Name - Harshita Singh
Sec - K
Univ Roll no - 2315000950
Class Roll no - (31)
Course / Branch - B.Tech (CSE)

ASSIGNMENT - 1 (Machine Learning)

Ans 1 - Used-car dataset → Regression or classification ?

- Task given : Predict whether a car will sell above it's listing price → this is a binary classification problem (Yes = Above price, No = Not above price).
- Justification :
  - Classification : The output is discrete (0/1). Model predicts probabilities (e.g., 0.8 chance the car sells above price).
  - Regression alterative : If instead we tried to predict the actual selling price as a continuous variable, then we'd use regression.
- Outputs & Metrics :
  - Classification : output = probability or class label.
    Metrics : Accuracy, Precision, Recall, F1-score, ROC-AUC.
  - Regression : Output = numerical price prediction.
    Metrics : RMSE, MAE, R^2.

Ans 2 - Handling Missing BMI & Glucose Values

- options :

1. Mean imputation : Replace missing values with the mean. Works if data is normally distributed without outliers.
2. Median imputation : Replace with the median. Better when distributions are skewed or when extreme outliers exist (common for BMI and glucose).
3. Dropping records : Appropriate only if :
   - Missing values are rare,
   - The records are missing completely at random (MCAR).

- clinical relevance :
  - BMI and glucose often have skewed distributions, with important clinical thresholds. Using median imputation preserves robustness.

- Mean imputation could distort results if a few patients have extreme values.
- Dropping should be avoided if missingness is systematic.

Ans 3 :- High Error / underfitting (high bias) - Model is performing poorly own or training.

Steps to improve :-
* Feature Engineering.
* Use a Stronger algo.
* Increase model capacity / complexity, tune, hyper parameters.
* Check data quality / labels, add more information feature.

Ans 4 :- One-hot Encoding Example

Input :

Name	Department
Alice	HR
Bob	Engineering
Charlie	HR
Dana	Sales

After one-hot encoding Department :

Name	Dept_Engineering	Dept_HR	Dept_Sales
Alice	0	1	0
Bob	1	0	0
Charlie	0	1	0
Dana	0	0	1

Harshafa Singh

Ans 5 (distances):

$d(E,A) = \sqrt{(6-8)^2 + (7-6)^2} = \sqrt{5} = 2.23$ pass

$d(E,B) = \sqrt{(6-5)^2 + (7-4)^2} = \sqrt{10} = 3.16$ fail

$d(E,C) = \sqrt{(6-7)^2 + (7-5)^2} = \sqrt{5} = 2.23$ pass

$d(E,D) = \sqrt{(6-3)^2 + (7-2)^2} = \sqrt{34} = 5.8$ fail

K = 3 nearest : A (pass), C (pass), B (fail).

prediction : pass (2 vs 1)

Ans 6 PCA on points (1,0), (0,1), (2,2), (4,4)

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	0	-0.75	-1.75	1.31	0.56	3.06
0	1	-1.75	-0.75	1.31	3.06	0.56
2	2	0.25	0.25	0.06	0.06	0.06
4	4	2.25	2.25	5.06	5.06	5.06
				7.74	8.74	8.74

mean $\bar{x} = 1.75$

$\bar{y} = 1.75$

Covariance Matrix :-

$$\begin{bmatrix} Cov(x) & Cov(x,y) \\ Cov(y,x) & Cov(y) \end{bmatrix}$$

$Cov(x) = \frac{1}{3} \times 8.74 = 2.91$

$Cov(x,y) = \frac{1}{3} \times 7.74 = 2.58 = Cov(y,x)$

$Cov(y) = \frac{1}{3} \times 8.74 = 2.91$

$$\text{Covariance Matrix} = \begin{bmatrix} 2.91 & 2.58 \\ 2.58 & 2.91 \end{bmatrix}$$

Eigen Values

For a symmetric matrix of the form $\begin{bmatrix} a & c \\ c & a \end{bmatrix}$, eigen values are $a+c$ and $a-c$.

Here, $a = 2.91$, $c = 2.58$

- $\lambda_1 = a + c = 2.91 + 2.58 = 5.5$
- $\lambda_2 = a - c = 2.91 - 2.58 = 0.33 \ (\approx 1/3)$

Interpretation / dimensionality reduction :
Total variance $= \lambda_1 + \lambda_2 = 5.83$. The first principal component ($\lambda_1 = 5.5$) captures $\approx 5.5/5.83 = 94.3\%$ of the variance
PCA reduces dimensionally by projecting data onto the top k principal components (here k=1) which retain the largest possible fraction of variance - projecting onto the first eigenvector keeps $\approx 94\%$ of the variability while reducing $2D \rightarrow 1D$.

Ans 7. Logistic regression (bank churn)

- Logistic function maps linear score z into probability :

$$p = \frac{1}{(1+e^{-z})}$$

- Positive coefficient for support calls → more calls = higher churn probability (odds ↑ by e^{coef}).

Harshita Singh

Ans 8 Confusion matrix metrics (TP = 50, FN = 20, FP = 15, TN = 85)

- Accuracy = 79.4 %
- Precision = 76.9 %
- Recall = 71.4 %
- Specificity = 85 %
- F1 = 74.1 %

In critical health monitor → Recall more important.

Ans 9 Smartphone resale price

- Task = predict exact resale price → Regression.
  - output : numeric price ; Metrics : RMSE, MAE, R^2.
- If predicting price ranges → classification.
  - Output : class labels ; Metrics : Accuracy, Precesion, Recall, F1.

Ans 10. Missing creatinine & hemoglobin

- Mean imputation : symmetric, no outliers.
- Median imputation : skewed data / outliers.
- Drop records : very few missing, MCAR.

Ans 11 Sentiment classifier (54% train, 56% val)

- Error type : Underfitting.
- Fix : richer features (TF-IDF, n-grams, embeddings) Stronger models (SVM, boosting, transformers), hyperparam tuning, more / better data.

Harshita Singh

Ans 12.

Item	Stationary	Office Supplies	Art
Pen	1	0	0
Notebook	0	1	0
Eraser	1	6	0
Marker	0	0	1

Ans 13. KNN ($k=3$)

distance → $P_1 = \sqrt{10}$, $P_2 \rightarrow \sqrt{10}$, $P_3 \rightarrow \sqrt{10}$, $P_4 = 5$

Nearest 3 : P_1, P_2, P_3 → pass → 2, fail = 1

→ predict pass.

Ans 14 Covar. Matrix → Eigen values show variance dir.

PCA keeps axis with max eigenvalue.

Ans 15 Sigmoid maps linear sum → prob.

Negative BMI coefficient = higher BMI + complications Prob.

Ans 16 Confusion Matrix (40, 10, 20, 130)

Accuracy = 85%.

Precision = 40 / 60 → 66.7%.

Recall = 40 / 50 → 80%.

Specificity = 130 / 150 → 86.7%.

F1 Score = 72.7%.

for fraud alerts → Recall more imp.

Ans 17. Classification (above / below median).

Regression only if predicting exact rent.

Metrics : Reg → RMSE, MAE.

Classification → Accuracy / F1.

Harshita Singh

Ans 18. Mean — Normal dism.
Median — Skewed / outliers.
Drop — Small missing set & not crucial clinically.

Ans 19. Error — Underfitting
Fix — Richer features, Stronger, models, tuning.

Ans 20

Title	Sci-fi	Romance	Drama
Inception	1	0	0
Titanic	0	1	0
Joker	0	0	1
Up	0	0	0

Ans 21 KNN (K=3)
Distance from X5(6,6):
$X_1 = \sqrt{2}$, $X_2 = \sqrt{13}$, $X_3 = \sqrt{10}$, $X_4 = \sqrt{12}$
Nearest 3: X_1 (pass), X_3, (pass)
X_2 (fail) → predict pass.

Ans 22 Covariance Eigenvalues → One large, One small.

PCA keeps axis with larger eigenvalues
→ max variance retained.

Ans 24. Accuracy → (45+125) / 200 → 85%.
Precision → 45 / 70 → 64.3%.
Recall → 45 / 50 → 90%.
Specificity → 125 / 150 → 83.3%.
F1 → 75%.
for fraud System ⟹ Recall more