Machine Learning Algorithms Explanation

1️⃣ Mall Customers – K-Means Clustering

Algorithm Used: K-Means Clustering (Unsupervised Learning)

Workflow Explanation:

1. **Load and Inspect Data**
2. Load dataset into a DataFrame.
3. Check column names, types, and missing values.
4. **Correlation Analysis**
5. Check relationships between numeric variables.
6. **Data Visualization**
7. Countplot for Gender distribution.
8. Countplot for Age.
9. Barplot for Annual Income vs Spending Score.
10. **Feature Selection**
11. Use Age, Annual Income, and Spending Score as features.
12. **K-Means Clustering**
13. Elbow Method to determine optimal number of clusters.
14. Fit final K-Means with the chosen `k`.
15. **Cluster Visualization**
16. 3D scatter plot colored by cluster.

Key Points: - Unsupervised learning – no target variable. - K-Means minimizes WCSS to group similar customers. - Useful for customer segmentation.

2️⃣ Temperatures – Simple Linear Regression

Algorithm Used: Simple Linear Regression (Supervised Learning)

Workflow Explanation:

1. **Load and Inspect Data**
2. Load temperature dataset.
3. Check shape and summary statistics.
4. **Data Visualization**
5. Scatter plot for JAN vs FEB temperatures.
6. Distribution plot for FEB.
7. **Data Preparation**
8. X = JAN temperatures, y = FEB temperatures.
9. Split data into training and testing sets.
10. **Train Linear Regression Model**
11. Fit a straight line: y = m*x + b.
12. **Prediction and Evaluation**
13. Predict FEB temperatures using test data.
14. Evaluate using MAE, MSE, RMSE.
15. **Visualization**
16. Bar chart comparing actual vs predicted.
17. Scatter plot with regression line.

Key Points: - Supervised learning – predict FEB from JAN. - Finds best-fit line minimizing error. - Error metrics assess model performance.

3️⃣ Heart Disease Prediction – Logistic Regression

Algorithm Used: Logistic Regression (Supervised Learning – Classification)

Workflow Explanation:

1. **Load and Inspect Data**
2. Load heart disease dataset.
3. Check column names, types, and missing values.
4. **Data Preprocessing**
5. Identify numeric columns.
6. Check and handle missing values.
7. **Train-Test Split**
8. Separate features (X) and target (y).
9. Split data into training (80%) and testing (20%).
10. **Train Logistic Regression Model**

```
lr_clf = LogisticRegression(solver='liblinear')
lr_clf.fit(X_train, y_train)
```

11. Uses the sigmoid function to map linear combination of features to probability.
12. Threshold 0.5 used to classify: 1 (disease), 0 (no disease).
13. **Prediction and Evaluation**
14. Predict on training and testing sets.
15. Evaluate with accuracy, confusion matrix, and classification report.

Key Points: - Supervised learning – binary classification. - Logistic regression models the probability of a class. - Output evaluated using classification metrics to ensure model performance.

4️⃣Admission Prediction – Decision Tree Classifier

Algorithm Used: Decision Tree (Supervised Learning – Classification)

Workflow Explanation:

1. **Load and Inspect Data**

2. Load dataset and check columns, types, and missing values.

3. **Data Preprocessing**

4. Identify numeric columns.
5. Check and handle missing values.

6. Scale features using StandardScaler.

7. **Train-Test Split**

8. Split data into training (80%) and testing (20%) sets.

9. **Train Decision Tree**

10. Recursively split features to maximize class purity.

11. **Model Evaluation**

12. Accuracy, classification report, confusion matrix.

13. **Training vs Testing**

14. Evaluate model on train and test sets to detect overfitting.

Key Points: - Supervised learning – predict target (admission yes/no). - Tree splits data using feature thresholds. - Scaled features improve performance.

Summary of Algorithms Across Programs

Program	Task Type	Algorithm	Key Steps
Mall Customers	Unsupervised	K-Means Clustering	Elbow method to find k, cluster by Age/ Income/Spending Score
Temperatures	Supervised (Regression)	Linear Regression	Predict FEB from JAN, fit line using MSE, evaluate with MAE/MSE/RMSE
Heart Disease	Supervised (Classification)	Logistic Regression	Binary classification, sigmoid function, evaluate with accuracy/confusion matrix
Admission Predict	Supervised (Classification)	Decision Tree	Split data, scale features, train tree, evaluate with accuracy/confusion matrix

Notes

- Ensure CSV files are available in the correct path.
- Install necessary libraries: pandas, numpy, matplotlib, seaborn, scikit-learn.
- Each algorithm demonstrates a complete ML workflow: data loading → preprocessing → modeling → evaluation.