FINAL EXAM COMP9321

Question 1

Steps

- Create a dataframe df_1 from given Dataset 1
- Create another dataframe df_2 from given Dataset 2
- Merge df_1 and df_2 on 'Device ID' and create a new dataframe df
- By using Describe() function of pandas library, check the stats, mean, the median of the dataframe.
- Having all the stats about null rows and columns, further actions will be taken to clean the dataframe and normalize all the values.
- Remove all the n/a or null rows from the dataframe.
- Assuming that most of the tickets are assigned to 'Morty', replace blank or dirty values of that column with 'Morty'.
- Assuming that the 'Quality Tested Date/Time' and 'Support Ticket Date/Time' columns are string converting the same into datetime using to_datetime() of pandas library for uniformity.
- Finally, on the cleaned dataframe we can scatter plot and plot heatmaps and histograms to get to know correlations and other insights which human eye can miss.
- Also, final dataframe can easily be converted into CSV for user readability and portability.
- Thus, insights can be drawn of this clean and uniform dataset/CSV file.

Question 2

A) K-means clustering is chosen by me to find different groups of customers. K-means algorithm creates different clusters and for this question our clusters are groups.

B)

Steps

- Select k random cluster centres and assign all the data points to the closest centre.
- Calculate the mean of the clusters and assign them as the new centre points.
- Assign the data points to the nearest centres.
- Repeat this process until there are no new centre points.

Suppose we take 2 new centre points. Centre 1 = 17, Centre 2 = 23.
Iteration 1:

Centre 1 = 17	Centre 2 = 23
A	C
B	E
D	
⇨ A + B + D / 3 ⇨ 18 + 7 + 12 / 3 ⇨ 12.33	⇨ C + E / 2 ⇨ 22 + 24 / 2 ⇨ 23

Iteration 2:

Centre 1 = 12.33	Centre 2 = 23
B	A
D	C
	E
⇨ B + D / 2 ⇨ 7 + 12 / 3 ⇨ 6.33	⇨ A + C + E / 2 ⇨ 18 + 22 + 24 / 2 ⇨ 21.33

Iteration 3:

Centre 1 = 6.33	Centre 2 = 21.33
B	A
D	C
	E
⇨ B + D / 2 ⇨ 7 + 12 / 3 ⇨ 6.33	⇨ A + C + E / 2 ⇨ 18 + 22 + 24 / 2 ⇨ 21.33

Since we don't observe any change in the centres now. We will finalize Centre1 as 6.33 and Centre 2 as 21.33.

C) We will take k = 2 as in initial to find the number of clusters and then gradually increase the value of k and then find the number of clusters in each iteration and select the value which generates the greatest number of clusters.

Question 3

N_1 = 10
N_2 = 90
N_3 = 10

- N_1 as 10 because they are the number of folds, as in how many times the iteration will take place
- N_2 as 90 because of it the training dataset on which the models will train for each fold
- N_3 as 10 because it is the test dataset on which trained models will provide the results for each fold

Now from each fold we will get 10 errors and taking the average from the errors we will get our desired result.

Question 4

True Positive = TP
True Negative = TN
False Positive = FP
False Negative = FN

Precision

- ⇨ (TP) / (TP + FP)
- ⇨ 8 / 8 + 2
- ⇨ **0.8**

Recall

- ⇨ (TP) / (TP + FN)
- ⇨ 8 / 8 + 12
- ⇨ **0.4**

F1 score

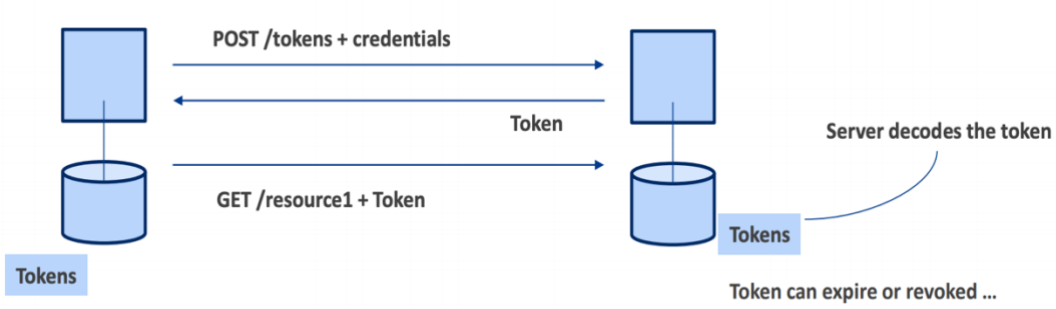
- ⇨ 2 * Precision * Recall / (Precision + Recall)
- ⇨ 2 * 0.8 * 0.4 / (0.8 + 0.4)
- ⇨ 0.64 / 1.2
- ⇨ **0.53333**

Question 5

A) A Token Base authentication system

Only for the first time, the user enters credentials and the server authorizes it and returns a token. This token is stored on the client-side and sent in every request that follows and no plain password. It must be used with TLS.

JWT (JSON Web Token) is an industry-standard which has majorly three parts and is self-signed and uses to authenticate itself. The server will authenticate the token without database lookup, and it manages statelessness and login password of the user is never revealed.



B) Rate limiting through API Keys

Rate limiting is another API security technique which prevents farming and manages costs so that the organization can track how many and from where the API is being used and can charge the client using the API accordingly.

It also controls DDoS Attack by expecting API keys in every request to the protected endpoint.

This will return 429 Too Many Requests HTTP response code if requests are coming in too quickly. The organization can also revoke the API key if the client violates the user agreement.

Question 6

Code

201 Created

Response Body


```
{
  "uri": "/orders/1101",
  "order_id": 1101,
  "creation_time": "2020-05-05 18:27:25",
  "drink": "latte"
}
```


Response Headers

content-length: 127
content-type: application/json
date: Tue, 05 May 2020 08:27:27 GMT
server: Werkzeug/0.16.1 Python/3.7.4

Explanation:
In this example, we are creating a resource on the successful execution of the given post request. Thus, returning the HTTP code 201 stands for created. Next returning uniform resource indicator (URI), order_id, creation_time and the drink in the Response Body.
Also, Response headers are returned which provides the information of the returned body such as the content-length, type and datetime and the server.
This facilitates the user who have minimal knowledge can also get his/her head around to understand the API response and structure.

Question 7

A) I have taken the KNN classification algorithm to solve this part of the exam.

- KNN is an instance map to points R^N.
- And this has the attributes less than 20 per instance.
- Also, we have an ample amount of training data and which is very fast.
- This never lose information and learn complex target functions.

Limitations:

- KNN is very slow at the query level.
- This classification can be easily fooled by irrelevant attributes.
- The bigger the dataset the better the performance.

B)

Steps

- Dataframes df_1 and df_2 area created from the given files.
- Both the dataframes are then merged on 'Identifier of people' and 'Identifier of people close to Device'.
- Using 'Duration of contacts (minutes)' and 'Timestamp' and new data 'Start time' and 'end time' extracted to put it the categorical column where it is divided into 'Morning', 'Evening', 'Afternoon' and 'Night' as values, dropping the same afterwards.
- Also, by using 'GPS Location' and 'Distance' and calculate and create the 'Area' column and put as weights, dropping the same afterwards.
- Put 'Infection Tested' in other categorical columns with 0 and 1 as values.
- Dropping other columns.
- By one-hot encoding we can find the clustering of the data.
- Thus, now we have cleaned data and can scatter plot or create heatmaps get more insights and correlation.
- And this dataframe can be exported in CSV file are better user readability and portability.

Question 8

A) *R* is not content-based because it predicted *B* but the not movie *C*. Therefore It must be *U-U* or *I-I* collaborative filtering. Also if we see the final prediction was *B* as the most popular movie. So, it clear that it is collaborative.

B) As we don't have any information about this new user U_2, it is not able to recommended by *R*. But if we have the information like interests of the user U_2 we could have used this to recommend movies.

   a. **User-User Collaborative filtering:**
      No recommendations on cold start will be given. Having 0.1 we can make *U-I* matrix to get which user have seen what all movies and user Cosine/Jaccard similarities to get the top similar users and then the most they have been watching.
   b. **Item-Item collaborative filtering:**
      We will similar items and also all the items mentioned above.