# DIFFERENCE BETWEEN CLUSTERING, CLASSIFICATION AND REGRESSION

## Claim:

Write difference between

clustering vs classification

Regression vs Classification

## Evidence:

Clustering is an unsupervised learning method that groups data based on similarities without predefined labels, while classification is a supervised learning method that assigns predefined labels to data points.

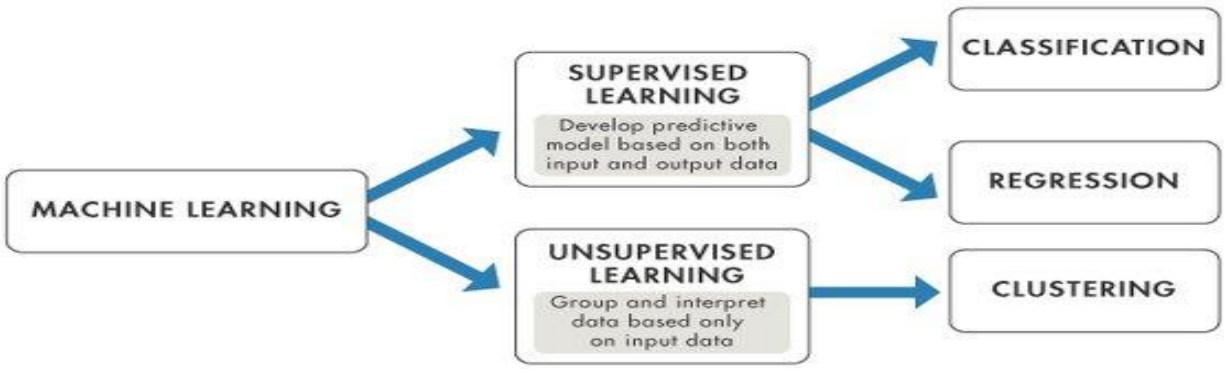
Regression predicts continuous values and focuses on modeling the relationship between variables, whereas classification predicts discrete labels and categorizes data points into predefined classes.

## Reasoning:

Data Mining: Concepts and Techniques" by Jiawei Han, Micheline Kamber, and Jian Pei, clustering techniques like K-means and hierarchical clustering are used to discover natural groupings in data, whereas "Pattern Recognition and Machine Learning" by Christopher M. Bishop explains how classification techniques such as Decision Trees and SVM require labeled training data to predict categories.

The Elements of Statistical Learning" by Trevor Hastie, Robert Tibshirani, and Jerome Friedman details regression methods like linear regression for predicting continuous outcomes, while "An Introduction to Statistical Learning" by Gareth James et al. explains classification techniques for assigning data points to categories, such as in email spam detection and handwritten digit recognition

## Differences Between Clustering, Classification, and Regression




### 1. Clustering vs. Classification

**Clustering**:

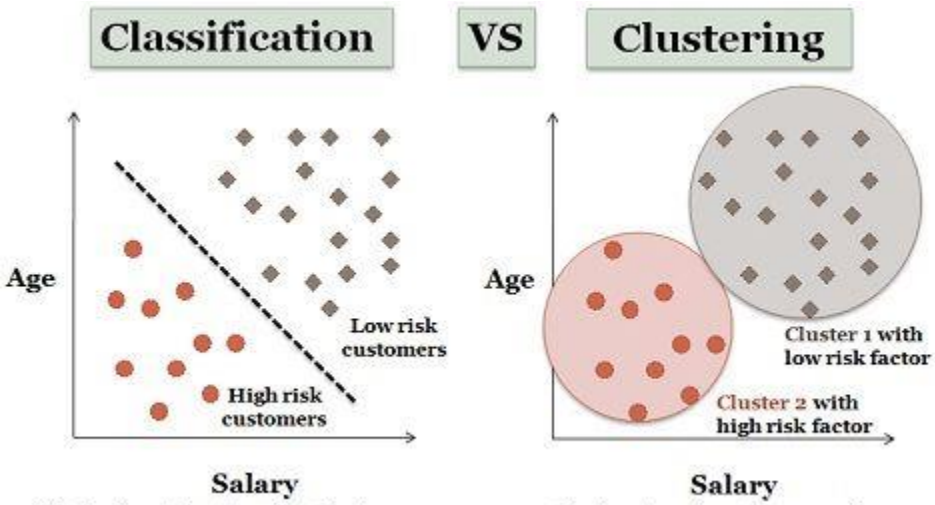
- **Definition**: Clustering is an unsupervised learning technique that groups data points into clusters based on similarity.
- **Goal**: Discover the inherent grouping in a dataset.
- **Data Labeling**: Does not require labeled data.
- **Examples**:

    - Market segmentation
    - Image compression
- **Techniques**:
    - K-means
    - Hierarchical clustering
    - DBSCAN

**Classification**:

- **Definition**: Classification is a supervised learning technique that assigns labels to data points based on predefined categories.
- **Goal**: Predict the category of new data points.
- **Data Labeling**: Requires labeled data for training.
- **Examples**:
    - Email spam detection
    - Handwritten digit recognition
- **Techniques**:
    - Decision Trees
    - Random Forest
    - Support Vector Machines (SVM)

**Diagram: Clustering vs. Classification**



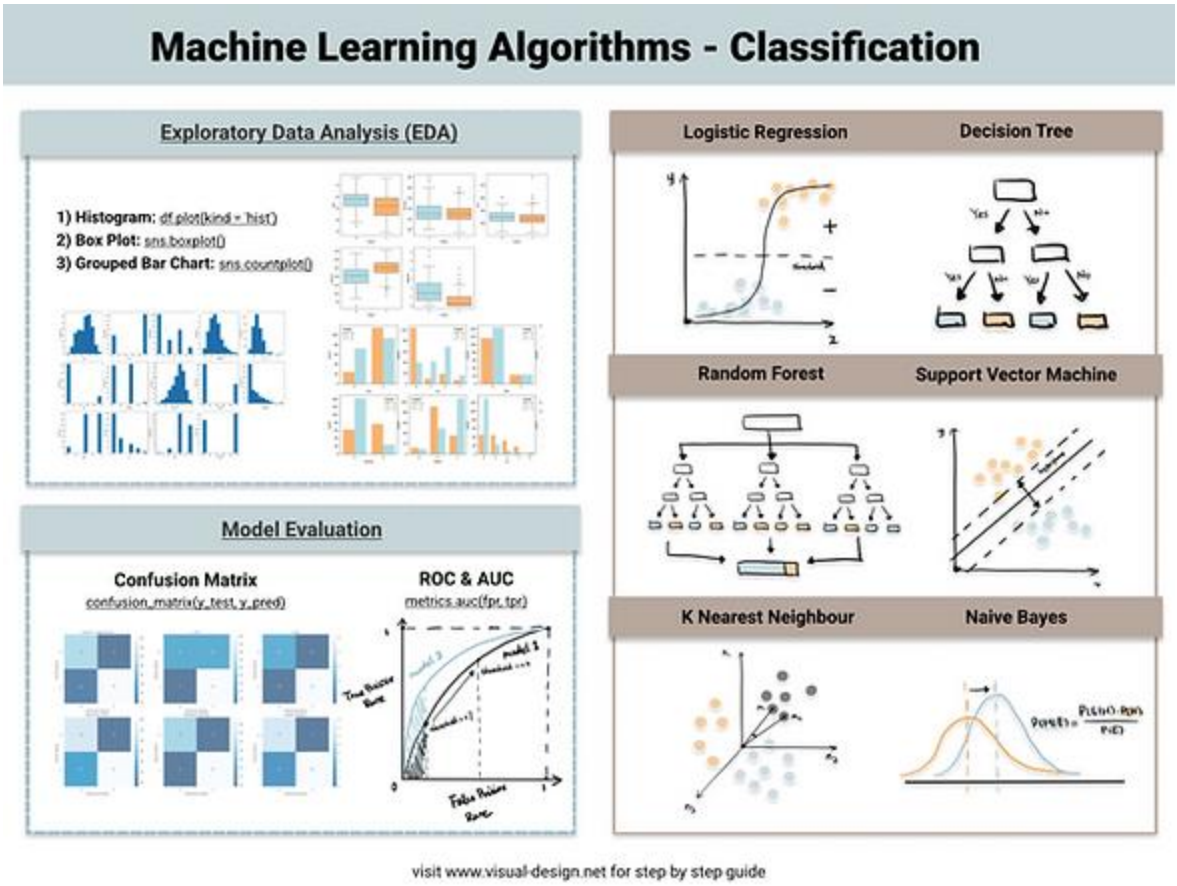
Risk classification for the loan payees on the basis of customer salary

| Parameter | CLASSIFICATION | CLUSTERING |
|---|---|---|
| Type | used for supervised learning | used for unsupervised learning |
| Basic | process of classifying the input instances based on their corresponding class labels | grouping the instances based on their similarity without the help of class labels |

| Parameter | CLASSIFICATION | CLUSTERING |
|---|---|---|
| Need | it has labels so there is need of training and testing dataset for verifying the model created | there is no need of training and testing dataset |
| Complexity | more complex as compared to clustering | less complex as compared to classification |
| Example Algorithms | Logistic regression, Naive Bayes classifier, Support vector machines, etc. | k-means clustering algorithm, Fuzzy c-means clustering algorithm, Gaussian (EM) clustering algorithm, etc. |

**Comparison Table: Clustering vs. Classification**

| Feature | Clustering | Classification |
|---|---|---|
| **Definition** | Unsupervised learning technique that groups data points based on similarity without predefined labels. | Supervised learning technique that assigns labels to data points based on predefined categories. |
| **Goal** | Discover the inherent grouping in a dataset. | Predict the category of new data points. |
| **Data Labeling** | Does not require labeled data. | Requires labeled data for training. |
| **Output** | Groups or clusters of similar data points. | Discrete labels (e.g., categories). |
| **Examples** | Market segmentation, Image compression, Customer segmentation | Email spam detection, Handwritten digit recognition, Disease diagnosis |
| **Techniques** | K-means, Hierarchical clustering, DBSCAN | Decision Trees, Random Forest, Support Vector Machines (SVM), Neural Networks |
| **Usage** | Data exploration, pattern recognition | Predictive modeling, categorization |
| **Evaluation** | Cluster quality metrics (e.g., silhouette score, Davies–Bouldin index) | Classification accuracy, precision, recall, F1 score |
| **Visualization** | Dendrograms, Cluster plots | Confusion matrix, ROC curves, Precision-Recall curves |
| **Scalability** | Can handle large datasets but may require more computational resources for complex algorithms. | Generally scalable with efficient algorithms like Random Forest or SVM. |




## 2. Regression vs. Classification

**Regression**:

- **Definition**: Regression is a supervised learning technique that predicts continuous values.
- **Goal**: Estimate the relationship between variables and predict continuous outcomes.
- **Output**: Continuous values (e.g., real numbers).
- **Examples**:
  - House price prediction
  - Temperature forecasting
- **Techniques**:
  - Linear Regression
  - Polynomial Regression
  - Support Vector Regression (SVR)

**Classification**:

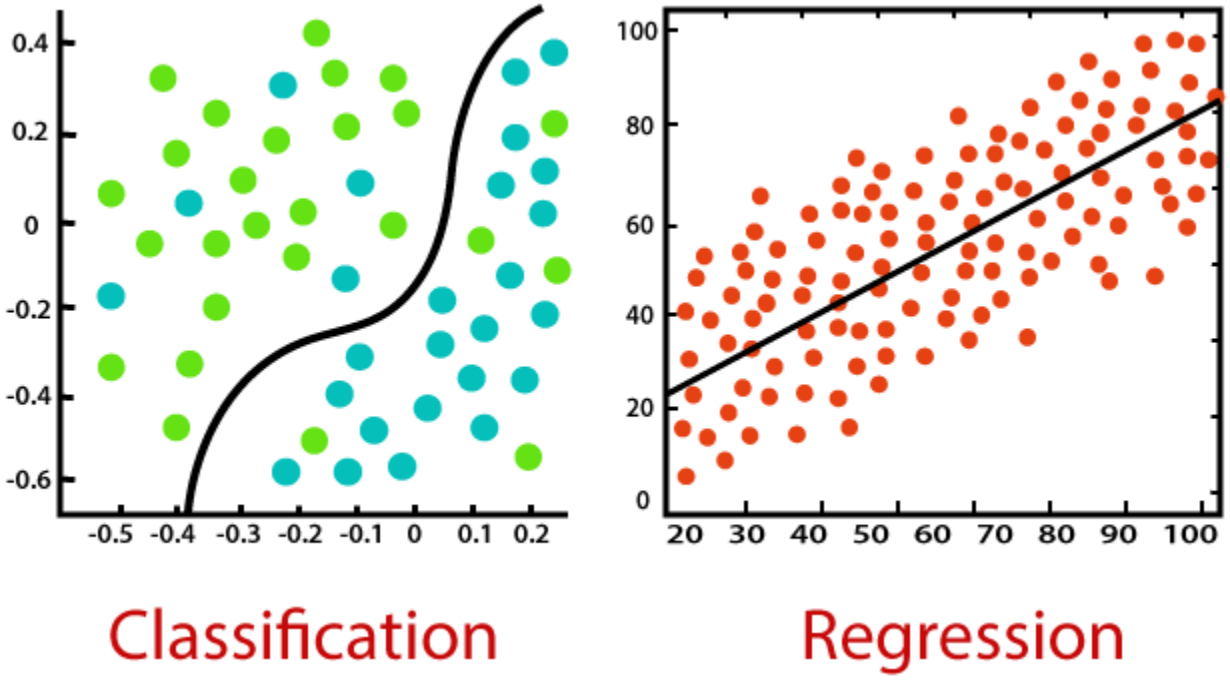
- **Definition**: Classification is a supervised learning technique that assigns labels to data points based on predefined categories.
- **Goal**: Predict the category of new data points.
- **Output**: Discrete labels (e.g., categories).
- **Examples**:
  - Email spam detection
  - Handwritten digit recognition

- **Techniques**:
  - Decision Trees
  - Random Forest
  - Support Vector Machines (SVM)

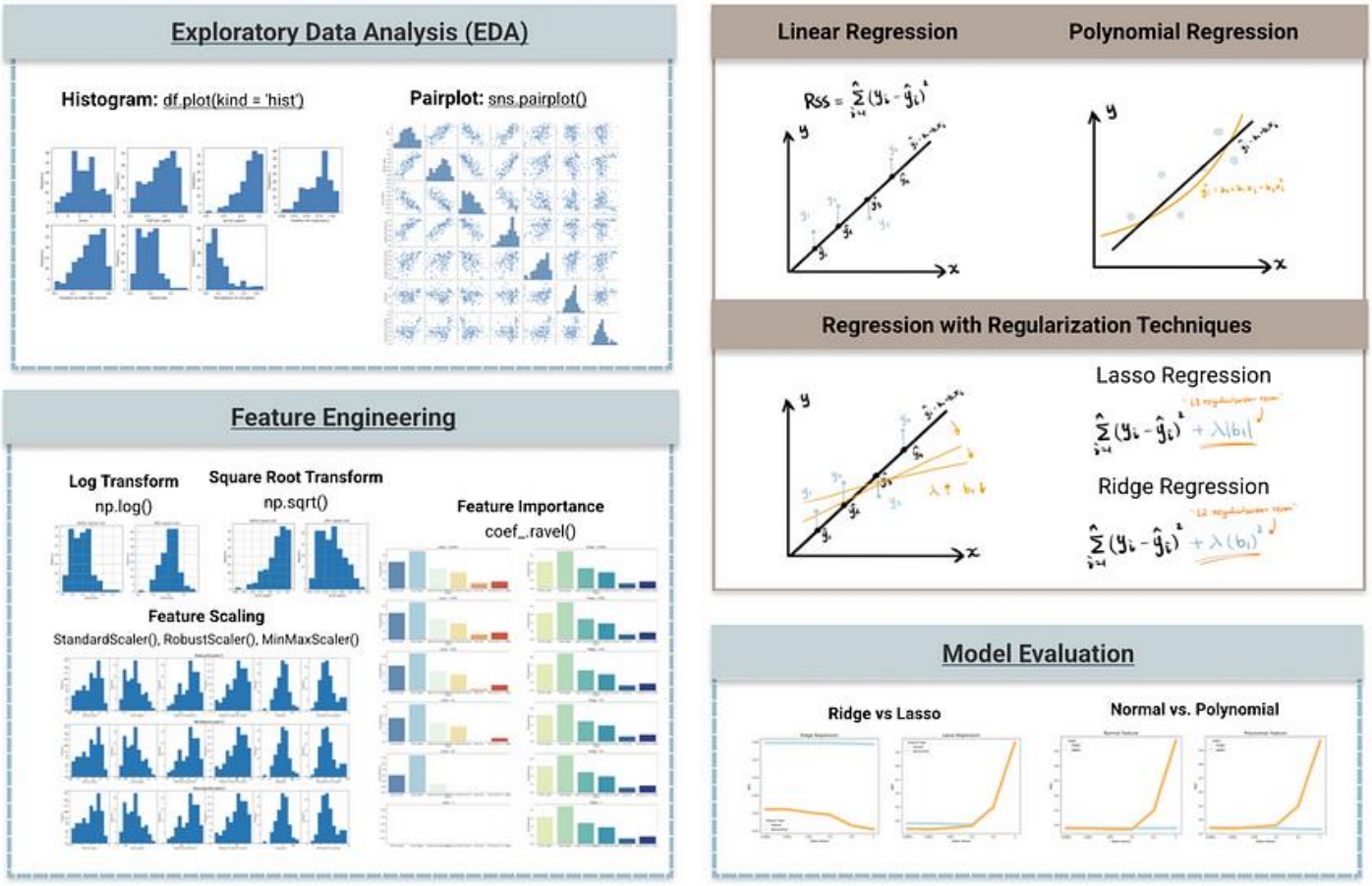
**Comparison Table: Regression vs. Classification**

| Feature | Regression | Classification |
|---|---|---|
| **Definition** | Predicts continuous values | Assigns labels to data points |
| **Goal** | Estimate relationships and predict outcomes | Predict the category of new data points |
| **Output** | Continuous values (e.g., real numbers) | Discrete labels (e.g., categories) |
| **Examples** | House price prediction, Temperature forecasting | Email spam detection, Handwritten digit recognition |
| **Techniques** | Linear Regression, Polynomial Regression, SVR | Decision Trees, Random Forest, SVM |
| **Data Labeling** | Requires labeled data for training | Requires labeled data for training |
| **Approach** | Supervised learning | Supervised learning |

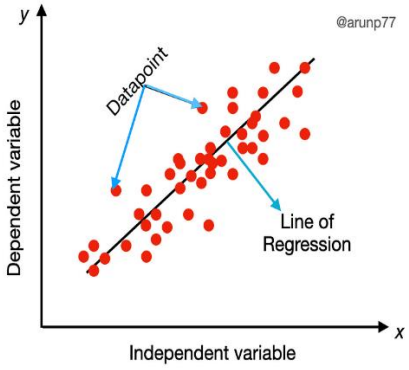
**Diagram: Regression vs. Classification**



# Machine Learning Algorithms - Regression




| Algorithm | Keyword | Diagram |
|---|---|---|
| Support Vector Machines (SVM) | Vector on Points | |
| Naïve Bayes | Probability Distribution | |
| Linear Regression<br>Logistic Regression | Straight Line<br>Logarithmic Line | |
| K-Means | Kernel (*central*) Mean | |
| K-Nearest Neighbour | Neighbouring Points | |
| Decision Trees | Tree Branches | |
| Neural Networks | Network with Layers of elements | input output hidden |




## Summary

Understanding the differences between clustering, classification, and regression is crucial in choosing the right approach for a given machine learning task. Clustering helps in discovering patterns without pre-labeled data, while classification and regression require labeled data to make predictions but differ in the nature of their outputs—discrete labels for classification and continuous values for regression.

This document provides a comprehensive overview, complete with diagrams to visually distinguish between these techniques.