Geek Coders

INNOVAITE

Presented by Shreya Suresh & Mukundan

Borcelle University

Session 2: Train Your First Model

Learn to build and train their first machine learning model, covering data preprocessing, model selection, training, and evaluation. This hands-on session is designed to give beginners real-world experience and a solid foundation in machine learning techniques.

Geek Coders

Overview

Geek Coders

Introduction

What is Machine Learning?

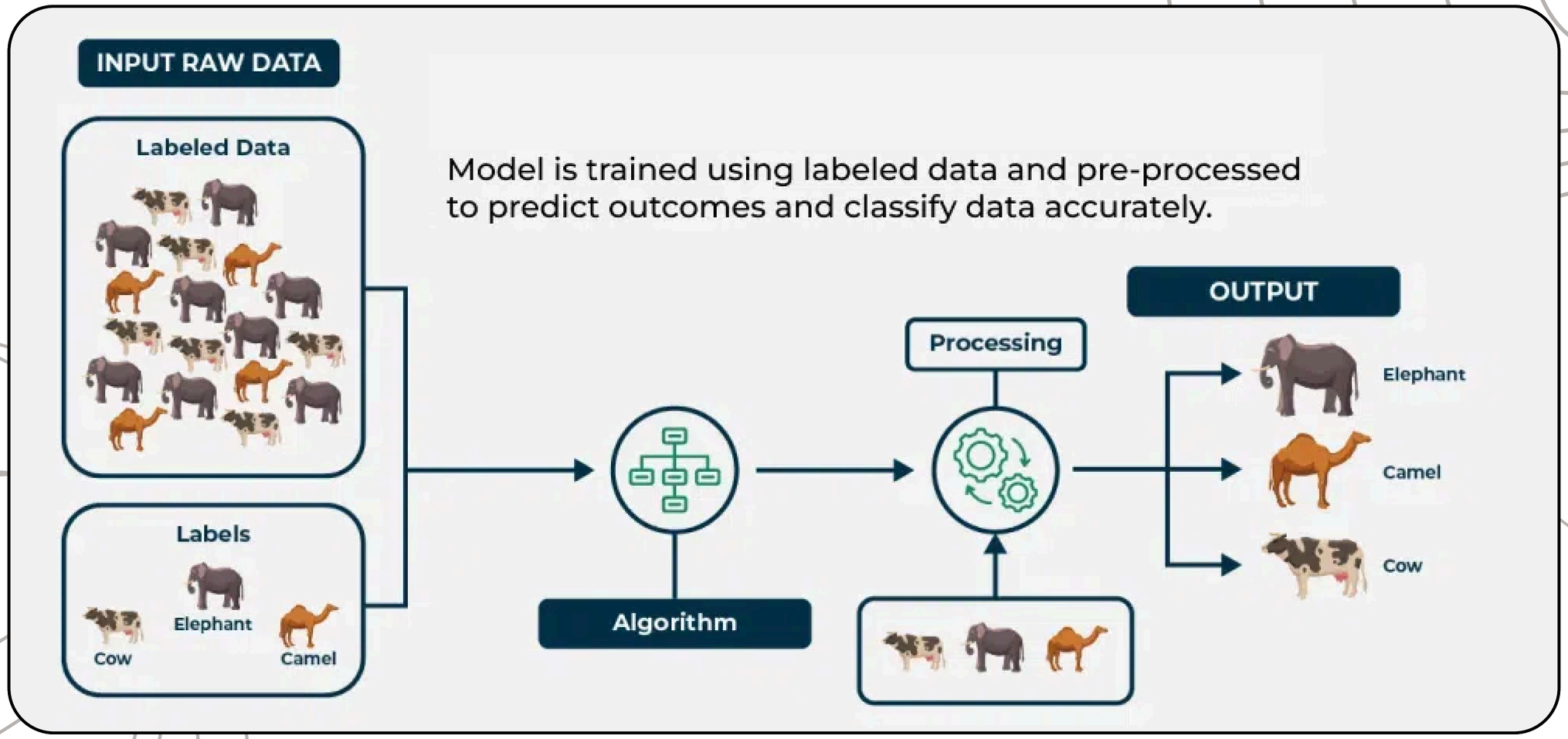
Machine Learning is a subset of Artificial Intelligence where computers are taught to learn from data and improve their performance over time without being explicitly programmed. Unlike traditional programming, where we write step-by-step instructions, ML uses algorithms to identify patterns in data and make predictions or decisions.

Geek Coders

Types of Machine Learning

Geek Coders

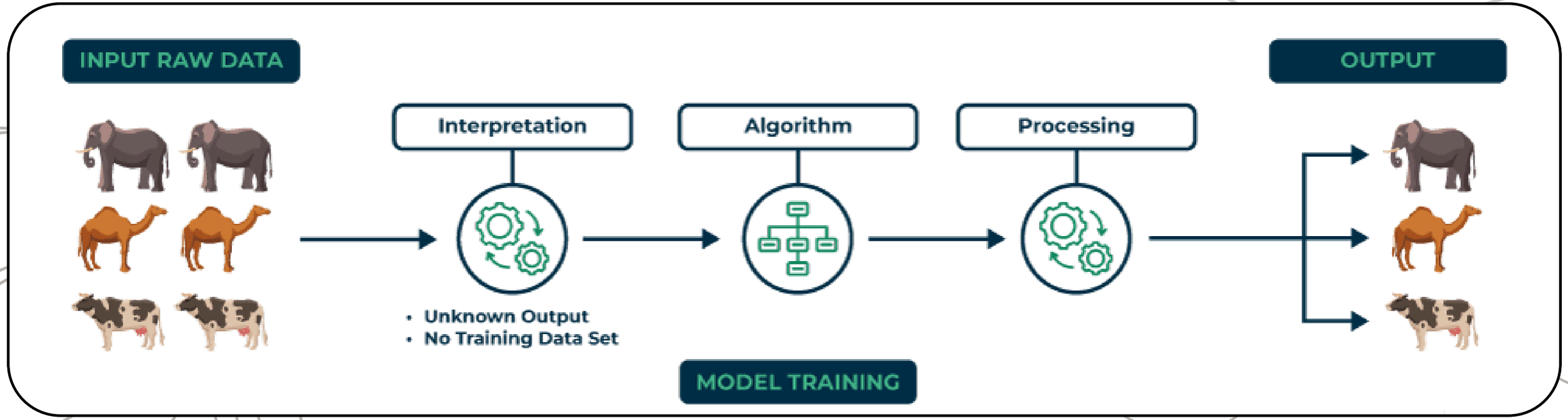
Supervised Learning:



Geek Coders

Unsupervised Learning

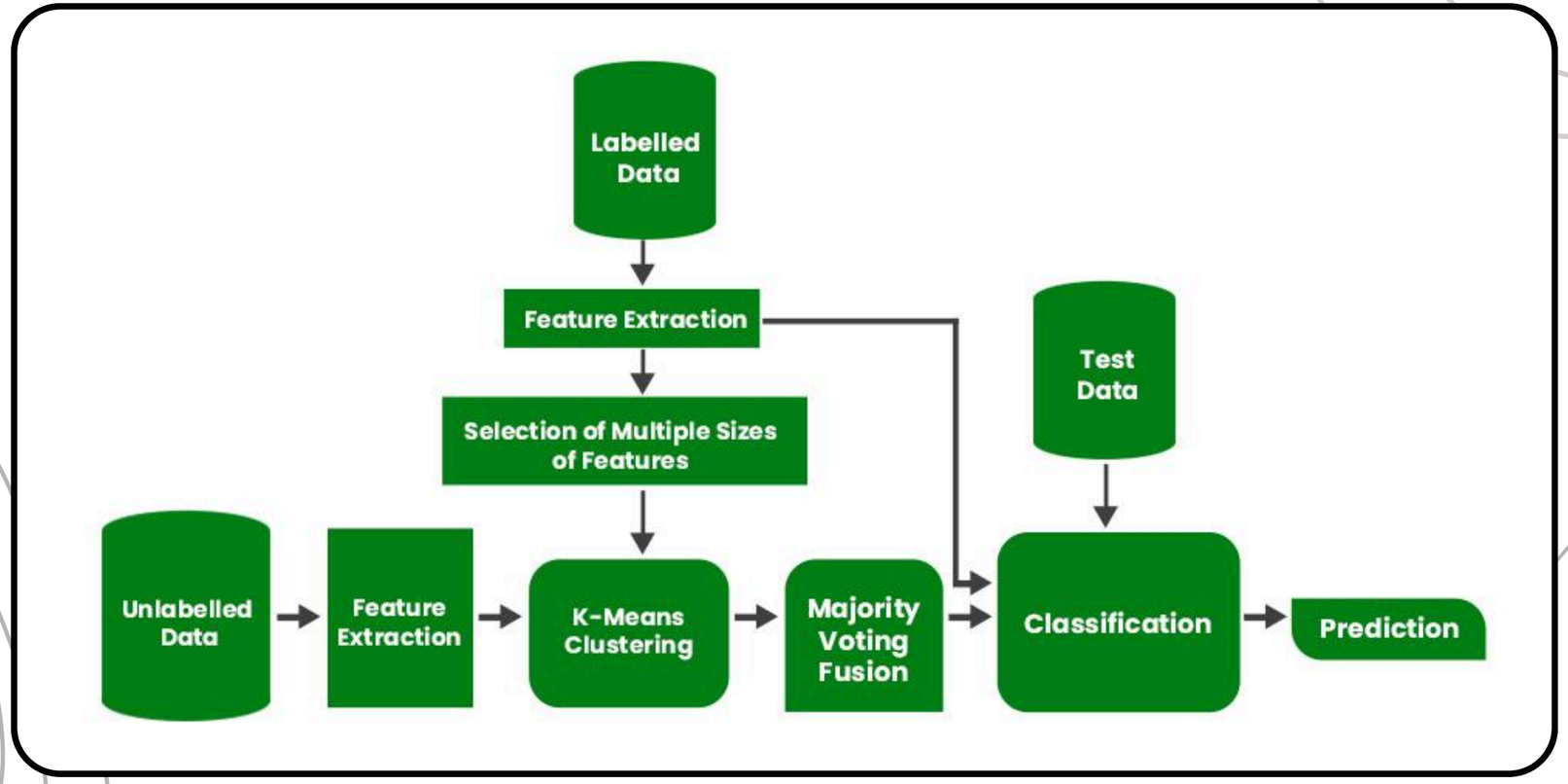
The algorithm identifies patterns in unlabeled data, like grouping customers based on purchasing behavior.



Geek Coders

Semi-supervised

Semi-supervised learning is particularly useful when there is a large amount of unlabeled data available, but it's too expensive or difficult to label all of it.



Geek Coders

Applications of Machine Learning

ML has revolutionized many industries

Here are a few applications:

1. **Healthcare:**
   Predicting diseases, analyzing medical images, and personalizing treatments.
2. **Finance:**
   Fraud detection, stock market predictions, and credit scoring.
3. **Retail:**
   Product recommendations and inventory optimization.
4. **Transportation:**
   Self-driving cars and route optimization.

Spotify use ML to recommend shows and music tailored to your preferences.

Geek Coders

Why Learn Machine Learning?

Why Learn Machine Learning?

The demand for ML professionals is growing exponentially as industries increasingly rely on data-driven solutions. Learning ML equips you with the skills to tackle real-world problems creatively and effectively.

In the next part of this workshop, you'll dive deeper into ML concepts, tools, and hands-on activities. Thank you for joining, and I hope this sparks your curiosity to explore ML further!

Geek Coders

Machine Learning Workflow

01 Collect and Explore Data

02 Data Preprocessing

03 Select and Train Model

04 Model Evaluation

05 Deployment

Collect and Explore Data

- Data Collection: Gather data from relevant sources (databases, APIs, sensors).
- Exploratory Data Analysis (EDA): Analyze distributions, trends, and relationships.
- Visualization: Use tools like Matplotlib, Seaborn for data insights.

Data Preprocessing

- Handle missing or inconsistent data.
- Scale/normalize numerical features and encode categorical variables.
- Feature selection or dimensionality reduction (e.g., PCA).
- Split data into training, validation, and test sets.

Select and Train Model

- Choose the algorithm based on the problem type (e.g., Linear Regression, Neural Networks).
- Train the model using the training set.
- Fine-tune hyperparameters using techniques like grid search.

Model Evaluation

- Evaluate model performance on validation/test set.
- Use appropriate metrics:
  - Regression: RMSE, MAE, R^2.
  - Classification: Accuracy, Precision, Recall, F1 Score, AUC.
- Detect overfitting/underfitting via cross-validation.

Deployment

- Deploy the model in production (e.g., API, web app).
- Ensure scalability and robustness for real-world use.
- Monitor the model's performance post-deployment.

Geek Coders

Regression and Classification

Geek Coders

Overview of Regression

Regression is a supervised learning technique used to model the relationship between a dependent variable (target) and one or more independent variables (features). It predicts continuous outcomes, such as sales, temperatures, or house prices.

Geek Coders

Overview of Classification

Classification is a supervised machine learning task where the goal is to predict the category or class of a given input based on its features.
Classification problems can be binary (two classes) or multi-class (more than two classes). It has applications in spam detection, medical diagnosis, and image recognition.

Geek Coders

Exploratory Data Analysis (EDA)

01 Introduction

02 Steps in EDA

03 Outcome of EDA

Geek Coders

Introduction

EDA involves analyzing and visualizing data to understand its structure and patterns.

Geek Coders

Steps In EDA

- Load the data, check dimensions, column names, data types, and preview the first few rows.
- Identify missing values and decide whether to drop or impute them.
- Analyze individual variables using histograms, bar charts, and summary statistics.
- Examine relationships between variables using scatter plots, box plots, and correlation matrices.
- Use heatmaps, pair plots, and other visualizations to uncover patterns and trends.

Geek Coders

Outcome Of EDA

EDA provides insights into the dataset, identifying patterns, relationships, outliers, and missing values. It ensures clean, well-prepared data, forming the foundation for accurate modeling.

Linear regression

Geek Coders

Linear Regression

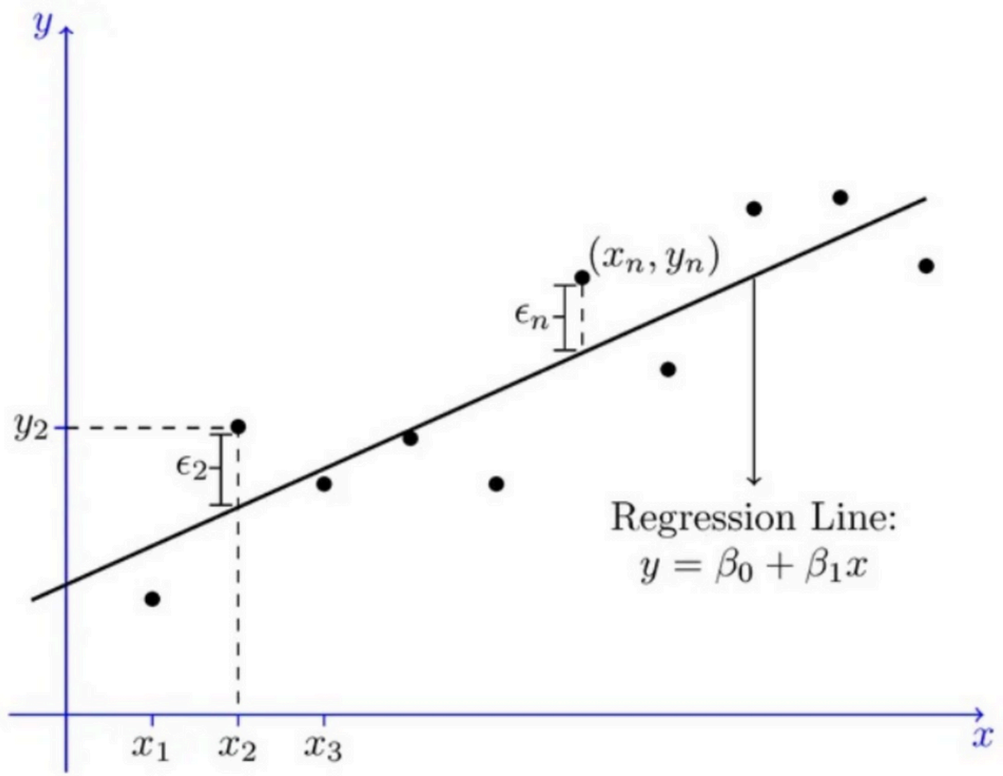
Linear regression is a supervised learning algorithm used to predict a continuous target variable (output) based on one or more input variables (features). It establishes a linear relationship between the input and output using the equation:

$$y=mx+c$$

In a multi-variable scenario:

$$y = w1x1 + w2x2 + ... + wnxn + b$$

Geek Coders



- **Applications: Predicting house prices, sales forecasting, and trend analysis.**

Geek Coders

Data Preprocessing

Geek Coders

Importance of Data Preprocessing

Data preprocessing improves data quality by handling missing values, outliers, and inconsistencies, ensuring reliable model performance. It also prepares data for analysis by transforming and normalizing features to enhance model accuracy and efficiency.

Geek Coders

Steps in Data Preprocessing

Data Cleaning:

- Handle missing values (e.g., imputation or removal).
- Remove duplicates and fix inconsistencies.

Data Transformation:

- Normalize or standardize numerical data.
- Encode categorical variables (e.g., one-hot encoding).

Feature Selection/Engineering:

- Select relevant features or create new ones.

Data Splitting:

- Divide the dataset into training, validation, and testing sets.

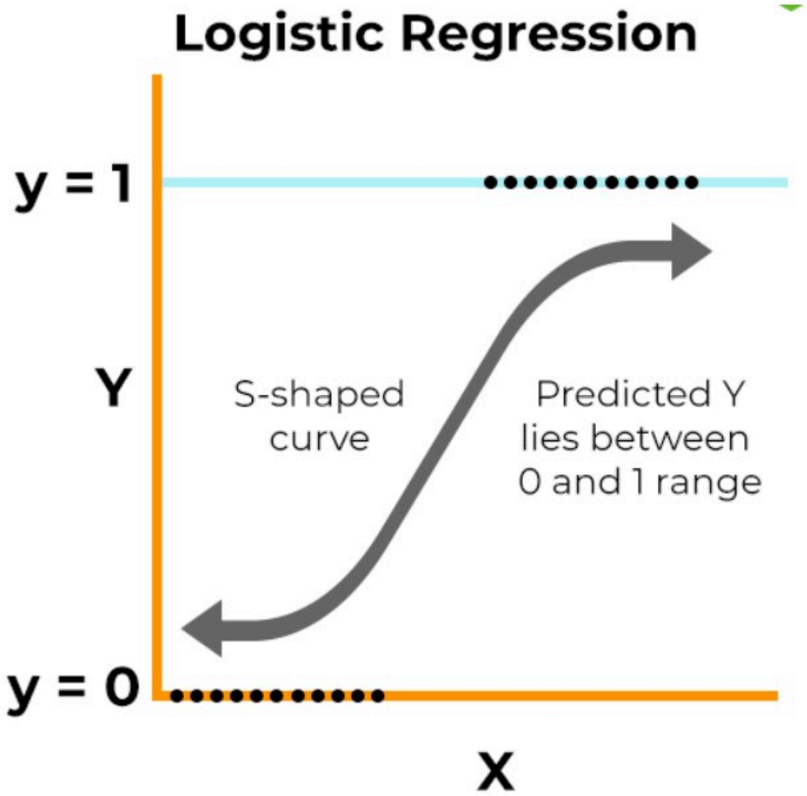
Logisctic regression

Geek Coders

How logistic regression works

Logistic Regression looks at the relationship between the input features (like age, income, or behavior) and the target outcome.
Instead of predicting a specific number (like in Linear Regression), it predicts the probability of the target being one category over the other.
The predicted probability is then converted into a class (e.g., churn or not churn) based on a threshold, usually 0.5.

Geek Coders



- **Simple and Effective:** Easy to use and works well for binary classification.
- **Interpretable**: Highlights how each feature impacts the outcome.
- **Probabilistic Outputs:** Provides confidence levels (e.g., "70% chance of churn").

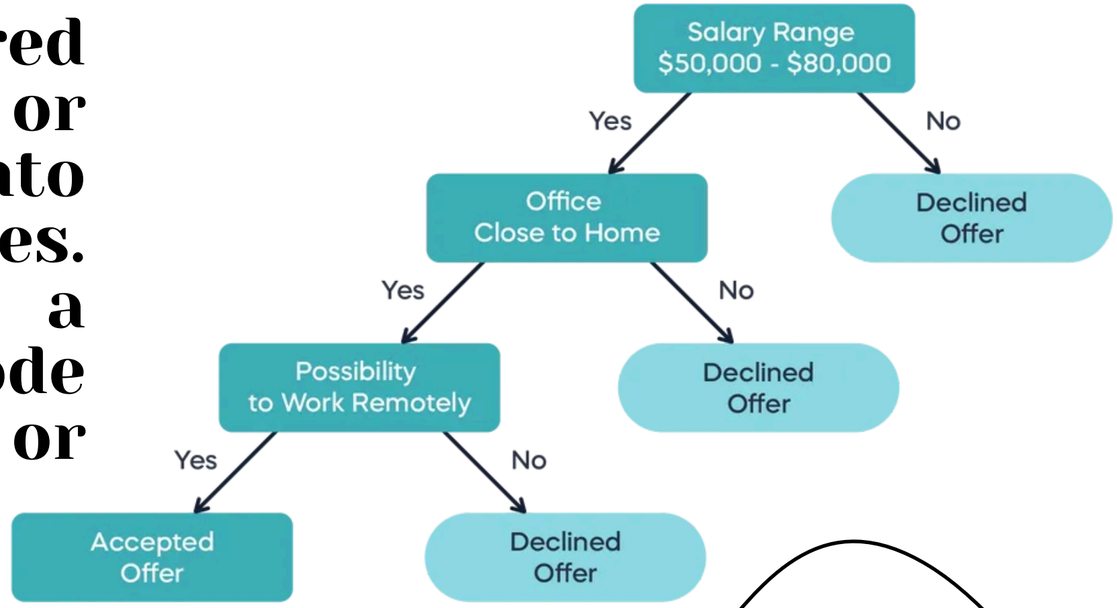
Geek Coders

Decision Trees

Geek Coders

What are Decision Trees?

A decision tree is a tree-structured model used for classification or regression, where data is split into branches based on feature values. Each internal node represents a decision, and each leaf node represents an outcome or prediction.



Geek Coders

Random Forest

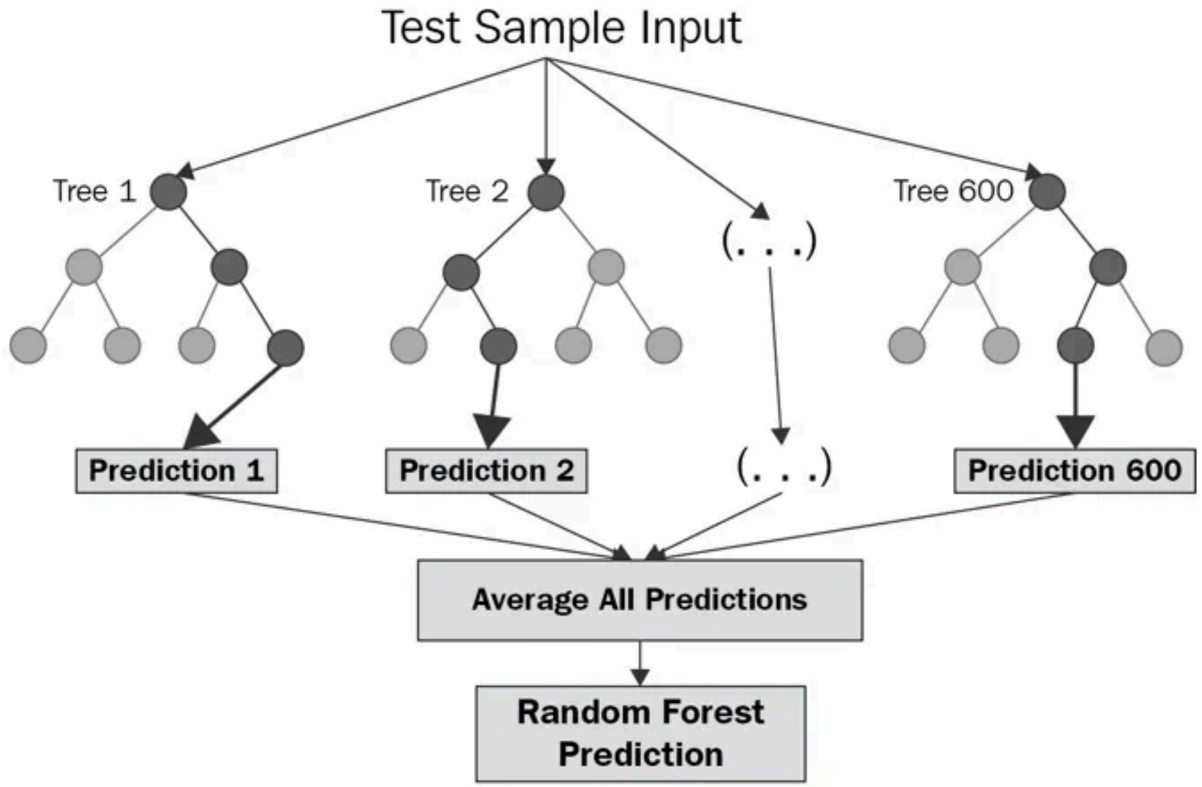
Geek Coders

What is Random Forest?

Random Forest is an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting. It works by averaging the predictions of individual trees for regression or using majority voting for classification tasks.

Geek Coders

What is Random Forest?



Geek Coders

SVM(Support Vector Machines)

What is SVM?

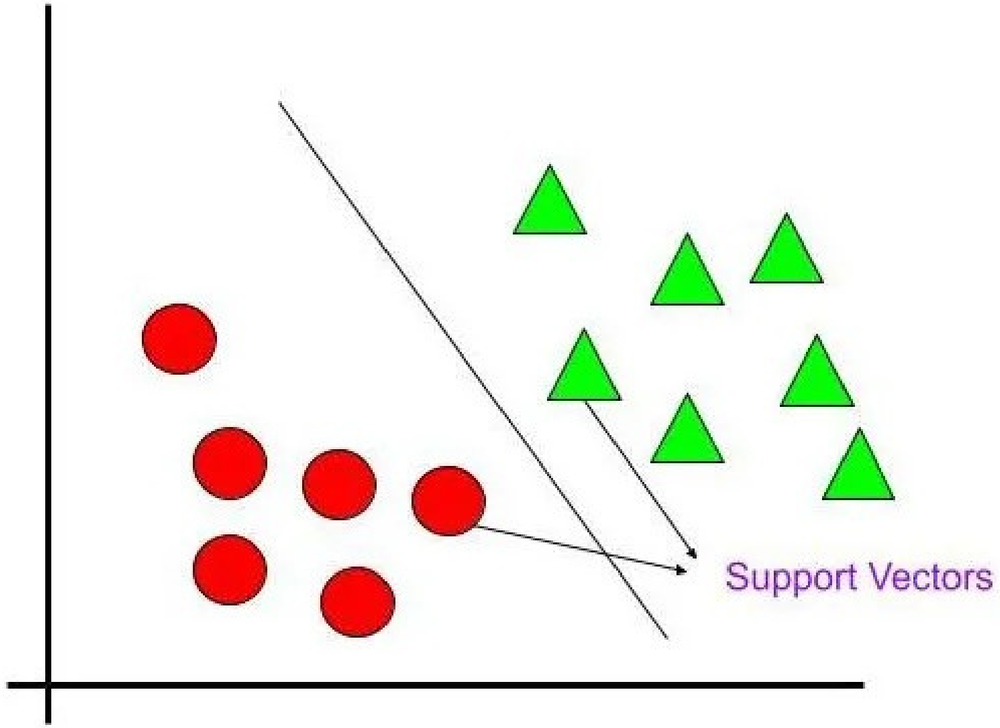
A supervised machine learning algorithm for classification and regression tasks. It separates data into classes using a hyperplane.

Key Concepts:

- Support Vectors: Points closest to the hyperplane that determine its position.
- Hyperplane: The decision boundary separating classes
- Margin: Distance between the hyperplane and the nearest data points. SVM maximizes this margin.

Applications

- Spam filtering
- Image recognition
- Sentiment analysis



Geek Coders

PCA (Principal Component Analysis)

Geek Coders

Purpose of PCA

Principal Component Analysis (PCA) simplifies high-dimensional data by reducing the number of features while retaining as much variance as possible. It identifies new axes (principal components) that maximize data variance, improving computational efficiency and reducing overfitting.

Geek Coders

K-Means Clustering

Geek Coders

K-means Clustering

Data Cleaning:

- Handle missing values (e.g., imputation or removal).
- Remove duplicates and fix inconsistencies.

Data Transformation:

- Normalize or standardize numerical data.
- Encode categorical variables (e.g., one-hot encoding).

Feature Selection/Engineering:

- Select relevant features or create new ones.

Data Splitting:

- Divide the dataset into training, validation, and testing sets.

Geek Coders

Neural Networks (NN)

Geek Coders

What are Neural Networks?

Inspired by the human brain for pattern recognition.
Composed of input, hidden, and output layers.

Image and speech recognition, natural language processing.
Power behind deep learning models like CNNs and RNNs.