Q1

You are working on a team that is tasked with developing a machine learning model to predict housing prices in a particular city. Your team has been given a dataset containing information about various houses in the city, such as the size of the house, the number of bedrooms, the year it was built, and the price it was sold for.

Sample Dataset:

Size (sq ft)	**Bedrooms**	**Year Built**	**Price (USD)**
1500	3	1990	250000
2000	4	1985	300000
1200	2	2000	200000
1800	3	1970	275000
2200	4	1988	350000
1400	2	1995	225000
1600	3	2005	275000
2400	4	1975	400000
1900	3	1998	325000
1700	2	1980	250000

Your task is to develop a model that predicts the price of the house based on given n features/parameters. The data split should be 70-30 with 70 being the Training data and 30 being the Testing data.

Q2

You are working for a telecommunications company and tasked with developing a machine learning model to predict customer churn. Customer churn occurs when a customer stops using the company's services, such as canceling their phone or internet plan.

Your team has been given a dataset containing information about various customers, such as their age, gender, location, monthly charges, and the services they are subscribed to. You will use this data to train a decision tree model to predict which customers are at the highest risk of churn.

Sample Dataset:

Customer ID	Age	Gender	Location	Monthly Charges	Internet Service	Phone Service	TV Service	Churn
1	35	Male	NY	50	Fiber optic	Yes	Yes	Yes
2	44	Female	CA	70	DSL	Yes	No	No
3	22	Male	TX	30	DSL	Yes	No	Yes
4	55	Female	NY	80	Fiber optic	Yes	Yes	No
5	33	Male	CA	45	DSL	Yes	No	No
6	20	Female	TX	25	DSL	Yes	No	Yes
7	68	Male	NY	100	Fiber optic	Yes	Yes	Yes
8	50	Female	CA	60	DSL	Yes	No	No
9	27	Male	TX	35	DSL	Yes	No	No
10	41	Female	NY	90	Fiber optic	Yes	Yes	Yes

Splitting of data into testing and training is up to you, but make sure that the training data is a bit bigger than the testing data.

Q3

You are working for a botanic garden and tasked with developing a machine learning model to classify Iris flowers based on their physical characteristics. The garden has collected data on three different types of Iris flowers: Iris setosa, Iris versicolor, and Iris virginica. Each flower has been measured for its sepal length, sepal width, petal length, and petal width.

Your team will use this data to train a K-nearest classifier model that can predict the type of Iris flower based on these four physical characteristics.

Sample Dataset:

ID	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
3	4.7	3.2	1.3	0.2	Iris setosa
4	7.0	3.2	4.7	1.4	Iris versicolor
5	6.4	3.2	4.5	1.5	Iris versicolor
6	6.9	3.1	4.9	1.5	Iris versicolor
7	6.5	3.0	5.2	2.0	Iris virginica
8	6.2	3.4	5.4	2.3	Iris virginica
9	5.9	3.0	5.1	1.8	Iris virginica

Again, the splitting of data into testing and training data sets is up to you.

Q4

You work for a financial institution and are tasked with developing a machine learning model to identify fraudulent credit card transactions. The institution has collected data on past transactions, including information such as transaction amount, location, and time, as well as whether or not the transaction was fraudulent.

Your team will use this data to train a Support Vector Machine (SVM) model that can predict whether a new transaction is fraudulent or not based on these features.

Sample Dataset:

ID	Amount	Location	Time	Fraudulent
1	100.00	USA	08:15	No
2	200.00	Canada	13:45	No
3	50.00	USA	19:30	Yes
4	75.00	Mexico	11:00	No
5	300.00	USA	15:20	Yes
6	150.00	USA	22:00	Yes
7	25.00	Canada	09:00	No
8	500.00	Mexico	17:30	Yes
9	80.00	USA	14:00	No

Again, the splitting of data into testing and training data sets is up to you.

Q4

You work for a marketing firm and are tasked with developing a customer segmentation strategy for a new client. The client has collected data on their customers, including demographic information such as age, gender, income, and education level, as well as purchase history data such as total spending, purchase frequency, and items purchased.

Your team will use this data to perform K-means clustering analysis to identify distinct customer segments based on their purchasing behavior and demographic characteristics.

Sample Dataset:

Customer ID	**Age**	**Gender**	**Income**	**Education**	**Total Spending**	**Purchase Frequency**	**Items Purchased**
1	24	Male	40000	College	500	10	20
2	35	Female	60000	College	1000	5	10
3	50	Male	80000	Graduate	2000	2	5
4	42	Female	70000	Graduate	1500	4	12
5	28	Male	45000	College	800	8	18
6	38	Female	90000	Graduate	2500	1	3
7	56	Male	120000	Postgrad	3000	1	5
8	33	Female	55000	College	1200	6	15
9	45	Male	65000	Graduate	1800	3	7

Follow the below steps to achieve required results:

- Encoding the categorical features (gender, education) and scaling the numerical features (age, income, total spending, purchase frequency, items purchased).
- Determine the optimal number of **K.**
- Examine the characteristics of each cluster to gain insights into customer behavior and preferences. For example, you may find that one cluster consists primarily of young, low-income customers who make frequent, small purchases, while another cluster consists of older, high-income customers who make infrequent, large purchases.