



Sri Lanka Institute of Information Technology

Assignment

Machine Learning - SE4060

Submitted by:

SID : IT15033338

Name : J. G. I. Madushanki

Date of submission:

18/04/2018

Problem

When students get marks for the exams lecturer can group students into different groups according to the marks they obtained in order to determine the students those who performed well in the exam and students those who didn't perform well in the exam. After grouping the students, lecturers / instructors can focus more on the students those who got low marks for the exam. Besides they can motivate the students those who got good marks in order to up keep their level of studies.

But it's too late to wait till students sit for the exam and get marks in order to identify the student's level of study. But in the normal lecturing process lecturer has to wait till students sit for an exam in order to identify students those who are performing well and students those who are not performing well.

But in this case we can get number of times a student raise his/her hand to answer questions and number of times students refer learning resources in the LMS (Learning Management System) like lecture notes and tutorials. These measures are effective factors that affect the final grade of the student.

So by using the number of times student raise hand in the classroom and number of times student refer learning materials in the LMS we can cluster students into different groups. So lecturer can decide to which groups he need to focus more. Besides lecturer can track the progress of the student by comparing the student's previous grouping with the current grouping.

Dataset

URL of the data set : <https://www.kaggle.com/c/student-academic-performance>

Downloaded from the kaggle web site.

Contains 480 instances under 16 attributes.

Related to the educational domain.

Attributes contain both numerical and categorical data.

File format : csv file(Comma Separated Value file)

Data set is collected from a Learning Management System(LMS) called Laboard 360 which is a multi agent LMS. It provides users the access to educational resources from any device with internet connection.

The dataset contains 480 student records based on 16 features. These features are also classified into three main categories as Demographic features like gender and nationality, Academic background features like educational stage, grade level and section, Behavioral features like raised hand in class, opening resources, answering survey by parents, and school satisfaction.

The dataset consists of 305 males and 175 females. The students belonging to different origins such as 179 students are from Kuwait, 172 students are from Jordan, 28 students from Palestine, 22 students are from Iraq, 17 students from Lebanon, 12 students from Tunis, 11 students from Saudi Arabia, 9 students from Egypt, 7 students from Syria, 6 students from USA, Iran and Libya, 4 students from Morocco and one student from Venezuela. In addition to that data set is collected through two educational semesters - 245 student records are collected during semester one while 235 student records are collected during semester two.

The data set includes a feature for the school attendance and students are classified into two categories based on their absence days. 191 students exceed 7 absence days and 289 students' absence days are under 7.

This dataset includes a feature for parent participation in the educational process. Parent participation feature have two sub features, Parent Answering Survey and Parent School Satisfaction. In the dataset 270 of the parents have answered survey and 210 have not, 292 of the parents have satisfied with the school and 188 have not.

Attribute	Description
Gender	Student's gender (Male or Female)
Nationality	Student's nationality (Kuwait, Lebanon, Egypt, Saudi Arabia, USA, Jordan, Venezuela, Iran, Tunis, Morocco, Syria, Palestine, Iraq, Libya)
Place of birth	Student's place of birth (Kuwait, Lebanon, Egypt, Saudi Arabia, USA, Jordan, Venezuela, Iran, Tunis, Morocco, Syria, Palestine, Iraq, Libya)
Educational stage	Educational level which the student belongs to (Lower level, Middle School, High School)
Grade levels	Grade student belongs to (G-01, G-02, G-03, G-04, G-05, G-06, G-07, G-08, G-09, G-10, G-11, G-12)
Section ID	Class room which the student belongs to (A, B, C)
Topic	Course topic (English, Spanish, French, Arabic, IT, Math, Chemistry, Biology, Science, History, Quran, Geology)
Semester	School year semester (First Semester, Second Semester)
Parent responsible for student	Parent who is responsible for the student (mom / father)
Raised hand	How many times the student raises his / her hand in classroom (value in between 1 and 100)
Visited resources	How many times the student visits a course content (value in between 0 and 100)
Viewing announcements	How many times the student checks the latest announcements (value in between 0 and 100)
Discussion groups	How many times the student participate on discussion groups (value in between 0 and 100)
Parents answering survey	Whether parents answered the surveys which was provided from school or not (Yes / No)
Parent school satisfaction	Whether parents satisfied with the school or not (Yes / No)
Student absence days	Number of absent days of each student (above-7, under-7)

Methodology

K means algorithm which is an unsupervised machine learning algorithm is used to perform clustering. As it is a unsupervised machine learning algorithm unlabeled data is provided as the input along with the number of clusters that we want to cluster the input data. Once unlabeled data is provided to the K means algorithm we will get data which are clustered among different predefined clusters. The K means algorithm uses iterative process in order to produce the final result.

In the initial stage of the K means algorithm centroids are randomly chosen. These centroids can be imaginary or real data points in the data set. Initial centroid location can be far extreme data points in the distribution or most closest data points in the distribution.

Algorithm operates in two iterative steps.

- 1) Data assignment step
- 2) Centroid update step

Data assignment step

Each centroid defines a cluster. In this step each data point is assigned to its closest centroid based on euclidean distance.

Centroid update step

In this step new centroid location is recalculated by calculating the mean of the all the data points assigned to that particular cluster.

K means algorithm iterates between these two steps until none of the data points change it's cluster.

Choosing K

Increasing the number of clusters will reduce the distance to data points. When increasing the number of clusters we will come up with the situation where only one data point belongs to a cluster. But this is not accurate. So that elbow method is used to find the optimal number of clusters.

In the elbow method a graph is drawn along with the number of clusters and sum of squared error. After plotting the graph value which belongs to the elbow of the graph is taken as the optimal number of clusters.

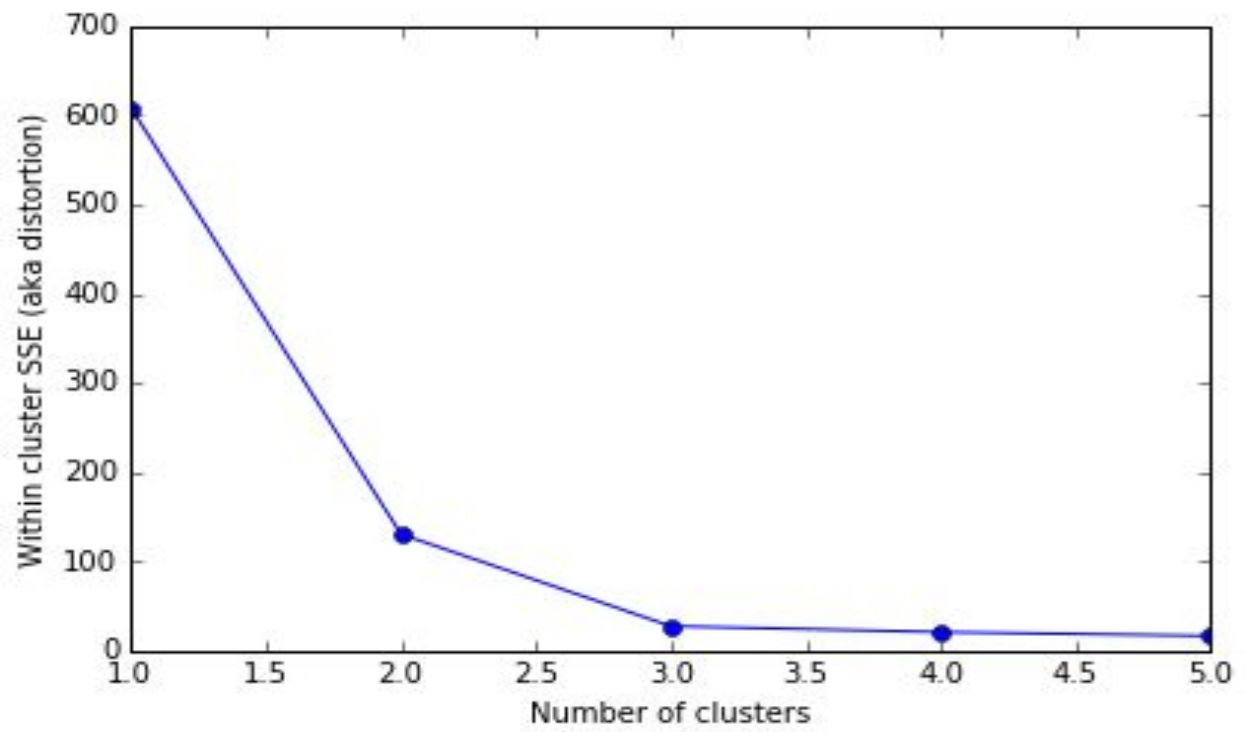
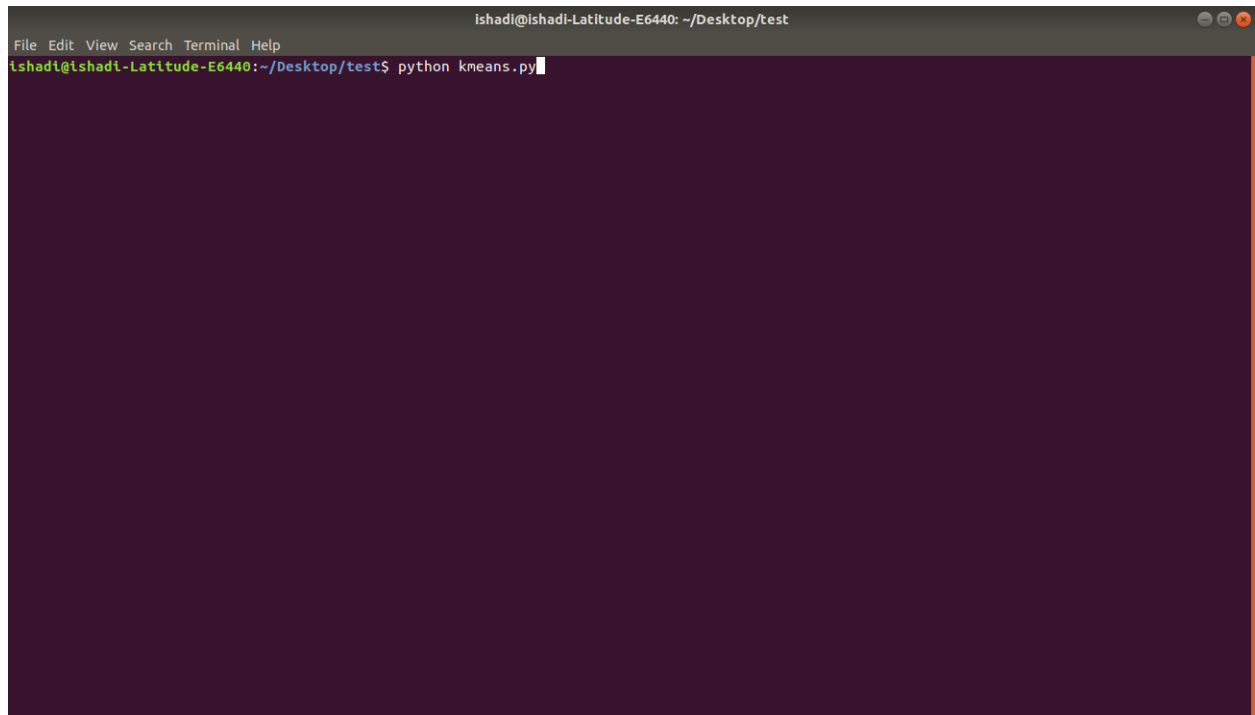


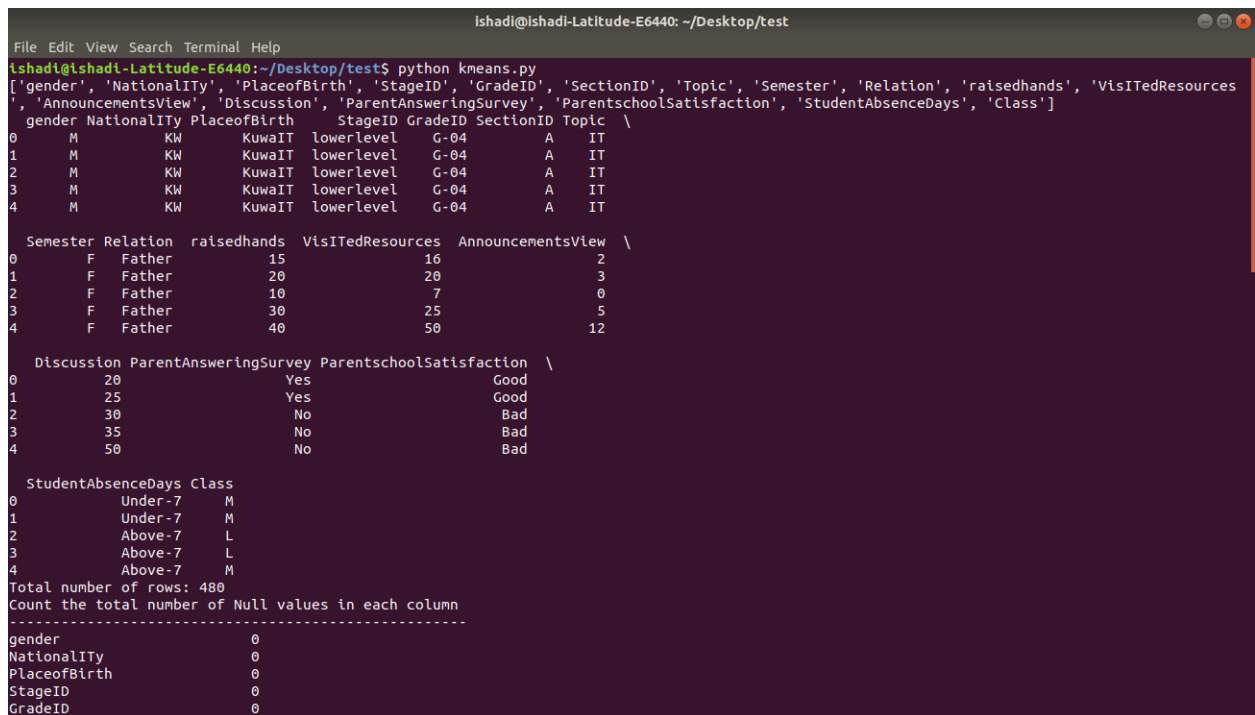
Figure 1 : Graph plotted with No of clusters and Sum of squared errors (Elbow method)

Results



```
ishadi@ishadi-Latitude-E6440: ~/Desktop/test
File Edit View Search Terminal Help
ishadi@ishadi-Latitude-E6440:~/Desktop/test$ python kmeans.py
```

Figure 2 : Running the python programme in the command prompt



```
ishadi@ishadi-Latitude-E6440: ~/Desktop/test
File Edit View Search Terminal Help
ishadi@ishadi-Latitude-E6440:~/Desktop/test$ python kmeans.py
['gender', 'Nationality', 'PlaceofBirth', 'StageID', 'GradeID', 'SectionID', 'Topic', 'Semester', 'Relation', 'raisedhands', 'VisITedResources', 'AnnouncementsView', 'Discussion', 'ParentAnsweringSurvey', 'ParentschoolSatisfaction', 'StudentAbsenceDays', 'Class']
gender Nationality PlaceofBirth StageID GradeID SectionID Topic \
0 M KW Kuwait lowerlevel G-04 A IT
1 M KW Kuwait lowerlevel G-04 A IT
2 M KW Kuwait lowerlevel G-04 A IT
3 M KW Kuwait lowerlevel G-04 A IT
4 M KW Kuwait lowerlevel G-04 A IT

Semester Relation raisedhands VisITedResources AnnouncementsView \
0 F Father 15 16 2
1 F Father 20 20 3
2 F Father 10 7 0
3 F Father 30 25 5
4 F Father 40 50 12

Discussion ParentAnsweringSurvey ParentschoolSatisfaction \
0 20 Yes Good
1 25 Yes Good
2 30 No Bad
3 35 No Bad
4 50 No Bad

StudentAbsenceDays Class
0 Under-7 M
1 Under-7 M
2 Above-7 L
3 Above-7 L
4 Above-7 M
Total number of rows: 480
Count the total number of Null values in each column
-----
gender 0
Nationality 0
PlaceofBirth 0
StageID 0
GradeID 0
```

Figure 3 : Running the python program in the command prompt and printing the outputs

```
ishadi@ishadi-Latitude-E6440: ~/Desktop/test
File Edit View Search Terminal Help
Total number of rows: 480
Count the total number of Null values in each column
-----
gender                0
Nationality            0
PlaceofBirth           0
StageID                0
GradeID                0
SectionID              0
Topic                  0
Semester               0
Relation               0
raisedhands            0
VisITedResources       0
AnnouncementsView      0
Discussion              0
ParentAnsweringSurvey  0
ParentschoolSatisfaction 0
StudentAbsenceDays     0
Class                  0
dtype: int64
('Total no: of rows and columns after dropping the rows which contain missing values ', (480, 17))
(480, 18)
A csv file called output.csv is generated
[16. 20. 7. 25. 50. 30. 12. 10. 21. 80. 88. 6. 1. 14. 70. 40. 30. 13.
15. 50. 60. 12. 21. 0. 2. 7. 19. 15. 85. 90. 80. 5. 19. 22. 11. 12.
6. 54. 0. 90. 13. 20. 12. 35. 33. 12. 10. 4. 80. 39. 14. 15. 90. 70.
50. 14. 5. 2. 60. 22. 10. 70. 90. 13. 5. 5. 10. 75. 69. 40. 30. 22.
2. 30. 0. 90. 70. 80. 3. 90. 15. 25. 5. 4. 70. 0. 12. 70. 12. 20.
8. 90. 70. 89. 44. 80. 60. 2. 3. 7. 90. 92. 6. 7. 12. 0. 26. 90.
12. 70. 88. 80. 5. 27. 2. 8. 80. 29. 35. 60. 12. 4. 90. 98. 6. 30.
9. 33. 10. 90. 9. 42. 3. 60. 80. 80. 80. 80. 80. 85. 60. 65. 75. 90.
10. 75. 75. 79. 55. 75. 80. 63. 91. 51. 50. 58. 50. 50. 51. 68. 89. 80.
82. 82. 72. 65. 82. 92. 52. 12. 62. 52. 22. 52. 62. 2. 52. 52. 42. 51.
70. 62. 75. 15. 35. 65. 15. 71. 71. 66. 25. 25. 91. 75. 75. 43. 65. 75.
15. 95. 90. 58. 5. 51. 10. 51. 31. 21. 41. 81. 90. 61. 81. 61. 50. 21.
41. 88. 61. 51. 69. 51. 61. 83. 81. 90. 11. 3. 84. 17. 42. 7. 72. 80.
60. 8. 10. 80. 80. 80. 80. 80. 20. 20. 92. 40. 94. 48. 40. 86. 6. 74.]
```

Figure 4 : Printing the outputs

```
ishadi@ishadi-Latitude-E6440: ~/Desktop/test
File Edit View Search Terminal Help
A csv file called output.csv is generated
[16. 20. 7. 25. 50. 30. 12. 10. 21. 80. 88. 6. 1. 14. 70. 40. 30. 13.
15. 50. 60. 12. 21. 0. 2. 7. 19. 15. 85. 90. 80. 5. 19. 22. 11. 12.
6. 54. 0. 90. 13. 20. 12. 35. 33. 12. 10. 4. 80. 39. 14. 15. 90. 70.
50. 14. 5. 2. 60. 22. 10. 70. 90. 13. 5. 5. 10. 75. 69. 40. 30. 22.
2. 30. 0. 90. 70. 80. 3. 90. 15. 25. 5. 4. 70. 0. 12. 70. 12. 20.
8. 90. 70. 89. 44. 80. 60. 2. 3. 7. 90. 92. 6. 7. 12. 0. 26. 90.
12. 70. 88. 80. 5. 27. 2. 8. 80. 29. 35. 60. 12. 4. 90. 98. 6. 30.
9. 33. 10. 90. 9. 42. 3. 60. 80. 80. 80. 80. 80. 85. 60. 65. 75. 90.
10. 75. 75. 79. 55. 75. 80. 63. 91. 51. 50. 58. 50. 50. 51. 68. 89. 80.
82. 82. 72. 65. 82. 92. 52. 12. 62. 52. 22. 52. 62. 2. 52. 52. 42. 51.
70. 62. 75. 15. 35. 65. 15. 71. 71. 66. 25. 25. 91. 75. 75. 43. 65. 75.
15. 95. 90. 58. 5. 51. 10. 51. 31. 21. 41. 81. 90. 61. 81. 61. 50. 21.
41. 88. 61. 51. 69. 51. 61. 83. 81. 90. 11. 3. 84. 17. 42. 7. 72. 80.
60. 8. 10. 80. 80. 80. 80. 80. 20. 20. 92. 40. 94. 48. 40. 86. 6. 74.
76. 26. 97. 17. 87. 99. 97. 34. 17. 97. 27. 94. 64. 84. 80. 70. 8. 28.
84. 81. 21. 20. 82. 62. 21. 31. 31. 41. 71. 60. 94. 87. 38. 39. 79. 79.
70. 71. 36. 24. 86. 64. 86. 64. 87. 74. 17. 14. 12. 34. 20. 44. 50. 44.
59. 64. 57. 84. 29. 34. 79. 64. 88. 84. 92. 90. 97. 95. 87. 80. 15. 10.
35. 20. 15. 10. 77. 85. 7. 2. 90. 80. 8. 7. 90. 80. 98. 89. 90. 80.
87. 92. 97. 82. 97. 82. 3. 4. 13. 9. 94. 89. 98. 97. 88. 87. 98. 90.
88. 98. 98. 88. 95. 70. 18. 58. 88. 81. 17. 2. 20. 9. 10. 2. 90. 82.
30. 22. 20. 12. 82. 83. 92. 93. 90. 77. 10. 0. 81. 88. 98. 90. 91. 98.
80. 82. 90. 92. 71. 80. 81. 89. 95. 91. 9. 8. 79. 88. 89. 90. 80. 72.
9. 7. 90. 92. 90. 96. 89. 92. 69. 62. 79. 77. 80. 82. 0. 2. 89. 89.
87. 79. 87. 88. 81. 83. 82. 82. 87. 86. 77. 76. 72. 76. 82. 84. 92. 84.
97. 98. 82. 78. 87. 88. 90. 91. 91. 90. 83. 81. 87. 77. 99. 96. 82. 93.
9. 6. 86. 82. 87. 88. 7. 4. 77. 74. 17. 14.]
[15. 20. 10. 30. 40. 42. 35. 50. 12. 70. 50. 19. 5. 20.
62. 30. 36. 55. 69. 70. 60. 10. 15. 2. 0. 8. 19. 25.
75. 30. 35. 4. 2. 8. 12. 10. 8. 45. 0. 50. 14. 19.
10. 30. 33. 20. 7. 70. 13. 29. 20. 39. 55. 49. 12. 16.
19. 5. 28. 27. 21. 50. 80. 17. 0. 13. 25. 65. 70. 39.
22. 29. 11. 19. 12. 50. 15. 20. 13. 80. 8. 8. 7. 7.
50. 1. 70. 19. 3. 5. 4. 80. 50. 55. 80. 100. 14. 6.
10. 50. 50. 70. 2. 1. 0. 0. 12. 70. 7. 90. 70. 77.
2. 25. 11. 0. 77. 25. 24. 60. 21. 0. 66. 70. 0. 12.
2. 55. 12. 70. 7. 80. 0. 12. 80. 70. 70. 60. 100. 100.]
```

Figure 5 : Printing the output on the command prompt


```
ishadi@ishadi-Latitude-E6440: ~/Desktop/test
File Edit View Search Terminal Help
97. 98. 82. 78. 87. 88. 90. 91. 91. 90. 83. 81. 87. 77. 99. 96. 82. 93.
9. 6. 86. 82. 87. 88. 7. 4. 77. 74. 17. 14.]
[ 15. 20. 10. 30. 40. 42. 35. 50. 12. 70. 50. 19. 5. 20.
62. 30. 36. 55. 69. 70. 60. 10. 15. 2. 0. 8. 19. 25.
75. 30. 35. 4. 2. 8. 12. 10. 8. 45. 0. 50. 14. 19.
10. 30. 33. 20. 7. 70. 13. 29. 20. 39. 55. 49. 12. 16.
19. 5. 28. 27. 21. 50. 80. 17. 0. 13. 25. 65. 70. 39.
22. 29. 11. 19. 12. 50. 15. 20. 13. 80. 8. 8. 7. 7.
50. 1. 70. 19. 3. 5. 4. 80. 50. 55. 80. 100. 14. 6.
10. 50. 50. 70. 2. 1. 0. 0. 12. 70. 7. 90. 70. 77.
2. 25. 11. 0. 77. 25. 24. 60. 21. 0. 66. 70. 0. 12.
2. 55. 12. 70. 7. 80. 0. 12. 80. 70. 70. 60. 100. 100.
10. 19. 10. 80. 10. 70. 100. 10. 60. 100. 80. 23. 100. 10.
70. 70. 70. 70. 22. 62. 82. 72. 70. 60. 55. 72. 51. 80.
60. 30. 40. 60. 20. 20. 50. 10. 60. 15. 80. 40. 60. 50.
85. 25. 10. 87. 85. 80. 75. 85. 23. 15. 95. 81. 53. 15.
92. 83. 27. 45. 15. 45. 25. 22. 29. 72. 67. 17. 27. 70.
27. 17. 87. 7. 17. 5. 27. 87. 96. 57. 77. 80. 62. 72.
87. 72. 2. 5. 73. 5. 51. 9. 19. 32. 32. 12. 52. 72.
72. 92. 72. 72. 22. 12. 70. 50. 80. 87. 70. 65. 15. 19.
69. 39. 59. 10. 80. 70. 80. 69. 10. 61. 21. 49. 70. 79.
19. 11. 10. 18. 90. 92. 42. 22. 95. 90. 72. 92. 82. 72.
74. 74. 95. 97. 40. 41. 51. 81. 71. 75. 49. 42. 90. 62.
90. 62. 98. 72. 18. 32. 10. 22. 11. 12. 15. 32. 65. 72.
95. 82. 25. 42. 55. 62. 78. 72. 60. 50. 60. 50. 60. 50.
10. 30. 24. 27. 10. 30. 80. 75. 40. 35. 10. 25. 10. 15.
70. 75. 78. 79. 16. 17. 40. 35. 14. 13. 24. 23. 20. 15.
5. 7. 10. 25. 30. 35. 32. 25. 72. 75. 20. 10. 90. 80.
80. 78. 10. 10. 98. 75. 10. 35. 10. 5. 10. 35. 20. 32.
10. 12. 11. 10. 69. 70. 89. 79. 15. 20. 4. 5. 88. 90.
86. 90. 78. 70. 80. 85. 88. 89. 39. 35. 76. 65. 96. 80.
2. 3. 50. 53. 70. 59. 78. 79. 10. 9. 98. 89. 88. 99.
82. 84. 70. 74. 90. 84. 80. 81. 10. 11. 85. 80. 95. 87.
85. 79. 80. 70. 80. 70. 89. 90. 69. 70. 75. 72. 77. 71.
87. 81. 15. 19. 25. 29. 45. 39. 85. 79. 90. 80. 80. 71.
72. 69. 85. 89. 80. 87. 15. 9. 81. 78. 80. 85. 2. 5.
50. 55. 30. 35.]
ishadi@ishadi-Latitude-E6440:~/Desktop/test$
```

Figure 6 : Printing the output on the command prompt

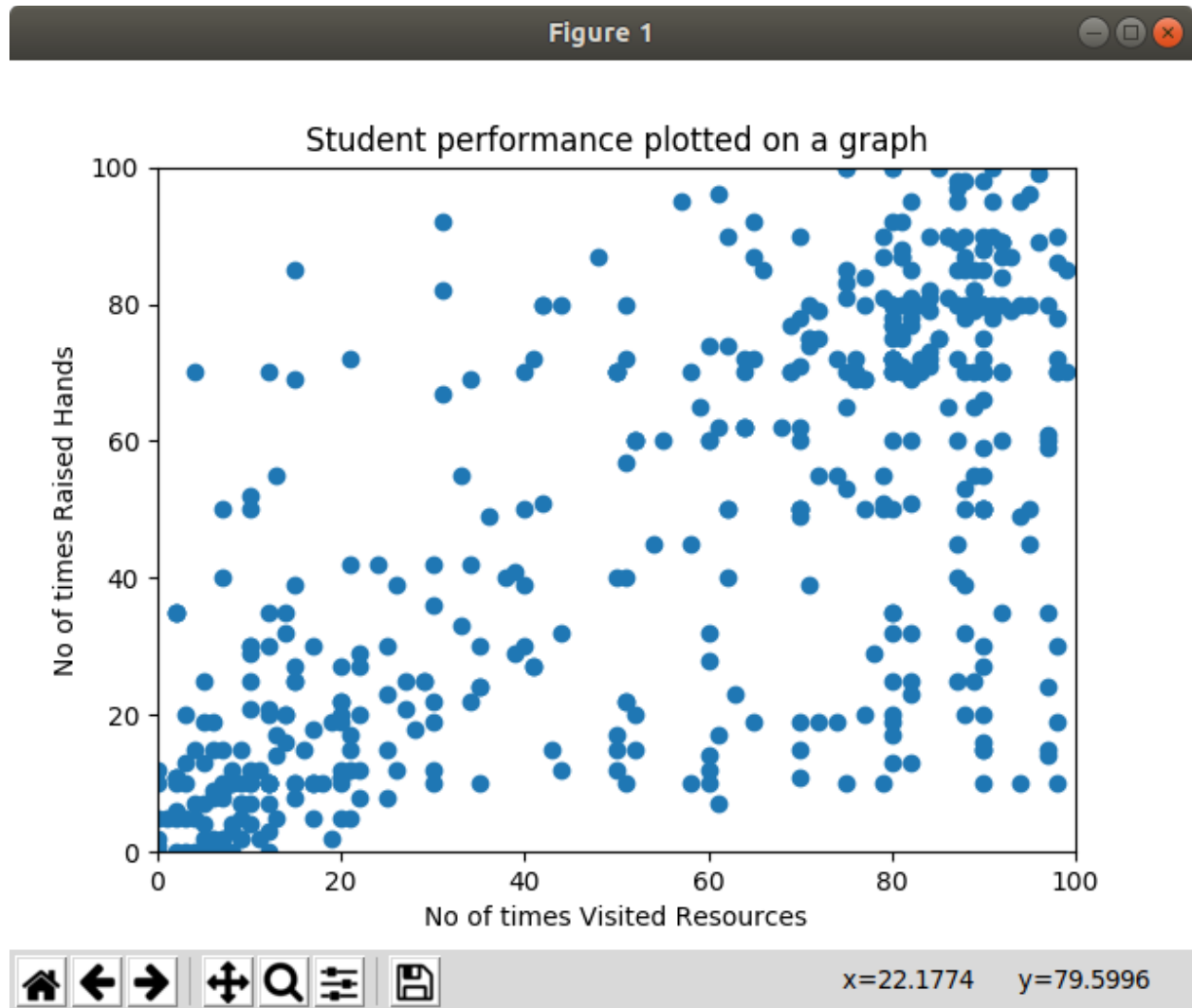


Figure 7: Data set plotted in a graph

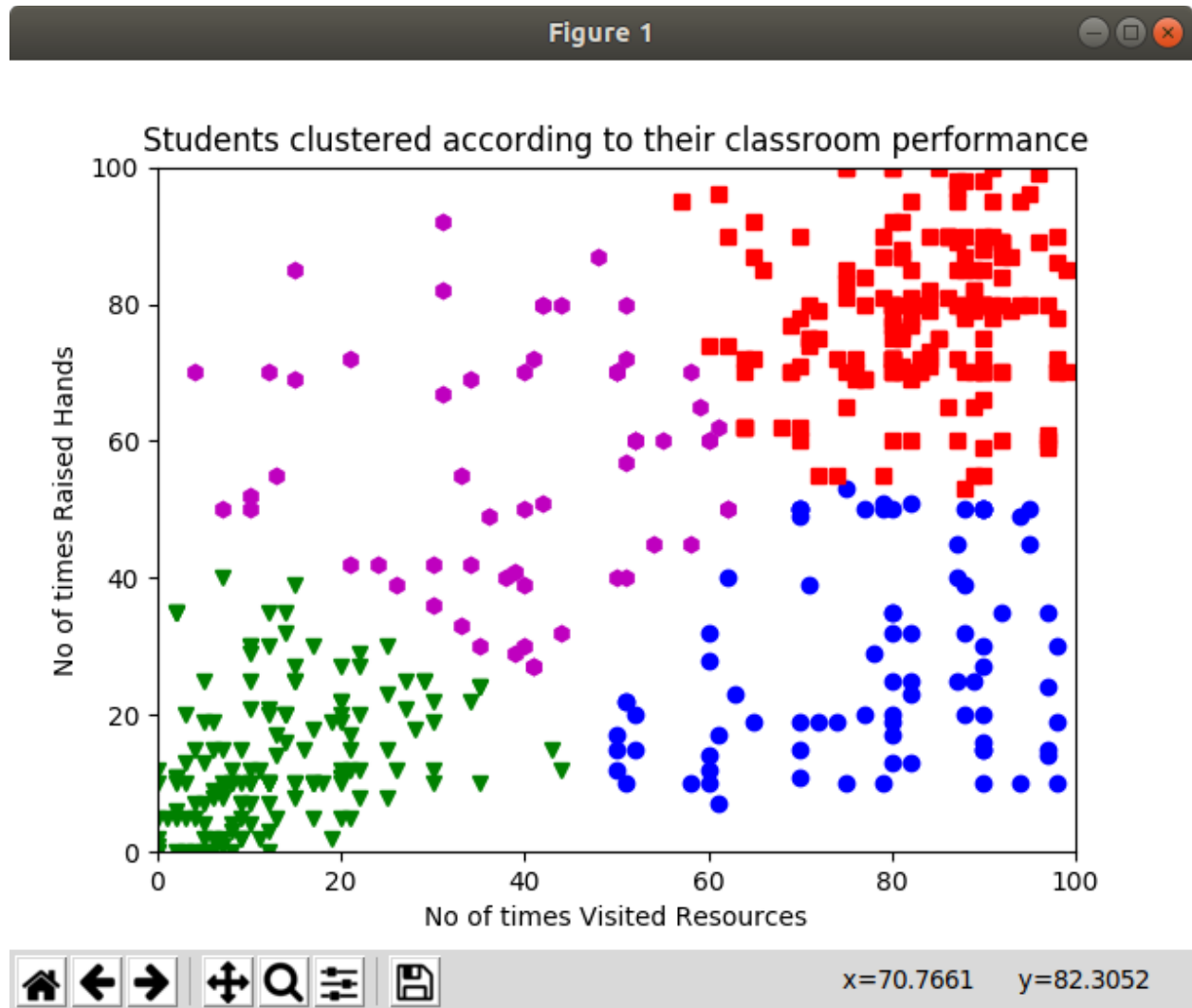


Figure 8 : Data set clustered among four clusters

Discussion

Limitations of k means

- User has to specify the number of clusters (K) at the initial beginning.
- K means algorithm can handle only numerical data. It can't handle categorical data.
- Outliers might affect the centroid location.

Future work

- Selecting Value of K (Number of clusters)

For some popular data sets like Iris data set K value is already defined. So that K value can be used to get accurate results from the k means algorithm. But for all the other data sets user needs to specify the K value. So performance of the clustering may be affected when selecting K value. So a proper method is required to determine the K value. Even though elbow method is used to find out the optimal number of clusters, it is required to perform clustering for a range of values of K. (eg: 1 to 200) Besides it is required to calculate sum of squared errors for each and every cluster and get the summation of sum of squared error in all the clusters in order to plot the graph. So it's a computationally expensive process. So it will degrade the performance of the algorithm.

So it is required to have a proper method to select the K value for k means algorithm

Appendix

```
#Importing the required libraries
import pandas as pd                # Provides DataFrame Object for data manipulation
import numpy as np                 # For fast mathematical functions
from sklearn.cluster import KMeans # Import K means algorithms from sklearn library
from sklearn import cluster
import matplotlib.pyplot as plt    # Matplotlib for plotting the graph

"-----"
# Data cleansing
"-----"

#link address to download the dataset
#https://www.kaggle.com/aljarah/xAPI-Edu-Data/data

data = pd.read_csv("xAPI-Edu-Data.csv")

# Print the headers
print(list(data))

# View the basic structure of data
print(data.head())

# Print the total no of rows
print("Total number of rows: {}".format(len(data)))

# ----- Dealing with missing values -----

# Replace NaN with an empty string or some other default value.
data.VisITedResources = data.VisITedResources.fillna("")
data.raisedhands = data.raisedhands.fillna("")

# Remove incomplete rows
data.dropna()

# count the number of Null values in each column
print("Count the total number of Null values in each column")
print("-----")
print(data.isnull().sum())

# Remove all the rows which has all NA values
data.dropna(how="all")

# No of rows and columns in the dataset after dropping the rows which contain the missing values
```

```

print("Total no: of rows and columns after dropping the rows which contain missing values ", data.shape)

# This tells Pandas that the column VNo of times visited Resources and No of raised hands needs to be a
integer value
data = pd.read_csv("xAPI-Edu-Data.csv", dtype={"VisITedResources": int})
data = pd.read_csv("xAPI-Edu-Data.csv", dtype={"raisedhands": int})

# Rename columns VisITedResources as Visited_Resources and raisedhands as Raised_Hands
data.rename(columns = {'VisITedResources':'Visited_Resources', 'raisedhands':'Raised_Hands'})

data = data.rename(columns = {'VisITedResources':'Visited_Resources', 'raisedhands':'Raised_Hands'})

# Save results to a new csv file
data.to_csv('cleanfile.csv')

dataSet = pd.read_csv('cleanfile.csv')
print(dataSet.shape)
dataSet.head()

# Required data is extracted to a csv file called output.csv based on the columns Visited_Resources and
Raised_Hands
df = pd.read_csv('cleanfile.csv')
header = ["Visited_Resources", "Raised_Hands"]
print('A csv file called output.csv is generated')
df.to_csv('output.csv', columns = header, header=0)

"-----"
# K - means implementation
"-----"

# Cleaned data file is imported
filename = "output.csv"

# No of times Visited_Resources are assigned to a variable called feature_1
feature_1 = np.genfromtxt("output.csv", usecols=[1], delimiter=',')

# No of times Raised_Hands are assigned to a variable called feature_2
feature_2 = np.genfromtxt("output.csv", usecols=[2], delimiter=',')

# feature_1 and feature_2 data is printed on the terminal
print feature_1
print feature_2

# Displaying the plot
plt.plot()

# X and Y limits of the plot

```

```

plt.xlim([0, 100])
plt.ylim([0, 100])
# Assigning X and Y labels
plt.xlabel('No of times Visited Resources')
plt.ylabel('No of times Raised Hands')

# Title of the plot
plt.title('Student performance plotted on a graph')

plt.scatter(feature_1, feature_2)

# Display the plot
plt.show()
# Create a new plot from data
plt.plot()
X = np.array(list(zip(feature_1, feature_2))).reshape(len(feature_1), 2)

# Colors for the input data plot
colors = ['b','g','r','m']          # b - Blue, g - Green, r - Red, m - Magenta

# Marker colors for the output data plot
markers = ['o', 'v','s','h']       # o - circle marker, v - triangle down marker, s - square marker, h -
hexagon marker

# No of clusters that we need to cluster the students according to marks
K = 4
kmeans_model = KMeans(n_clusters=K).fit(X)
plt.plot()
for i, l in enumerate(kmeans_model.labels_):
plt.plot(feature_1[i], feature_2[i], color=colors[l], marker=markers[l],ls='None')

# X and Y limits of the plot is defined
plt.xlim([0, 100])
plt.ylim([0, 100])

# Title of the graph
plt.title('Students clustered according to their classroom performance')

# Assigning X and Y variables
plt.xlabel('No of times Visited Resources')
plt.ylabel('No of times Raised Hands')

# Displaying the plot
plt.show()

```

