1 GENERAL ML QUESTIONS

1.1 IDENTIFY A LEARNING PROBLEM

[7 points] Describe the movie recommendation learning problem by stating as precisely as possible the task, performance measure, and training experience.

Task- Recommending the movies from collection of movies datasets for a user based on his preferences

Performance Measure- percent of movies correctly recommended for a user

Training- Collection of movies, their rating, Genres, Actor/ Actress, Director, Production House

1.2 MACHINE LEARNING TYPES

• Predict the stock market index Supervised algorithm – regression we want to map input to a continuous output

• Identify whether or not an alumni is going to donate to UTD :- Supervised algorithm - Classification

Supervised machine algorithm as they will be used to learn to classify new observations

• Recommend online courses that are better taken together:- Unsupervised algorithm – others (Clustering) as it does not have labeled outputs, so its goal is to infer the natural structure present within a set of data points.

• Segment customers based on social-demographic attributes :- Unsupervised algorithm - others as it does not have labeled outputs, so its goal is to infer the natural structure present within a set of data points

Suppose we want to predict whether a flight on a particular day will be canceled based on three factors: whether there is a snowstorm, where it is an official holiday, and whether it is a long-distance (≥ 4 hours) flight. Table 2.1 shows the training examples. Please answer the following three questions.

No. Snowstorm Holiday Long Distance Canceled

No.	Snowstorm	Holiday	Long Distance	Canceled
1	T	F	T	No
2	T	F	T	Yes
3	F	F	T	No
4	T	F	T	No
4	F	F	F	No
6	F	F	F	Yes
7	T	F	F	Yes
8	F	F	F	Yes

Table 2.1: Flight Data Training Examples

1. If we chose “Holiday” as the root of a decision tree, what would be the effects? Explain

in terms of information gain. [4 points]

We can not select “Holiday” as the root of decision tree as all the observations belong to same category and we cannot split further. hence it does not represent the whole dataset.

Entropy (Holiday, F) = 1 as all are from same category

Entropy (Flight dataset) = 1 (Yes =4, No= 4)

Information Gain (Holiday) = 1-1 =0

2. If “Holiday” is not proper as the root node, which attribute will you choose as the root

node of the decision tree? Explain your reasons with necessary calculations. [6 points]

The maximum information gain is for Long Distance. The information gain for Snow Strom and Holiday is 0. Hence the Root node should be Long distance as it has information gain is 0.2

Below table shows the calculation-

		Canceled flight					
		Yes	NO	Total	Entropy	Total obs	Avg entrophy
Long dista	T	1	3	4	0.8112781	8	0.405639062
	F	3	1	4	0.8112781		0.405639062
Information Gain		0.188721876					

		Canceled flight					
		Yes	NO	Total	Entropy	Total obs	Avg entrophy
SnowStorm	T	2	2	4	1	8	0.5
	F	2	2	4	1		0.5
Information Gain		0					

3. Describe your complete decision tree in words (sample format: If there is a snowstorm, the flight will be canceled). [5 points]

if there is long distance and snowstorm, it is most likely that flight will not be cancelled.

if there is long distance and no snowstorm, then flight will not be cancelled.

If there is no long distance and no snow storm, it is most likely that flight will be cancelled.

if there is no long distance and there is snowstorm, then flight will be cancelled.