Experiment 1.1

Student Name: Kanishk Soni **UID: 20BCS9398**
Branch: BE-CSE **Section/Group: 20BCS-DM_708B**
Semester: 6th **Date of Performance: 06-03-2023**
Subject Name: Data Mining Lab **Subject Code: 20CSP-376**

1. Aim:

Demonstration of pre-processing on .arff file using R Programming.

2. Objective:

To represent the creation of file using R Studio and displaying the pattern on Weka Tool forfurther extraction and analysis of knowledge.

3. Code and Output:

Program Code:

```
#creating variables to store similar kind of data also known as vectorsroll <- 1:7
studs <- c("AAA","BBB","CCC","DDD","EEE","FFF","GGG")
marks <- c(44,49,37,41,29,32,45)
status <- c("P", "P", "F", "P", "F", "F", "P")

#converting vectors into factors
status_factor <- factor(status)

#creating a data frame out of declared vectors
df <- data.frame(roll, studs, marks, status_factor, stringAsFactors=FALSE)print(df)
```

```
#checking the class of vectors
class(studs)
class(marks)

print(status_factor)

#checking for the factor variable if it is a factor
print(is.factor(status_factor))
#checking the levels of the factor
levels(status_factor)
#summarizing the factor variable to check the counts of level
summary(status_factor)

print(class(df))
#str() shows the structure of the objects created in r. it is an alternativeto display the summary of the objects
print(str(df))
print(summary(df))

#loading up the RWeka library into the session
library(RWeka)
#writing an arff file
write.arff(df,file="D:\\College\\Sem6\\DM_Lab\\Ex1_df.arff")
```

Console:

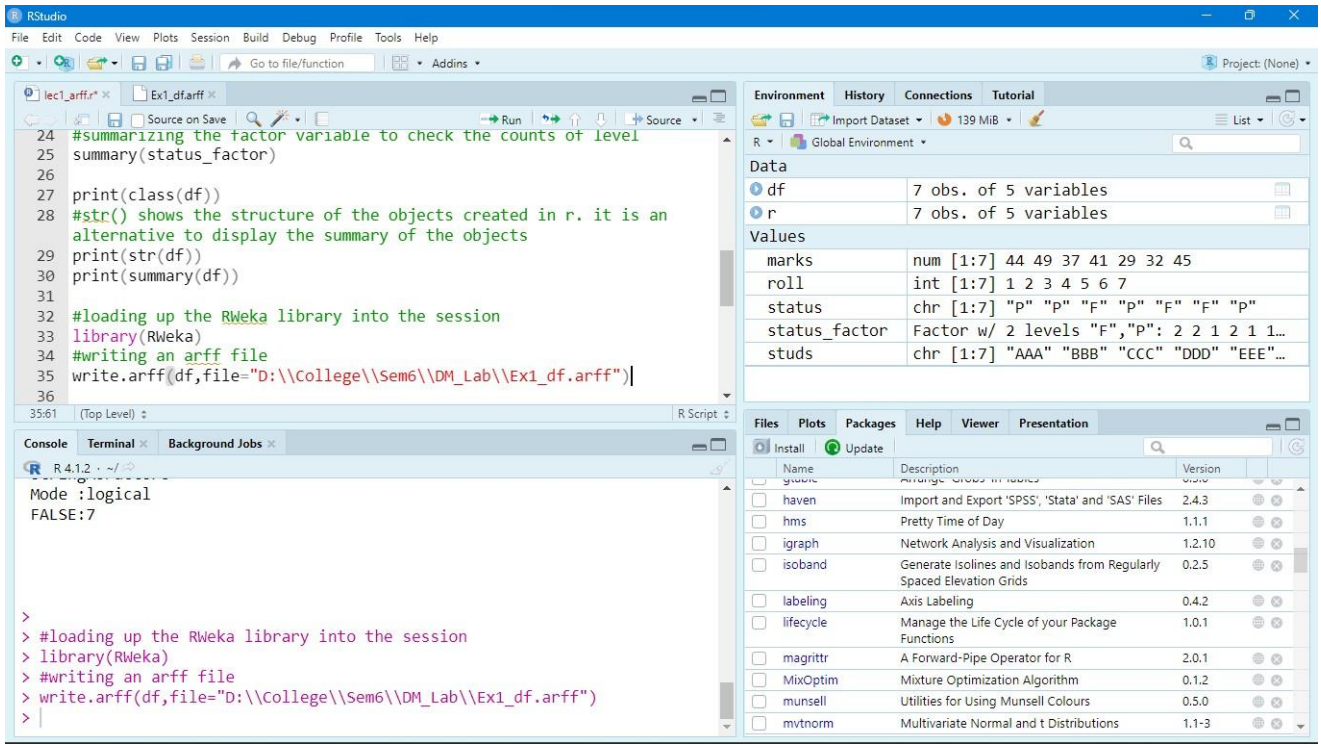
```
> #creating variables to store similar kind of data also known as vectors
> roll <- 1:7
> studs <- c("AAA","BBB","CCC","DDD","EEE","FFF","GGG")
> marks <- c(44,49,37,41,29,32,45)
> status <- c("P", "P", "F", "P", "F", "F", "P")
>
> #converting vectors into factors
```

```
> status_factor <- factor(status)

>
> #creating a data frame out of declared vectors
> df <- data.frame(roll, studs, marks, status_factor, stringAsFactors=FALSE)
> print(df)
  roll studs marks status_factor stringAsFactors
1    1   AAA    44             P           FALSE
2    2   BBB    49             P           FALSE
3    3   CCC    37             F           FALSE
4    4   DDD    41             P           FALSE
5    5   EEE    29             F           FALSE
6    6   FFF    32             F           FALSE
7    7   GGG    45             P           FALSE
>
> #checking the class of vectors
> class(studs)
[1] "character"

> class(marks)
[1] "numeric"

>
> print(status_factor)
[1] P P F P F F P
Levels: F P

>
> #checking for the factor variable if it is a factor
> print(is.factor(status_factor))
[1] TRUE

> #checking the levels of the factor
> levels(status_factor)[1]
"F" "P"

> #summarizing the factor variable to check the counts of level
> summary(status_factor)
F P
3 4

>
> print(class(df))
```

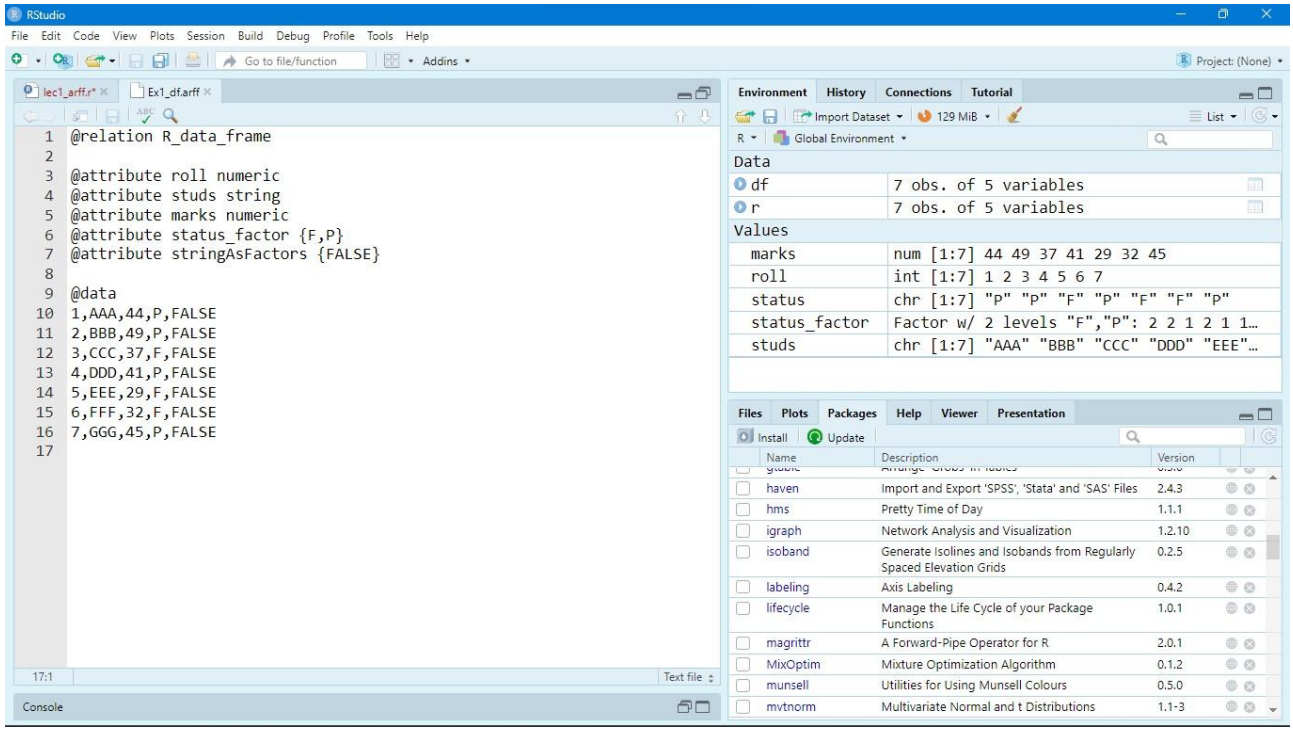
```
[1] "data.frame"
> #str() shows the structure of the objects created in r. it is an alternative to
display the summary of the objects
> print(str(df))
'data.frame': 7 obs. of  5 variables:
 $ roll           : int  1 2 3 4 5 6 7
 $ studs          : chr  "AAA" "BBB" "CCC" "DDD" ...
 $ marks          : num  44 49 37 41 29 32 45
 $ status_factor  : Factor w/ 2 levels "F","P": 2 2 1 2 1 1 2
 $ stringAsFactors: logi FALSE FALSE FALSE FALSE FALSE FALSE ...NULL

> print(summary(df))
      roll            studs               marks          status_factor
 Min.   :1.0   Length:7           Min.   :29.00   F:3
 1st Qu.:2.5   Class  :character   1st Qu.:34.50   P:4
 Median :4.0   Mode   :character   Median :41.00
 Mean   :4.0                      Mean   :39.57
 3rd Qu.:5.5                      3rd Qu.:44.50
 Max.   :7.0                      Max.   :49.00
 stringAsFactors
 Mode :logical
 FALSE:7


> #loading up the RWeka library into the session
> library(RWeka)
> #writing an arff file
> write.arff(df,file="D:\\College\\Sem6\\DM_Lab\\Ex1_df.arff")
```

Output:

RStudio Window:



ARFF File:



4. Conclusion:

ARFF stands for Attribute-Relation File Format. It is an ASCII text file that describes a listof instances sharing a set of attributes. ARFF files have two distinct sections. The first section is the Header Section and it is followed by the Data Section.
The header section contains various information related to the dataset like the name ofthe relation, columns, and type of columns. The header section contains 2 parts Table/ relation and attribute part.
Data section is used to represents the data or entries for available columns. (According to the order in header section data would be inserted).
Data section starts with @data, and this section must be added after Header section. Only single record can be written in single line.