**PROFILING FOR AUTHENTICATION AND AUTHORIZATION**
**CS 773 Data Mining and Security Course Project**

**Submitted**

**Team**

**(UIN 01103978)**

## ABSTRACT:

Data Mining is a process of predicting outcomes based on the patterns and correlations within large data sets. Using a broad range of techniques, we are not only confined to reduce risks in a business but also we can use this information to increase revenues, cut costs, improve customer relationships and much more.

Current project revolves around the user and analysis of the user data to extract useful information and to develop "User Profile". For this project we were provided the data of 19 Users and their daily routine in a department. Example data includes but not limited to the login and logout times of the user, Machines, Printers and Files accessed, Emails sent/received and Resources used. The aim of the project is to develop user profile information based on various patterns such as LogIn and LogOut time of a user, Programmes and Files accessed, Machines and Printers used etc.

Data provided was strictly analyzed using some of the Data Mining Techniques such as Association Rules and Clustering techniques. Additional information provided helped us in creating various user profiles based on the patterns. Users were mapped into clusters based on the user similarities. Assumptions made during this process strictly follow data mining and security concepts. Data provided in Excel was divided based on the column "record type" and was loaded into SQL Server. For this purpose we have used Microsoft SQL Server 2017. K-Means Clustering is done using a data mining tool Weka 3.8.

## INTRODUCTION:

User Profiles for given 19 users were created based on:

i) LogIn Pattern
ii) Program Access Pattern
iii) File Access Pattern
iv) Printer Usage Pattern
v) E-mail Pattern.
vi) Machine Usage Pattern

The data is categorized into three record types with each record type providing a different kind of data useful to develop a login profile for each user. Let's take a look at the given data.

*Type 1 Records:*

Type 1 Records provide us the data related to user logins and logouts.

Various attributes provided in type 1 data are listed below.

1. Record Type
2. User
3. Machine
4. Date
5. Login time
6. logout time
7. Average number of user processes at any time
8. Maximum number of user processes
9. Total keyboard characters typed
10. CPU use (in seconds) by user processes.

*Type 2 Records:*

Type 2 records helps us in identifying the resource usage for the given user.

Various attributes in the type2 records are listed below.

1. Record Type
2. User
3. Machine
4. Date
5. Start time
6. Program
7. Execution time
8. File R – Read/File RW (Read write)/File W (write);
9. Printer
10. Pages printed.

*Type 3 Records:*

Type 3 records help us in identifying the email usage of the users.

Various attributes in type 3 records are listed below.

1. Record Type
2. User
3. Machine
4. Date
5. Start time
6. E-mail Program
7. E-mail address
8. Received (R )/Sent(S)
9. Bytes
10. Attachments

## OBSERVATIONS:

## LogIn Pattern:

For Type 1 records we were provided with LogIn and LogOut times for which we have assigned numerical values by categorizing the given times with 24 hour format starting from 00:00:00 ending with 23:59:59 and assigned each category a numerical value to draw useful conclusions for employees based on LogIn and LogOut times.

We ran some queries on SSMS 2017 and came to some conclusions on creating User Profile based on LogIn Patterns.

**Query 1:**

Query below represents how many days a user have logged in.

```
SELECT [User_Id],
COUNT(*) AS [Total Number of Days User Logged IN]
FROM dbo.DataMiningProject
WHERE Record_Id = 1
GROUP BY User_Id
```

**Query 2:**

Query to Identify Login and Logout Times: Considering normal working hours between 8:00AM and 5:00 PM we have got LogIn and LogOut Times for the users. If the user had logged in between '080000' and '085959' and Loggedout between '170000' and'175959' then we assumed that he is logging in during normal working hours. If this is not the case then the user has logged in or logged out multiple times.

```
SELECT DISTINCT User_Id
, (CASE WHEN Login_Time BETWEEN '080000' AND '085959' THEN 9
        ELSE ''
        END) AS Login_Time
, (CASE WHEN Logout_Time BETWEEN '170000' AND '175959' THEN 18
        ELSE ''
        END) AS LogOut_Time
FROM dbo.DataMiningProject
WHERE Record_Id = 1
GROUP BY User_Id, Login_Time, Logout_Time
```

**Query 3:**

Below query provides us the data where we will know how many times a user logged in during weekdays/weekends.

```
;WITH [Days]
AS
(SELECT  User_Id, (CASE WHEN Date IN ('090608','090708','091308','091408'
                                      , '092008', '092108','092708', '092808')
            THEN 'Weekend'
            ELSE 'Weekday'
        END) AS 'Dayoftheweek'
FROM dbo.DataMiningProject
WHERE Record_Id = 1
)

SELECT User_Id, [Dayoftheweek], COUNT(*) AS [Total Login]
FROM Days
GROUP BY User_Id, [Dayoftheweek]
```

Based on above queries we have created Table1 (LogIn Patterns).

Table1: Login Patterns

| UserId | Tot Num. of days User Logged in | Login Time | Logout Time | Weekday total Login | Weekend Total login | Num. of times user logging IN during Working hours | Num. of times user logging in before/after Working hours | Num. of times user logging OUT during Working hours | Num. of times user logging OUT before/after Working hours |
|---|---|---|---|---|---|---|---|---|---|
| U01 | 22 | 9 | 18 | 22 | 0 | 0 | 22 | 22 | 0 |
| U02 | 22 | Multiple Times | Multiple Times | 22 | 0 | 7 | 15 | 12 | 10 |
| U03 | 22 | 9 | 18 | 22 | 0 | 0 | 22 | 22 | 0 |
| U04 | 22 | Multiple Times | Multiple Times | 22 | 0 | 6 | 16 | 12 | 10 |
| U05 | 22 | 9 | 18 | 22 | 0 | 0 | 22 | 22 | 0 |
| U06 | 22 | Multiple Times | Multiple Times | 22 | 0 | 6 | 16 | 12 | 10 |
| U07 | 22 | 9 | 18 | 22 | 0 | 0 | 22 | 22 | 0 |
| U08 | 22 | Multiple Times | Multiple Times | 22 | 0 | 6 | 16 | 12 | 10 |
| U09 | 22 | 9 | 18 | 22 | 0 | 0 | 22 | 22 | 0 |
| U10 | 22 | Multiple Times | Multiple Times | 22 | 0 | 6 | 16 | 12 | 10 |
| U11 | 22 | 9 | 18 | 22 | 0 | 0 | 22 | 22 | 0 |
| U12 | 22 | Multiple Times | Multiple Times | 22 | 0 | 8 | 14 | 20 | 2 |
| U13 | 22 | Multiple Times | Multiple Times | 22 | 0 | 8 | 14 | 20 | 2 |
| U14 | 22 | Multiple Times | Multiple Times | 22 | 0 | 8 | 14 | 20 | 2 |
| U15 | 21 | Multiple Times | Multiple Times | 21 | 0 | 8 | 13 | 20 | 1 |
| U16 | 22 | Multiple Times | Multiple Times | 22 | 0 | 8 | 14 | 20 | 2 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| U17 | 22 | Multiple Times | Multiple Times | 22 | 0 | 8 | 14 | 20 | 2 |
| U18 | 8 | Multiple Times | Multiple Times | 1 | 7 | 1 | 7 | 7 | 1 |
| U19 | 8 | Multiple Times | Multiple Times | 1 | 7 | 1 | 7 | 7 | 1 |

Using Weka tool K-Means Clustering was run and following observations were derived from the output. By looking at the LogIn patterns of 19 users it was concluded that 17 Users work during weekdays and 2 work on the weekends. Of the 17 Users working during Weekdays 6 Users LogIn and LogOut during regular working hours whereas the other 11 Users LogIn and LogOut Multiple times. Hence there were 3 Clusters that was derived from the data.

```
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 18.814439548546694

Initial starting points (random):

Cluster 0: U10,22,MultipleTimes,MultipleTimes,22,0,6,16,12,10
Cluster 1: U07,22,9,18,22,0,0,22,22,0
Cluster 2: U18,8,MultipleTimes,MultipleTimes,1,7,1,7,7,1

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                        Cluster#
Attribute                                    Full Data          0          1          2
                                              (19.0)     (11.0)      (6.0)      (2.0)
=============================================================================================================
UserId                                            U01        U02        U01        U18
TotNumdaysUserLoggedin                        20.4737    21.9091         22          8
LoginTime                              MultipleTimes MultipleTimes          9 MultipleTimes
LogoutTime                             MultipleTimes MultipleTimes         18 MultipleTimes
WeekdaytotalLogin                           19.7368    21.9091         22          1
WeekendtotalLogin                            0.7368          0          0          7
numberoftimesuserloggingINduringWorkinghours  4.2632     7.1818          0          1
numberoftimesuserlogginginbeforeafterWorkinghours 16.2105 14.7273       22          7
numberoftimesuserloggingOUTduringWorkinghours 17.1579    16.3636         22          7
numberoftimeuserloggingOUTbeforeafterWorkinghours 3.3158  5.5455          0          1



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      11 ( 58%)
1       6 ( 32%)
2       2 ( 11%)
```

| Cluster | Users | Derivation based on |
|---|---|---|
| Cluster 0 | U02,U04,U06,U08,U10, U12,U13,U14,U15,U16,U17 | Weekday Users with Multiple LogIn |
| Cluster 1 | U01,U03, U05,U07, U09, U11 | Weekday Users with regular LogIn |
| Cluster 2 | U18, U19 | Weekend Users |

## Program Access Pattern:

For Type 2 records we were provided with Machine used and Programs executed during weekdays and weekends. Besides that we were also provided the data with programs executed during working and non-working hours. Below are the queries used to derive Table 2 (Program Access Pattern)

**Query 4:**

If a user has used a single Machine for all of the executions then we have identified the machine else we have declared that the user have used more than a single machine and hence assigned "More than One machine Used".

```
SELECT DISTINCT User_Id,Machine_Id
FROM dbo.DataMiningProject
WHERE Record_Id = 2
ORDER BY User_Id
```

**Query 5:**

Below query gives us the total number of programs executed by a user

```
SELECT User_Id
, COUNT(Program) As Total_Programs_Executed
FROM dbo.DataMiningProject
WHERE Record_Id = 2
GROUP By User_Id
ORDER BY User_Id
```

**Query 6:**

Below Query provides us data with the programs executed during working hours

```
SELECT User_Id, COUNT(*) AS Programs_Executed_During_Working_Hours
FROM dbo.DataMiningProject
WHERE Record_Id = 2
AND Start_Time BETWEEN '080000' AND '170000'
GROUP BY User_Id
ORDER BY User_Id
```

**Query 7:**

Below Query provides total execution time by each user.

```
SELECT User_Id, SUM(CAST(Execution_Time AS INT)) AS Total_Execution_Time
FROM dbo.DataMiningProject
WHERE Record_Id = 2
GROUP BY User_Id
ORDER BY User_Id
```

From Queries 4-7 we have derived a table

Table2: Program Access Pattern

| User_Id | Machines Used to execute the programs | Total number of programs executed | Number of programs executed during working hours | Number of programs executed after working hours | Total Execution Time |
|---|---|---|---|---|---|
| U01 | M01 | 19 | 18 | 1 | 11910 |
| U02 | M02 | 28 | 27 | 1 | 28310 |
| U03 | M03 | 19 | 18 | 1 | 11910 |
| U04 | M04 | 28 | 27 | 1 | 28310 |
| U05 | M05 | 19 | 18 | 1 | 11910 |
| U06 | M06 | 28 | 27 | 1 | 28310 |
| U07 | M07 | 19 | 18 | 1 | 11910 |
| U08 | M08 | 28 | 27 | 1 | 28310 |
| U09 | M09 | 19 | 18 | 1 | 11910 |
| U10 | More than one machine used | 28 | 27 | 1 | 28310 |
| U11 | More than one machine used | 19 | 18 | 1 | 11910 |
| U12 | More than one machine used | 19 | 15 | 4 | 11910 |
| U13 | More than one machine used | 28 | 23 | 5 | 18300 |
| U14 | More than one machine used | 19 | 15 | 4 | 11910 |

| | | | | | |
|---|---|---|---|---|---|
| U15 | More than one machine used | 19 | 15 | 4 | 11910 |
| U16 | M16 | 13 | 12 | 1 | 8550 |
| U17 | M19 | 13 | 12 | 1 | 8550 |
| U18 | M18 | 10 | 9 | 1 | 6720 |
| U19 | M19 | 10 | 9 | 1 | 6720 |

Using Weka tool K-Means Clustering was run and following observations were derived from the output. By looking at the Program access patterns of 19 users it was concluded that 6 Users executed 28 programs and 9 Users have executed 19 programs. There was only 1 unique user who accessed Unique machine (M18) even though the total number of programs executed and the execution time was same with U19. Hence there were 4 Clusters that was derived from the data.

```
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 27.27600334492465

Initial starting points (random):

Cluster 0: U10,Morethanonemachineused,28,27,1,28310
Cluster 1: U07,M07,19,18,1,11910
Cluster 2: U18,M18,10,9,1,6720
Cluster 3: U17,M19,13,12,1,8550

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                        Cluster#
Attribute                                    Full Data          0          1          2          3
                                                (19.0)     (6.0)      (9.0)      (1.0)      (3.0)
=============================================================================================================================
UserId                                             U01        U02        U01        U18        U16
MachinesUsedtoexecutetheprograms  Morethanonemachineused Morethanonemachineused Morethanonemachineused        M18        M19
TotalNumberofProgramsExecuted                 20.2632         28         19         10         12
NumberofProgramsExecutedDuringWorkingHours    18.5789    26.3333         17          9         11
NumberofProgramsExecutedAfterWorkingHours      1.6842     1.6667          2          1          1
TotalExecutionTime                         15662.1053  26641.6667      11910       6720       7940



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       6 ( 32%)
1       9 ( 47%)
2       1 (  5%)
3       3 ( 16%)
```

| Cluster | Users | Derivation based on |
|---|---|---|
| Cluster 0 | U02,U04,U06,U08,U10,U13 | Number of Programs Executed (28) |
| Cluster 1 | U01,U03,U05,U07,U09,U11,U12,U14,U15 | Number of Programs Executed (19) |

| | | |
|---|---|---|
| Cluster 2 | U18 | Unique Machine Accessed |
| Cluster 3 | U16,U17,U19 | Execution Time |

## File Access Pattern:

For Type 2 records we were provided with Machine used and Files accessed during weekdays and weekends along with Number of files opened with Read/Write access. Below are the queries used to derive Table 3 (File Access Pattern)

**Query 8:**

Below query provides us the number of files accessed by each user.

```
SELECT User_Id, COUNT(File_Name) AS Total_Number_Of_Files_Accessed
FROM dbo.DataMiningProject
GROUP BY User_Id
ORDER BY User_Id
```

**Query 9:**

Below query provides us the total number of files accessed during weekday and weekend.

```
SELECT User_Id
     , COUNT(File_Name) AS Total_Number_Of_Files_Accessed_During_Weekday
FROM dbo.DataMiningProject
WHERE DATE NOT IN ('090608','090708','091308','091408','092008','092108'
,'092708'
,'092808')
AND Record_Id = 2
GROUP BY User_Id
ORDER BY User_Id
```

**Query 10:**

Below query provides us the number of files opened by access type.

```
SELECT User_Id,Data_Access, COUNT(*) AS Access_By_Type
FROM dbo.DataMiningProject
WHERE Record_Id = 2
GROUP BY User_Id,Data_Access
ORDER BY User_Id
```

Table 3: File Access Pattern

| User_Id | Machines Used to access the files | Total Number of Files Accessed | Number of files accessed during weekdays | Number of files accessed during weekends | Number of Files opened with Read Access | Number of Files opened with read write access |
|---|---|---|---|---|---|---|
| U01 | M01 | 19 | 19 | 0 | 11 | 8 |
| U02 | M02 | 28 | 28 | 0 | 16 | 12 |
| U03 | M03 | 19 | 19 | 0 | 11 | 8 |
| U04 | M04 | 28 | 28 | 0 | 16 | 12 |
| U05 | M05 | 19 | 19 | 0 | 3 | 16 |
| U06 | M06 | 28 | 28 | 0 | 7 | 21 |
| U07 | M07 | 19 | 19 | 0 | 3 | 16 |
| U08 | M08 | 28 | 28 | 0 | 7 | 21 |
| U09 | M09 | 19 | 19 | 0 | 3 | 16 |
| U10 | More than one machine used | 28 | 28 | 0 | 7 | 21 |
| U11 | More than one machine used | 19 | 19 | 0 | 19 | 0 |
| U12 | More than one machine used | 19 | 19 | 0 | 19 | 0 |
| U13 | More than one machine used | 28 | 28 | 0 | 25 | 3 |
| U14 | More than one machine used | 19 | 19 | 0 | 19 | 0 |
| U15 | More than one machine used | 19 | 19 | 0 | 19 | 0 |
| U16 | M16 | 13 | 13 | 0 | 12 | 1 |
| U17 | M19 | 13 | 13 | 0 | 12 | 1 |
| U18 | M18 | 10 | 0 | 10 | 9 | 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| U19 | M19 | 10 | 0 | 10 | 9 | 1 |

Using Weka tool K-Means Clustering was run and following observations were derived from the output. By looking at the File access patterns of 19 users it was concluded that 6 Users accessed 28 files hence fall into Cluster0. Based on the files opened with Data access type (Read/Write) 3 Users fall into Cluster1. Based on Number of files accessed during weekdays and weekends 2 Uses fall into Cluster2 and 8 of them fall into Cluster3. Hence there were 4 Clusters that was derived from the data.

```
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 27.843092333352075

Initial starting points (random):

Cluster 0: U10,Morethanonemachineused,28,28,0,7,21
Cluster 1: U07,M07,19,19,0,3,16
Cluster 2: U18,M18,10,0,10,9,1
Cluster 3: U17,M19,13,13,0,12,1

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                        Cluster#
Attribute                                    Full Data          0          1          2          3
                                               (19.0)     (6.0)      (3.0)      (2.0)      (8.0)
=============================================================================================================================
UserId                                             U01        U02        U05        U18        U01
MachinesUsedtoaccesstheprograms  Morethanonemachineused Morethanonemachineused        M05        M18 Morethanonemachineused
TotalNumberofFilesAccessed                    20.2632         28         19         10       17.5
NumberofFilesAccessedDuringWeekdays           19.2105         28         19          0       17.5
NumberofFilesAccessedDuringWeekends            1.0526          0          0         10          0
NumberofFilesopenedwithReadAccess             11.9474         13          3          9      15.25
NumberofFilesopenedwithReadWriteAccess         8.3158         15         16          1       2.25



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       6 ( 32%)
1       3 ( 16%)
2       2 ( 11%)
3       8 ( 42%)
```

| Cluster | Users | Derived Based on |
|---|---|---|
| Cluster 0 | U02,U04,U06,U08,U10,U13 | Total Number of Files Accessed |
| Cluster 1 | U05,U07,U09 | Number of Files opened with Data Access Type |
| Cluster 2 | U18,U19 | Number of Files accessed during weekdays |
| Cluster 3 | U01,U03,U11,U12,U14,U15,U16,U17 | Number of Files accessed during weekends |

## Printer Access Pattern:

Along with other 2 patterns Printer access was also provided for Type 2 records. Hence we have derived total number of pages printed along with number of pages printed during weekday and weekend. Below are the queries used to derive Table 4 (Printer Access Pattern)

**Query 11:**

Below query provides us the data with unique printer used by the user.

```
SELECT DISTINCT User_Id,Printer_Id
FROM dbo.DataMiningProject
WHERE Record_Id = 2
AND Printer_Id <> ''
```

**Query 12:**

Below query provides us the data with total number of pages printed.

```
SELECT User_Id, SUM(CAST(Pages_Printed AS INT)) AS Total_Number_Of_Pages_Printed
FROM dbo.DataMiningProject
WHERE Record_Id = 2
GROUP BY User_Id
ORDER BY User_Id
```

**Query 13:**

Below query provides us the data with total number of pages printed during weekday and weekend.

```
SELECT User_Id
     , SUM(CAST(Pages_Printed AS INT)) AS Total_Number_Of_Pages_During_Weeday
FROM dbo.DataMiningProject
WHERE Record_Id = 2
AND DATE NOT IN ('090608','090708','091308','091408','092008','092108'
,'092708'
,'092808')
GROUP BY User_Id
ORDER BY User_Id
```

Table 4: Printer Access Pattern

| UserId | Machine Used for printing pages | Printer_Id used for Printing | Total number of pages Printed | Total pages printed during weekdays | Total pages printed during weekends |
|---|---|---|---|---|---|
| U01 | M01 | PR1 | 142 | 142 | 0 |
| U02 | M02 | PR1 | 160 | 160 | 0 |
| U03 | M03 | PR1 | 142 | 142 | 0 |
| U04 | M04 | PR1 | 160 | 160 | 0 |

| | | | | | |
|---|---|---|---|---|---|
| U05 | M05 | PR2 | 142 | 142 | 0 |
| U06 | M06 | PR2 | 160 | 160 | 0 |
| U07 | M07 | PR2 | 142 | 142 | 0 |
| U08 | M08 | PR2 | 160 | 160 | 0 |
| U09 | M09 | PR2 | 142 | 142 | 0 |
| U10 | More than one machine used | PR2 | 160 | 160 | 0 |
| U11 | More than one machine used | PR3 | 142 | 142 | 0 |
| U12 | More than one machine used | PR3 | 142 | 142 | 0 |
| U13 | More than one machine used | PR4 | 202 | 202 | 0 |
| U14 | More than one machine used | PR4 | 142 | 142 | 0 |
| U15 | More than one machine used | PR4 | 142 | 142 | 0 |
| U16 | M16 | PR4 | 92 | 92 | 0 |
| U17 | M19 | PR6 | 92 | 92 | 0 |
| U18 | M18 | PR5 | 56 | 0 | 56 |
| U19 | M19 | PR6 | 56 | 0 | 56 |

Using Weka tool, K-Means Clustering was run and following observations were derived from the output. By looking at the Printer access patterns of 19 users it was concluded that 10 have printed pages during the weekend hence fall into Cluster0. Based on the pages printed during the weekday 3 Users fall into Cluster1. Based on the unique machine and printer used 1 User fall into Cluster 1 and 3 of them fall into Cluster3. Hence there were 4 Clusters that was derived from the data.

```
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 35.05760563459477

Initial starting points (random):

Cluster 0: U10,Morethanonemachineused,PR2,160,160,0
Cluster 1: U07,M07,PR2,142,142,0
Cluster 2: U18,M18,PR5,56,0,56
Cluster 3: U17,M19,PR6,92,92,0

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                    Cluster#
Attribute                                Full Data          0                1                2                3
                                            (19.0)     (10.0)           (5.0)            (1.0)            (3.0)
=============================================================================================================================
UserId                                         U01        U02              U01              U18              U16
MachinesUsedtoPrintPages  Morethanonemachineused Morethanonemachineused  M01              M18              M19
PrinterIdusedforPrinting                       PR2        PR2              PR2              PR5              PR6
TotalNumberofPagesPrinted                 135.5789        157              142               56               80
TotalPagesPrintedDuringWeekdays           129.6842        157              142                0          61.3333
TotalPagesPrintedDuringWeekends             5.8947          0                0               56          18.6667



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      10 ( 53%)
1       5 ( 26%)
2       1 (  5%)
3       3 ( 16%)
```

| Cluster | Users | Derived Based on |
|---|---|---|
| Cluster 0 | U02,U03,U04,U05,U06,U07,U08,U09,U10,U13 | Pages printed during weekend. |
| Cluster 1 | U01,U11,U12,U14,U15 | Pages printed and Pages printed during weekday.. |
| Cluster 2 | U18 | Unique Machine and Printer Used |
| Cluster 3 | U16,U17,U19 | How |

## Email Pattern:

For Type 3 records we were provided with Machine used and Emails sent/received along with the attachments in the email. Based on this data some queries are written to gain knowledge on the pattern. Below are the queries used to derive Table 5 (Email Pattern)

**Query 14:**

Below query provides us the data with Unique Email Program Id for a user.

```
SELECT User_Id, COUNT(Email) AS Total_Number_Of_Emails
FROM dbo.DataMiningProject
WHERE Record_Id = 3
GROUP BY User_Id
ORDER BY User_Id
```

**Query 15:**

Below query provides us the data with total number of Emails for each user.

```
SELECT User_Id,Email, COUNT(Email) AS Total_Number_Of_Emails
FROM dbo.DataMiningProject
WHERE Record_Id = 3
GROUP BY User_Id, Email
ORDER BY User_Id
```

**Query 16:**

Below query provides us the data with total number of attachments for each user.

```
SELECT User_Id,Email_Status,SUM(Attachments) AS Total_Number_Of_Attachments
FROM dbo.DataMiningProject
WHERE Record_Id = 3
GROUP BY User_Id, Email_Status
ORDER BY User_Id
```

**Query 17:**

Below query provides us the data with total number of Emails during weekdays and weekends for each user.

```
SELECT User_Id,COUNT(Email) AS Total_Number_Of_Emails_During_Weekdays
FROM dbo.DataMiningProject
WHERE Record_Id = 3
AND DATE NOT IN ('090608','090708','091308','091408','092008','092108'
,'092708'
,'092808')
GROUP BY User_Id
ORDER BY User_Id
```

**Table 5: Email Pattern**

| UserId | Machine Id | Email program | Total number of emails | Email ID used more number of times | Email Status | Number of attachments sent/received | Email usage during week days | Email usage during weekends | Is mom@icare.com used for sending/ receiving emails |
|---|---|---|---|---|---|---|---|---|---|
| U01 | M01 | E1 | 11 | jones@pqr.com | Sent | 10 | 11 | 0 | No |
| U02 | M02 | E1 | 12 | jones@pqr.com | Sent | 10 | 12 | 0 | Yes |
| U03 | M03 | E1 | 11 | jones@pqr.com | Sent | 10 | 11 | 0 | No |
| U04 | M04 | E1 | 12 | jones@pqr.com | Sent | 10 | 12 | 0 | Yes |
| U05 | M05 | E1 | 11 | jones@pqr.com | Sent | 10 | 11 | 0 | No |
| U06 | M06 | E1 | 12 | smith@abc.org | Received | 10 | 12 | 0 | Yes |
| U07 | M07 | E1 | 11 | smith@abc.org | Sent | 10 | 11 | 0 | No |
| U08 | M08 | E1 | 12 | smith@abc.org | Received | 10 | 12 | 0 | Yes |
| U09 | M09 | E3 | 11 | smith@abc.org | Sent | 10 | 11 | 0 | No |
| U10 | More than one machine used | E3 | 12 | smith@abc.org | Received | 10 | 12 | 0 | Yes |
| U11 | More than one machine used | E1 | 11 | xyz@sai.org | Sent | 10 | 11 | 0 | No |
| U12 | More than one machine used | E1 | 11 | xyz@sai.org | Sent | 10 | 11 | 0 | No |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| U13 | More than one machine used | E1 | 11 | xyz@sai.org | Sent | 10 | 11 | 0 | No |
| U14 | More than one machine used | E1 | 11 | xyz@sai.org | Sent | 10 | 11 | 0 | No |
| U15 | More than one machine used | E1 | 11 | bob@xyz.com | Received | 10 | 11 | 0 | No |
| U16 | M16 | E1 | 11 | bob@xyz.com | Received | 10 | 11 | 0 | No |
| U17 | M19 | E4 | 11 | bob@xyz.com | Received | 10 | 11 | 0 | No |
| U18 | M18 | E5 | 11 | bob@xyz.com | Received | 10 | 0 | 11 | No |
| U19 | M19 | E4 | 11 | bob@xyz.com | Received | 10 | 0 | 11 | No |

Using Weka tool K-Means Clustering was run and following observations were derived from the output. By looking at the Email patterns of 19 users it was concluded that 5 Users have same number of emails hence fall into Cluster0. Based on Email usage during weekdays and weekends 9 Users fall into Cluster1 and Cluster2 respectively.. Based on Email Id used more number of times 3 users fall into Cluster3. Hence there were 4 Clusters that was derived from the data.

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 40.0

Initial starting points (random):

Cluster 0: U10,Morethanonemachineused,E3,12,smith@abc.org,Received,10,12,0,Yes
Cluster 1: U07,M07,E1,11,smith@abc.org,Sent,10,11,0,No
Cluster 2: U18,M18,E5,11,bob@xyz.com,Received,10,0,11,No
Cluster 3: U17,M19,E4,11,bob@xyz.com,Received,10,11,0,No

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                    Cluster#
Attribute                                Full Data          0                  1                  2                  3
                                           (19.0)        (5.0)              (9.0)              (2.0)              (3.0)
=============================================================================================================================
UserId                                         U01        U02                U01                U18                U15
MachinesUsedtoPrintPages  Morethanonemachineused        M02 Morethanonemachineused                M18 Morethanonemachineused
EmailProgram                                   E1         E1                 E1                 E4                 E1
TotalNumberofEmails                       11.2632         12                 11                 11                 11
EmailIDUsedManyTimes              jones@pqr.com  smith@abc.org      xyz@sai.org        bob@xyz.com        bob@xyz.com
EmailStatus                                 Sent       Received           Sent               Received           Received
NumberofAttachmentsSsentReceived              10         10                 10                 10                 10
EmailUsageDuringWeekdays                  10.1053         12                 11                  0                 11
EmailUsageDuringWeekends                   1.1579          0                  0                 11                  0
EmailIdUsedforSendingReceivingEmails           No        Yes                 No                 No                 No



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       5 ( 26%)
1       9 ( 47%)
2       2 ( 11%)
3       3 ( 16%)
```

| Cluster | Users | Derived Based on |
|---|---|---|
| Cluster 0 | U02,U04,U06,U08,U10 | Total Number of Emails |
| Cluster 1 | U01,U03,U05,U07,U09,U11,U12,U13,U14 | Email Usage During Weekdays |
| Cluster 2 | U18,U19 | Email Usage During Weekends |
| Cluster 3 | U15,U16,U17 | Email id Used bob@xyz.com and total number of emails is 11 |

## Machine Usage Pattern:

Machine usage pattern is not for a particular record type but indeed applies to all record types. From the data provided we have derived the user that have used the machine, whether the machine was shared or not. Besides these we have identified whether the machine was used during weekdays or weekends. Based on this data some queries are written to gain knowledge on the pattern. Below are the queries used to derive Table 6 (Machine Usage Pattern)

**Query 18:**

Below query provides us the unique machine used by the user. If there are more than one user then we have identified the machine as a multiple user machine.

```
SELECT DISTINCT Machine_Id,User_Id
FROM dbo.DataMiningProject
ORDER By Machine_Id
```

**Query 19:**

Below query represents whether the machine was accessed during weekdays or weekends.

```
SELECT Machine_Id
     , (CASE WHEN COUNT(*) <> 0 THEN 'Y'
             ELSE 'N'
        END) AS Machine_Accessed_During_Weekday
FROM dbo.DataMiningProject
WHERE DATE NOT IN ('090608','090708','091308','091408','092008','092108'
,'092708'
,'092808')
GROUP BY Machine_Id
ORDER BY Machine_Id
```

**Query 20:**

Below query provides us the data whether the machine was shared or not.

```
SELECT Machine_Id
     , (CASE WHEN COUNT(*) <> 0 THEN 'Y'
             ELSE 'N'
        END) AS Machine_Accessed_During_Weekend
FROM dbo.DataMiningProject
WHERE DATE IN ('090608','090708','091308','091408','092008','092108'
,'092708'
,'092808')
GROUP BY Machine_Id
ORDER BY Machine_Id
```

**Table 6: Machine Usage Pattern**

| Machine Used | Machine Used By | Shared (Y/N) | Machine Accessed During Weekdays (Y/N) | Machine Accessed During Weekends |
|---|---|---|---|---|
| M01 | U01 | N | Y | N |
| M02 | U02 | N | Y | N |
| M03 | U03 | N | Y | N |
| M04 | U04 | N | Y | N |

| | | | | |
|---|---|---|---|---|
| M05 | U05 | N | Y | N |
| M06 | U06 | N | Y | N |
| M07 | U07 | N | Y | N |
| M08 | Multiple users | Y | Y | N |
| M09 | U09 | N | Y | N |
| M10 | U10 | N | Y | N |
| M11 | U11 | N | Y | N |
| M12 | U12 | N | Y | N |
| M13 | U13 | N | Y | N |
| M14 | U14 | N | Y | N |
| M15 | U15 | N | Y | N |
| M16 | U16 | N | Y | N |
| M18 | U18 | N | Y | Y |
| M19 | Multiple users | Y | Y | Y |
| M21 | Multiple users | Y | Y | N |
| M22 | Multiple users | Y | Y | N |
| M23 | Multiple users | Y | Y | N |
| M24 | Multiple users | Y | Y | N |
| M25 | Multiple users | Y | Y | N |
| M26 | Multiple users | Y | Y | N |
| M27 | Multiple users | Y | Y | N |
| M28 | Multiple users | Y | Y | N |
| M29 | Multiple users | Y | Y | N |
| M30 | Multiple users | Y | Y | N |

Using Weka tool K-Means Clustering was run and following observations were derived from the output. By looking at the Machine usage patterns of 19 users it was concluded that 25 Machines were accessed only during weekdays and hence fall into Cluster0. Based on the unique user using the machine the other 3 clusters were formed with a single user in each cluster. Hence there were 4 Clusters that was derived from the data.

```
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 51.0

Initial starting points (random):

Cluster 0: M10,U10,N,Y,N
Cluster 1: M07,U07,N,Y,N
Cluster 2: M12,U12,N,Y,N
Cluster 3: M16,U16,N,Y,N

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute                  Full Data           0           1           2           3
                             (28.0)      (25.0)      (1.0)       (1.0)       (1.0)
=============================================================================================
MachineUsed                     M01         M01         M07         M12         M16
MachineUsedBy           Multipleusers Multipleusers     U07         U12         U16
IsShared                        N           N           N           N           N
MachineAccessedDuringWeekdays   Y           Y           Y           Y           Y
MachineAccessedDuringWeekends   N           N           N           N           N



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      25 ( 89%)
1       1 (  4%)
2       1 (  4%)
3       1 (  4%)
```

| Cluster | Machines | Derived Based on |
|---|---|---|
| Cluster 0 | M01,M02,M03,M04,M05,M06,M08,M09,M10,M11,M13,M14,M15,M18,M19,M21,M22,M23,M24,M25,M26,M27,M28,M29,M30. | Machines accessed only during weekdays |
| Cluster 1 | M07 | Machine is used by Unique user (U07) |
| Cluster 2 | M12 | Machine is used by Unique user (U12) |
| Cluster 3 | M16 | Machine is used by Unique user (U16) |

## Association Rules:

We would like to make a rule that we need at least a Support of 4 and an Accuracy of 75%

*From Login Pattern:*

1) If Login_time category = 9
   Logout_time category = 18
   Number of days user logged in = 22.
   For the above rule
   Coverage = 6
   Support = 6 ,
   Accuracy = 100%

2) If Number of time User logging in during working hours = 8
   Number of time User logging in before/after working hours = 14
   For the above rule
   Coverage = 5
   Support = 5 ,
   Accuracy = 100%

*From Program Access Pattern:*

3) If Total Programs Executed = 19
   Number of Programs executed during working hours = 18
   Total Execution Time = 11910
   For the above rule
   Coverage = 6
   Support = 6 ,
   Accuracy = 100%

*From File Access Pattern:*

4) If Machine Used to access the file = More than one
   Total Number of Files Accessed = 19
   Total Number of Files Accessed during weekdays = 19
   For the above rule
   Coverage = 4
   Support = 4 ,
   Accuracy = 100%

*From Printer Access Pattern:*

5) If Printer Used to print the pages = PR2
   Total Number of pages printed = 142
   For the above rule
   Coverage = 6
   Support = 3 ,
   Accuracy = 50%

But according to our rule we want to get a Support of at least 4 and an accuracy of at least 75%
So by pruning total number of pages from the rule we can accomplish our criteria
After pruning we get
   Coverage = 6

Support = 6
Accuracy = 100%

*From Email Pattern:*

6) If Email Program = E1
   Email Id used = xyz@sai.org
   Total Number of Attachments= 10
   For the above rule
   Coverage = 14
   Support = 4
   Accuracy = 28.6%

But according to our rule we want to get a Support of at least 4 and an accuracy of at least 75%
So by pruning Email Id Used we can accomplish our criteria
After pruning we get
Coverage = 14
Support = 14
Accuracy = 100%

7) If Email Status = Sent
   Email Id used = jones@pqr.com
   Total Number of Emails = 11
   For the above rule
   Coverage = 5
   Support = 3
   Accuracy = 60%

But according to our rule we want to get a Support of at least 4 and an accuracy of at least 75%
So by pruning total number of emails we can accomplish our criteria
After pruning we get
Coverage = 5
Support = 5
Accuracy = 100%

*From Machine Usage Pattern:*

8) If Machine Accessed During Weekdays = Y
   Machine Accessed During Weekends = N
   Shared = Y
   For the above rule
   Coverage = 30
   Support = 12
   Accuracy = 42.85%

But according to our rule we want to get a Support of at least 4 and an accuracy of at least 75%
So by pruning the attribute Shared we can accomplish our criteria
After pruning we get
Coverage =30
Support =28
Accuracy = 93.3%

## Overfitting and Outliers:

***Overfitting:***

Overfitting plays a crucial role in determining a predictive model with the given data. When a predictive model has too much over fitting then the accuracy is too low for new data that is not seen during the training data which was used earlier. This leads to a misconception on predictive modeling as we fail to predict next trend because of less accuracy.

There are several ways to reduce the Outfitting problem such as Cross Validation, Training with more data, Remove features, Bagging and Boosting etc. Besides these there is another well known principle called as **"Occam's razor"** which states that if there are two models with same accuracy then we adopt simple model because it's more like to be generalised .

***Outliers:***

An outlier is an observation point that is distant from other observations. An outlier may be variability in measurement or it may indicate experimental errors. An outlier if ignored can cause serious problems in statistical analysis.Thus Outliers contains useful information about abnormal characteristics of the systems. Analysing the outliers help us in deriving most useful insights through predictive models.

## Anomalies In Dataset:

First glance of the data in the SQL Server Management Studio (SSMS) by writing some initial queries we have figured out that there are some anomalies in the data and major anomaly is that it has been found that there are Users who never logged into the machine but accessed the resources. Below queries and the results provides us the information about the anomalies.

Below query gives us all the Machines used by the user at least once in Type 1 Records. From the results it has been found that there were 3 Machines which were never used by any user. The machines are M15,M17 and M20

```
SELECT DISTINCT Machine_Id
FROM dbo.DataMiningProject
WHERE Record_Id = 1
ORDER BY Machine_Id
```

It has been found in Type 2 Record that the resources from above mentioned Machines are accessed by users without logging in to the machines.

```
SELECT [User_Id]
      , Machine_Id
      , Login_Time
      , Logout_Time
      , [File_Name]
      , Data_Access,Printer_Id
FROM dbo.DataMiningProject
WHERE Record_Id = 2
AND Machine_Id IN ('M15','M17','M20')
```

Results from above query

| | User_Id | Machine_Id | Login_Time | Logout_Time | File_Name | Data_Access | Printer_Id |
|---|---|---|---|---|---|---|---|
| 1 | U15 | M15 | NULL | NULL | F0100 | R | PR4 |
| 2 | U15 | M15 | NULL | NULL | F0200 | R | PR4 |
| 3 | U15 | M15 | NULL | NULL | F0300 | R | PR4 |

It has been found in Type 3 Record that the resources from above mentioned Machines are accessed by users without logging in to the machines.

```
SELECT [User_Id]
      , Machine_Id
      , Login_Time
      , Logout_Time
      , [File_Name]
      , Data_Access,Printer_Id
FROM dbo.DataMiningProject
WHERE Record_Id = 3
AND Machine_Id IN ('M15','M17','M20')
```

110 %

Results | Messages

| | User_Id | Machine_Id | Login_Time | Logout_Time | File_Name | Data_Access | Printer_Id |
|---|---|---|---|---|---|---|---|
| 1 | U15 | M15 | NULL | NULL | NULL | NULL | NULL |
| 2 | U15 | M15 | NULL | NULL | NULL | NULL | NULL |

## Security Measures:

Above mentioned anomalies clearly indicates that there is a security breach in the system. Major anomaly is that the user accessed the resources without logging into the machine. This is a sign that if there are any external resources that would want to have access to the data in the system then it will be possible which is a major threat to the department.

Of various measures that can be taken to prevent the security breach below are some of the solutions:

1. Auditing the resource and data usage on a frequent basis to identify any unauthorised usage.
2. Revoking the access levels to the users who moved out of the organisation with immediate effect.
3. Using Multi Factor Authentication (MFA) for Ex: User login credentials and passcode to the mobile phone to make sure right user is having right access levels.
4. Access levels to be checked frequently so that specific user will have only required permissions on a role basis .

By adapting security features we can have the utmost satisfaction of the Client/Organisation ensuring highest level of data protection and privacy .

## Conclusions:

After a thorough analysis of the data using various data mining techniques such as Association Rules (Pruning), K-Means clustering we have come to a few conclusions on User Profile patterns.

LogIn Patterns:
Based on Login pattern we came to a conclusion that 17 Users fall into weekday resources and 2 (U18,U19) fall into weekend resources.
One other pattern that was found was a few users (U01,U03,U05,U07,U09,U11) Login and Logout at particular times of the day but the others login and logout multiples times.

Program Access Pattern:
Based on K-Means Clustering on Program Access Pattern we have concluded that of 19 Users 6 of them have executed 28 programs whereas 9 have executed 19 programs.
By looking at the data in the table it was also concluded that more number of programs were executed during working hours.

File Access Pattern:
Form the data in the table we have observed that the resources U18 and U19 are the only 2 Users that have accessed the files during the weekend whereas all other users accessed the files during the weekdays.
Based on number of times files were opened with read/write access we can conclude that U06,U08 and U10 are the users who have updated the files more number of times.

Printer Access Pattern:
Based on K-Mean Clustering and also looking at the data we have figured out that Printer "PR2" was used more number of times.
It was also obvious that User (U13) has printed more number of pages (202).

Email Pattern:
Based on the data we have figured out that xyz@sai.org and smith@abc.org are the 2 emails that have sent more number of emails.
Also one other observation was that U02,U04,U06,U08 and U10 have sent.received personal emails.

Machine Usage Pattern;
Of the 27 Machines available 25 of them have been used during weekdays and 2 on weekends. We have also observed that of 27 machines 12 have been used by Multiple users whereas 15 were used by unique user.

## References:


**https://www.datascience.com/blog/k-means-clustering**
**https://www-users.cs.umn.edu/~kumar001/dmbook/ch6.pdf**
**https://www.youtube.com/watch?v=TtBgfXmIDHQ&t=5s**
**https://simple.wikipedia.org/wiki/Occam%27s_razor**
**https://en.wikipedia.org/wiki/Outlier**