Internal Assessment Test 1- March 2025

Sub:	**Machine learning-1**			Sub Code:	**BCS602**	Branch:	CSE- AIML
Date:	**25/03/2025**	Duration:	90min's	Max Marks: 50	Sem/Sec:	VI CSE-AIML	OBE

	SCHEME AND SOLUTIONS	MARKS	CO	RBT
1a)	Differentiate between the Supervised and Unsupervised machine learning techniques by taking suitable examples.	5	CO1	L2

Scheme: -

Difference with example -2.5 Marks each

Solution: -

Supervised	Unsupervised
Labelled data	Unlabeled data
Used for classification and regression	Used for clustering
Assigns labels or categories	Used for grouping
Example+Algorithms	Example+Algorithms

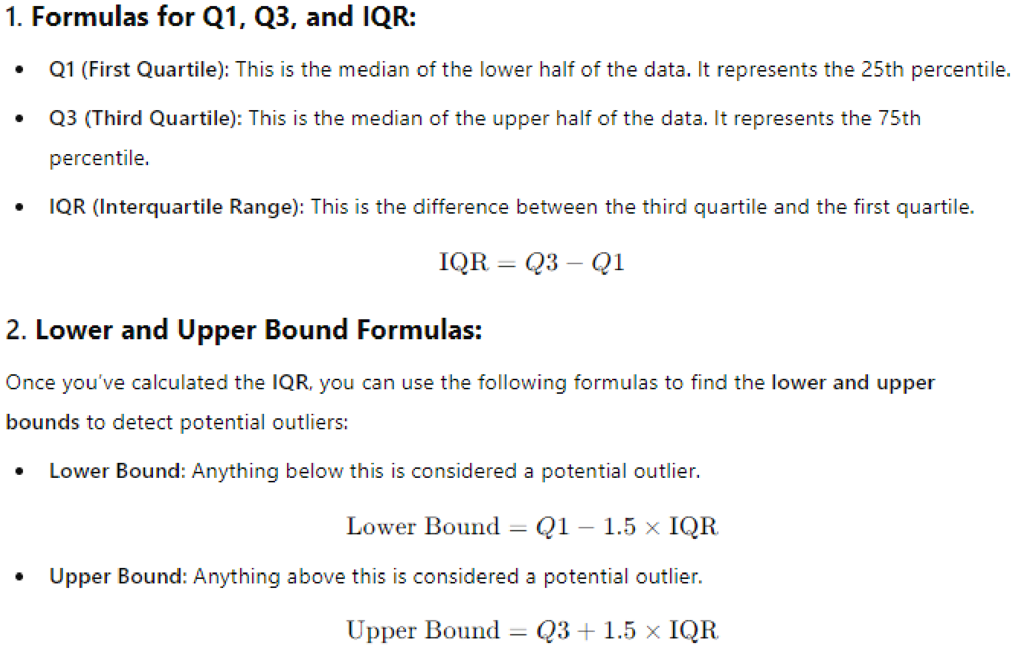
		MARKS	CO	RBT
1.(b)	What is an outlier? Explain the technique to determine an outlier from the following- 1,3,6,8,12,14,18,20,24,30,50.	5	CO1	L3

Scheme: -

Outlier-1Mark
Interquartile Range and formula for computing outlier-2Marks
Problem solving-2Marks

Solution: -

Outlier is a datapoint which is significantly different from other datapoints in a dataset.



1. Formulas for Q1, Q3, and IQR:

- **Q1 (First Quartile)**: This is the median of the lower half of the data. It represents the 25th percentile.
- **Q3 (Third Quartile)**: This is the median of the upper half of the data. It represents the 75th percentile.
- **IQR (Interquartile Range)**: This is the difference between the third quartile and the first quartile.

$$\text{IQR} = Q3 - Q1$$

2. Lower and Upper Bound Formulas:

Once you've calculated the **IQR**, you can use the following formulas to find the **lower and upper bounds** to detect potential outliers:

- **Lower Bound**: Anything below this is considered a potential outlier.

$$\text{Lower Bound} = Q1 - 1.5 \times \text{IQR}$$

- **Upper Bound**: Anything above this is considered a potential outlier.

$$\text{Upper Bound} = Q3 + 1.5 \times \text{IQR}$$

Problem Solution: There is no outliers(LB=-21 UB=51)

		MARKS	CO	RBT
2(a)	Explain the various types of categorical data and numerical data.	5	CO1	L2

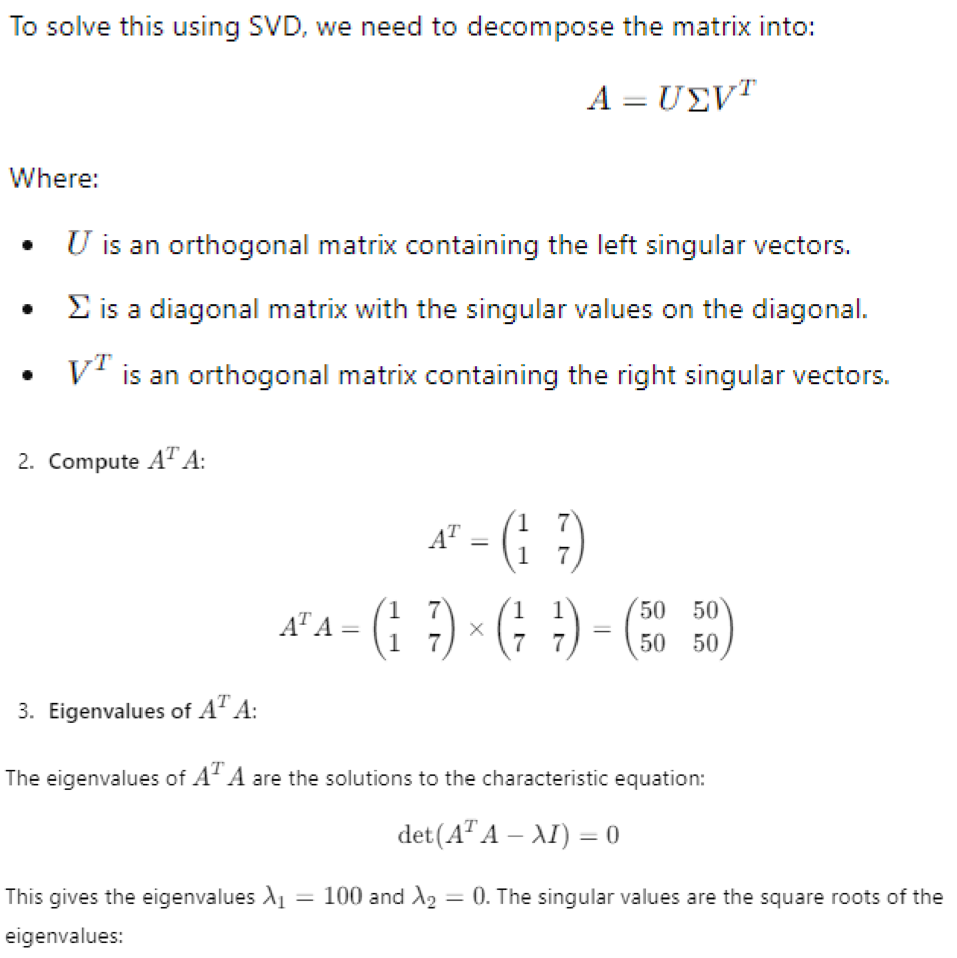
	Scheme and Solution: Categorical data types-Nominal and Ordinal Type with example 2.5Marks Numerical data types-Interval and Ratio data with example 2.5Marks			
2(b)	Explain the measures of skewness and kurtosis. **Scheme :-** Skewness, its types-2.5Marks Kurtosis-2.5Marks **Solution :-** **Skewness -**shows the direction and degree of symmetry. Positive and negative skewness. **Coefficient of skewness** $\text{Skewness} = \dfrac{3(\bar{x} - \text{Median})}{s}$ **Kurtosis-** • It indicates the peaks of data • A measure of whether the data is heavily tailed or light tailed w.r.t normal distribution • High or low kurtosis	5	CO1	L2
3(a)	Explain the role and different types of probability distributions in machine learning. **Scheme:** Normal Distribution-2Marks Rectangular Distribution-1Mark Exponential Distribution -1Mark Binomial Distribution -2Marks Poisson Distribution -1Mark Bernoulli Distribution -1Mark-1Mark **Solution:** Write about each distribution with the necessary formulas.	8	CO2	L2
3(b)	What is the curse of dimensionality? **Scheme –** Definition 2Marks **Solution:-** As the dimensionality increases the model complexity also increases. So it may effect the expected prediction and accuracy	2	CO2	L1
4a)	Apply SVD on the following matrix $\begin{bmatrix} 1 & 1 \\ 7 & 7 \end{bmatrix}$	5	CO2	L3

Scheme:-

Compute the three matrices for SVD Decomposition -4Marks

Eigen Value computation-1Mark

Solution:-



To solve this using SVD, we need to decompose the matrix into:

$$A = U\Sigma V^T$$

Where:

- U is an orthogonal matrix containing the left singular vectors.
- Σ is a diagonal matrix with the singular values on the diagonal.
- V^T is an orthogonal matrix containing the right singular vectors.

2. **Compute A^TA:**

$$A^T = \begin{pmatrix} 1 & 7 \\ 1 & 7 \end{pmatrix}$$

$$A^TA = \begin{pmatrix} 1 & 7 \\ 1 & 7 \end{pmatrix} \times \begin{pmatrix} 1 & 1 \\ 7 & 7 \end{pmatrix} = \begin{pmatrix} 50 & 50 \\ 50 & 50 \end{pmatrix}$$

3. **Eigenvalues of A^TA:**

The eigenvalues of A^TA are the solutions to the characteristic equation:

$$\det(A^TA - \lambda I) = 0$$

This gives the eigenvalues $\lambda_1 = 100$ and $\lambda_2 = 0$. The singular values are the square roots of the eigenvalues:

$$\sigma_1 = 10, \quad \sigma_2 = 0$$

$$U = \begin{pmatrix} 0.14 & -0.99 \\ 0.99 & 0.14 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 10 & 0 \\ 0 & 0 \end{pmatrix}$$

$$V^T = \begin{pmatrix} 0.70 & -0.70 \\ 0.70 & 0.70 \end{pmatrix}$$

4b) Explain the PCA algorithm with the necessary steps. 5 CO2 L2

Scheme:-

Each step carries 1 mark.

Solution:-

1.The mean is subtracted from the dataset(x-m). This is to transform the dataset with zero mean.

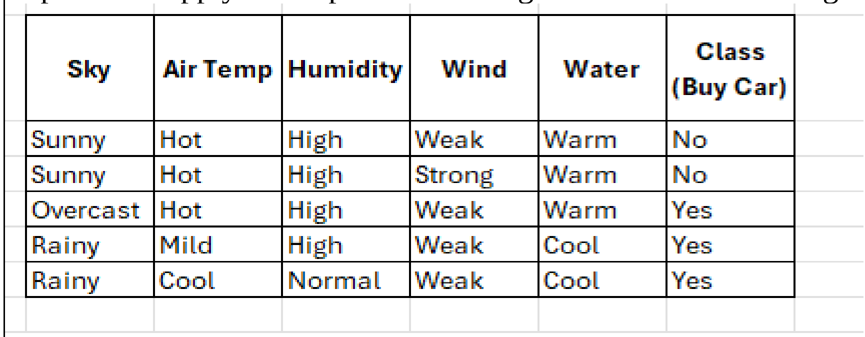
2.The covariance of dataset is obtained C.

3.Eigen values and Eigen vectors of the covariance matrix is calculated.

4.The eigen values and sorted and top eigen vectors are selected as feature vector. Obtain the transpose of feature vector,A.

5.Obtain the PCA transform y=A*(x-m)

5a) Explain and apply the steps of Find-S algorithm on the following dataset, 5 CO2 L3



Sky	Air Temp	Humidity	Wind	Water	Class (Buy Car)
Sunny	Hot	High	Weak	Warm	No
Sunny	Hot	High	Strong	Warm	No
Overcast	Hot	High	Weak	Warm	Yes
Rainy	Mild	High	Weak	Cool	Yes
Rainy	Cool	Normal	Weak	Cool	Yes

Scheme:
Steps-4Marks
Final Solution -1Mark

Solution
Initalize the hypothesis
Generalize using the first positive example
For other positive instances,
Check the attribute value if its same like in hypothesis retain the value otherwise change the hypothesis value to ?
Ignore all negative instances

The final solution expected is <? ? ? Weak ?>

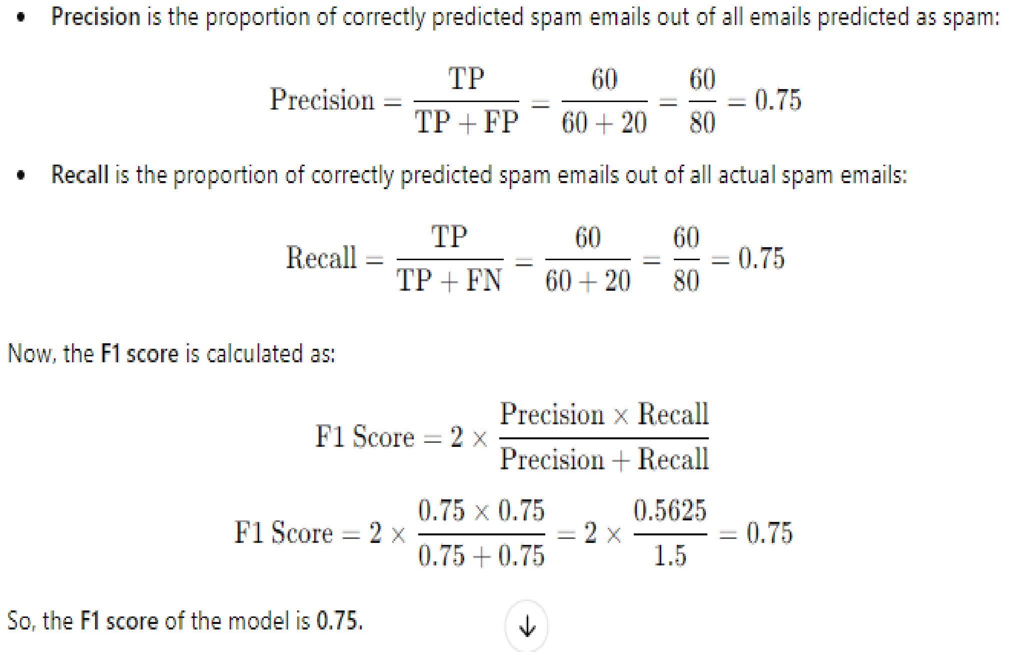
5b) You have a test dataset containing 200 emails.The model classified 80 emails as spam in that 60 emails are correctly predicted as spam.The remaining 120 emails were classified as non spam in that 100 mails the model correctly predicted them as not spam. Calculate the accuracy and F1 Score of the model. 5 CO2 L3

Scheme:-
Confusion Matrix=1Mark
Accuracy-2Marks
F1Score -2Marks

Solution:
True Positives (TP): The number of emails correctly predicted as spam (60).
False Positives (FP): The number of emails incorrectly predicted as spam, but they are actually not spam (80 - 60 = 20).
True Negatives (TN): The number of emails correctly predicted as non-spam (100).
False Negatives (FN): The number of emails incorrectly predicted as non-spam, but they are actually spam (120 - 100 = 20).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total emails}}$$

$$\text{Accuracy} = \frac{60 + 100}{200} = \frac{160}{200} = 0.80$$



- **Precision** is the proportion of correctly predicted spam emails out of all emails predicted as spam:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} = \frac{60}{60+20} = \frac{60}{80} = 0.75$$

- **Recall** is the proportion of correctly predicted spam emails out of all actual spam emails:

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} = \frac{60}{60+20} = \frac{60}{80} = 0.75$$

Now, the **F1 score** is calculated as:

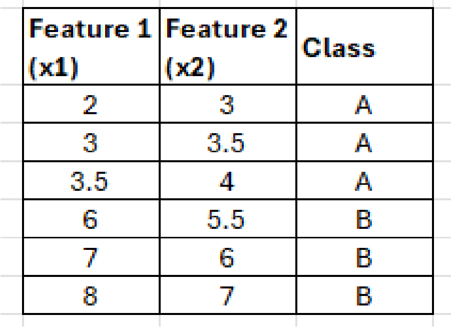
$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{F1 Score} = 2 \times \frac{0.75 \times 0.75}{0.75 + 0.75} = 2 \times \frac{0.5625}{1.5} = 0.75$$

So, the **F1 score** of the model is **0.75**.

6a) Apply K-NN and Weighted K-NN on the test instance(4.5,5,?) using the following dataset (take K=3),

10 CO3 L3



Feature 1 (x1)	Feature 2 (x2)	Class
2	3	A
3	3.5	A
3.5	4	A
6	5.5	B
7	6	B
8	7	B

Scheme:

KNN -4Marks
Weighted KNN -6 Marks

Solution:
Find the Euclidean distance from datapoints and test instance

KNN:-
The distances in ascending order:

1. **(3.5, 4)** → Distance = 1.41 → Class: A
2. **(6, 5.5)** → Distance = 1.58 → Class: B
3. **(3, 3.5)** → Distance = 2.12 → Class: A
4. **(7, 6)** → Distance = 2.69 → Class: B
5. **(2, 3)** → Distance = 3.2 → Class: A
6. **(8, 7)** → Distance = 4.03 → Class: B

For **K-NN**, the class prediction is based on the majority class of the K nearest neighbors.
For K = 3, the nearest neighbors are:

1. (3.5, 4) → Class: A

2. (6, 5.5) → Class: B
3. (3, 3.5) → Class: A

The majority class among these 3 nearest neighbors is **A** (2 out of 3 neighbors are class A).
Thus, the predicted class for the test instance (4.5,5)(4.5, 5)(4.5,5) using **K-NN** is **A**.

Weighted KNN
Find the Euclidean distance of the data points and test instance
Select the least three distances and find inverse distance

Instance	Distance	Inverse Distance	class
1	2.121	0.471	A
2	1.414	0.707	A
3	1.58	0.633	B

Sum of Inverse=1.811

Instance	Distance	Inverse Distance	Weighted Sum	class
1	2.121	0.471	.2603	A
2	1.414	0.707	.3905	A
3	1.58	0.633	.3490	B

Sum of Weighted class A=.6508
Sum of weighted class B=.3490
As per Weighted KNN label is Class A

Faculty Signature CCI Signature HOD Signature