CS262: FUNDAMENTALS OF DATA SCIENCE
FINALS

What is machine learning?

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

Recommendation engines are a common use case for machine learning. Other popular uses include fraud detection, spam filtering, malware threat detection, business process automation (BPA) and Predictive maintenance.

Why is machine learning important?

Machine learning is important because it gives enterprises a view of trends in customer behavior and business operational patterns, as well as supports the development of new products. Many of today's leading companies, such as Facebook, Google and Uber, make machine learning a central part of their operations. Machine learning has become a significant competitive differentiator for many companies.

What are the different types of machine learning?

Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. The type of algorithm data scientists chooses to use depends on what type of data they want to predict.

Supervised learning: In this type of machine learning, data scientists supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.
Unsupervised learning: This type of machine learning involves algorithms that train on unlabeled data. The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined.
Semi-supervised learning: This approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labeled training data, but the model is free to explore the data on its own and develop its own understanding of the data set.
Reinforcement learning: Data scientists typically use reinforcement learning to teach a machine to complete a multi-step process for which there are clearly defined rules. Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task. But for the most part, the algorithm decides on its own what steps to take along the way.

How does supervised machine learning work?

Supervised machine learning requires the data scientist to train the algorithm with both labeled inputs and desired outputs. Supervised learning algorithms are good for the following tasks:
Binary classification: Dividing data into two categories.
Multi-class classification: Choosing between more than two types of answers.
Regression modeling: Predicting continuous values.

Ensembling: Combining the predictions of multiple machine learning models to produce an accurate prediction.

How does unsupervised machine learning work?

Unsupervised machine learning algorithms do not require data to be labeled. They sift through unlabeled data to look for patterns that can be used to group data points into subsets. Most types of deep learning, including neural networks, are unsupervised algorithms. Unsupervised learning algorithms are good for the following tasks:

Clustering: Splitting the dataset into groups based on similarity.
Anomaly detection: Identifying unusual data points in a data set.
Association mining: Identifying sets of items in a data set that frequently occur together.
Dimensionality reduction: Reducing the number of variables in a data set.

How does semi-supervised learning work?

Semi-supervised learning works by data scientists feeding a small amount of labeled training data to an algorithm. From this, the algorithm learns the dimensions of the data set, which it can then apply to new, unlabeled data. The performance of algorithms typically improves when they train on labeled data sets. But labeling data can be time consuming and expensive. Semi-supervised learning strikes a middle ground between the performance of supervised learning and the efficiency of unsupervised learning.

Some areas where semi-supervised learning is used include:

Machine translation: Teaching algorithms to translate language based on less than a full dictionary of words.
Fraud detection: Identifying cases of fraud when you only have a few positive examples.
Labelling data: Algorithms trained on small data sets can learn to apply data labels to larger sets automatically.

How does reinforcement learning work?

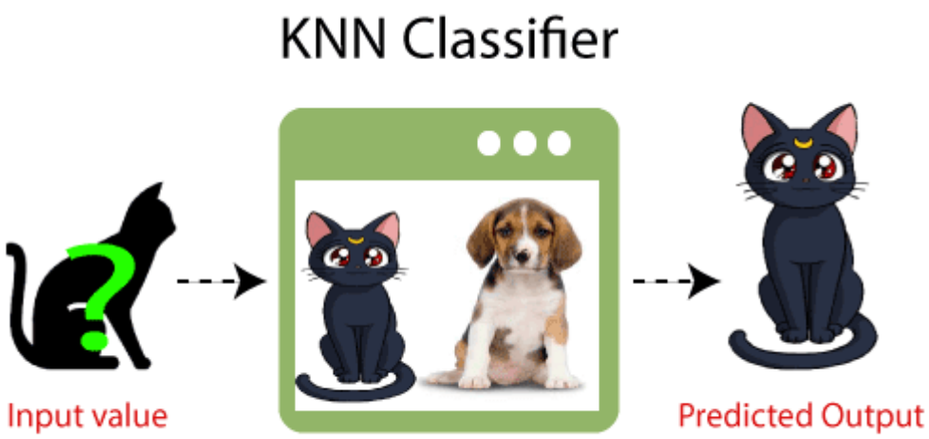
Reinforcement learning works by programming an algorithm with a distinct goal and a prescribed set of rules for accomplishing that goal. Data scientists also program the algorithm to seek positive rewards -- which it receives when it performs an action that is beneficial toward the ultimate goal -- and avoid punishments -- which it receives when it performs an action that gets it farther away from its ultimate goal. Reinforcement learning is often used in areas such as:

Robotics: Robots can learn to perform tasks the physical world using this technique.
Video gameplay: Reinforcement learning has been used to teach bots to play a number of video games.
Resource management: Given finite resources and a defined goal, reinforcement learning can help enterprises plan out how to allocate resources.

REFERENCE: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML#:~:text=Machine%20learning%20(ML)%20is%20a,to%20predict%20new%20output%20values.

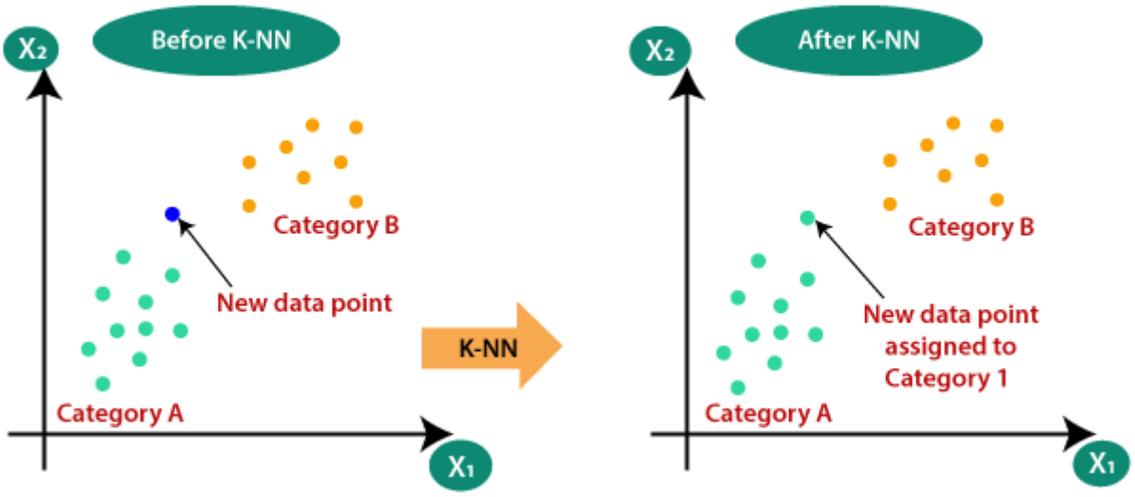
K-NEAREST NEIGHBOR (KNN)

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. ○ K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
- **Example:** Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.



Why do we need a K-NN Algorithm?

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:



Advantages of KNN Algorithm:

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

Disadvantages of KNN Algorithm:

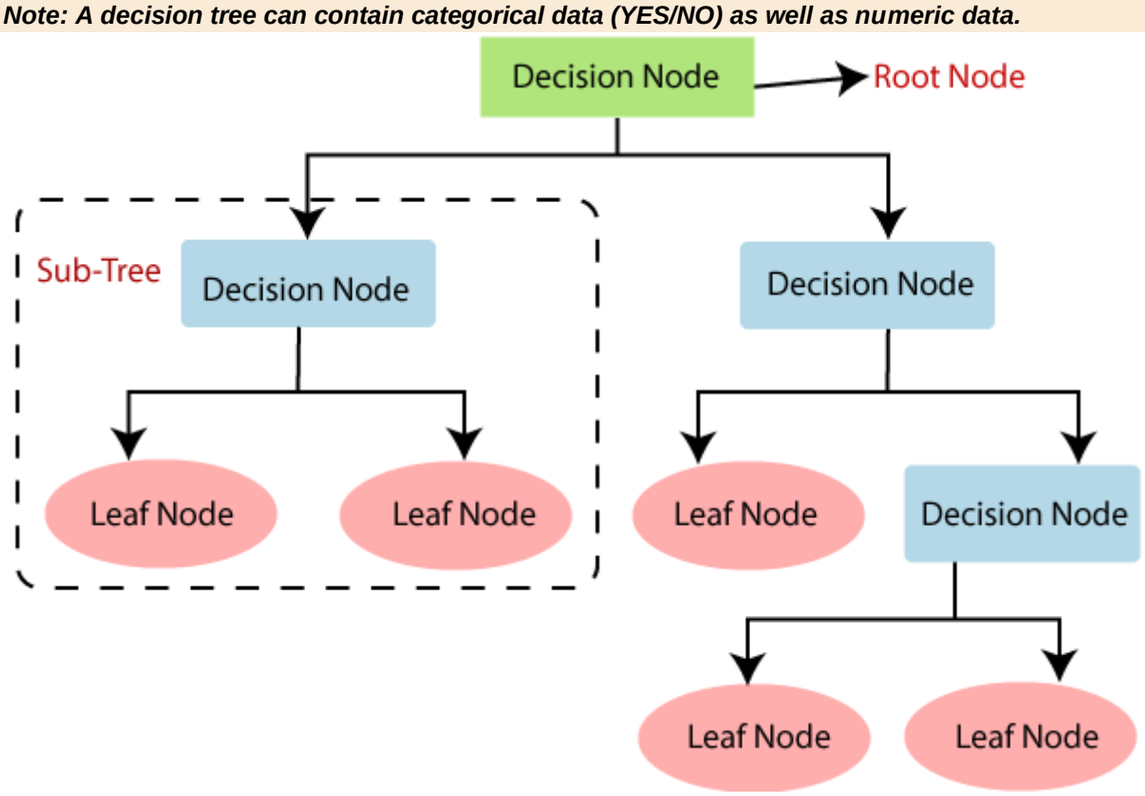
- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training sample

REFERENCE: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

DECISION TREE CLASSIFICATION ALGORITHM

- Decision Tree is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a treestructured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**
- In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.
- ***It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.***
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- In order to build a tree, we use the **CART algorithm,** which stands for **Classification and Regression Tree algorithm.**
- A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.
- Below diagram explains the general structure of a decision tree:

Note: A decision tree can contain categorical data (YES/NO) as well as numeric data.



Why use Decision Trees?

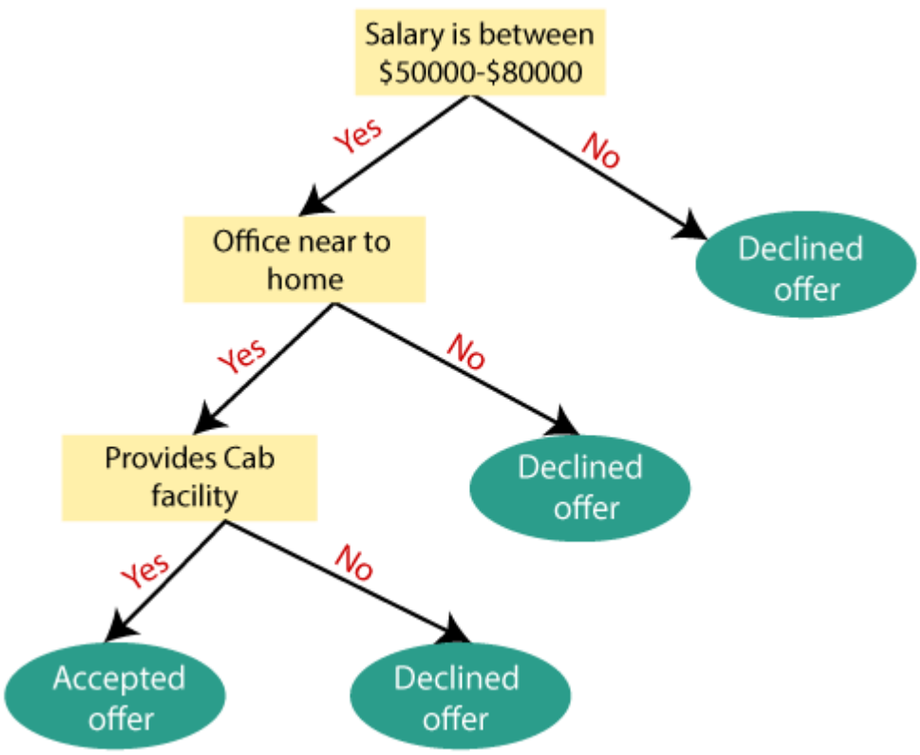
There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree-like structure.

Decision Tree Terminologies

- **Root Node:** Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
- **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
- **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.
- **Branch/Sub Tree:** A tree formed by splitting the tree.
- **Pruning:** Pruning is the process of removing the unwanted branches from the tree.
- **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

Example: Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not. So, to solve this problem, the decision tree starts with the root node (Salary attribute by ASM). The root node splits further into the next decision node (distance from the office) and one leaf node based on the corresponding labels. The next decision node further gets split into one decision node (Cab facility) and one leaf node. Finally, the decision node splits into two leaf nodes (Accepted offers and Declined offer). Consider the below diagram:



Advantages of the Decision Tree

- It is simple to understand as it follows the same process which a human follow while making any decision in real-life.

- It can be very useful for solving decision-related problems. ○ It helps to think about all the possible outcomes for a problem. ○ There is less requirement of data cleaning compared to other algorithms.

Disadvantages of the Decision Tree

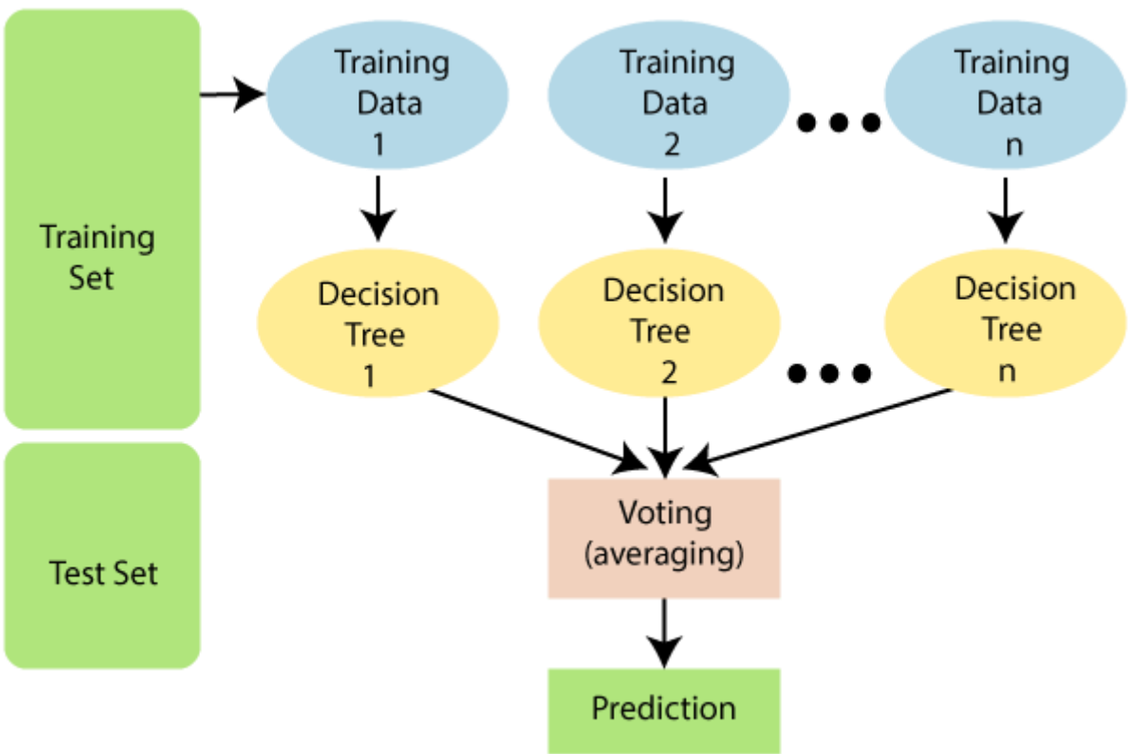
- The decision tree contains lots of layers, which makes it complex. ○ It may have an overfitting issue, which can be resolved using the **Random Forest algorithm.**
- For more class labels, the computational complexity of the decision tree may increase.

REFERENCE: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm
RANDOM FOREST ALGORITHM

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.*

As the name suggests, ***"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."*** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



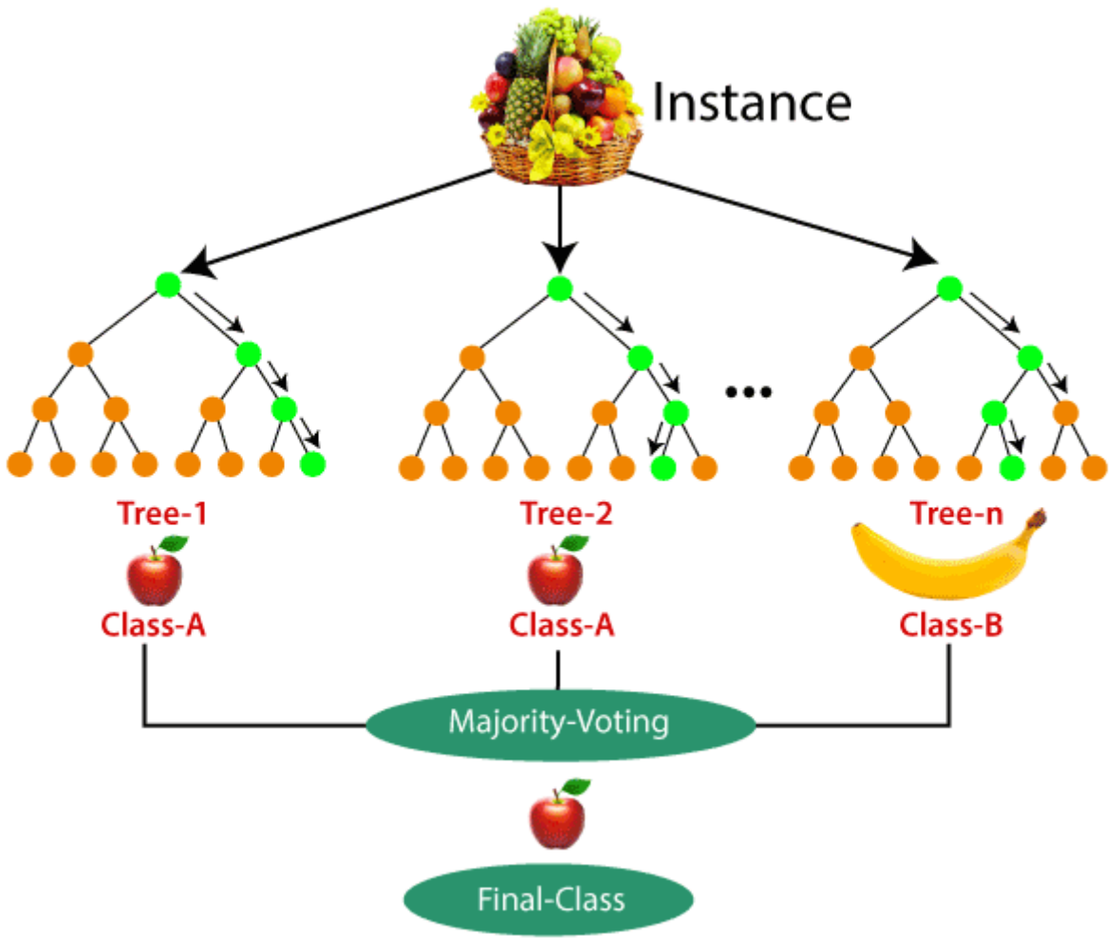
Why use Random Forest?

Below are some points that explain why we should use the Random Forest algorithm:

<=”” li=””> ○ It takes less training time as compared to other algorithms. ○ It predicts output with high accuracy, even for the large dataset it runs efficiently.

○ It can also maintain accuracy when a large proportion of data is missing.

Example: Suppose there is a dataset that contains multiple fruit images. So, this dataset is given to the Random forest classifier. The dataset is divided into subsets and given to each decision tree. During the training phase, each decision tree produces a prediction result, and when a new data point occurs, then based on the majority of results, the Random Forest classifier predicts the final decision. Consider the below image:



Applications of Random Forest

There are mainly four sectors where Random forest mostly used:

1. **Banking:** Banking sector mostly uses this algorithm for the identification of loan risk.
2. **Medicine:** With the help of this algorithm, disease trends and risks of the disease can be identified.
3. **Land Use:** We can identify the areas of similar land use by this algorithm.
4. **Marketing:** Marketing trends can be identified using this algorithm.

Advantages of Random Forest

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue.

Disadvantages of Random Forest

- Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

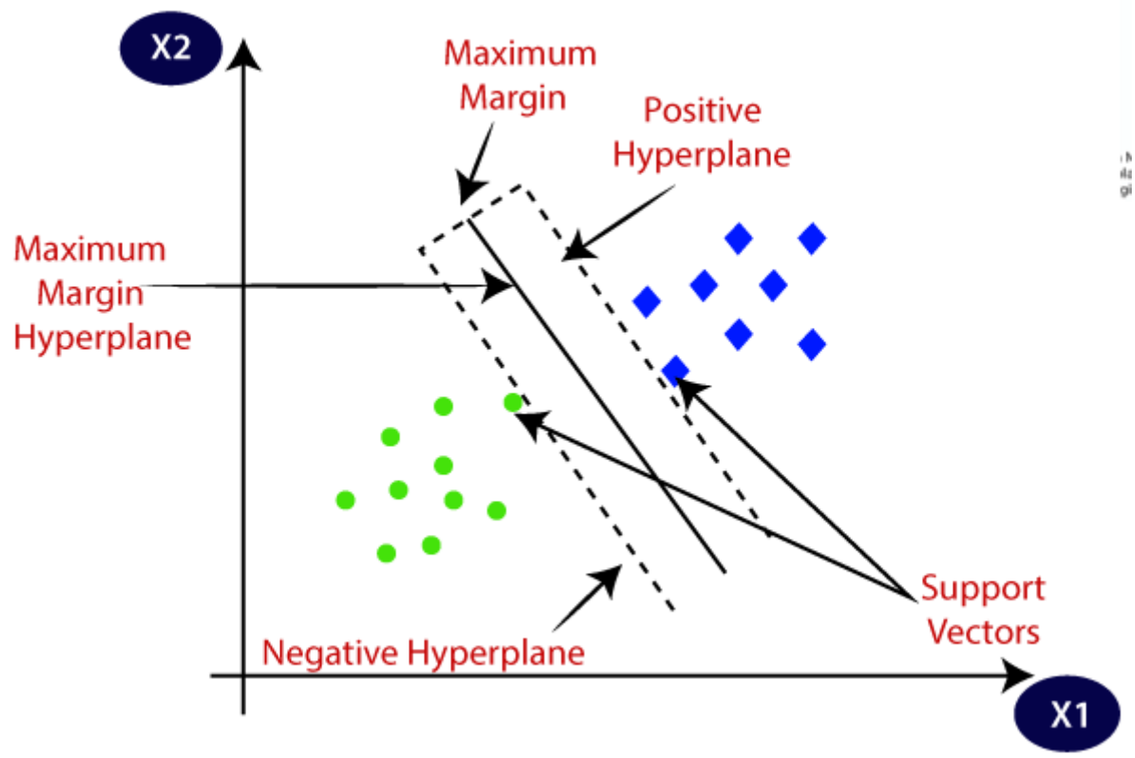
REFERENCE: https://www.javatpoint.com/machine-learning-random-forest-algorithm

SUPPORT VECTOR MACHINE ALGORITHM

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate ndimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



SVM can be of two types:

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

Hyperplane and Support Vectors in the SVM algorithm:

Hyperplane: There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

REFERENCE: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

NEURAL NETWORKS

Neural networks in machine learning refer to a set of algorithms designed to help machines recognize patterns without being explicitly programmed. They consist of a group of interconnected nodes. These nodes represent the neurons of the biological brain.

The basic neural network consists of:

- The input layer
- The hidden layer
- The output layer

Neural networks in machine learning use mathematical or computational models to process information. These neural networks are usually non-linear, which allows them to model complex relationships between data input and output and find patterns in a dataset.

The application of neural networks in machine learning tends to take one of these three broad categories:

- Classification whereby a neural network can recognize patterns and sequences
- Functional approximation and regression analysis
- Data processing including clustering and filtering data

Using neural networks for machine learning has some advantages including:

- They store information on the entire network, meaning that the neural network can continue functioning even if some information is lost from one part of the neural network.
- Once neural networks are trained with a quality data set, they save on costs and time as they take a shorter time to analyze data and present results. They are also less prone to errors, especially if they are trained with high-quality data.
- Neural networks provide quality and accuracy in results

What Type of Algorithm is a Neural Network?

Several types of neural networks exist today. These neural networks are classified based on their density, layers, structure, data flow, and depth activation filters among other features. We are going to focus on three types of neural networks.

- **Convolutional neural network (CNN)**

A convolutional neural network (CNN) is a deep learning algorithm specifically designed to process image data. Convoluted neural networks are used in image recognition and processing.

The neural networks in a CNN are arranged similarly to the frontal lobe of the human brain, a part of the brain responsible for processing visual stimuli.

The convolutional neural network consists of:

A convolutional layer,

A pooling layer,

A fully connected input layer,

A fully connected layer, and

A fully connected output layer

Unlike multi-level perceptrons, CN is sophisticated enough to apply filters that capture the spatial and temporal dependencies in an image. This way, CNN can provide higher accuracy, even with high-resolution images.

The applications of convolutional neural networks include in

Image recognition

Video recognition

Image classification

Medical image analysis

Image segmentation

Natural language processing (NLP)

Recommender systems

Advantages of a Convolutional Neural Network

The advantages of convolutional neural networks include:

- They can detect important features without human supervision
- It has the highest accuracy amongst image detection algorithms
- It is easy to understand and implement

- **Recurrent neural network (RNN)**

A recurrent neural network (RNN) is an artificial neural network that uses sequential or time-series data to solve problems in speech recognition and language translation. RNNs have been used in:

Language translation

Natural language processing

Speech recognition

Image captioning

Due to their relevance in speech-related tasks, RNNs have been used in applications such as:

Google translate

Siri

Speech synthesis

Voice search

Brain computer-interfaces

Handwriting recognition

Music composition

- **Deep Neural Network (DNN)**

A deep neural network (DNN) is an artificial neural network consisting of multiple layers between the input and output layers. These layers could be recurrent neural network layers or convolutional layers making DNN's a more sophisticated machine learning algorithm. DNNs are capable of recognizing sound, creative thinking, recognizing voice commands, and analysis.

How does it work

DNN is a type of machine learning algorithm that learns through repetitive action from many samples. When you feed a computer with a piece of information, the DNN sorts the data based on its elements, for example, the pitch of a sound.

The data is passed through successive layers until it can accurately determine the type of sound made in the data. The model then receives feedback on the correct answer which strengthens its learning process.

Advantages of Deep Neural Network (DNN)

Deep Neural networks also have its own benefits of use:

- DNNs are capable of learning the nonlinear mapping between inputs and outputs, and the underlying structure of the input data vectors
- DNNs are capable of self-learning
- They are scalable

Examples

Deep neural networks have been used in applications such as self-driving cars, smartphones, drones, and

games.

REFERENCE: https://omdena.com/blog/types-of-neural-network-algorithms-in-machinelearning/#:~:text=Neural%20networks%20in%20machine%20learning%20refer%20to%20a%20set%20of, neurons%20of%20the%20biological%20brain.