Q1

We could see from the dataset1 that there are many empty values in the fourth row, and my option is to delete the entire row. And from dataset2, ‘the ticket handled by’ column of the last row also missing value, and we can just replace by using “Morty”. Because of lack of resources, we don’t want to lose this row,

And then we also could use the pandas to select the useful column for the company, which respectively are Quality Tested Date/Time, operator, Support Ticket Date/time, Ticket Handled by

Based on the same deviceID.

Q2

1. K-means algorithm**.** Because it is a classification problems and we could just use classify the cusomers who purchase different number of items.
2. According to the set clustering number K, randomly select K clustering centers

Evaluate the distance between each sample and the cluster center. If the sample is closer to the i-th cluster center, it is considered to belong to the i-th cluster.

Calculate the average (Mean) position of the samples in each cluster, and move the cluster center to this position.

Repeat the above steps until the position of each cluster center no longer changes.

So that means whenever there are new customers which are being statistic, we could use that algorithm to group them, and get the predicted item which they buy.

3. Firstly, We could use different distance formula to find the best group, and we also could use different score to evaluate our models , such as MSE.

Q3

N1 = 10, N2 = 90, N3 = 10

Because it is 10-fold cross-validation, we can get N1 = 10.

We should separate the dataset into two parts for the cross validation in order to prevent the overfit, and the validation set has 90 examples and testing set has 10. So that means N2 = 90, N3 = 10.

Q4

tp = number of true positives = 8

fp = number of false positives = 2

tn = number of true negatives= 12

fn = number of false negatives = 11

Precision score = tp / (tp + fp) =0.8

recall score = tp / (tp + fn) = 0.4

F1-score score = 2 * (precision * recall) / (precision + recall) = 2 * 0.8 *0.4/ 0.8+0.4 = 0.53333

Q5

1. JWT. Because it is a Token-based method. First the token is self contained. And it also use Keyed-Hash Message Authentication Code (HMAC), which provides integrity in security.
2. API Keys Method. Because API provider can use it to limit the usage rate, for example 8 calls per second.

Q6.

```
HTTP/1.1 200 OK
Server: Apache
Date: Sat,31 Dec 2005 23:59:59 GMT
Content-Type: text/html; charset=ISO-8859-1
Date: Wed, 20 May 2020 09:54:43 GMT
Keep-Alive: timeout=5, max=1000
Connection: Keep-Alive
Age: 3472
Date: Wed, 20 May 2020 010:54:43 GMT
X-Cache-Info: caching
Content-Length: 122
```

response format
/coffeeOrders

```
{
    “Id”: “1”;
    “type”: ”latte”,
    “extra”: “no”,
    “server”: “peter”
    Payment:{
        “time”: “2020-05-20”,
        “card”: “20392948”
    }
}
```

The json document contains the id of the order and what kind of drinks and the waiter/waitresses and also the payment method, which contains the date and card information.

Q7

1. I will use the decision trees.

Because we need to rapidly have the numberic prediction, and therefore we can choose some key attributes of the dataset to predict who are the people at risk.

Limitation:

unstable, sometimes a small change in the data can lead to a large change. And they are often relatively inaccurate caused by overfitting.

2. Distance/Duration of contact/GPS Location Timestamp/age

I will change the time into timestamp format and put it into models and the gps location should be changed into one number. If the mse is too large, I maybe abandon the gps location.

Q8

1. I think it is the User-based collaborative filtering and Item-based collaborative filtering.

Content-based recommender system is not possible because user U1 is interested in the time period 2000s, the director D2 and the genre Comedy. But the type of b is 2010,comedy,D1. The favorite content is different.

As for User-based collaborative filtering, because the rating of B is 500, what means it has liked by a lot of people, some of them may be u1's friend. So it can be understand that the system just recommend his friend's favorite movie.

As for Item-based collaborative filtering, the recommendation system will calculate items by using people's rating of those items. In this case, some other movies like C have high rating which is similar to movie B, so B is recommended to u1.

2. If U2 has rated some movies, the recommendation system will recommend U2 a movie that is because using the Item-based collaborative filtering can recommend similar items.

But if U2 has not rated movies but his/her friend who has the same interests with him/her, our recommendation system will also recommend U2 a movie because the User-based collaborative filtering will provide the movies liked by others.

Otherwise R cannot recommend a movie because u1 has zero relation with movies.