Context and Perspective

Richard works for a large online retailer. His company is launching a next-generation eReader soon, and they want to maximize the effectiveness of their marketing. They have many customers, some of whom purchased one of the company's previous generation digital readers. Richard has

noticed that certain types of people were the most anxious to get the previous generation device, while other folks seemed to content to wait to buy the electronic gadget later. He's wondering what makes some people motivated to buy something as soon as it comes out, while others are less

driven to have the product. Richard's employer helps to drive the sales of its new eReader by offering specific products and services for the eReader through its massive web site—for example, eReader owners can use the company's web site to buy digital magazines, newspapers, books, music, and so forth. The company also sells thousands of other types of media, such as traditional printed books and electronics of every kind. Richard believes that by mining the customers' data regarding general consumer behaviors on the web site, he'll be able to figure out which customers will buy the new eReader early, which ones will buy next, and which ones will buy later on. He hopes that by predicting when a customer will be ready to buy the next-gen eReader, he'll be able to time his target marketing to the people most ready to respond to advertisements and promotions

Organizational Understanding

Richard wants to be able to predict the timing of buying behaviors, but he also wants to understand how his customers' behaviors on his company's web site indicate the timing of their purchase of the new eReader. Richard has studied the classic diffusion theories that noted scholar

and sociologist Everett Rogers first published in the 1960s. Rogers surmised that the adoption of a new technology or innovation tends to follow an 'S' shaped curve, with a smaller group of the most enterprising and innovative customers adopting the technology first, followed by larger

groups of middle majority adopters, followed by smaller groups of late adopters. Those at the front of the blue curve are the smaller group that are first to want and buy the technology. Most of us, the masses, fall within the middle 70-80% of people who eventually acquire the technology. The low end tail on the right side of the blue curve are the laggards, the ones who eventually adopt. Consider how DVD players and cell phones have followed this curve. Understanding Rogers' theory, Richard believes that he can categorize his company's customers into one of four groups that will eventually buy the new eReader: Innovators, Early Adopters, Early Majority or Late Majority. These groups track with Rogers' social adoption theories on the diffusion of technological innovations, and also with Richard's informal observations about the speed of adoption of his company's previous generation product. He hopes that by watching the customers' activity on the company's web site, he can anticipate approximately when each person will be most likely to buy an eReader. He feels like data mining can help him figure out which activities are the best predictors of which category a customer will fall into. Knowing this, he can time his marketing to each customer to coincide with their likelihood of buying

Data Understanding

Richard has engaged us to help him with his project. We have decided to use a decision tree model in order to find good early predictors of buying behavior. Because Richard's company does all of its business through its web site, there is a rich data set of information for each customer,

including items they have just browsed for, and those they have actually purchased. He has prepared two data sets for us to use. The training data set contains the web site activities of customers who bought the company's previous generation reader, and the timing with which they

bought their reader. The second is comprised of attributes of current customers which Richard hopes will buy the new eReader. He hopes to figure out which category of adopter each person in the scoring data set will fall into based on the profiles and buying timing of those people in the

training data set. In analyzing his data set, Richard has found that customers' activity in the areas of digital media and books, and their general activity with electronics for sale on his company's site, seem to have a lot in common with when a person buys an eReader. With this in mind, we have worked with Richard to compile data sets comprised of the following attributes:

- User_ID: A numeric, unique identifier assigned to each person who has an account on the company's web site.

- Gender: The customer's gender, as identified in their customer account. In this data set, it is recorded a 'M' for male and 'F' for Female. The Decision Tree operator can handle non-numeric data types.

- Age: The person's age at the time the data were extracted from the web site's database. This is calculated to the nearest year by taking the difference between the system date and the person's birth date as recorded in their account.

- Marital_Status: The person's marital status as recorded in their account. People who indicated on their account that they are married are entered in the data set as 'M'. Since the web site does not distinguish single types of people, those who are divorced or widowed are included with those who have never been married (indicated in the data set as 'S').

- Website_Activity: This attribute is an indication of how active each customer is on the company's web site. We used the web site database's information which records the duration of each customers visits to the web site to calculate how frequently, and for how long each time, the customers use the web site. This is then translated into one of three categories: Seldom, Regular, or Frequent.

- Browsed_Electronics_12Mo: This is simply a Yes/No column indicating whether or not the person browsed for electronic products on the company's web site in the past year.

- Bought_Electronics_12Mo: Another Yes/No column indicating whether or not they purchased an electronic item through Richard's company's web site in the past year.

 Bought_Digital_Media_18Mo: This attribute is a Yes/No field indicating whether or not the person has purchased some form of digital media (such as MP3 music) in the past year and a half. This attribute does not include digital book purchases.

 Bought_Digital_Books: Richard believes that as an indicator of buying behavior relative to the company’s new eReader, this attribute will likely be the best indicator. Thus, this attribute has been set apart from the purchase of other types of digital media. Further, this attribute indicates whether or not the customer has ever bought a digital book, not just in the past year or so.

 Payment_Method: This attribute indicates how the person pays for their purchases. In cases where the person has paid in more than one way, the mode, or most frequent method of payment is used. There are four options:

 Bank Transfer—payment via e-check or other form of wire transfer directly from the bank to the company.

 Website Account—the customer has set up a credit card or permanent electronic funds transfer on their account so that purchases are directly charged through their account at the time of purchase.

 Credit Card—the person enters a credit card number and authorization each time they purchase something through the site.

 Monthly Billing—the person makes purchases periodically and receives a paper or electronic bill which they pay later either by mailing a check or through the company web site’s payment system.

eReader_Adoption: It consists of data for customers who purchased the previous-gen eReader. Those who purchased within a week of the product’s release are recorded in this attribute as ‘Innovator’. Those who purchased after the first week but within the second or third weeks are entered as ‘Early Adopter’. Those who purchased after three weeks but within the first two months are ‘Early Majority’. Those who purchased after the first two months are ‘Late Majority’.

Data Preparation:

Import Data:

Import both the training and scoring datasets into RapidMiner.

Data Cleaning:

Handle missing values if any.

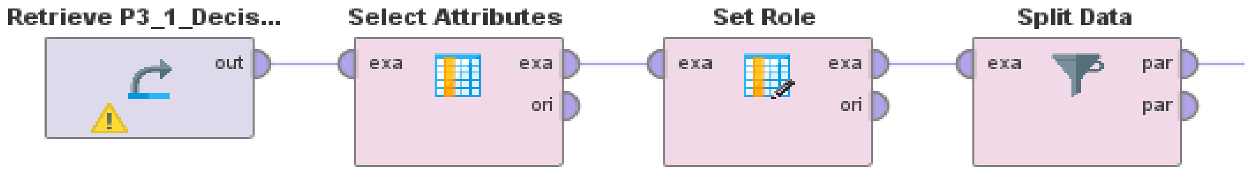
Encode categorical variables (Gender, Marital_Status, Website_Activity, Payment_Method) into numerical values.

Feature Selection:

Analyze the importance of features using techniques like Information Gain, and select relevant features for modeling.

Data Splitting:

Split the training data into training and validation sets for model evaluation.



Modeling:

Decision Tree Model:

Use the Decision Tree operator in RapidMiner.

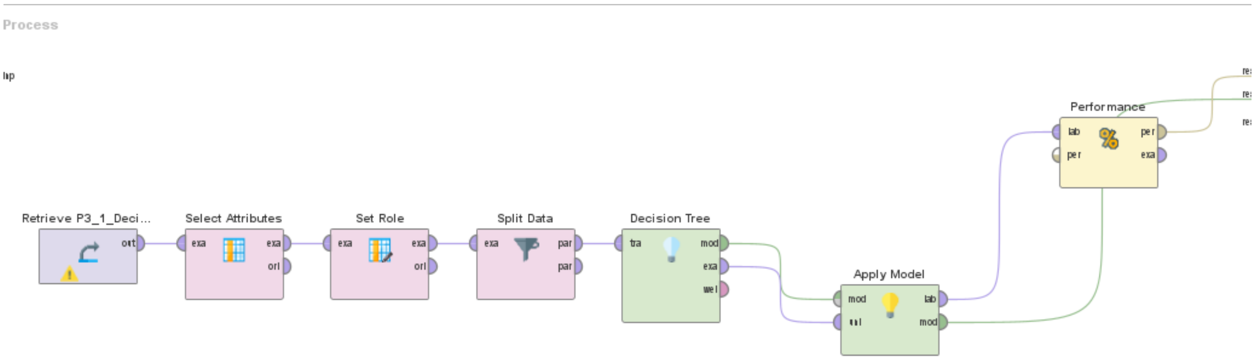
Train the model using the training data.

Optimize hyperparameters if necessary, like tree depth, minimum leaf size, etc.

Evaluate Different Algorithms:

Experiment with different tree algorithms (e.g., C4.5, CART, CHAID) to see which one gives the best results.

Compare their performance using appropriate evaluation metrics.



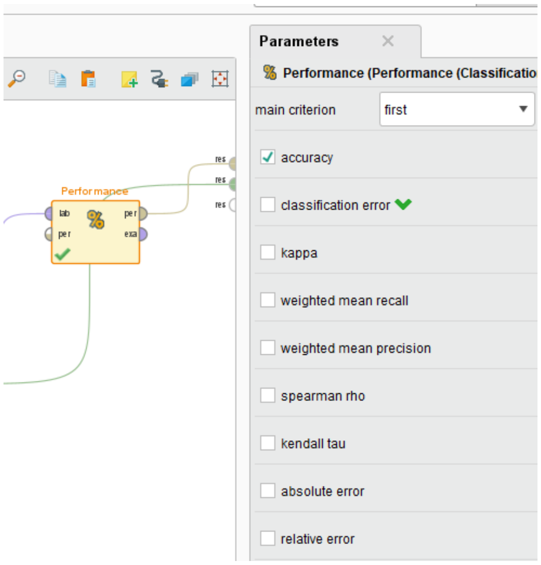
Evaluation:

Model Evaluation:

Evaluate the decision tree model using metrics like accuracy, precision, recall, and F1-score on the validation set.

Fine-tuning:

If the model performance is not satisfactory, consider fine-tuning the hyperparameters or trying different algorithms.



Parameters

Performance (Performance (Classificatio

main criterion first

☑ accuracy

□ classification error ✓

□ kappa

□ weighted mean recall

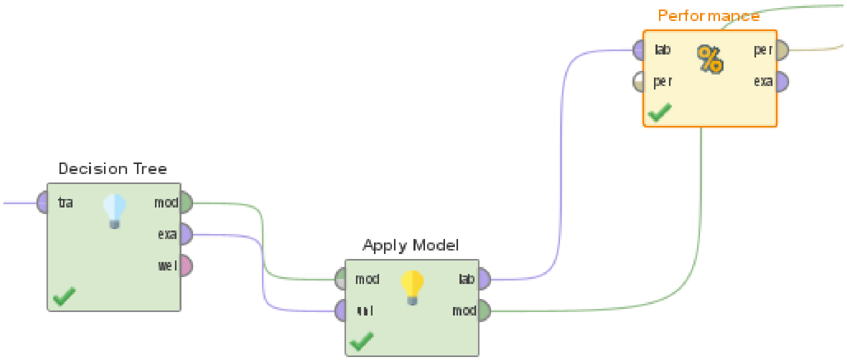
□ weighted mean precision

□ spearman rho

□ kendall tau

□ absolute error

□ relative error



Deployment:

Scoring:

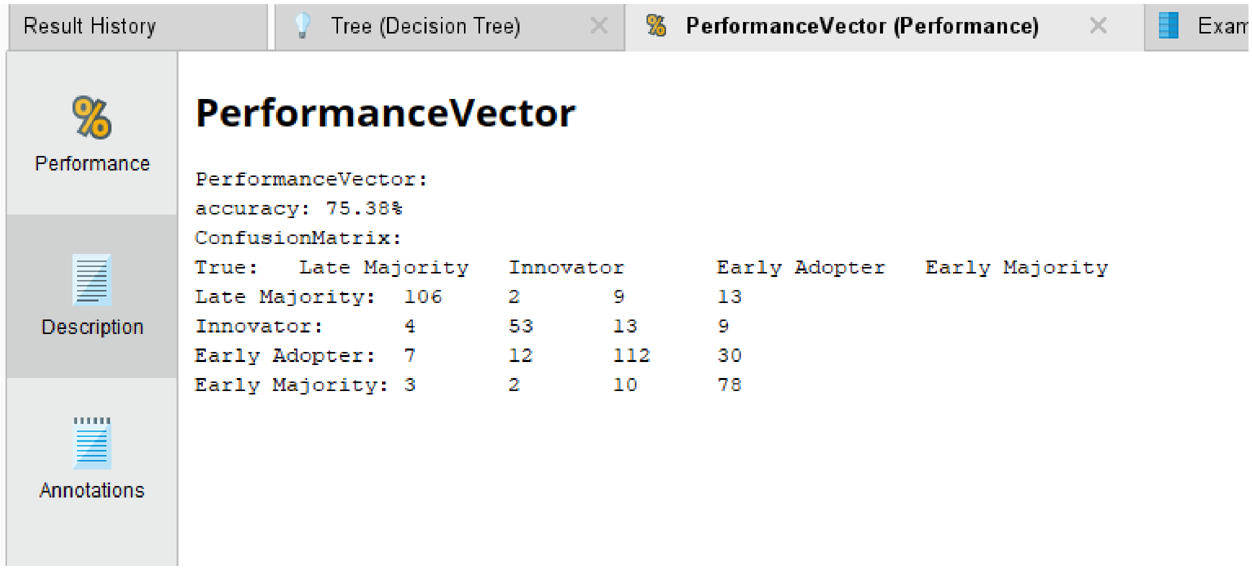
Use the trained model to predict the eReader adoption categories for the scoring dataset (new customers).

Interpretation:

- **Late Majority:**
  - **True Positives (106):** The model correctly predicted 106 customers as "Late Majority."
  - **False Negatives (24):** The model incorrectly predicted 24 customers as other categories when they were actually "Late Majority."
- **Innovator:**
  - **True Positives (53):** The model correctly predicted 53 customers as "Innovators."
  - **False Negatives (26):** The model incorrectly predicted 26 customers as other categories when they were actually "Innovators."
- **Early Adopter:**

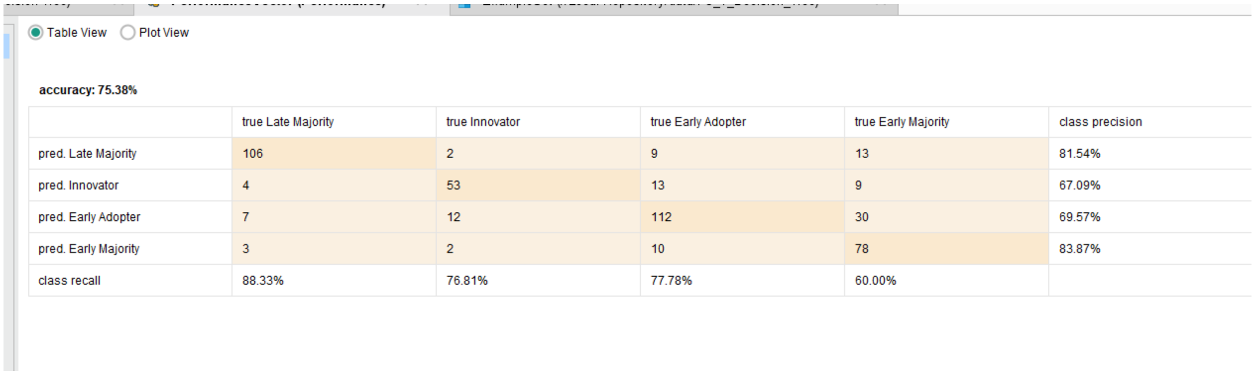
    - **True Positives (112):** The model correctly predicted 112 customers as "Early Adopters."
    - **False Negatives (42):** The model incorrectly predicted 42 customers as other categories when they were actually "Early Adopters."

- **Early Majority:**
    - **True Positives (78):** The model correctly predicted 78 customers as "Early Majority."
    - **False Negatives (15):** The model incorrectly predicted 15 customers as other categories when they were actually "Early Majority."



PerformanceVector

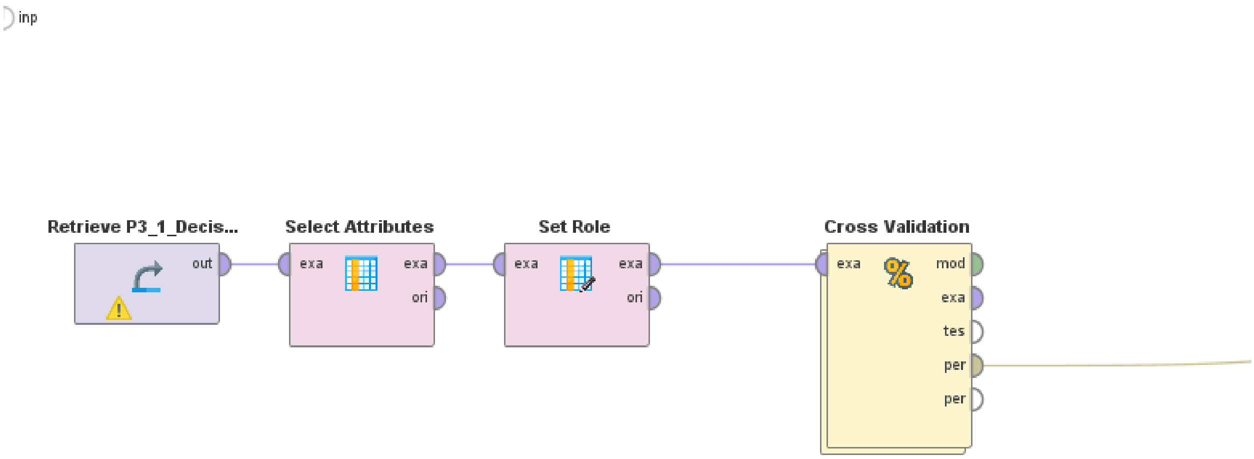
```
PerformanceVector:
accuracy: 75.38%
ConfusionMatrix:
True:   Late Majority   Innovator       Early Adopter   Early Majority
Late Majority:  106     2       9       13
Innovator:      4       53      13      9
Early Adopter:  7       12      112     30
Early Majority: 3       2       10      78
```

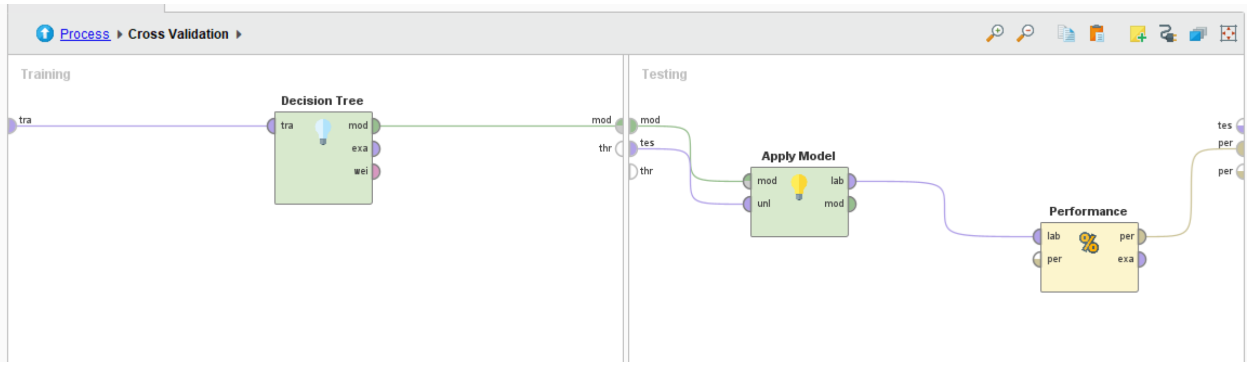


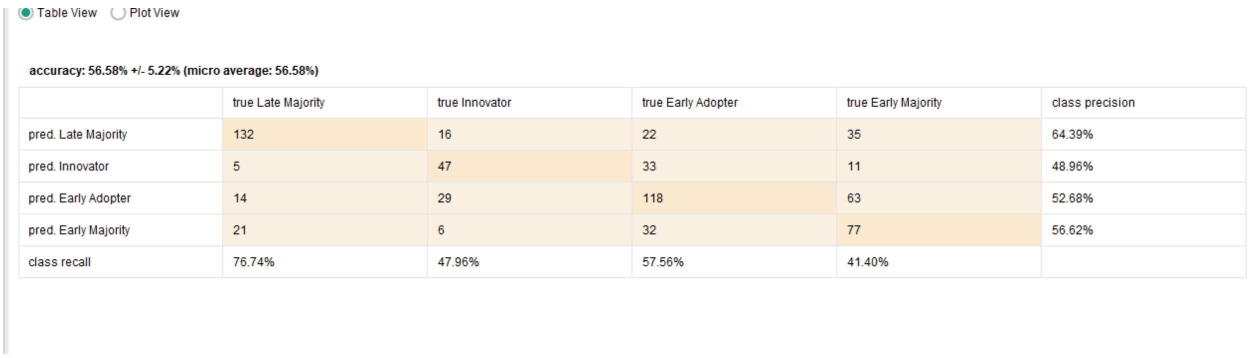
accuracy: 75.38%

	true Late Majority	true Innovator	true Early Adopter	true Early Majority	class precision
pred. Late Majority	106	2	9	13	81.54%
pred. Innovator	4	53	13	9	67.09%
pred. Early Adopter	7	12	112	30	69.57%
pred. Early Majority	3	2	10	78	83.87%
class recall	88.33%	76.81%	77.78%	60.00%	

Cross validation:

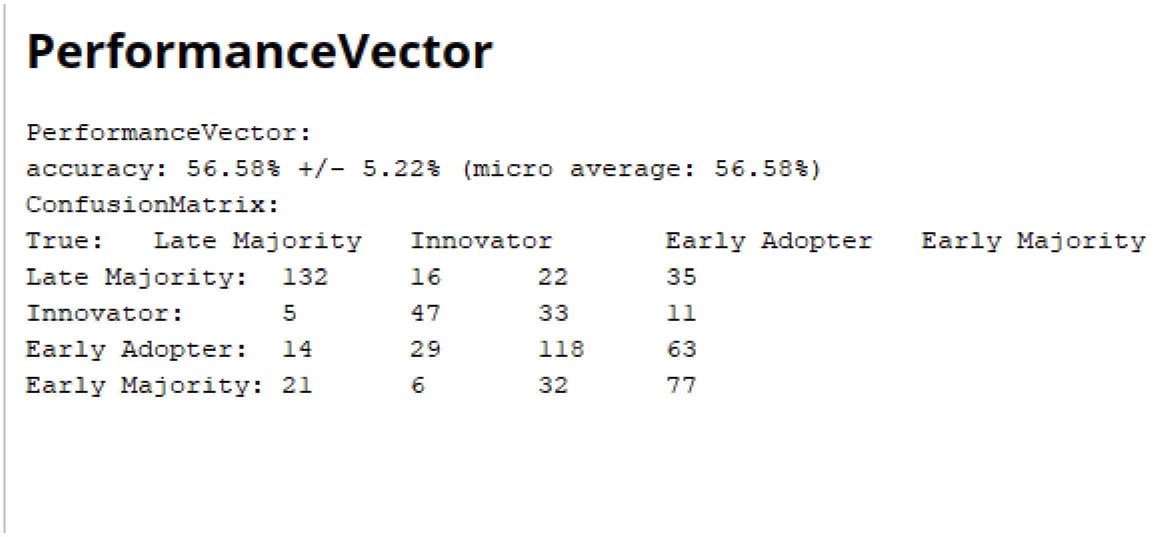

inp
Retrieve P3_1_Decis...
out
Select Attributes
exa
exa
ori
Set Role
exa
exa
ori
Cross Validation
exa
mod
exa
tes
per
per


Process
Cross Validation
Training
tra
Decision Tree
tra
mod
exa
wei
mod
thr
Testing
mod
tes
thr
Apply Model
mod
unl
lab
mod
Performance
lab
per
per
exa
tes
per
per



accuracy: 56.58% +/- 5.22% (micro average: 56.58%)

	true Late Majority	true Innovator	true Early Adopter	true Early Majority	class precision
pred. Late Majority	132	16	22	35	64.39%
pred. Innovator	5	47	33	11	48.96%
pred. Early Adopter	14	29	118	63	52.68%
pred. Early Majority	21	6	32	77	56.62%
class recall	76.74%	47.96%	57.56%	41.40%	



PerformanceVector

```
PerformanceVector:
accuracy: 56.58% +/- 5.22% (micro average: 56.58%)
ConfusionMatrix:
True:    Late Majority   Innovator       Early Adopter   Early Majority
Late Majority:   132     16      22      35
Innovator:       5       47      33      11
Early Adopter:   14      29      118     63
Early Majority:  21      6       32      77
```

Accuracy:

The model's accuracy is 56.58%, with a micro-average of 56.58%. The micro-average considers the total true positives, true negatives, false positives, and false negatives across all classes and calculates the accuracy.

Interpretation:

- **Late Majority:**
  - **True Positives (132):** The model correctly predicted 132 customers as "Late Majority."
  - **False Negatives (73):** The model incorrectly predicted 73 customers as other categories when they were actually "Late Majority."
- **Innovator:**
  - **True Positives (47):** The model correctly predicted 47 customers as "Innovators."
  - **False Negatives (49):** The model incorrectly predicted 49 customers as other categories when they were actually "Innovators."
- **Early Adopter:**

  - **True Positives (118):** The model correctly predicted 118 customers as "Early Adopters."
  - **False Negatives (92):** The model incorrectly predicted 92 customers as other categories when they were actually "Early Adopters."
- **Early Majority:**
  - **True Positives (77):** The model correctly predicted 77 customers as "Early Majority."
  - **False Negatives (58):** The model incorrectly predicted 58 customers as other categories when they were actually "Early Majority."