AI/ML - Test 3 - M. Rachal.

1) What is 'training set' and 'test set' in a Machine Learning Model, Give examples.

In the Machine Learning, when we need to analyze the data, the data preprocessing should be done, before we go to training and testing the model. the dataset should be divided as x and y, ~~[illegible] should be independent variables and [illegible] be [illegible] the dependent variables~~, x will be the independent variable, dropping the target variable and y will be the target variable, From sklearn we access the model selection and import the train_test_split, which the preprocessed data and named as x and y. ~~[illegible]~~ The dataset split the data to training set and testing with the parameters of test size, random state shuffle. Through train_test_split we can build ML models and predict the target values with Actual target values.

Ex:
```
from sklearn.model_selection import train_test_split.
    x_train, x_test, y_train, y_test = train_test_split (x, y, test_size=0.3
                                                        random_state=1)

print(x_train.shape)
Print(x_test.shape)
print(y_train.shape)
print(y_test.shape)
```

(2). How missing and corrupted data is handled in dataset?

When we load the dataset, the data has been collected without any proper preprocessing, it will be messy data to analyze and do prediction. So we need to preprocess the data by finding the missing values, nan values and outliers. we also need to ensure that by finding how much Nan values or missing values present in the data. It can be handled by dropping the null values or impute it with the amount of ~~data~~ null values present in the data.

we can drop by.

ex: df.drop(['class'], inplace = True)

or we can impute it by '0' or '1' or by mean, median and modes.

=> df.fillna.

Difference b/w precision and recall.

(3). **Precision:**

The total no. of predicted positive values.

$$\text{precision} = \frac{\text{True positive}}{TP + FP}$$

Recall:

the total no. of Actual positive values.

$$\text{Recall} = \frac{\text{True Positive}}{TP + FN}$$

(4). **Support vectors in SVM?**

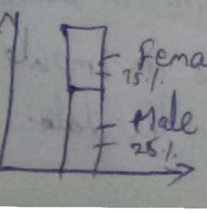
In Support Vector Machines the support vectors are the data points that are nearer to the hyper plane. which the hyper planes are help to avoid the outlier and reduce the over fitting of data. The maximum datapoints nearer to the hyperplane that lies under the boundary lines are support vectors.

(5) Significance of hue, size, style.

Hue:

Hue is the parameter of Seaborn that helps to differentiate the different class of categorical feature.

for ex sns.barplot (df, hue = df ['Gender'])



Size: size is the parameter of seaborn, it helps to handle the size of image that has been visualized.

style: style is the parameter of seaborn, it handles the the plot like background color, line style as '--' '-.' or darkgrid, ticks for style for column name in visualization.

6. Colors are Effectively used in data visualization.

- First visualizing the data stands important than calculating the data with built in functions.
- Colors plays essential role in data visulization to understand the data more effectively.
- the colours parameters helps in identify Highest ranking values when we visualize the rating for different brands.
- In heatmap 'c-map' helps in identify the correlation of the data, by the transparency on light to dark color difference

7. Characteristics of Effective data visualization:

- Through line plot we can identify the trends of data based on high range and low range.
- Through the boxplot we can able to understand the outliers of the data and IQR range.
- Through 'KDE' we can able to identify the skewness of the data, how the data has been distributed, whether we need to do data transformation or we have normal distribution
- Through the 'marker' parameter such as '*' '+', 'x' we can identify the distribution of data points with the markers.
- with the 'hue' parameter it helps in decision making to focus more on the large.

8) Ecommerce or Entertainment how reccommendation system are working:

Example: Netflix

If we watched "slice of life" or "horror" genre movies. Next we log in again to OTT platform, the home page give "Top 10 reccommadation of horror movies' which you might be interested in, by based on the movie you watched lastly or recently.

9. Real life Example for Clustering

cluster 1: small family, high spenders.
cluster 2: Larger family, high spenders
cluster 3: Small family, low spenders
cluster 4: Large family, low spenders.

10 When the value of K increases, there will be few elements in the clusters, which the smaller the value will be results in under-clusters and the larger the value can cause over clusters. To avoid this we should pick the K at the spot at sum of squared distance to flatten out and form an elbow.