

PRACTICAL-1

Aim: Demonstration of preprocessing on dataset student.arff

Description :

Preprocessing the student.arff dataset involves loading the data, handling missing values, and encoding categorical variables. Additionally, normalization and feature selection are applied to prepare the dataset for analysis. Finally, the dataset is split into training and testing sets to evaluate model performance.

Solution :

Part :- 1

- ① Relation student
- ② attribute id numeric
- ③ attribute name string
- ④ attribute gender {male, female}
- ⑤ attribute hobby string
- ⑥ attribute age numeric
- ⑦ data
 - 1, Rabit, male, writing, 19
 - 2, gerutam, male, cricket, 20
 - 3, hitesh, male, reading, 20
 - 4, hina, female, dancing, 20
 - 5, khushree, female, Singing, 20



**SILVER OAK
UNIVERSITY**
EDUCATION TO INNOVATION

Part :- 2.

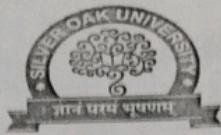
Step 1 => Open The weka Software

Step 2 => Click on The 'Explorer' Button

Step 3 => Click on The 'open file' Button

Step 4 => Choose your ".arff" file

Step 5 => Analyse your output.



SILVER OAK
UNIVERSITY
EDUCATION TO INNOVATION

Part :- 3

Out Put
====

Age

Statistic	Value
minimum	19
maximum	20
Mean	19.6
Std Dev	0.548



**SILVER OAK
UNIVERSITY**
EDUCATION TO INNOVATION

Post Practical Questions:

1) Which of the following is the first step in preprocessing the student.arff dataset?

- A) Encoding categorical variables B) Normalization
 C) Loading the data D) Splitting the dataset into training and testing sets

2) What is one common technique for handling missing values in the student.arff dataset?

- A) Removing rows with missing values B) Adding new rows with random values
C) Ignoring the missing values D) Duplicating rows with missing values

3) Which method is used to convert categorical variables into a numerical format?

- A) Standardization B) One-hot encoding
C) Min-Max scaling D) Imputation

References :

<https://www.geeksforgeeks.org/data-mining/>

<https://www.simplilearn.com/what-is-data-mining-article>

Conclusion:

Signature with Date of Completion	
-----------------------------------	--

PRACTICAL-2

Aim: Demonstration of preprocessing on dataset labor.arff

Description :

Preprocessing the labor.arff dataset involves loading the data, handling missing values, and encoding categorical variables. Additionally, normalization and feature selection are applied to prepare the dataset for analysis. Finally, the dataset is split into training and testing sets to evaluate model performance.

Solution :

Part I:-

- ① Relation Labor
- ② attribute id numeric
- ③ attribute name string
- ④ attribute contact numeric
- ⑤ attribute age numeric
- ⑥ attribute gender {male, female}
- ⑦ attribute wage numeric
- ⑧ data
 - 1, dhruv, 7112131415, 30, male, 8000
 - 2, abdul, 2122232425, 32, male, 7000
 - 3, Salman, 3132333435, 37, male, 7500
 - 4, Salma, 4142434445, 35, female, 6500
 - 5, ruli, 5152535455, 32, female, 6000



SILVER OAK
UNIVERSITY
EDUCATION TO INNOVATION

Part 2 :-

steps for weka

step 1 => open the weka software

step 2 = click on the 'Explorers' button

step 3 => click on the 'open file' button

step 4 => choose your '.arff' file

step 5 => Analyse your output

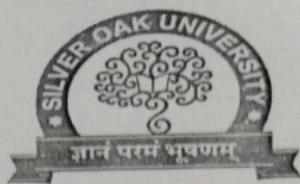


Part :- 3

Output.

Name :- wuge

Statistic	Value
minimum	6000
Maximum	8000
mean	7000
Std Dev	790.569



**SILVER OAK
UNIVERSITY**
EDUCATION TO INNOVATION

Post Practical Questions:

1) Why is normalization applied to numerical features during preprocessing?

- A) To delete irrelevant features B) To standardize the range of numerical features
C) To convert numerical data into categorical data D) To increase the size of the dataset

2) What is the purpose of splitting the student.arff dataset into training and testing sets?

- A) To reduce the size of the dataset B) To encode categorical variables
 C) To evaluate model performance D) To handle missing values

References :

<https://www.geeksforgeeks.org/data-mining/>

<https://www.simplilearn.com/what-is-data-mining-article>

Conclusion:

Signature with Date of Completion	
Marks out of 10	



Practical 3

Aim : Demonstration of Association rule process on dataset contact-lenses.arff using Apriori algorithm

Description :

The association rule process begins by loading the contact-lenses.arff dataset, which contains information on various attributes related to contact lens prescriptions. The **Apriori algorithm** is then applied to identify frequent itemsets by iteratively exploring subsets of the data, ensuring that the support threshold is met. After frequent itemsets are identified, the algorithm generates strong association rules that satisfy minimum confidence levels, highlighting relationships between different attributes.

Solution :

Part :- 1

- ① relation Contact-lenses
- ② attribute age { young, Pre-Presbyopic, Presbyopic }
- ③ attribute spectacle-prescription { myope, hypermetropic }
- ④ attribute astigmatism { no, yes }
- ⑤ attribute tear-prod-rate { reduced, normal }
- ⑥ attribute Contact-lenses { soft, hard, none }

⑦ data

young, myope, no, reduced, soft

pre-Presbyopic, hypermetropic, yes, normal, hard

Presbyopic, myope, no, reduced, none.



- young, hypermetropic, yes, normal, hard,
- pre-presbyopic, myope, no, reduced, none
- Presbyopic, hypermetropic, yes, normal, soft
- pre-presbyopic, myope, yes, reduced, soft
- young, myope, yes, reduced, none
- presbyopic, hypermetropic, no, reduced, none
- young, myope, no, reduced, none

Part :- 2

Step = 1 \Rightarrow open the weka software

1.1 \Rightarrow click on Explorer Button

1.2 \Rightarrow click on the open file Button

1.3 \Rightarrow choose your .arff file

Step 2 \Rightarrow select 'Associate' option from toolbar

2.1 \Rightarrow click on "choose" button

2.2 \Rightarrow choose 'aPriori' Algorithm

Step 3 \Rightarrow click on "start" button

Step 4 \Rightarrow , Analyse your output

Part :- 3

Output :-

A Priori

minimum support : 0.35 (3 instances)

minimum metric <Confidence> : 0.9

Number of cycles performed : 13

Generated sets of large itemsets :-

size of set of large itemsets L(1) :- 10

size of set of large itemsets L(2) :- 19

size of set of large itemsets L(3) : 17

size of set of large itemsets L(4) : 2

Best rules found :-

1. age = young S \Rightarrow Spectacle - Prescip = myope 5
2. Contact -lenses = soft S \Rightarrow Spectacle - Prescip = myope 5
3. Spectacle - Prescip = hy permetropic 6
 \Rightarrow age = Presbyopic 4



Post Practical Questions:

- 1) What is the first step in the association rule process using the Apriori algorithm on the contactlenses.arff dataset?

A) Generating association rules
C) Identifying frequent itemsets
 B) Loading the dataset
D) Evaluating the rules

- 2) Which algorithm is used to identify frequent itemsets in the contactlenses.arff dataset?

A) K-means
 C) Apriori
B) Decision Tree
D) SVM

- 3) What criterion must be met for itemsets to be considered frequent in the Apriori algorithm?

A) Minimum confidence
 C) Minimum support
B) Maximum confidence
D) Maximum support

References :

<https://www.geeksforgeeks.org/data-mining/>

<https://www.simplilearn.com/what-is-data-mining-article>

Conclusion:

Signature with Completion	Date of
Marks out of 10	



Practical 4

Aim: Demonstration of Association rule process on dataset test.arff using apriori algorithm.

Description :

The association rule process begins by loading the contact-lenses.arff dataset, which contains information on various attributes related to contact lens prescriptions. The Apriori algorithm is then applied to identify frequent itemsets by iteratively exploring subsets of the data, ensuring that the support threshold is met. After frequent itemsets are identified, the algorithm generates strong association rules that satisfy minimum confidence levels, highlighting relationships between different attributes.

Solution :

Part :- I

- @ relation test
- @ attribute milk {yes, no}
- @ attribute bread {yes, no}
- @ attribute jelly {yes, no}
- @ attribute butter {yes, no}
- @ attribute sugar {yes, no}
- @ attribute egg {yes, no}

@ data

- yes, yes, no, yes, no, no
- yes, yes, no, no, no, yes
- no, no, yes, no, yes, no
- yes, yes, no, yes, yes, yes
- yes, yes, no, no, yes, yes
- yes, no, yes, yes, yes, no
- no, no, no, yes, yes, yes
- no, yes, yes, no, no, no
- yes, yes, no, yes, yes, no
- no, yes, no, yes, no, yes



SILVER OAK
UNIVERSITY
EDUCATION TO INNOVATION

Part :- 2.

Step :- 1 \Rightarrow open The weku Software

1.1 \Rightarrow Click on The "Explorer" Button

1.2 \Rightarrow Click on The "open file" Button

1.3 \Rightarrow choose your ".arff" file.

step :- 2 \Rightarrow select "Associate" option from toolbar

2.1 \Rightarrow Click on "choose" button

Choose appropriate Algorithm
(Here "Apriori" Algorithm)

step 3 \Rightarrow click on "start" Button

step 4 \Rightarrow Analyse your output.



Pcfst :- 3 Output

A Priori

minimum support : 0.35 (3 instances)

minimum metric < Confidence > : 0.9

Number of cycles performed : 13

Generated sets of large itemsets:

size of set of large itemsets L(1) : 12

size of set of large itemsets L(2) : 24

size of set of large itemsets L(3) : 16

size of set of large itemsets L(4) : 3

Best rule(s) found :-

1. Egg = yes \Rightarrow jelly = no 5

2. milk = yes jelly = no 5 \Rightarrow bread = yes 5

3. milk = yes bread = yes \Rightarrow jelly = no 5

4. sugar = no \Rightarrow bread = yes 4

5. bread = yes butter = yes \Rightarrow jelly = no 4

6. bread = yes egg = yes 4 \Rightarrow jelly = no 4

7. bread = no 3 \Rightarrow sugar = yes 3

8. jelly = yes 3 \Rightarrow egg = no 3

9. bread = yes sugar = yes 3 \Rightarrow milk = yes 3

10. milk = yes egg = yes 3 \Rightarrow bread = yes 3



**SILVER OAK
UNIVERSITY**
EDUCATION TO INNOVATION

Post Practical Questions:

1) After identifying frequent itemsets, what is the next step in the Apriori algorithm?

- A) Normalizing the data B) Splitting the dataset
 C) Generating strong association rules D) Encoding categorical variables

2) What is the main purpose of evaluating the generated association rules in the test.arff dataset?

- A) To convert numerical data to categorical data B) To understand patterns and correlations
C) To reduce the size of the dataset D) To remove outliers

References :

<https://www.geeksforgeeks.org/data-mining/>

<https://www.simplilearn.com/what-is-data-mining-article>

Conclusion:

Signature with Date of Completion	
Marks out of 10	