INF 3202 - Data Mining and Business Intelligence

Data Mining Algorithms

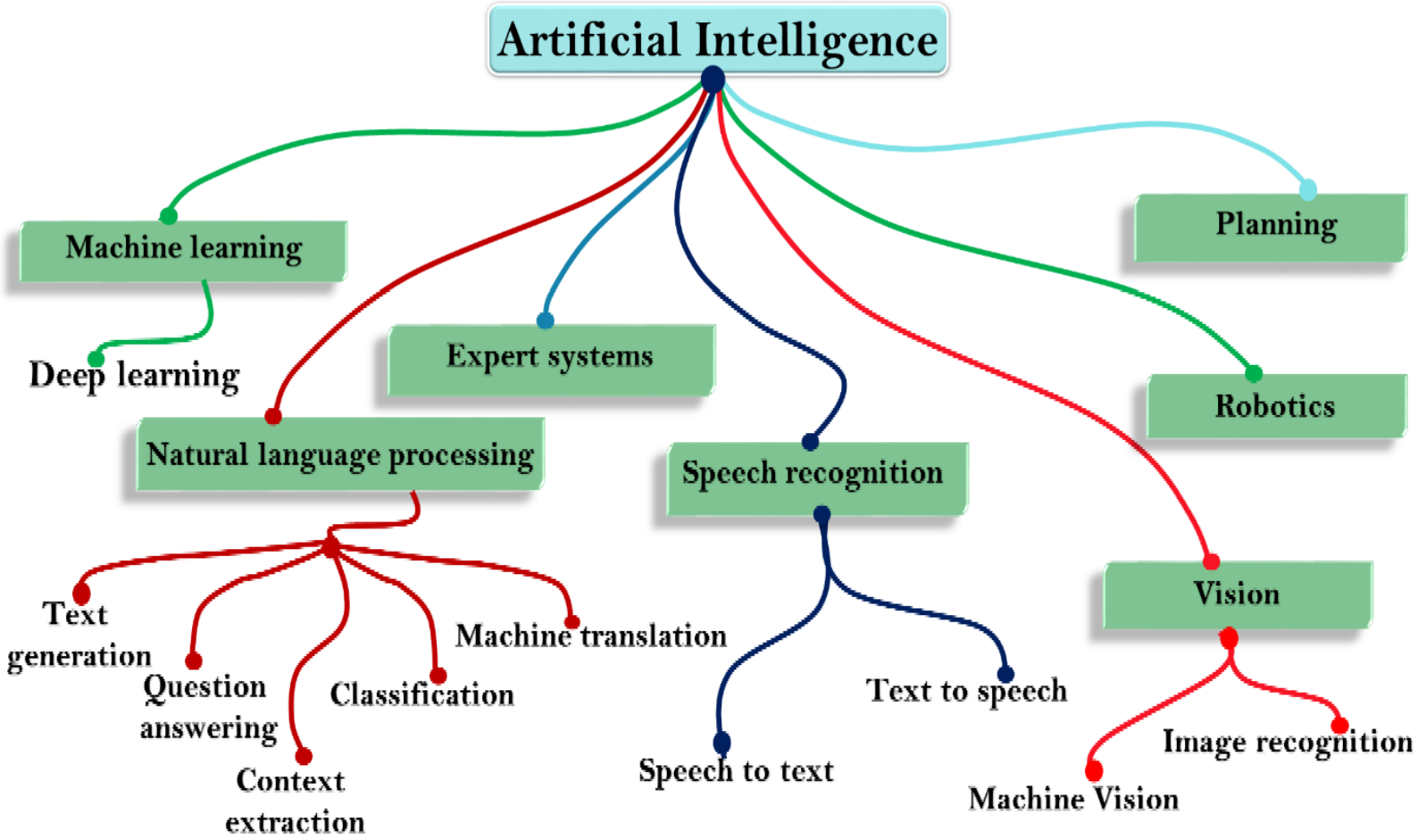
By

Dr. Omar Haji Kombo

Introduction

- Data mining algorithms are computational procedures used to discover patterns, correlations, or relationships within large datasets.
- These algorithms sift through massive amounts of data to extract useful information, trends, or insights that might not be immediately apparent.

Artificial Intelligence



Machine Learning Approaches to Data Mining

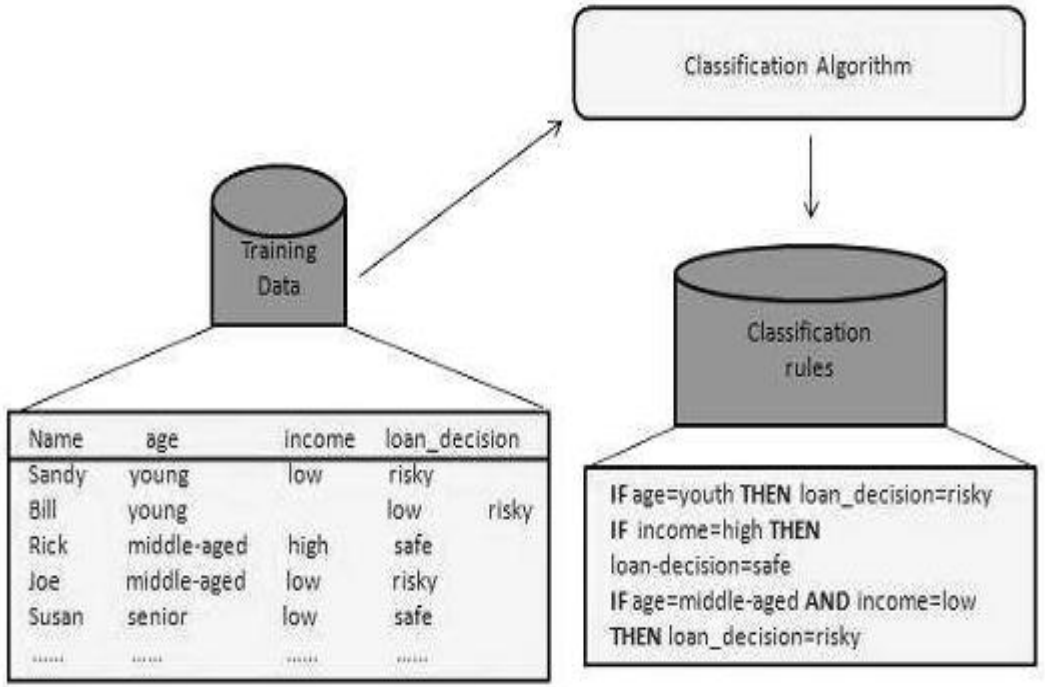
Machine learning methods are fundamental in extracting meaningful patterns and insights from large datasets in various domains.

These approaches include:

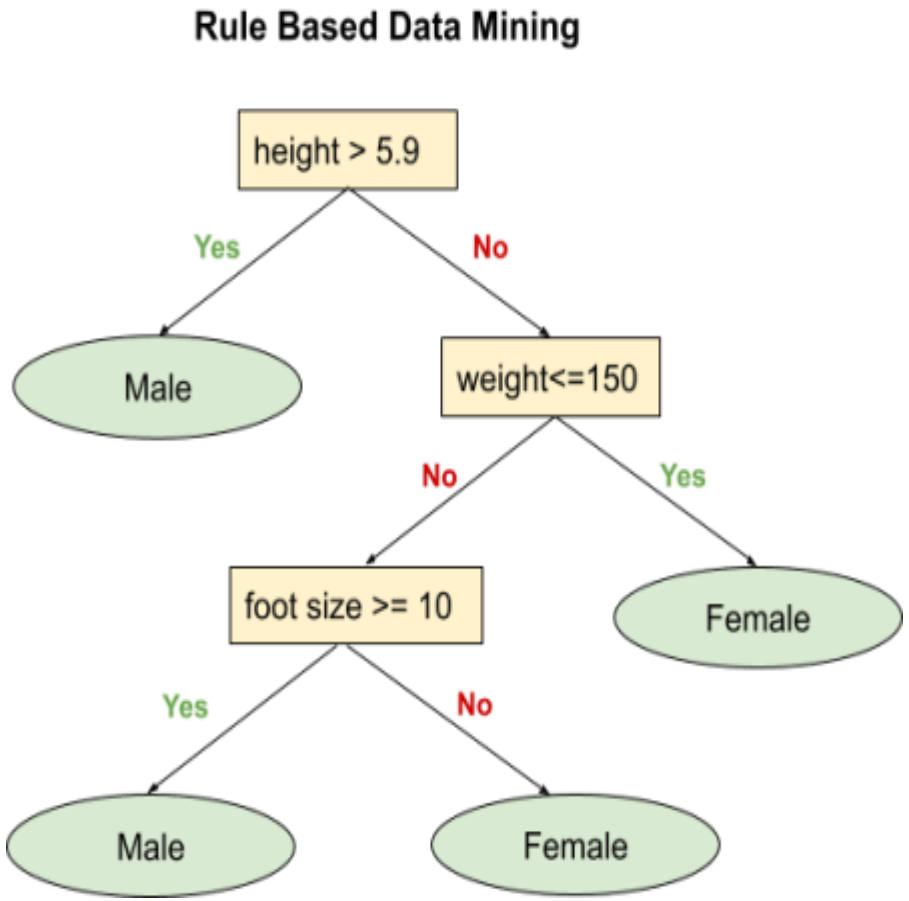
- Classification: Categorizing data into predefined classes.
- Regression: Predicting continuous values.
- Clustering: Grouping similar data points into clusters.
- Association Rule Learning: Finding interesting relationships or patterns between variables.
- Anomaly Detection: Identifying rare or unusual data points that differ significantly from the majority of the data.
- Dimensionality Reduction: Reducing the number of features or variables in a dataset.
- Reinforcement Learning: Learning optimal actions through trial and error by interacting with an environment.

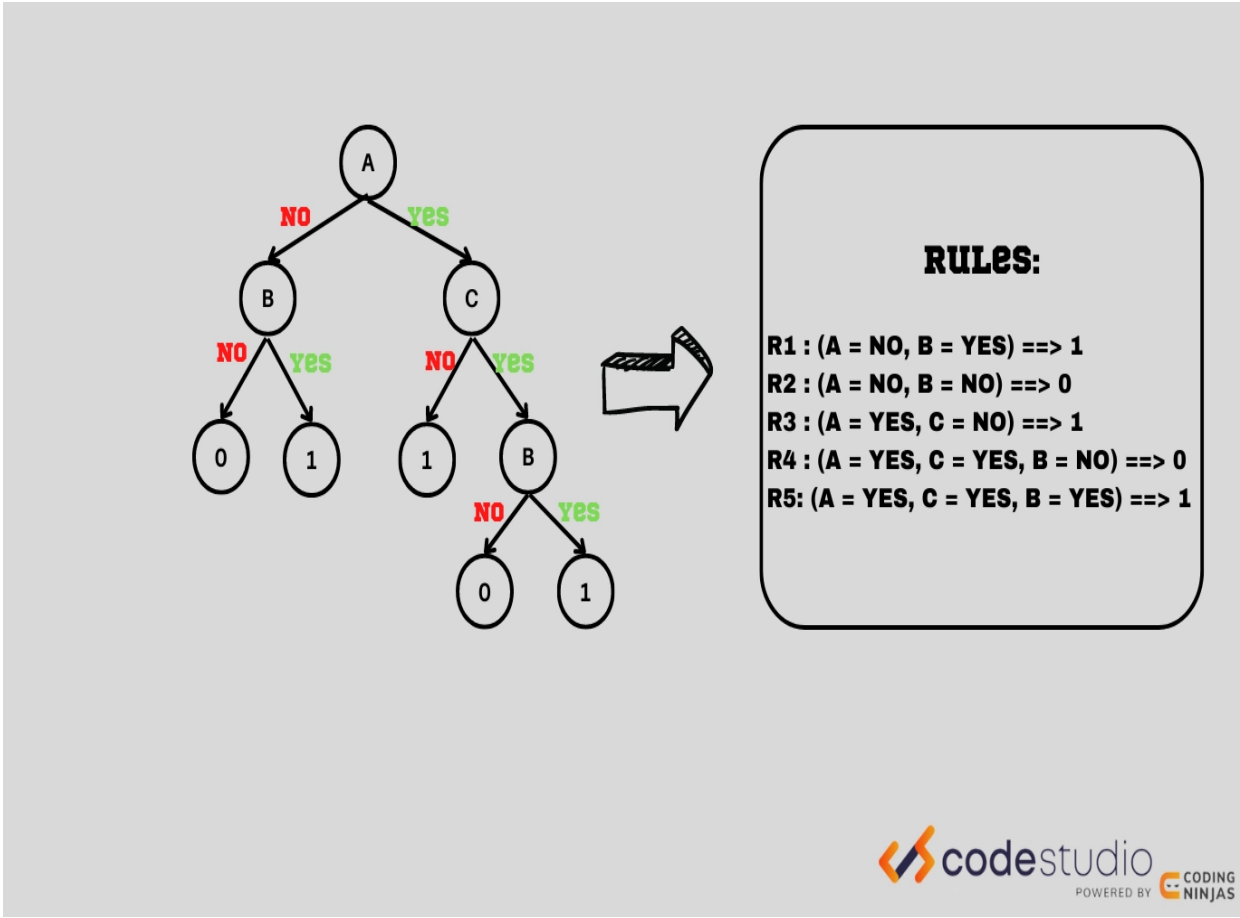
Rule-based Data Mining

- Rule-based classification in data mining is a technique in which class decisions are taken based on various "if...then… else" rules. Thus, we define it as a classification type governed by a set of IF-THEN rules. We write an IF-THEN rule as:
- **"IF condition THEN conclusion."**



Example of Rude-based Data Mining





Types of AI Algorithms

- Supervised learning
- Semi-supervised learning
- Reinforced learning
- Unsupervised learning

Supervised Algorithms

These algorithms learn from **labeled data**, where each input is associated with a corresponding output label. Examples include:

- Linear Regression
- Logistic Regression
- Decision Trees
- Random Forest
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Neural Networks

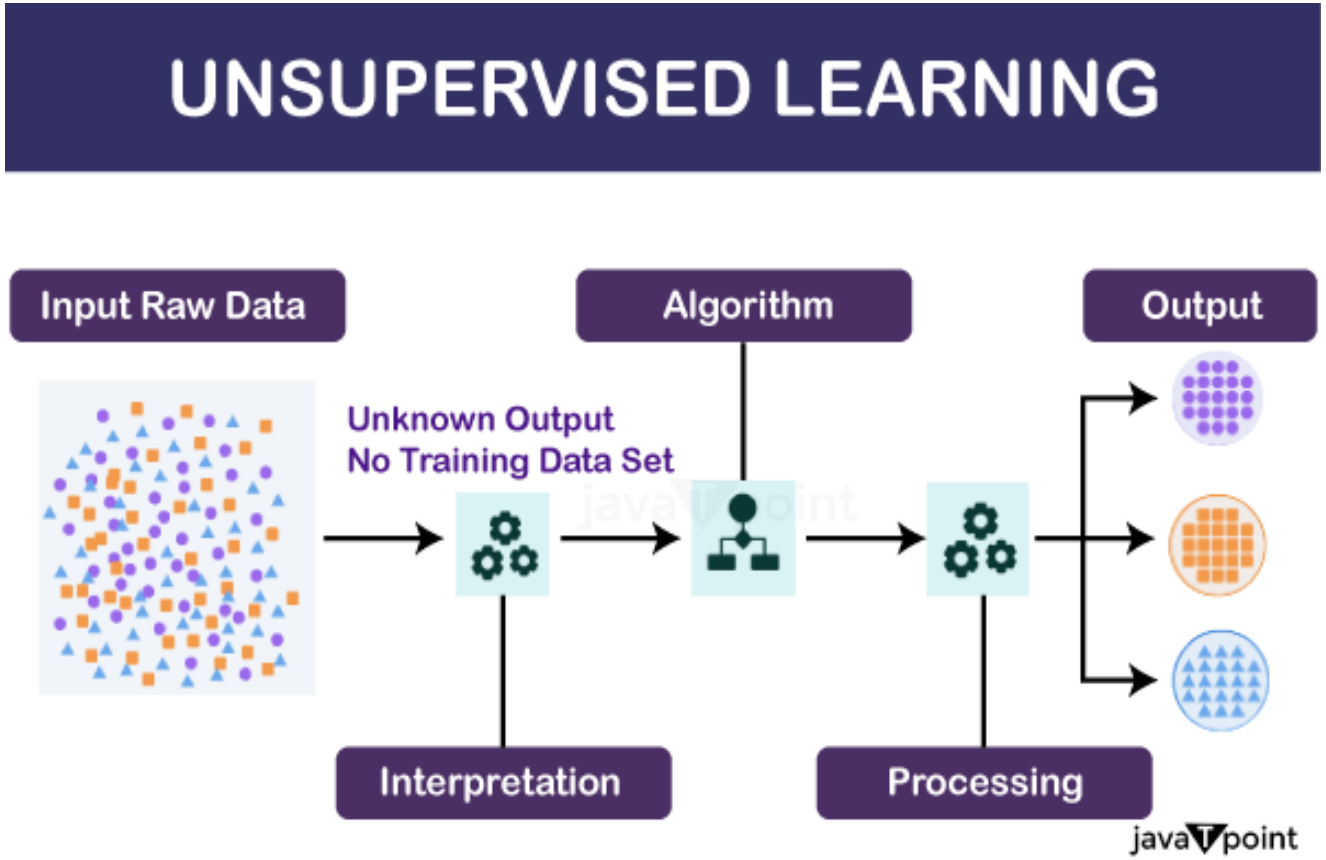
- Applications: Image Classification, Text Classification, Regression Analysis, Recommendation Systems, Natural Language Processing (NLP), Time Series Forecasting, Healthcare, Finance.

Unsupervised Learning Algorithms

- These algorithms learn from unlabeled data, seeking to find hidden patterns or structures within the data. Examples include:
  - K-means Clustering
  - Hierarchical Clustering
  - Principal Component Analysis (PCA)
  - Independent Component Analysis (ICA)
  - Generative Adversarial Networks (GANs)
  - Self-Organizing Maps (SOM)
- Applications: Fraud detection, Intrusion detection, Detecting unusual network activity that could indicate a security breach, Equipment failure prediction, etc.

Unsupervised Learning

The algorithm iteratively processes the data to identify patterns, clusters, or structures that exist within it. During training, the algorithm adjusts its parameters to optimize a given objective function, such as minimizing intra-cluster distance in clustering algorithms or maximizing data variance in dimensionality reduction techniques.



Semi-Supervised Learning Algorithms:

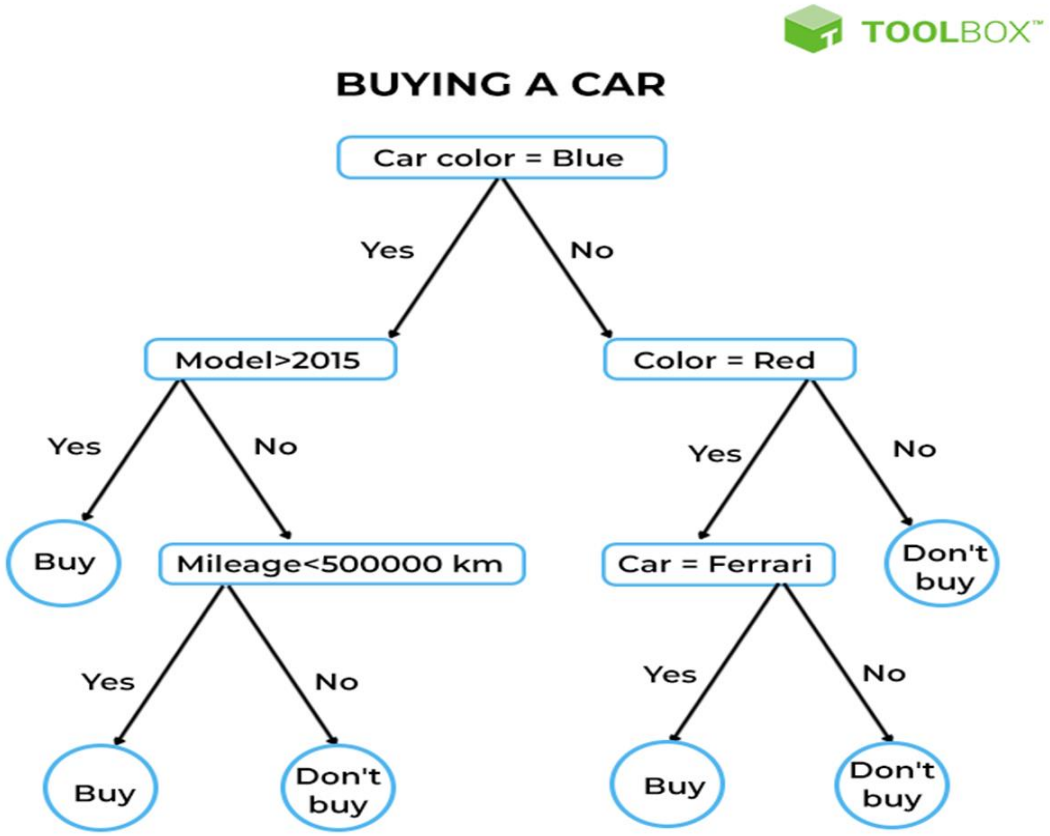
- These algorithms utilize a combination of labeled and unlabeled data for training. They leverage the available labeled data along with the unlabeled data to improve learning accuracy and generalization. Examples include:
  - Label Propagation
  - Co-Training
  - Self-Training
  - Tri-Training

Reinforcement Learning Algorithms

- These algorithms learn through interaction with an environment by taking actions and receiving rewards or penalties based on those actions. Examples include:
  - Q-Learning
  - Deep Q-Networks (DQN)
  - Policy Gradient Methods
  - Actor-Critic Methods
  - Proximal Policy Optimization (PPO)
  - Monte Carlo Tree Search (MCTS)

Decision trees algorithm

- Decision trees are supervised machine learning operations that model decisions, outcomes, and predictions using a flowchart-like tree structure.



Text Mining and Natural Language Processing (NLP) Algorithms

- These algorithms process and analyze natural language data, enabling machines to understand, interpret, and generate human language. Examples include:
  - Techniques for analyzing unstructured text data
  - Word Embeddings (Word2Vec, GloVe)
  - Recurrent Neural Networks (RNNs)
  - Long Short-Term Memory Networks (LSTMs)
  - Transformer Models (BERT, GPT)
  - Named Entity Recognition (NER)
  - Sentiment Analysis

Computer Vision Algorithms

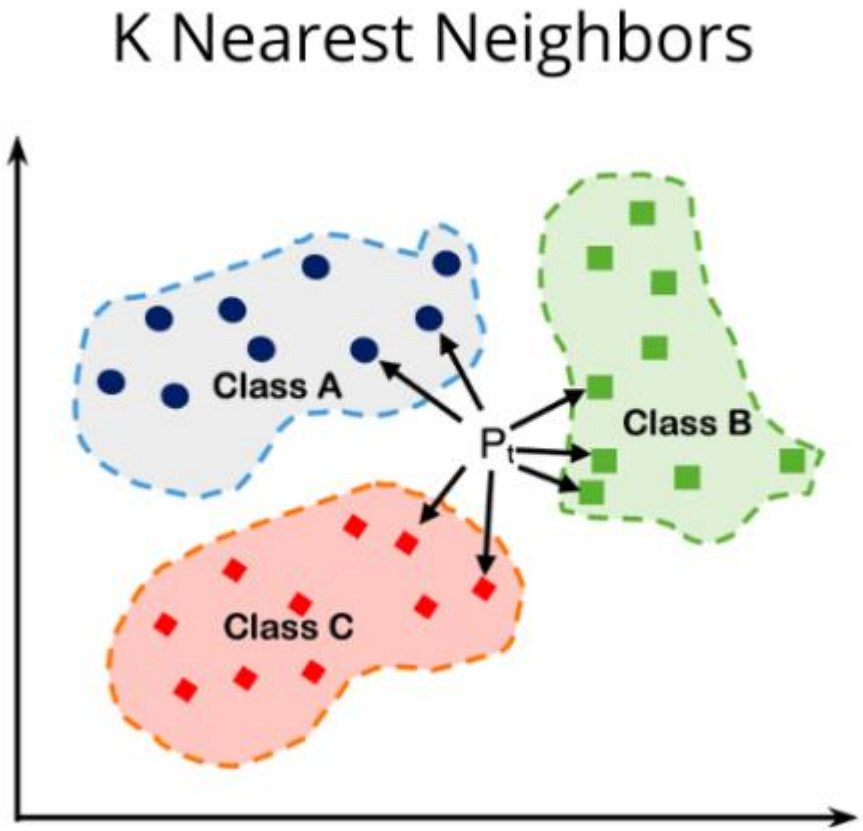
- These algorithms analyze and interpret visual data from the real world, enabling machines to perceive, understand, and interpret images and videos. Examples include:
  - Convolutional Neural Networks (CNNs)
  - Object Detection (YOLO, SSD)
  - Image Classification
  - Image Segmentation
  - Facial Recognition
  - Optical Character Recognition (OCR)

Evolutionary Algorithms

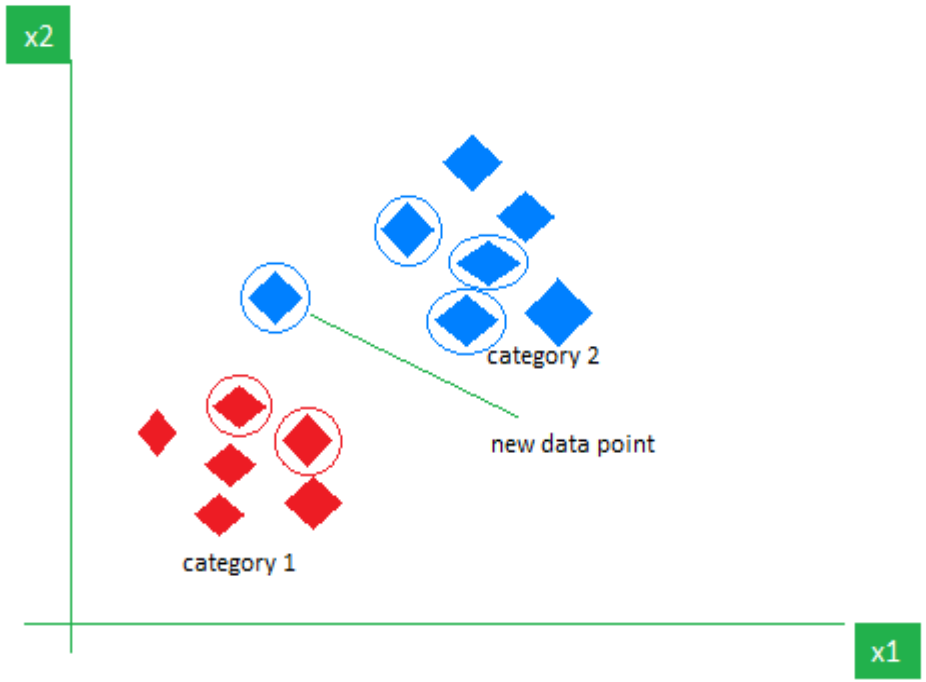
- These algorithms are inspired by biological evolution and natural selection processes to find optimal solutions to complex problems. Examples include:
  - Genetic Algorithms
  - Genetic Programming
  - Evolution Strategies
  - Differential Evolution
  - Particle Swarm Optimization (PSO)
  - Ant Colony Optimization (ACO)

K-Nearest Neighbors

- The **K-Nearest Neighbors (KNN) algorithm** is a supervised machine learning method employed to tackle classification and regression problems. Evelyn Fix and Joseph Hodges developed this algorithm in 1951, which was subsequently expanded by Thomas Cover. The article explores the fundamentals, workings, and implementation of the KNN algorithm.



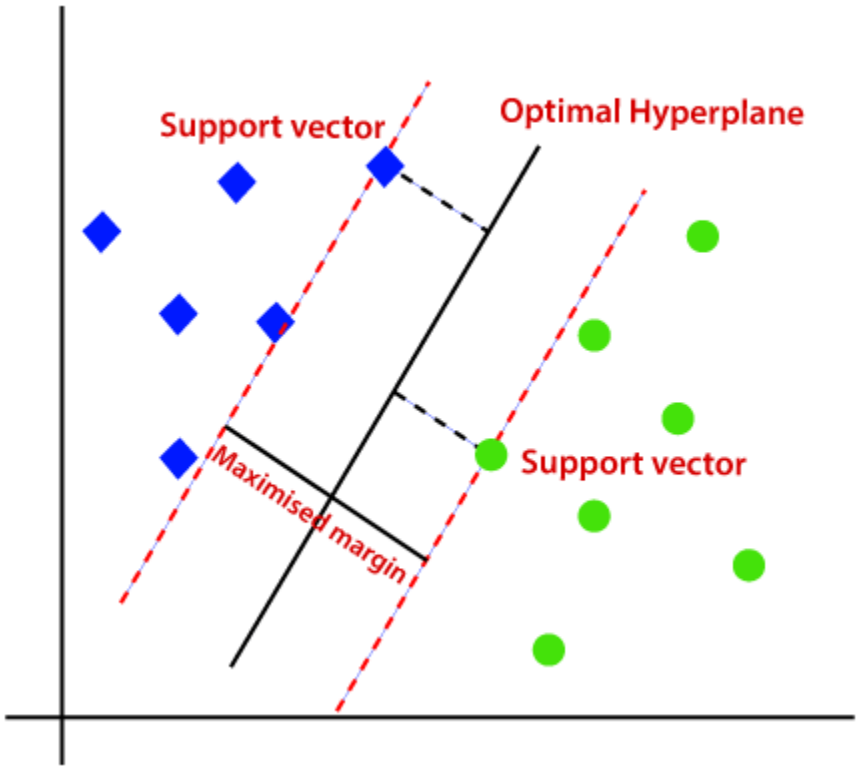
K-Nearest Neighbors



ID	Height	Age	Weight
1	5	45	77
2	5.11	26	47
3	5.6	30	55
4	5.9	34	59
5	4.8	40	72
6	5.8	36	60
7	5.3	19	40
8	5.8	28	60
9	5.5	23	45
10	5.6	32	58
11	5.5	38	?

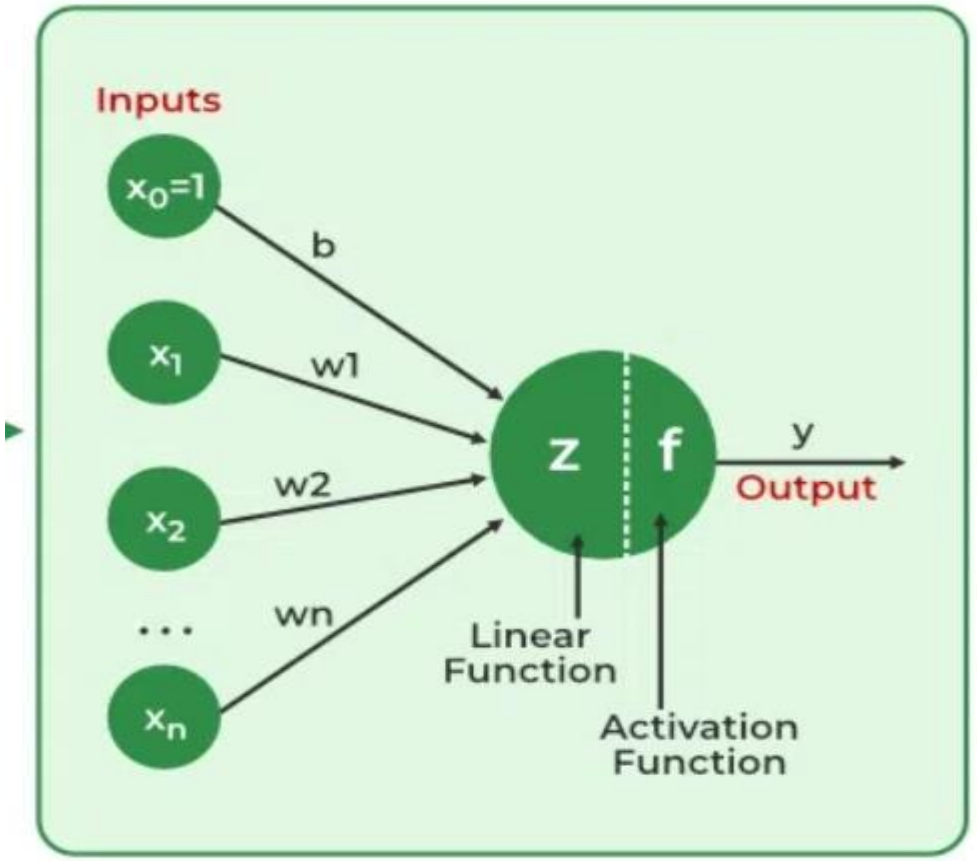
Support Vector Machines (SVM)

- SVM is a versatile machine learning algorithm utilized for linear or nonlinear classification, regression, and outlier detection tasks.
- It is widely employed in diverse applications like text and image classification, spam detection, gene expression analysis, face detection, and anomaly detection due to its adaptability and efficiency with high-dimensional data and nonlinear relationships.
- SVM excels in finding the maximum separating hyperplane between different classes within the target feature, making it highly effective for various classification tasks.



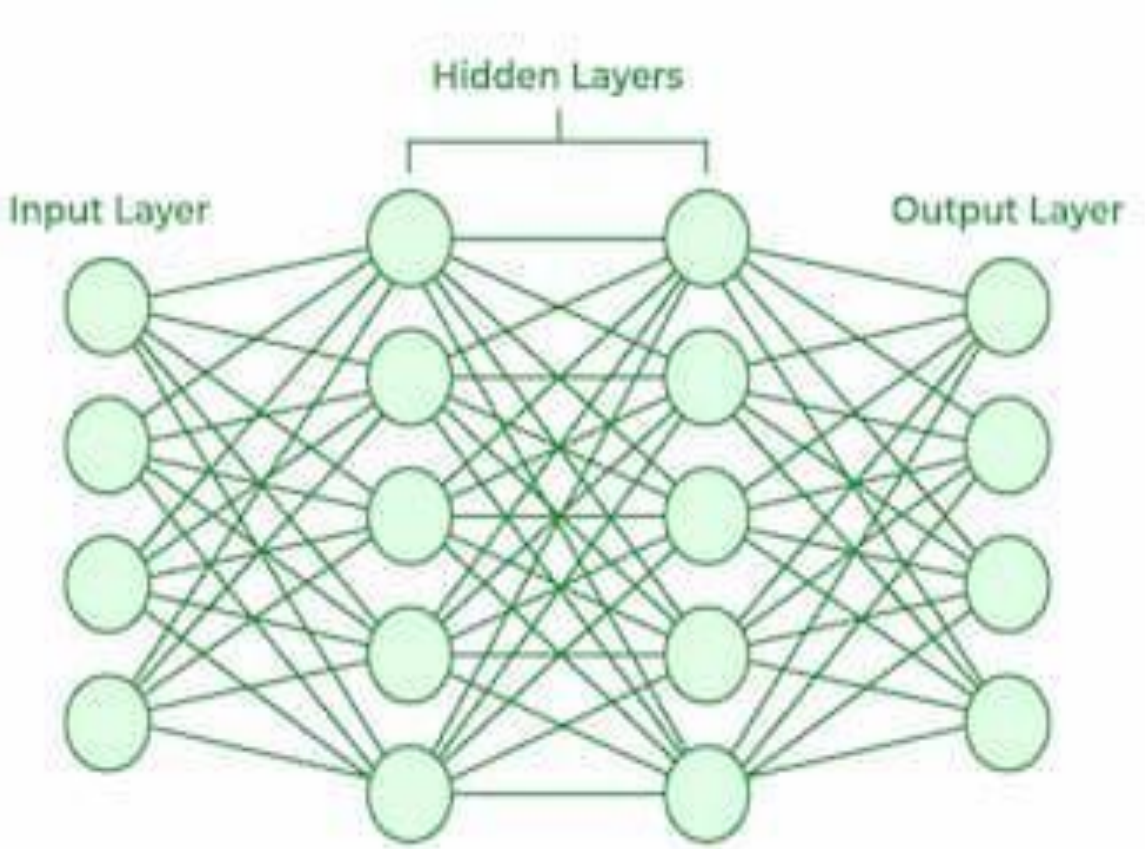
Neural Networks

- Neural Networks are computational models that mimic the complex functions of the human brain.
- The neural networks consist of interconnected nodes or neurons that process and learn from data, enabling tasks such as pattern recognition and decision making in machine learning.
- The article explores more about neural networks, their working, architecture and more



Neural Networks

The example of neural network is for email classification, with input features such as email content, sender details, and subject. These inputs are weighted and processed through hidden layers, enabling the network to learn patterns indicative of spam or legitimate emails. Through training, the network refines its weights via backpropagation, improving its ability to differentiate between spam and non-spam emails. Ultimately, the output layer produces a binary prediction (1 for spam, 0 for not spam), demonstrating the effectiveness of neural networks in tasks like email filtering.

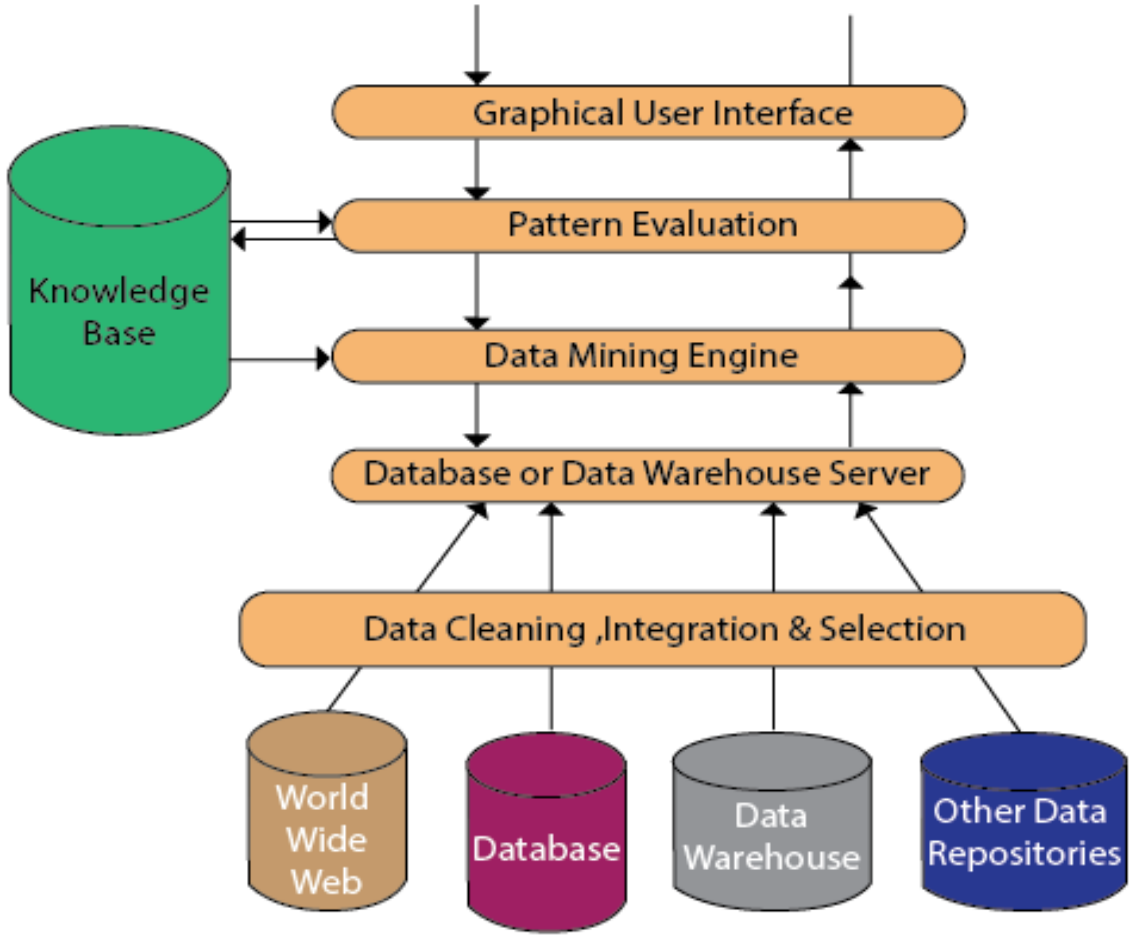


Ethical and Legal Considerations

- Privacy concerns and data protection regulations
- Ethical issues in data collection, analysis, and usage
- Ensuring fairness and transparency in data-driven decision making

Data Mining Architecture

Database: source of structured data
Data warehouse: easily accessed set of historical records from multiple sources
OLAP Cubes: more concise data summaries
Web Logs: text document that contains a record of all activity related to a specific web server over a defined period of time
Social Media Platforms: Analyzing tweets' sentiment towards particular brands
External Data Sources: Other external sources include government/industry data & surveys, online resources like blogs, etc.



Data Source

- The actual source of data is the Database, data warehouse, World Wide Web (WWW), text files, and other documents.
- You need a huge amount of historical data for data mining to be successful. Organizations typically store data in databases or data warehouses.
- Data warehouses may comprise one or more databases, text files spreadsheets, or other repositories of data.
- Sometimes, even plain text files or spreadsheets may contain information. Another primary source of data is the World Wide Web or the internet.

Data Preprocesses

- Before passing the data to the database or data warehouse server, the data must be cleaned, integrated, and selected.
- As the information comes from various sources and in different formats, it can't be used directly for the data mining procedure because the data may not be complete and accurate.
- So, the first data requires to be cleaned and unified. More information than needed will be collected from various data sources, and only the data of interest will have to be selected and passed to the server.
- These procedures are not as easy as we think. Several methods may be performed on the data as part of selection, integration, and cleaning.

Data Warehouse Server and Data Mining Engine

Data Warehouse Server

- The database or data warehouse server consists of the original data that is ready to be processed.
- Hence, the server is cause for retrieving the relevant data that is based on data mining as per user request.

Data Mining Engine

- The data mining engine is a major component of any data mining system.
- It contains several modules for operating data mining tasks, including association, characterization, classification, clustering, prediction, time-series analysis, etc.
- In other words, we can say data mining is the root of our data mining architecture.
- It comprises instruments and software used to obtain insights and knowledge from data collected from various data sources and stored within the data warehouse.

Pattern Evaluation Module

- The Pattern evaluation module plays a key role in examining patterns by applying a threshold value.
- It works in conjunction with the data mining engine to narrow down the search to interesting patterns.
- This component typically employs stake measures in collaboration with data mining modules to target the search towards intriguing patterns, possibly using a stake threshold to sift through discovered patterns.
- Alternatively, depending on the implementation of data mining techniques, the pattern evaluation module may be integrated with the mining module. For effective data mining, it's highly recommended to integrate the evaluation of pattern stake as deeply as possible into the mining process to restrict the search to only the most intriguing patterns.

Graphical User Interface

- The graphical user interface (GUI) module communicates between the data mining system and the user.
- This module helps the user to easily and efficiently use the system without knowing the complexity of the process.
- This module cooperates with the data mining system when the user specifies a query or a task and displays the results.

Knowledge Base

- The knowledge base is helpful in the entire process of data mining. It might be helpful to guide the search or evaluate the stake of the result patterns.
- The knowledge base may even contain user views and data from user experiences that might be helpful in the data mining process.
- The data mining engine may receive inputs from the knowledge base to make the result more accurate and reliable. The pattern assessment module regularly interacts with the knowledge base to get inputs, and also update it.

Thank you