ASSIGNMENT-2

Q.1. Construct a Linear Regression Model for the following data set. Predict the price for a house with area 2200 sq ft and 3 bedrooms.

Size (sq ft)	Number of Bedrooms	Price (in $)
1000	2	300,000
1500	3	450,000
2000	3	500,000
2500	4	600,000
3000	4	700,000

Q.2. Construct a Linear Regression Model for the following data set. Predict the price for a car with age 3 years and mileage 50,000 miles.

Age (years)	Mileage (miles)	Price (in $)
1	10,000	25,000
3	30,000	19,000
5	50,000	15,000
7	70,000	12,000
10	100,000	8,000

Q.3. Actual and predicted values are given for 10 students in the following data set.

STUDENTS	ACTUAL DATA	PREDICTED DATA
1	PASS	PASS
2	PASS	FAIL
3	FAIL	PASS
4	FAIL	FAIL
5	PASS	PASS
6	FAIL	FAIL
7	PASS	PASS
8	PASS	FAIL
9	FAIL	PASS
10	FAIL	FAIL

(a) Find the recall and precision for the above dataset.

(b) Find the F-1 score and confusion matrix for the above dataset.

Q.4. Use Logistic Regression to predict the result (pass or fail) of a student based on number of study hour. Use the model to predict whether a student will pass if the number of study hours is

(a) 4 (b) 6

Hours Studied	Passed (1 = Yes, 0 = No)
1	0
2	0
3	0
4	1
5	1
6	1
7	1

Use the Threshold Function as $\boldsymbol{Thresh}(\boldsymbol{x}) = \begin{cases} \mathbf{0}, & \boldsymbol{x} < \mathbf{0.5} \\ \mathbf{1}, & \boldsymbol{x} \geq \mathbf{0.5} \end{cases}$ and $\boldsymbol{w} = (\boldsymbol{w_0}, \boldsymbol{w_1}) = (\mathbf{-6}, \mathbf{1.5})$.

Q.5. Construct a Decision Tree for predicting whether a student will pass an exam based on the features: Study Hours and Number of Homework Completed.

Study Hours	Homework Completed	Passed Exam (Target)
A1	B2	0
A2	B1	0
A2	B3	1
A3	B1	0
A3	B2	1
A4	B4	1

Q. 6. Suppose we have the following dataset. In this dataset, there are four attributes. And on the basis of these attributes, make a Decision Tree.

Age	**Competition**	**Type**	**Profit**
Old	Yes	software	Down
Old	No	software	Down
Old	No	hardware	Down
Mid	yes	software	Down
Mid	yes	hardware	Down
mid	No	hardware	Up
mid	No	software	Up
new	yes	software	Up
new	No	hardware	Up
new	No	software	Up

Q.7. Derive the Mean Squared Error function in matrix form for the Linear Regression Model, i.e. Derive $L(w) = \frac{1}{m}(Xw - y_{true})^T(Xw - y_{true})$.

Q.8. Derive the minimum value for the MSE $L(w)$ using gradient. i.e. Derive $w = (X^TX)^{-1}X^Ty_{true}$.

Q. 9. Use K-Means Algorithm (up to the second iteration) to cluster the following eight points into three clusters: A(2, 10), B(2, 5), C(8, 4), D(5, 8), E(7, 5), F(6, 4), G(1, 2), H(4, 9) with A(2, 10), D(5, 8) and G(1, 2) as the initial cluster centroids. Consider the following two distance functions in two different cases.

(a) The distance function between two points $a = (x_1, y_1)$ and $b = (x_2, y_2)$ is defined as $d(a, b) = |x_1 - x_2| + |y_1 - y_2|$.

(b) The distance function between two points $a = (x_1, y_1)$ and $b = (x_2, y_2)$ is defined as $d(a, b) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$.

Q.10. The dataset of pass or fail in an exam of 6 students is given in the table.

Study Hour (X)	29	16	33	28	29	30
Pass (1), Fail (0)	0	0	1	0	1	1

Use logistic regression as classifier, calculate the probability of pass for the student who studied 30 hours, where $\boldsymbol{w = (w_0, w_1) = (-64, 2)}$.