raisoni EDUCATION

Experiment No : 1

* Title : Exploratory Data Analysis and visualization of Relationship in a Given data set.

* Aim : To perform Exploratory Data Analysis (EDA) on a given dataset and visualize relationship between variables to uncover pattern trends and insights.

* Theory :

Exploratory Data Analysis (EDA) is an approach to analyzing dataset to summarize their main characteristics often using visual methods. It helps in :

- Understanding data distribution and missing values.
- Identifying correlations and relationship between features.
- Detecting anomalies and outliers.

Key Techniques in EDA :

- Summary statistic : Mean, Median, standard deviation etc

- Data visualization : Histogram, scatter plots, pair plots, box plots, heat map etc.

raisoni EDUCATION

- Feature correlation analysis : Using correlation matrices and scatter plots.

Visualization Techniques in EDA

EDA heavily relies on visual methods to better understand data relationships and distributions.

1. Histogram : Show the distribution of a single variable.

2. Boxplots : Help detect outliers and compare distribution across categories.

3. Scatterplot : Show relationship between two numerical variable.

4. Heatmaps : Represent correlation matrices of identify strong relationships.

5. Pair plots : Provide a quick overview of relationship between multiple numerical variables.

- Importance of EDA

1) Helps in understanding datasets
2) Detect data quality issues like missing value.
3) Aids in feature selection by identifying relevant variable.

raisoni EDUCATION

Algorithm :

- Step 1 : Load and inspect the dataset
- step 2 : Perform data cleaning (handle missing values, duplicates, outliers)
- Step 3 : Generate summary statistics (mean, median, min. max value)
- step 4 : visualize relationship using various plots (scatterplot, heatmap, pair plots box plots.)
- Step 5 : Interpret the result and derive insight.

Result :

- summary statistics of the dataset
- observation from visualizations (e.g., correlation strength, presence of outliers, distribution patterns).
- key insight derived from EDA

Conclusion :

Hence, we sucessfully perform the EDA and visualize the relationship for given dataset.

raison EDUCATION

Experiment No 2

* Title : Exploratory Data Analysis, Visualization and Linear Regression Model Training on the Housing price dataset.

* Aim : To perform Exploratory Data Analysis (EDA) on the Housing price dataset visualize relationship between variables, train a linear Regression Model and evaluate its performance.

* Theory :

- Exploratory Data Analysis (EDA) :
EDA is a crucial step in data analysis that helps understand the dataset before applying any machine learning model. The main objectives of EDA include :

  - Identifying pattern, trends and relationships between variables.
  - Detecting missing values, outliers, and anamolies.
  - Summarizing data distribution and correlations.

- * Linear Regression :
Linear Regression is a supervised learning algorithm used for predicting continous values.

raisoni EDUCATION

The relationship between the dependent variable (housing price) and independent variables (features like square footage, number of bedrooms, location, etc.) is modeled using the equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_n + \cdots + \beta_n X_n + \varepsilon$$

where,

Y = Dependent variable
X = independent variable
β = Coefficient
ε = Error term

- Key Points of Linear Regression:

1. Linearity: Relation between independent & Dependent variable is linear.

2. Indepence: Observation are independent for each other.

3. Homoscedasticity: Constant variance of residual (errors)

4. No multicolinearity: Independent variables should not be highly correlated.

raisoni EDUCATION

* Algorithm :

1. Load the dataset and perform an initial inspection.
2. Perform EDA :
   - Summary statistics and missing values analysis.
   - Visualizations (histogram, scatter plots, heatmaps).
3. Feature Engineering :
   - Handle missing data
   - Encode categorical variables.
   - Normalize / standardize numerical features if needed
4. Train - Test split : Divide the dataset into training and testing sets.
5. Train linear Regression Model on training Data.
6. Evaluate Model performance using MSE, RMSE and R^2 score.

raisoni EDUCATION

Result :

- Summary statistics of the dataset
- Insight from EDA visualization
- Model performance metrics (MSE, RMSE, R^2 score).

Conclusion

- Key observations from the dataset
- Find the strength of relationship between features and housing price
- Performance of linear Regression model.
- Possible improvement

Experiment No. 4

Title: Exploratory Data Analysis (EDA) and Ridge Regression Model training on a Given Dataset

Aim: To perform Exploratory Data Analysis (EDA) on a given dataset, visualize relationships between variable and a Ridge Regression model and evaluate its performance

Theory:

- Ridge Regression:
Ridge Regression is a type of linear regression that includes an L2 regularization term to prevent over fitting by penalizing large coefficients. It modifies the cost function by adding a regularization term:

$$J(\beta) = \sum (y_i - \hat{y}_i)^2 + \lambda \sum \beta_i^2$$

where:
1) $J(\beta)$: is the cost function
2) λ (alpha): is the regularization parameter (higher λ shrink coefficient)
3) β represents the model coefficients.

raisoni EDUCATION

Model Evaluation Metrics

1. Mean squared Error (MSE)

MSE measures the average square difference between actual & predicted value.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

2. Root Mean squared Error (RMSE)

RMSE is simply the square root of MSE. It is used to bring the error value back to the original unit of dependent variable.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

3. R-squared (R^2) score

R^2, also called the coefficient of determination, measures how well the independent variable

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

raisoni EDUCATION

Algorithm :

1. load and inspect the dataset
2. Perform EDA
   - Visualize feature distributions
   - Identify correlations
   - Detect and handle missing values or outliers.
3. Feature Engineering:
   - Encode categorical variables
   - Scale numerical features
4. Train - Test - split : split the dataset into training and testing sets.
5. Train Ridge Regression model with different regularization strengths.
6. Evaluate model performance using MSE, RMSE and R^2 score.

Results :

- Observation from ~~Eda~~ EDA visualizations
- Model performance metric (MSE, RMSE, R^2 score) for Ridge regression.

Conclusion :

Ridge Regression performance and comparison with other models. It impact of regularization on model performance.

raison EDUCATION

Experiment No: 4

* Title: Exploratory Data analysis, visualization and Lasso Regression model Training on a Given Dataset.

* Aim: To perform EDA on a given dataset, visualize relationships between variables, train a Lasso Regression model and evaluate its performance.

* Theory:

- Exploratory Data Analysis (EDA)

EDA is the process of analyzing and summarizing datasets using statistical & visualization techniques to uncover insights before applying machine learning models.

1- Lasso Regression:

Lasso (Least Absolute Shrinkage and selection Operator) Regression is a linear regression technique that include L1 regularization, which helps with both feature selection and preventing overfitting.

$$J(\beta) = \sum (y_i - \hat{y}_i)^2 + \lambda \sum |\beta_j|$$

Where,
'J'(β)' is the cost function
'λ' (alpha) is the regularization parameter
'β' represent the model coefficients.

raisoni EDUCATION

Key points of lasso regression

- Unlike Ridge regression, which shrinks coefficients close to zero, Lasso can shrink some coefficients exactly to zero.
- Useful when dealing with high dimensional data.

Algorithm:

step 1: Load and inspect the dataset

step 2: Perform EDA
1. Visualize feature distribution
2. Identify correlations using heatmap.
3. Detect and handle missing values or outliers

step 3: Feature Engineering:
- Encode categorical variables
- Scale numerical features.

step 4: Train - Test split Divide the dataset into training and testing sets.

step 5: Train Lasso Regression Model using different values of α (alpha)

step 6: Evaluate model performance using MSE, RMSE and R^2 score.

raisoni EDUCATION

Results :

- Feature relationships and correlation analysis.
- Model performance metrics (MSE, RMSE, R^2 score) for Lasso Regression
- Important features selected by Lasso regression.

Conclusion :

The effectiveness of Lasso regression in selecting important features. Model performance evaluation and possible improvements. Future steps tunning value, trying alternative model like Ridge regression.

raisoni EDUCATION

Experiment No : 5

Title : Exploratory Data Analysis, visualization and logistic Regression Model Training on the Iris Dataset.

Aim : To perform Exploratory Data Analysis on the Iris dataset visualize relationship between variables train a logistic regression model and evaluate its performance in classifying iris species.

Theory :

Logistic Regression :

Logistic Regression is a supervised machine learning algorithm used for classification tasks. It models the probability that a given input belong to a specific class using sigmoid function

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}}$$

Where

- "$P(y=1|x)$" is the probability of an instance belonging to class 1
- 'β' represent the model coefficients
- x represent the model coefficients
- The sigmoid function ensure the output is between 0 & 1.

raisoni EDUCATION

Model Evaluation Metrics

- Accuracy : Measures overall correctness
- Precision, Recall and F1 score : Evaluate class specific performance
- Confusion Matrix : Visualize correct and incorrect classifications.

Types of Logistic Regression

- Binary logistic Regression : Two classes (e.g Yes / No, 0/1)
- Multinominal logistic Regression : More than two classes (e.g Red, Blue, Green)
- Ordinal logistic Regression : Ordered categories (e.g Low, medium, high)

Algorithm :

1. Load the dataset and check its structure.
2. Perform EDA :
   - Check class distribution
   - Visualize feature distributions.
   - Analyze feature relationships using pair plots & heat maps.
3. Data Preprocessing :
   - Handle missing values (if any)
   - Encode categorical variables

- Split data into training and testing sets.

4. Train a logistic Regression model.

5 Evaluate model Performance using accuracy confusion matrix and classification report.

Results:

- EDA Findings:
  1.) Sepal length and petal length are highly correlated.
  2.) Setosa is distinctly separable from the other two species.
- Model performance:
  1. Accuracy
  2. Confusion Matrix
  3. Classification Report.

Conclusion:

The model achieved high accuracy indicating good performance. Setosa was the easiest to classify, while versicolor and virginica had some misclassification.

raisoni EDUCATION

Experiment No 6:

Title : Exploratory Data Analysis, visualization and logistic Regression Model Training on the user Dataset.

Aim : To perform Exploratory Data Analysis (EDA) on the User dataset, visualize relationship between variables, train a logistic Regression model and evaluate its performance in classifying ~~your~~ user behaviour.

Theory :
Logistic regression is a fundamental machine learning algorithm used for classification problems. Unlike linear regression, which predicts continuous values, logistic regression predict probabilites and maps them to binary (0 or 1) or multiclass outputs.

Logistic Regression uses the sigmoid function to transform linear outputs into probabilities between 0 and 1

Sigmoid function formula:

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

where :
- $z = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$
- $\sigma(z)$ represents probability belong to class 1

If $P(y=1) \geq 0.5$ → classify as 1.
If $P(y=1) < 0.5$ → classify as 0.

The sigmoid function ensure outputs are mapped into probability range (0 to 1)

Cost function for logistic Regression

$$J(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i) \right]$$

Where,

- y_i is actual class
- $\hat{y}_i$ is the predicted probability for class 1
- The function penalizes incorrect predictions more if the model is highly confident but wrong

* Algorithm :

1. Load and inspect the dataset

2. Perform EDA :
   - Analyze feature distributions using histograms and boxplots.
   - Identify correlation using a heatmap
   - Visualize relationships between numerical variable with pair plots.

3. Preprocess Data :
   - Handle missing values.

- Encode categorical variables
- Scale numerical feature
- Split data into training and testing sets.

4. Train Logistic Regression Model using the preprocessed dataset.

5. Evaluate Model Performance using accuracy confusion matrix, classification report and AUC score.

Results:

- EDA Findings:
  1. The dataset has no missing values
  2. certain features may have strong correlations affecting classification.
  3. The target variable, is imbalanced balanced.
- Model Performance:
  1. Accuracy
  2. Confusion Matrix
  3. Precision, Recall & F1-score
  4. ROC Curve & AUC

Conclusion:

The logistic Regression model provided good classification accuracy. The ROC curve and AUC indicate how well the model separates classes.

raisoni EDUCATION

Experiment No : 7

* Title : Exploratory Data Analysis, Visulization and Naive Bayes classification on the Wine Dataset.

* Aim To perform Exploratory Data Analysis (EDA) on the wine dataset, visualize relationships between variables train a Naive Bayes classifier and evaluate its performance in classifying wine categories.

* Theory :

Exploratory Data Analysis (EDA)

EDA is a process that involves understanding dataset charecteristics using summary statistics and visualizations.

- checking the distribution of variables.
- Identifying patterns, relationships and corredations between features.
- Detecting missing values and outliers.
- Preparing data for machine learning models.

° Naive Bayes Classifier

Naive Bayes is a probabilistic classification algorithm based on Bayes Theorem, assuming feature independence

raisoni EDUCATION

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

where,

- $P(A|B)$ is the probability of class A given feature B.
- $P(B|A)$ is the probability of feature B given class A.
- $P(A)$ is the prior probability of class A.
- $P(B)$ is the probability of feature B occuring.

- Types of Naive Bayes classifiers:

1. Gaussian Naive Bayes : Assume normal distribution of features.
2. Multinomial Naive Bayes : Suitable for discrete feature control.
3. Bernoulli Naive Bayes : Used for binary feature data.

For wine dataset, Gaussian Naive Bayes is appropriate since feature are continous.

- Model Evaluation Metrics:

1. Accuracy
2. Precision
3. Confusion Matrix

raisoni EDUCATION

Result :

- EDA Findings :
  1. The dataset contains no missing values.
  2. Certain feature show strong correlations.
  3. Some classes may have over-lapping feature.

Conclusion :

The Naive Bayes model achieved good accuray on the test set. standardization of features improved model performance. The confusion matrix revealed some misclassifications, especially in overdapping feature space.

raisoni EDUCATION

Experiment No: 8

* Title : Exploratory Data Analysis visualization and Decision Tree classification on the Titanic Dataset.

* Aim : To perform Exploratory Data Analysis (EDA) on the Titanic dataset, visualize relationship between features, train a Decision Tree classifier and evaluate its performance in predicting survival.

* Theory :

Exploratory Data Analysis (EDA)

EDA helps in understanding data characteristics using statistical summaries and visualizations.

Decision Tree classifier

A decision tree is a supervised learning algorithm used for classification and regression tasks. It splits the dataset into branches based on feature conditions following an "if-else" structure.

The Gini index and Entropy are common criteria for splitting:

- Gini Index: Measures impurity.
- Entropy: Measures information gain at each split.

raisoni EDUCATION

A Decision Tree classification data by traversing branches from root to leaf nodes, making it interpretable but prone to overfitting.

- Model Evaluation metrics
  1. Accuracy : Overall correctness of prediction
  2. Precision, Recall, and F1-score : Class-specific performance evaluation
  3. Confusion Matrix : Breakdown of correctness and incorrect classification
  4. Feature Importance : Identifies influential feature in decision making

* Algorithm :
  1. Load and inspect the dataset
  2. Perform EDA on the dataset
  3. Data Preprocessing, Handel missing values, convert categorical into numerical etc.
  4. Train a Decision tree classifier using entropy / Gini criteria.
  5. Evaluate Model Performance using accuracy confusion Matrix and feature importance.

raisoni EDUCATION

Result :

EDA Findings :

- Gender : Females had a higher survival rate.
- Passanger class : Higher class passangers had a better survival rate.
- Feature Correlations : Pclass, Fare and age were key predictors.

Conclusion :

The decision tree model provided good accuracy in predicting survival. Feature importance analysis revealed that gender class and fare were crucial.

raison EDUCATION

Experiment No. 9

* Title : Exploratory Data Analysis, visualization and K-Nearest Neighbors (KNN) classification on the Digits Dataset.

* Aim : To perform Exploratory Data Analysis (EDA) on the Digits dataset, visualize relationships between feature, train a K-Nearest Neighbors between feature train a K-Nearest Neighbor classifier and evaluate its performance in recognizing handwritten digits.

* Theory :

K-Nearest Neighbors (KNN) classifier

KNN is non-parametric, supervised learning algorithm that classifies a data point based on the majority class of its k nearest neighbors.

Key properties :

- Distance metric : Common choice include Euclidean, Manhattan and Minkowski distance.
- K-value selection : A small K may lead to overfitting while a large K may oversmath class boundaries.

- Weighted voting : closer neighbor can be given higher importance

Mathematically, the Euclidean distance between two points $X_1 (x_1, x_2, \ldots, x_n)$ and $X_2 (y_1, y_2, \ldots, y_n)$ is

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Model Evaluation metrics:

1. Accuracy
2. Precision, Recall & F1-Score
3. Confusion matrix

* Algorithm:

1. Load and inspect the dataset

2. Perform EDA:
   1. Visualize sample digit images
   2. Reduce dimensions using PCA for visualization

3. Data Preprocessing:
   1. Normalize pixel values.
   2. Split data into training and testing sets.

4. Train a KNN classifier and optimize K value.

5. Evaluate model performance using accuracy confusion matrix and classification report.

Results :

- CDA Findings :
  1. Each digit is represented as an 8×8 pixel image.
- PCA visulization shows clustering pattern among digits
- Feature scaling improves distance-based classification.

Conclusion :

The KNN model performed well in recognizing handwritten digits. Featuring scaling was crucial for better performance.

EDUCATION

Experimental No : 11

* Title : Exploratory Data Analysis, visualization and K means clustering on the Iris Dataset

* Aim : To perform EDA and visualize the relationships among features in the iris dataset, apply the K-mean clustering algorithm and evaluate the clustering performance.

* Theory :

K-Mean clustering

K Mean is an unsupervised machine learning algorithm used to group similar data points into K clusters based on feature similarity.

Steps :

1. Choose K (number of clusters)
2. Initialize K cluster Centroid randomly
3. Assign each data point to the nearest centroid
4. Recalculate centroid based on mean of assigned points.
5. Repeat steps 3-4 until Centroid stabalize

Mathematical Concept:

Uses Euclidean distance to measure similarity between data points.

Evaluation metrics:

1. Inertia: Total distance of points from cluster center.
2. Silhouette Score: How well seperated and cohesive the clusters are (-1 to 1)
3. Comparison with True Labels: Although unsupervised, we can compare clusters with actual species labels for evaluation.

Algorithm:

1. Load the dataset and perform exploratory analysis.
2. Visualize distributions and pairwise relationships.
3. Preprocess data: scale features for better clustering performance.
4. Determine Optimal k using Elbow method and silhouette score.
5. Train K-Means Model & Predict Cluster.

6. Compare predicted clusters with actual species.

7. Visualize clustered data and centroids using PCA.

* Result :
- cluster formed : 3 clusters corresponding closely to 3 actual species.
- silhouette score : Approximately, indicating good seperation.
- visualize : PCA plots clearly show well seperated clusters.
- comparision : K-Means performed well without supervision in approximating true species labels.

* Conclusion : K-Means was successful in clustering the Iris data into meaningful group. EDA helped understand data distribution and relationship PCA enables visual inspection of clustering effectiveness.

raison EDUCATION

Experiment No: 10

Title: Exploratory Data Analysis, Visualization and clustering with K-means on an Income Dataset

Aim:

To perform EDA and visualize relationship in the Income dataset, apply the K-means clustering algorithm to segment data based on income related pattern and evaluate clustering performance using appropriate metrics.

Theory:

K-Means Clustering

K-Means is an unsupervised learning algorithm used for clustering data into K distinct groups based on similarity

Key steps in K-Means:

1. Choose the number of clusters (k)
2. Initialize K centroids randomly.
3. Assign each point to nearest ~~data~~ centroid
4. Update centroid by calculating the mean of assigned points
5. Repeat 3 and 4 step until convergence

Distance Metric :

Typically uses Euclidean distance

Evaluation Metrics :

1. Inertia (within - cluster sum of squared) : Lower is better

2. Silhouette Score : Measure how similar an object is to its own cluster vs other cluster. (range -1 to 1)

3. Elbow Method : Helps determine the optimal number of clusters (k)

K - Mean is not a regression technique ; it is used for clustering not for predicting a continous targel variable.

Algorithm :

1. Load & inspect the dataset

2. EDA :
   - Plot distribution of income spending score, age
   - Use scatter plot and pair plots

to visualize feature relationships

3. Data Preprocessing:
   1. Handle missing values (if any)
   2. Encode categorical variables
   3. Scale / normalize the features

4. K-means clustering:
   - Determine optimal k using Elbow method and / or Silhouette score.
   - Fit the k-means model on selected features.

5. Evaluate Clustering:
   - Visualize cluster using 2D or 3D scatter plots
   - Print and silhouette score.

* Result:
- cluster formed: K Means formed 4 distinct clusters based on income, age, and spending score.
- Inertia: (displayed from code)
- Silhouette : " -
- visualization: PCA and scatter plot show well seperated groups indicating meaningful segmentation.

raisoni EDUCATION

* Conclusion :

K means effectively clustered individuals based on income-related data. PCA helped visualize high dimensional clusters in 2D. Silhouette score confirmed the clustering quality. Cluster can be used for customer segmentation, targeted marketing or socioeconomic analysis.