

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST-2 EXAMINATION – DECEMBER 2022

M.Tech 1st Semester (Introduction to Data Science)

COURSE CODE: 22M11CI112

MAX. MARKS: 25

COURSE NAME: Introduction to Data Science

COURSE COORDINATOR: Dr. Simran Setia

COURSE CREDITS: 3

MAX. TIME: 1.5 HRS

*Note: All questions are compulsory. Carrying mobile phone during examinations will be treated as a case of unfair means.*

Q1. Suppose the following data is collected to judge whether there is a relationship between political party preference and opinion on tax reform. Perform hypothesis testing on the given data to check the relationship between political party preference and opinion on tax reform. 6 marks [CO5]

Opinion-> Party	Favor	Indifferent	Opposed
Democrat	138	83	64
Republican	64	67	84

Q2. Calculate the probabilities for the following scenarios: [CO5]

- A. In a neighborhood, 90% of the children were falling sick due to flu and 10% due to measles. The probability for observing rashes for measles is 0.95 and for flu is 0.08. If a child develops rashes, find the probability of the child having flu. 2 marks
- B. The average number of acres burnt by the forest fires in New Mexico is 4300 acres per year with a standard deviation of 750 acres. The distribution of the number of acres burnt is normal over different years. What is the probability that in a given year an area of forest between 2500 and 4200 acres will be burnt? 2 marks
- C. Suppose you are playing a game of darts. The probability of success is 0.4. What is the probability that you will hit the bull's eye on the third try? 2 marks

Q3. Suppose a data scientist has been delegated the task of predicting the median house price of houses in a particular district. The data given for the above problem comprises of following features:

- A. Longitude
- B. Latitude
- C. Median age of housing
- D. Total rooms
- E. Total bedrooms

- F. Population of area
- G. Median income
- H. Ocean Proximity

In order to predict the median housing price, the data scientist performs the following steps:

- He reads the data set and splits it into training and testing data set with 80:20 ratio.
- He then applies linear regression model with X as features and Y as median housing price.
- He obtains an RMSE of around 0.

Evaluate the procedure followed by the data scientist. You may add/delete/modify the steps given in the above procedure. 8 marks [CO6]

Q4. Encode the following data set (only the columns Gender and Remarks) using an appropriate technique/techniques. 5 marks [CO3]

Employee_ID	Gender	Remarks
45	Male	Nice
78	Female	Good
56	Female	Great
12	Male	Great
7	Female	Nice