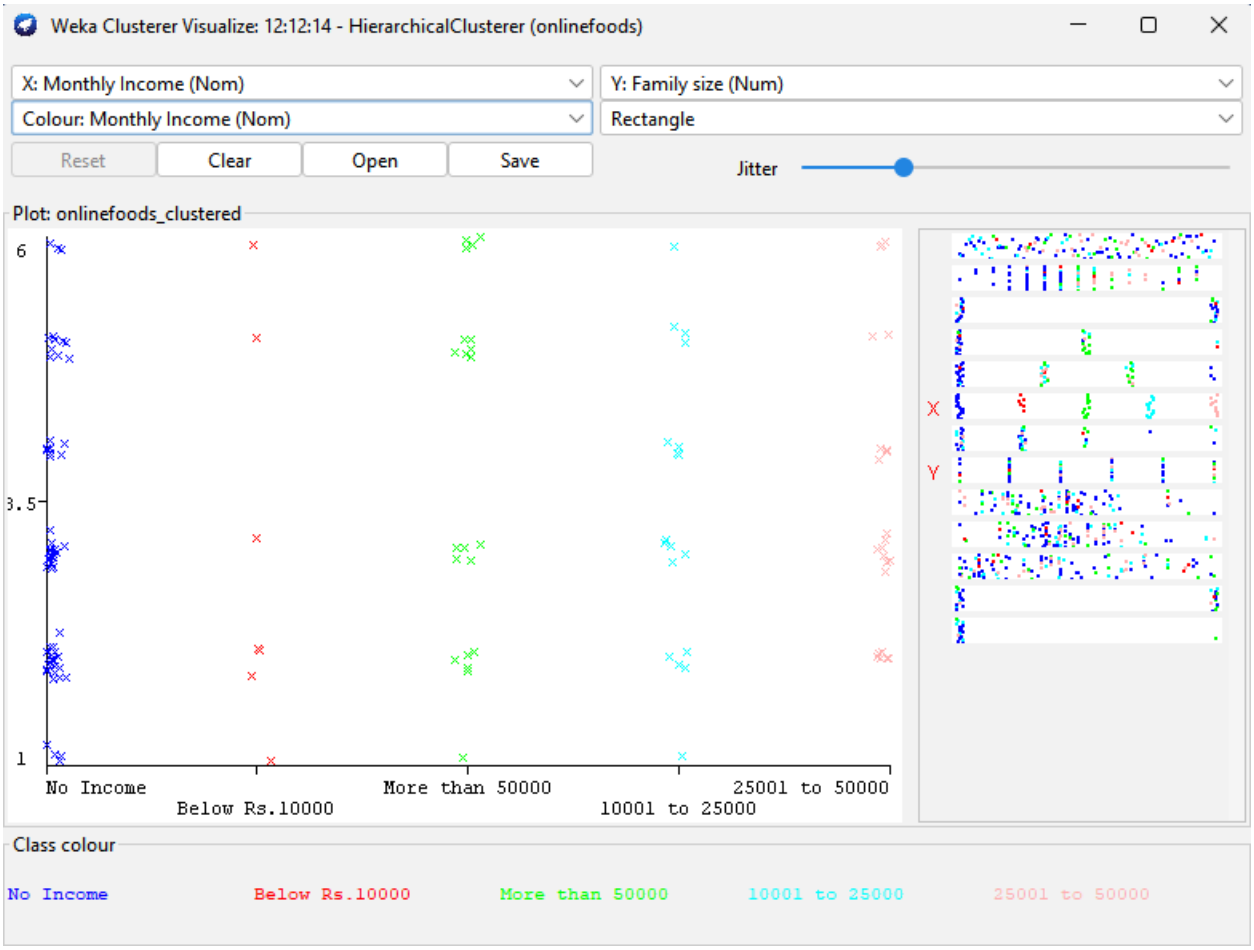


จัดกลุ่มเพศ(Gender)ของผู้ใช้งานและประวัติการศึกษาที่มี(Educational Qualification) ด้วย FiltheredClusterer



จัดกลุ่มรายได้(monthly income)ของผู้ใช้งานและขนาดครอบครัวที่มี(family size)

ด้วย Hierarchical Clusterer



d. ให้วิเคราะห์หาลักษณะของ Cluster แต่ละกลุ่ม

Simple Kmean

1. จากการใช้ Clusterer ในการ จัดกลุ่มอายุ(Age)ของผู้ใช้งานและสถานะอาชีพ(Occupation) ด้วย Simple Kmean แสดงให้เห็นว่า ในกลุ่มของอาชีพทั้ง 4 ที่มีการแสดงออกมานั้นมี Student, Employee, Self-Employee, House Wife ซึ่งในอายุตั้งแต่ 18 ถึง 25+ ส่วนใหญ่เป็น Student และ Employeeมีอายุที่สูงกว่า Student ประมาณ 19-33 ในขณะที่ Self Employeeมีตั้งแต่20-30 สุดท้าย House Wife มีราวๆอายุ 30+ ซึ่งมีข้อมูลที่น้อย

Filthered Clusterer

2. จากการใช้ Clusterer ในการจัดกลุ่มเพศ(Gender)ของผู้ใช้งานและสถานะ(Educational Qualifications) ด้วย Filthered Cluster แสดงให้เห็นว่า ในกลุ่มของระดับการศึกษาทั้ง 5 ที่มีการแสดงออกมานั้นมี Post Graduate, Graduate, Ph.d, Unducated, School ซึ่งในข้อมูลได้แสดงว่าเพศที่มีการระบุแต่ละเพศมีระดับการศึกษาเท่าใดมากที่สุดตามเพศ

Hierarchical Clusterer

3. จากการใช้ Cluster ในการจัดกลุ่มรายได้(Monthly Income) และ ขนาดครอบครัว(Family size) ด้วย Hierarchical Clusterer แสดงให้เห็นว่า ในกลุ่ม ขนาดครอบครัวและ รายได้ส่วยใหญ่นั้นแสดงออกมาว่าขนาดที่เยอะ มีรายได้เป็นเท่าไหร่ หรือ ขนาดรายได้เท่าไหร่มีรายได้ income เท่าไหร่ ซึ่งส่วนใหญ่ข้อมูลรวมจัวอยู่ที่ ขนาดครอบครัวประมาณ1-2 ไม่มีรายได้ อาจจะเป็นนักศึกษา

Association Rule (20 คะแนน)

a. ให้นักศึกษาเลือก Data Set พร้อมคำอธิบายตัวข้อมูล

ชุดข้อมูลที่นักศึกษาเลือกคือชุดข้อมูลของ **onlinefoods** จากเว็บไซต์ Kaggle เป็นชุดข้อมูล excel หรือ .csv ซึ่งชุดข้อมูลดังกล่าวเป็นคอลเลกชันที่ครอบคลุมที่ประกอบด้วยรายการที่สมจริง 389รายการซึ่งรวบรวมอย่างพิถีพิถันเพื่อรวบรวมแพลตฟอร์มการสั่งอาหารออนไลน์ในช่วงระยะเวลาหนึ่ง ประกอบด้วยคุณลักษณะต่างๆ ที่เกี่ยวข้องกับอาชีพ ขนาดครอบครัว ผลตอบรับ ฯลฯ ข้อมูลภายในคอลัมน์มีดังนี้ https://www.kaggle.com/datasets/sudarshan24byte/online-food-dataset

- Demographic Information:
- Age: Age of the customer.
- Gender: Gender of the customer.
- Marital Status: Marital status of the customer.
- Occupation: Occupation of the customer.
- Monthly Income: Monthly income of the customer.
- Educational Qualifications: Educational qualifications of the customer.
- Family Size: Number of individuals in the customer's family.
- Location Information:
- Latitude: Latitude of the customer's location.
- Longitude: Longitude of the customer's location.
- Pin Code: Pin code of the customer's location.
- Output: Current status of the order (yes, no).

b. Data Preparation รายละเอียดขั้นตอนของการเตรียมข้อมูลก่อนจะนำไปใช้ในการพัฒนา ด้วย Machine Learning

6. กำหนดเป้าหมายของ Data set นี้คือการจัดจำแนกชุดข้อมูลรวมถึงการ clean data, transformation และ combining data

7. ดาวน์โหลดข้อมูลจาก Kaggle.com โดยได้ดาวน์โหลดเป็นไฟล์ .csv ซึ่งข้อมูลมีทั้งหมด 389 ข้อมูล และมี Attribute 11 attribute โดยไฟล์มีชื่อว่า **onlinefood** ทำการ clean ข้อมูลที่มีเช่นข้อมูลที่ใช้งานไม่ได้ หรือข้อมูลที่ไม่สะอาดอ่านได้

8. ทำการแปลงข้อมูลเช่นการทำ Scaling, Normalization, หรือ Encoding ข้อมูลเพื่อให้ข้อมูลพร้อมนำเข้าสู่ Model ได้อย่างเหมาะสม

9. นำชุดข้อมูลที่ได้ Clean แล้วมารวมกันหรือรวมข้อมูลจากหลายแหล่งเพื่อเตรียมสำหรับการใช้งานใน Machine Learning Model

10. นำชุดข้อมูลไปใช้กับโปรแกรมหรือซอฟแวร์

c. กำหนด Minimum Support 5% และ Minimum Confidence ที่ 60% (หรือสามารถกำหนดได้เอง

ตามความเหมาะสมของข้อมูล)

Untitled1* × project2.R* ×

Source on Save

```
1  install.packages("arules")
2  library(arules)
3
4
5  onlinefood <- read.csv("C:/Users/Admin/Downloads/datasci+prj/onlinefoods.csv", header = TRUE, sep = ",")
6
7  rules <- apriori(onlinefood)
8  summary(rules)
9  inspect(rules)
10
11 rules1 <- apriori(onlinefood, parameter = list(minlen = 2, maxlen = 10, supp = 0.05, conf = 0.6))
12 summary(rules1)
13 inspect(rules1)
14
```

```
      Educational.Qualifications=Graduate,
      Output=Yes}                          => {Marital.Status=Single}                0.06701031  0.9629630 0.06958763 1.3941404    26
[7140] {Age=[18,23),
      Marital.Status=Single,
      Occupation=Student,
      Monthly.Income=No Income,
      Educational.Qualifications=Graduate,
      Output=Yes}                          => {Gender=Male}                          0.06701031  0.6666667 0.10051546 1.1651652    26
[7141] {Age=[18,23),
      Gender=Male,
      Marital.Status=Single,
      Monthly.Income=No Income,
      Educational.Qualifications=Graduate,
      Output=Yes}                          => {Occupation=Student}                   0.06701031  1.0000000 0.06701031 1.8743961    26
[7142] {Age=[18,23),
      Gender=Male,
      Marital.Status=Single,
      Occupation=Student,
      Educational.Qualifications=Graduate,
      Output=Yes}                          => {Monthly.Income=No Income}             0.06701031  0.9629630 0.06958763 1.9980194    26
[7143] {Age=[18,23),
      Gender=Male,
      Marital.Status=Single,
      Occupation=Student,
      Monthly.Income=No Income,
      Output=Yes}                          => {Educational.Qualifications=Graduate}  0.06701031  0.7027027 0.09536082 1.5403878    26
[7144] {Gender=Male,
      Marital.Status=Single,
      Occupation=Student,
      Monthly.Income=No Income,
      Educational.Qualifications=Graduate,
      Output=Yes}                          => {Age=[18,23)}                          0.06701031  0.6842105 0.09793814 2.8241881    26
[7145] {Age=[23,25),
      Marital.Status=Single,
      Occupation=Student,
      latitude=[12.96,12.98),
      longitude=[77.58,77.61),
      Pin.code=[5.6e+05,5.6002e+05)}       => {Output=Yes}                           0.05927835  0.9583333 0.06185567 1.2353267    23
[7146] {Age=[23,25),
      Occupation=Student,
      latitude=[12.96,12.98),
      longitude=[77.58,77.61),
      Pin.code=[5.6e+05,5.6002e+05),
      Output=Yes}                          => {Marital.Status=Single}                0.05927835  1.0000000 0.05927835 1.4477612    23
[7147] {Age=[23,25),
      Marital.Status=Single,
      latitude=[12.96,12.98),
      longitude=[77.58,77.61),
      Pin.code=[5.6e+05,5.6002e+05),
      Output=Yes}                          => {Occupation=Student}                   0.05927835  1.0000000 0.05927835 1.8743961    23
[7148] {Age=[23,25),
      Marital.Status=Single,
      Occupation=Student,
      latitude=[12.96,12.98),
      Pin.code=[5.6e+05,5.6002e+05),
      Output=Yes}                          => {longitude=[77.58,77.61)}              0.05927835  0.9200000 0.06443299 2.7042424    23
```

d. สร้าง Strong Association Rule ที่มีConfidence มากไปน้อย จำนวน 10-20 กฎ

```r
install.packages("arules")
library(arules)

# อ่านไฟล์ CSV
onlinefood <- read.csv("C:/Users/Admin/Downloads/datasci+prj/onlinefoods.csv", header = TRUE, sep = ",")

# สร้าง Strong Association Rule
rules <- apriori(onlinefood, parameter = list(supp = 0.05, conf = 0.6, minlen = 2, maxlen = 10, target = "rules"))

# เลือกกฎที่มี Confidence มากไปน้อย
rules_sorted <- sort(rules, by = "confidence", decreasing = TRUE)

# แสดง 10-20 กฎแรก
inspect(rules_sorted[0:20])
```

```
> # แสดง 10-20 กฎแรก
> inspect(rules_sorted[0:20])
     lhs                                                                  rhs                          support    confidence coverage   lift     count
[1]  {Monthly.Income=More than 50000, longitude=[77.61,77.76]}         => {Age=[25,33]}                0.05670103 1          0.05670103 2.269006 22
[2]  {Occupation=Student, Output=No}                                   => {Marital.Status=Single}      0.05927835 1          0.05927835 1.447761 23
[3]  {Age=[18,23), Family.size=[3,4)}                                  => {Output=Yes}                 0.06443299 1          0.06443299 1.289037 25
[4]  {Age=[18,23), latitude=[12.87,12.96)}                             => {Output=Yes}                 0.09020619 1          0.09020619 1.289037 35
[5]  {Age=[18,23), longitude=[77.61,77.76]}                            => {Output=Yes}                 0.06701031 1          0.06701031 1.289037 26
[6]  {Age=[18,23), Pin.code=[5.6005e+05,5.6011e+05]}                   => {Output=Yes}                 0.07731959 1          0.07731959 1.289037 30
[7]  {Age=[18,23), Pin.code=[5.6002e+05,5.6005e+05)}                   => {Marital.Status=Single}      0.07731959 1          0.07731959 1.447761 30
[8]  {Age=[18,23), longitude=[77.58,77.61)}                            => {Marital.Status=Single}      0.07731959 1          0.07731959 1.447761 30
[9]  {Age=[18,23), Family.size=[4,6]}                                  => {Marital.Status=Single}      0.09536082 1          0.09536082 1.447761 37
[10] {Age=[18,23), Family.size=[4,6]}                                  => {Output=Yes}                 0.09536082 1          0.09536082 1.289037 37
[11] {Age=[18,23), Gender=Female}                                      => {Marital.Status=Single}      0.09278351 1          0.09278351 1.447761 36
[12] {Age=[18,23), Educational.Qualifications=Post Graduate}           => {Occupation=Student}         0.08762887 1          0.08762887 1.874396 34
[13] {Age=[18,23), Educational.Qualifications=Post Graduate}           => {Marital.Status=Single}      0.08762887 1          0.08762887 1.447761 34
[14] {Age=[18,23), Educational.Qualifications=Post Graduate}           => {Output=Yes}                 0.08762887 1          0.08762887 1.289037 34
[15] {Age=[18,23), Monthly.Income=No Income}                           => {Occupation=Student}         0.18298969 1          0.18298969 1.874396 71
[16] {Monthly.Income=No Income, Family.size=[1,3)}                     => {Occupation=Student}         0.15979381 1          0.15979381 1.874396 62
[17] {Occupation=Student, latitude=[12.96,12.98)}                      => {Marital.Status=Single}      0.18298969 1          0.18298969 1.447761 71
[18] {Occupation=Student, Pin.code=[5.6002e+05,5.6005e+05)}            => {Marital.Status=Single}      0.17783505 1          0.17783505 1.447761 69
[19] {Occupation=Student, longitude=[77.58,77.61)}                     => {Marital.Status=Single}      0.21391753 1          0.21391753 1.447761 83
[20] {Monthly.Income=No Income, Educational.Qualifications=Post Graduate} => {Occupation=Student}      0.25515464 1          0.25515464 1.874396 99
>
```