DATA MINING PROJECT REPORT

KREETHI MISHRA
AP19110010424
CSE-C

INTRODUCTION:

Data is the most powerful weapon in today's world. With technological advancement in the field of data science and artificial intelligence, machines are now empowered to make decisions for a firm and benefit them. Here is where data mining comes into the picture. This technique helps businesses and firms to analyze valuable user data to their benefit.

What is Data Mining?

The method of extracting useful information to identify patterns and trends in the form of useful data that allows businesses and huge firms to analyze and make decisions from huge sets of data is called Data Mining. In layman's terms, Data Mining is the process of recognizing hidden patterns in the information extracted from the user or data which is relevant to the company's business, and passing it through various data wrangling techniques for categorization into useful data, which is collected and stored in particular areas such as data warehouses, efficient analysis, data mining algorithms, which helps them decision making and other data requirements which benefits them in cost-cutting and generating revenue.

Data mining uses complex mathematical algorithms to perform data segmentation and evaluation of the probability of future decisions for the business. There are also many other terms, appearing in some articles and documents, carrying a similar or slightly different meaning, such as knowledge extraction from databases, knowledge extraction, data archeology, data dredging, data analysis, and so on (Chen et al,1996).Several typical kinds of knowledge can be discovered through KDD,including association rules, characteristic rules, classification rules,clustering, evolution, and deviation analysis (Chen et al, 1996).In the age of information there computers produce huge amount of data beyond our expectation. Data becomes too large or Complex for Classical or Manual Analysis. The number of records can be counted in millions or billions. Data in database becomes high dimensional too many fields/features/attributes exist.High rate of growth through logging or automatic data collection exist,data increase due to variety of technology is available to process it barcode, and scanners, satellites, digital camera, server logs, etc can be mentioned.Technology becomes cheap, fast and with much storage capability store data in databases, data warehouses, and digital libraries. It is not only the quality and quantity of the data; the format available is also another issue to deal with.

For this project I attempted to use KMeans Clustering to cluster Universities into two groups, Private and Public.

Mechanism Used

K Means Clustering is an unsupervised learning algorithm that tries to cluster data based on their similarity. Unsupervised learning means that there is no outcome to be predicted, and the algorithm just tries to find patterns in the data. In k means clustering, we have the specify the number of clusters we want the data to be grouped into. The algorithm randomly assigns each observation to a cluster, and finds the centroid of each cluster. Then, the algorithm iterates through two steps: Reassign data points to the cluster whose centroid is closest. Calculate new centroid of

each cluster. These two steps are repeated till the within cluster variation cannot be reduced any further. The within cluster variation is calculated as the sum of the euclidean distance between the data points and their respective cluster centroids.

Explanation of the Mechanism:

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.
The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

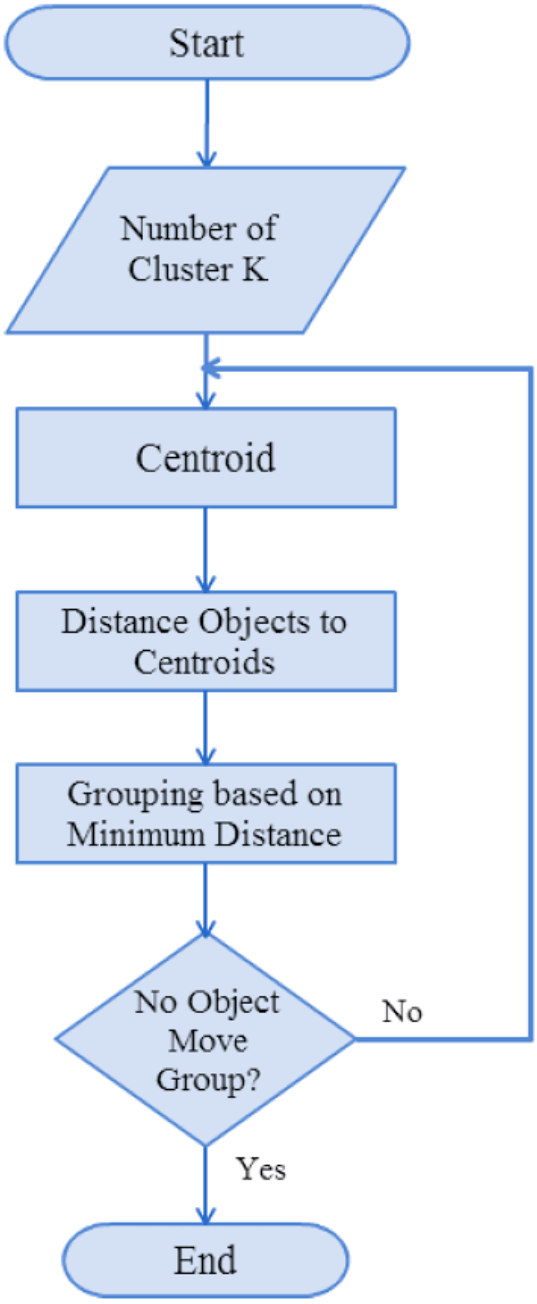
Hence each cluster has data points with some commonalities, and it is away from other clusters.

print('K Means ALGORITHIM STEPS :')

Image('0232A118-EF98-410D-B996-6A0769F9712F.png',height = 300)

K Means ALGORITHIM STEPS :

Out[ ]:



Implementation Of Mechanism Used With Python

Import Libraries

In [1]:

```
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

Create some Data

In [2]:

```
from sklearn.datasets import make_blobs
```

In [3]:

```
# Create Data
data = make_blobs(n_samples=200, n_features=2,
                  centers=4, cluster_std=1.8,random_state=101)
```

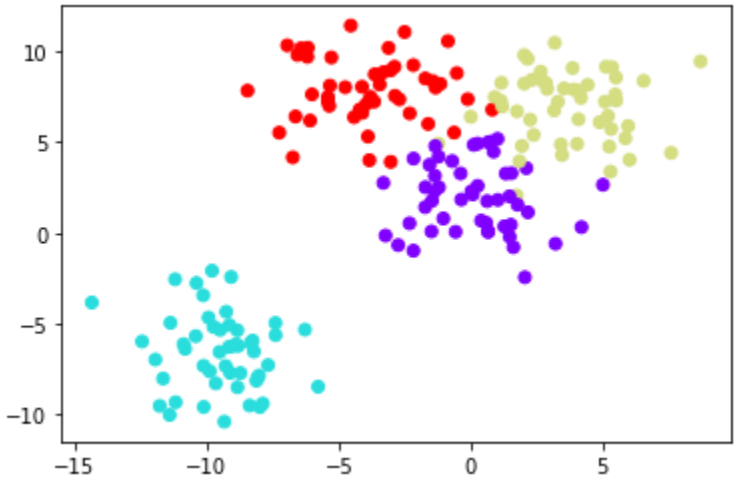
Visualize Data

In [4]:

```
plt.scatter(data[0][:,0],data[0][:,1],c=data[1],cmap='rainbow')
```

Out[4]:

<matplotlib.collections.PathCollection at 0x2209f65bfa0>



Creating the Clusters

In [5]:

```
from sklearn.cluster import KMeans
```

In [6]:

```
kmeans = KMeans(n_clusters=4)
```

In [7]:

```
kmeans.fit(data[0])
```

Out[7]:

```
KMeans(n_clusters=4)
```

In [8]:

```
kmeans.cluster_centers_
```

Out[8]:

```
array([[-4.13591321,  7.95389851],
       [-9.46941837, -6.56081545],
       [ 3.71749226,  7.01388735],
       [-0.0123077 ,  2.13407664]])
```

In [9]:

```
kmeans.labels_
```

Out[9]:

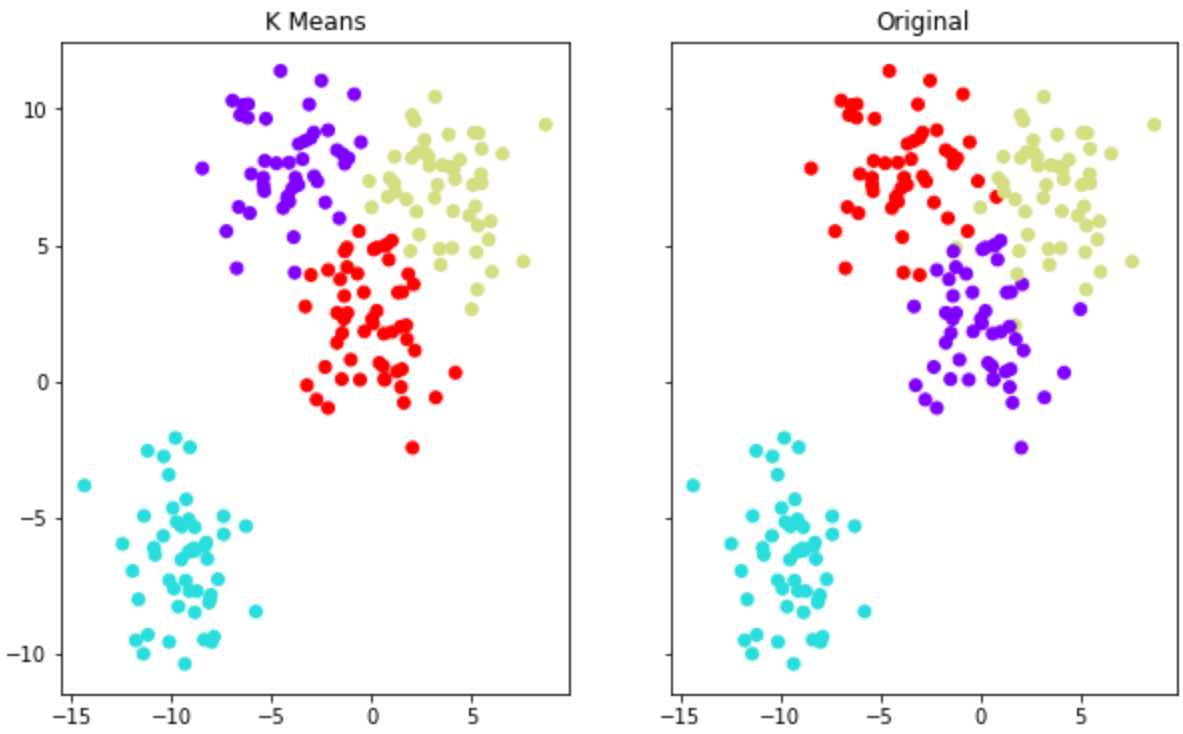
```
array([0, 2, 3, 2, 2, 1, 2, 3, 2, 3, 0, 3, 2, 2, 0, 3, 2, 3, 1, 0, 1, 3,
       3, 1, 0, 1, 1, 3, 2, 2, 0, 1, 2, 3, 3, 0, 1, 1, 1, 3, 1, 0, 0, 0,
       3, 2, 0, 3, 1, 3, 3, 0, 2, 3, 1, 0, 3, 3, 0, 2, 1, 2, 1, 0, 2, 3,
       1, 2, 2, 1, 2, 3, 1, 3, 1, 2, 2, 3, 0, 3, 3, 1, 2, 1, 3, 3, 3, 0,
       3, 1, 1, 1, 1, 3, 3, 1, 2, 0, 1, 2, 3, 1, 3, 3, 2, 3, 1, 2, 1, 1,
       2, 0, 0, 2, 1, 2, 0, 0, 2, 0, 3, 0, 3, 0, 3, 2, 0, 3, 1, 0, 0, 0,
       3, 1, 1, 0, 2, 0, 2, 3, 1, 2, 1, 0, 0, 2, 3, 1, 0, 0, 0, 0, 3, 2,
       3, 0, 2, 2, 2, 3, 2, 3, 3, 0, 1, 0, 3, 2, 0, 3, 2, 3, 0, 2, 3, 0,
       2, 2, 1, 2, 0, 1, 1, 0, 1, 1, 1, 1, 1, 3, 1, 2, 2, 0, 1, 3, 2, 2,
       1, 3])
```

In [10]:

```
f, (ax1, ax2) = plt.subplots(1, 2, sharey=True,figsize=(10,6))
ax1.set_title('K Means')
ax1.scatter(data[0][:,0],data[0][:,1],c=kmeans.labels_,cmap='rainbow')
ax2.set_title("Original")
ax2.scatter(data[0][:,0],data[0][:,1],c=data[1],cmap='rainbow')
```

Out[10]:

```
<matplotlib.collections.PathCollection at 0x220a1c6c790>
```



note, the colors are meaningless in reference between the two plots.

PROBLEM:

This project focuses on investigating the application of KDD to explore and discover patterns within the University dataset. An overview of the research field and the specifics of this particular research are presented below.

For this project I attempted to use KMeans Clustering to cluster Universities into two groups, Private and Public.

Note, I actually have the labels for this data set, but I did NOT use them for the KMeans clustering algorithm, since that is an unsupervised learning algorithm.

When using the K Means algorithm under normal circumstances, it is because you don't have labels. In this case I used the labels to try to get an idea of how well the algorithm performed, but you won't usually do this for Kmeans.

The DataSet used for the project:

I used a data frame with 777 observations on the following 18 variables.

- Private A factor with levels No and Yes indicating private or public university
- Apps Number of applications received
- Accept Number of applications accepted
- Enroll Number of new students enrolled
- Top 10 Perc Pct. new students from top 10% of H.S. class
- Top 25 Perc Pct. new students from top 25% of H.S. class
- F.Undergrad Number of full time undergraduates
- P.Undergrad Number of part time undergraduates
- Outstate Out-of-state tuition
- Room.Board Room and board costs
- Books Estimated book costs
- Personal Estimated personal spending
- PhD Pct. of faculty with Ph.D.’s
- Terminal Pct. of faculty with terminal degree
- S.F.Ratio Student/faculty ratio
- perc.alumni Pct. alumni who donate
- Expend Instructional expenditure per student
- Grad.Rate Graduation rate

Output:

Evaluation

Creating a new column for df called 'Cluster', which is a 1 for a Private school, and a 0 for a public school.

In [118]:
```
def converter(cluster):
    if cluster=='Yes':
        return 1
    else:
        return 0
```

In [119]:
```
df['Cluster'] = df['Private'].apply(converter)
```

In [122]:
```
df.head()
```

Out[122]:

	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio
Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1
Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2
Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9
Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7
Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9

In [123]:
```
from sklearn.metrics import confusion_matrix,classification_report
print(confusion_matrix(df['Cluster'],kmeans.labels_))
print(classification_report(df['Cluster'],kmeans.labels_))
```

```
[[138  74]
 [531  34]]
              precision    recall  f1-score   support

           0       0.21      0.65      0.31       212
           1       0.31      0.06      0.10       565

avg / total       0.29      0.22      0.16       777
```

Explanation of the output :

I have used a confusion matrix and classification report to see how well the Kmeans clustering worked. By the end I can conclude that it was better considering the algorithm is purely using the features to cluster the universities into 2 distinct groups which happens in our real life.

Applications of Data Mining

- Financial Analysis: The banking and finance industry relies on high-quality and processed, reliable data. In the finance industry, data can be used for a variety of purposes, like portfolio management, predicting loan payments, and determining credit ratings.
- Telecommunication Industry: With the advent of the internet the telecommunication industry is expanding and growing at a fast pace. Data mining can help important industry players to improve their service quality to compete with other businesses.
- Intrusion Detection: Network resources can face threats and actions of the cybercriminals can intrude on their confidentiality. Therefore, detection of intrusion has proved as a crucial data mining practice. It enables association and correlation analysis, aggregation techniques, visualization, and query tools, which can efficiently detect any anomalies or deviations from normal behavior.
- Retail Industry: The established retail business owner maintains sizable quantities of data points covering sales, purchasing history, delivery of goods, consumption, and customer service. Database management has improved with the arrival of e-commerce marketplaces and emerging new technologies.
- Spatial Data Mining: Geographic Information Systems and many other navigation applications utilize data mining techniques to create a secure system for vital information and understand its implications. This new emerging technology includes the extraction of geographical, environmental, and astronomical data, extracting images from outer space.

Conclusion

Data mining is a composite discipline that can represent a variety of methods or techniques used in different analytic methods that helps firms and organizations to make efficient business decisions and benefit them, they perform this by types of questions and using various levels of user input or rules to arrive at a decision. In this way, user data can be used intelligently for the benefit of the firm.