MODELING AND ANALYSIS.

Models defined as an abstract representation of reality.

Data modeling refers to the architecture that allows data analysis to use data in decision-making processes.

Data modeling is the architecture that makes analysis possible.

Data modeling illustrates relationships between data types and finds ways to group and organize data by establishing formats and attributes.

"A data model can be compared to a roadmap, an architect's blueprint or any formal diagram that facilitates a deeper understanding of what is being designed,"

Companies must build models around business needs, translate business needs into data structures, create concrete database designs and be ready to evolve as businesses change.

Types of data models

Conceptual Data Model

The conceptual data model describes the database at a very high level and is useful to understand the needs or requirements of the database helps to provide a clear and concise overview of the data objects, their relationships, and the business rules that define them.

Example is the entity/relationship model (ER model). The E/R model specializes in entities, relationships, and even attributes that are used by database designers.

Logical Data Model

A logical data model is a representation of data objects, their relationships, and their attributes independent of technology or physical implementation. It is an abstract view of the data, focusing on how the data is organized and how it relates to other data.

The main goal of a logical data model is to capture the essential data elements and logical structures needed for the organization's business operations. This model helps in understanding the business processes and provides a clear understanding of data sources, data requirements, and data flow.

A logical data model is created by first gathering the requirements from stakeholders and subject matter experts. These requirements are then translated into a conceptual model, which is a high-level representation of the business entities and their relationships.

Physical Data Model

A physical data model refers to an actual implementation of a database system. It details the specific structures and characteristics of the database schema that will be used to store, manage, and retrieve data. A physical data model is the technical blueprint of the database and includes information such as the table names, column names, data types, primary and foreign key constraints, indexes, and other implementation-specific details.

The physical data model is created by mapping the logical data model onto the target DBMS. This involves determining how entities and relationships from the logical data model will be represented as tables, columns, and constraints in the physical data model. The physical data model also takes into account the performance and scalability requirements of the system.

DATA ANALYSIS

Data analysis is about using data and information to drive business decisions,

Types of data analysis

Common data analysis approaches include:

- **Statistical analysis:** The process of collecting large volumes of data and using statistics and data analysis techniques to identify trends, patterns and insights.
- **Inferential analysis:** A subtype of statistical analysis that generates conclusions about a large group by analyzing data from smaller data samples of that group.
- **Diagnostic analysis:** An analytical process that focuses on why things happen and seeks to identify the root causes by analyzing data and identifying patterns, trends and correlations between variables.
- **Data mining:** The practice of scanning through large data sets to identify patterns and relationships to find solutions to specific problems.
- **Predictive analysis:** Uses specific data, known as features, to predict future trends and events. Predictive analytics tools leverage machine learning and AI technology to drive complex predictive analysis algorithms.
- **Prescriptive analysis:** A type of data analytics and data mining that uses historical data to recommend the best course of action to achieve a desired outcome.

N.B A data model that is well-designed is the foundation to creating business intelligence and data warehouse applications that result to a significant business value.

It is the key to success in Business Intelligence (BI).

Effective data modeling results in transforming data into an enterprise information resource that is rational, far-reaching and present. Data is transformed from operational or source systems into a data for analysis.

BUSINESS INTELLIGENCE(BI)

(BI) is a technology-driven process for analyzing data. It is used to measure performance progress toward business goals, perform quantitative analysis, report and share data, identify customer insights, and much more.

Business Intelligence enables businesses to organize, analyze and contextualize business data from all around the company.

Business intelligence defined as a set of mathematical models and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision making processes

Business intelligence (BI) refers to capabilities that enable organizations to make better decisions, take informed actions, and implement more-efficient business processes.

Business intelligence combines business analytics, data mining, data visualization, data tools and infrastructure, and best practices to help organizations make more data-driven decisions.

BI capabilities allow the following

- Collect up-to-date data from your organization
- Present the data in easy-to-understand formats (such as tables and graphs)
- Deliver data in a timely fashion to the employees in your organization

Business Intelligence Applications

BI for the Finance Industry

Business intelligence techniques can help firms spot emerging trends and patterns, monitor the effectiveness of their products and services, mitigate the risk of unknowns, and build new investment strategies.

Banks, and firms in the financial services sector can automate the end-end loan application and processing process. Using relevant BI apps, customer data can be collected from different data sources, stored, and made accessible to the relevant stakeholders. Doing so prevents silos and enhances transparency in the process.

Similarly, financial firms can leverage Business Intelligence applications to mitigate the risks involved in many of their products and services. Based on customer transactions' history, earning capacity, and financial holdings, firms can estimate the risk of granting loans or other debt instruments like credit cards.

BI for Manufacturing

Business Intelligence apps help stakeholders gain a complete picture of their supply chain. The analytics derived from such tools help businesses make reliable forecasts and manage inventory. Doing so reduces the lead time and enhances the capability of the firm to meet just-in-time demands.

BI tools help monitor the downtime of employees and the machinery and equipment in a manufacturing firm. The data helps them undertake predictive maintenance and ensure continuous, uninterrupted operation. Based on the data available, firms can monitor the manufacturing cycle of a product and work out ways to reduce the length of the cycle.

BI for Hospitality

Hospitality businesses deal with a wide range of customers, and retaining them depends on the degree of personalization you offer in the services. Businesses in the hospitality industry depend on business intelligence to make reliable forecasts about occupancy rates. Such forecasts are made while considering several attributes like seasonal changes, trends, and patterns, and so they help businesses maximize their revenue.

BI for Retail

Retail firms can use Business Intelligence apps to collect and access data about demographics, sales, purchase history, etc. By processing this information, they can make reliable forecasts about customer demand and manage inventory accordingly. Also, businesses can leverage customer behavior information to personalize their offerings.

BI can help retail businesses track their marketing campaign’s performance and make necessary changes. They can leverage the features of BI applications to launch and manage stores in multiple locations.

Sales forecasting

Businesses use business intelligence to forecast sales performancecbetter planning resources, inventory levels, and prices. The insights gained from analyzing large amounts of historic internal or market data can help inform more intelligent decision-making.

Market analysis

Business intelligence provides an efficient way for companies to gain insight into their target markets and understand consumer preferences more accurately. With BI tools, companies can analyze current trends, competitor activity, and customer feedback to understand their market better.

Customer relationship management

Business intelligence solutions can help companies build stronger customer relationships by providing insights into customer behavior. Companies can use data from surveys and other sources to get an idea of what customers want or need from their products or services, allowing them to adjust their offerings accordingly.

Business intelligence architectures/ components.

The main components of business intelligence are

i. Data warehouse,
ii. Business analytics
iii. Technology

Data warehouse holds data obtained from internal sources as well as external sources. The internal sources include various operational systems.

Business analytics creates a report as and when required through queries and rules i.e Data mining an important aspect of business analytics.

Analytics: This is where businesses and enterprises turn their data into insights. By using data analytics, businesses can gain a better understanding of their customers, their operations, and the trends that are shaping their industry.

Technology: This is the software and hardware that organizations use to support their BI efforts. This can include everything from data warehouses and ETL tools to reporting and visualization platforms.

Business intelligence benefits

1. Improve data accuracy
2. Make better decisions more quickly
3. Improve mission-critical outcomes
4. Share data across business functional areas
5. Gain better visibility into financial and operational information
6. Identify and reduce inefficiencies
7. Eliminate waste, fraud, and abuse
8. Improve productivity
9. Boost return on investment, while cutting total cost of ownership
10. Enhance transparency and service at all levels

Data Mining & Business Analytics.

Data mining discovering interesting patterns of data from data bases.

Data mining, also known as knowledge discovery is the process of uncovering patterns and other valuable information from large data sets.

Data mining is the process of searching and analyzing a large batch of raw data in order to identify patterns and extract useful information.

Companies use data mining software to learn more about their customers. It can help them to develop more effective marketing strategies, increase sales, and decrease costs.

Data mining relies on effective data collection, warehousing, and computer processing.

Data Mining Techniques

Data mining uses algorithms to convert large collections of data into useful output.

Common types of data mining techniques include association rules, classification, clustering, decision trees, K-Nearest Neighbor, neural networks, and predictive analysis.

i) Association rules, search for relationships between variables.

Association rule mining finds interesting associations and relationships among large sets of data items. T

his rule shows how frequently a item set occurs in a transaction.

Association rule mining is a technique used to identify patterns in large data sets.

It involves finding relationships between variables in the data and using those relationships to make predictions or decisions.

The goal of association rule mining is to uncover rules that describe the relationships between different items in the data set.

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases.

For example, consider a dataset of transactions at a grocery store. Association rule mining could be used to identify relationships between items that are frequently purchased together. For example, the rule "If a customer buys bread, they are also likely to buy milk" is an association rule that could be mined from this data set.

We can use such rules to inform decisions about store layout, product placement, and marketing efforts.

Use Cases of Association Rule Mining:

Market Basket Analysis

This involves analyzing the items customers purchase together to understand their purchasing habits and preferences.

For example, a retailer might use association rule mining to discover that customers who purchase bread are also likely to purchase Milk. We can use this information to optimize product placements and promotions to increase sales.

Customer Segmentation

Association rule mining can also be used to segment customers based on their purchasing habits.

For example, a company might use association rule mining to discover that customers who purchase certain types of products are more likely to be younger.

Fraud Detection

You can also use association rule mining to detect fraudulent activity. For example, a credit card company might use association rule mining to identify patterns of fraudulent transactions, such as multiple purchases from the same merchant within a short period of time.

Association Rule Mining Algorithms

Apriori Algorithm

It is designed to work on the databases that contain transactions. This algorithm uses a breadth-first search and Hash Tree to calculate the item set efficiently.

It is mainly used for market basket analysis and helps to understand the products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.

Eclat Algorithm

Eclat algorithm (**Equivalence Class Transformation)** This algorithm uses a depth-first search technique to find frequent item sets in a transaction database.

It performs faster execution than Apriori Algorithm.

F-P Growth Algorithm

The F-P growth algorithm stands for **Frequent Pattern**, and it is the improved version of the Apriori Algorithm. It represents the database in the form of a tree structure that is known as a frequent pattern or tree. The purpose of this frequent tree is to extract the most frequent patterns.

ii)Classification

Classification is a task in data mining that involves assigning a class label to each instance in a dataset based on its features.

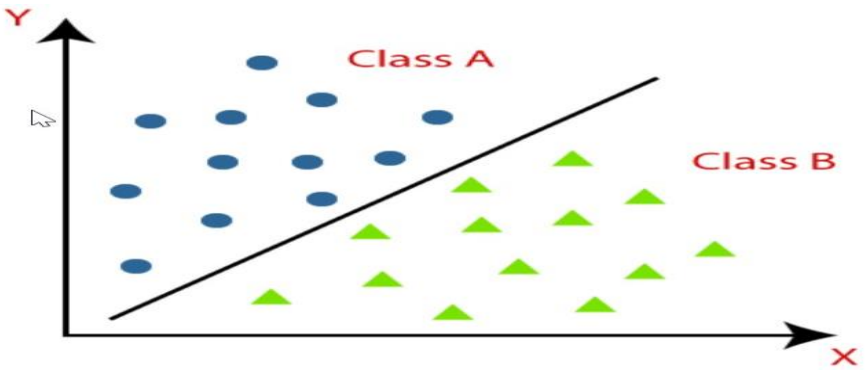
The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data.

Classification predicts the category the data belongs to.

Unlike regression, the output variable of Classification is a category, not a value, such as "Green or Blue", "fruit or animal", etc. Since the Classification algorithm is a Supervised learning technique, hence it takes labeled input data, which means it contains input with the corresponding output.

The goal of classification is to build a model that accurately predicts the class labels of new instances based on their features.

Example In the below diagram, there are two classes, class A and Class B. These classes have features that are similar to each other and dissimilar to other classes.



Examples of ML Classification Algorithms:

- **Linear Models**
  - Logistic Regression
  - Support Vector Machines
- **Non-linear Models**
  - K-Nearest Neighbours
  - Naïve Bayes
  - Decision Tree Classification

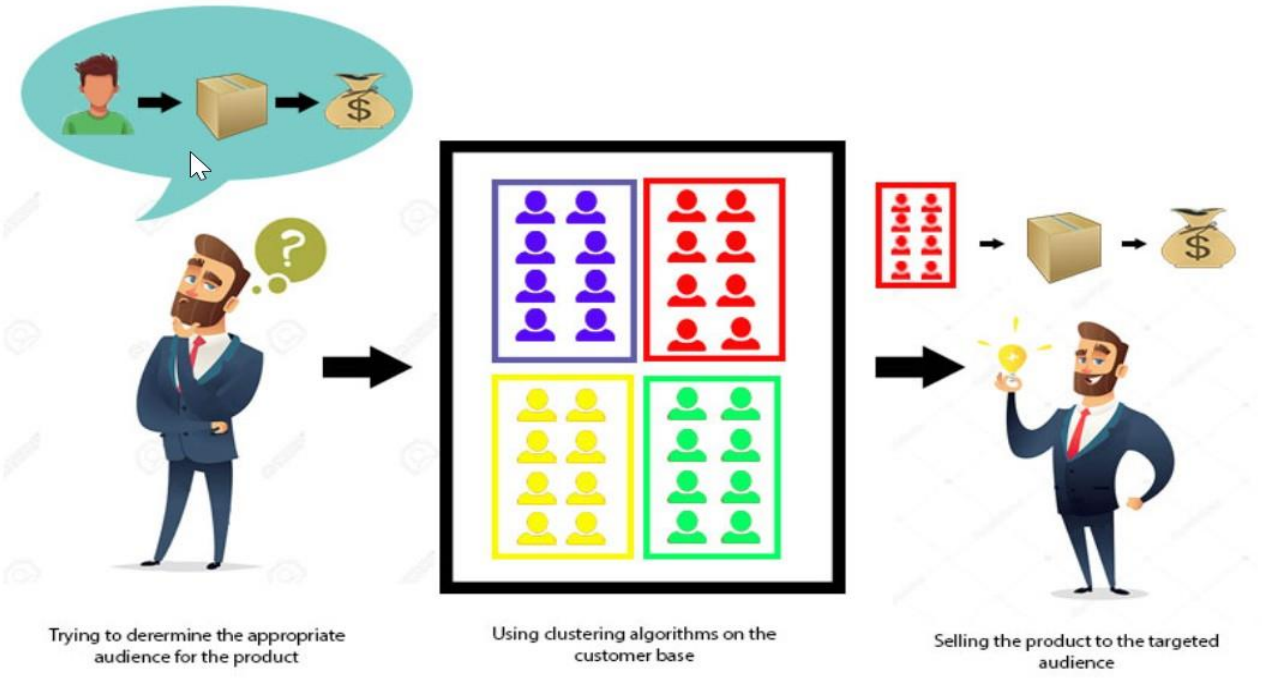
- Random Forest Classification

iii) Clustering is similar to classification. However, clustering identifies similarities between objects, then groups those items based on what makes them different from other items.

Clustering is the process of dividing the entire data into groups (also known as clusters) based on the patterns in the data.

Clustering algorithms in data mining will help to split data into several subsets. Each subset has data like one another.

Example, A market manager with a new product to sell. The product would bring enormous profit, as long as it is sold to the right people. So, how can the manager tell who is best suited for the product from the company's huge customer base?



Applications of Data Mining Cluster Analysis

Customer Segmentation
Used to group customers with similar behavior, preferences, and purchasing patterns to create more targeted marketing campaigns.

Image Segmentation
Used to segment images into different regions based on their pixel values, which can be useful for tasks such as object recognition and image compression.

Anomaly Detection
Used to identify outliers or anomalies in datasets that deviate significantly from normal behavior.

Text Mining
Used to group documents or texts with similar content, which can be useful for tasks such as document summarization and topic modeling.

Biological Data Analysis
Used to group genes or proteins with similar characteristics or expression patterns, which can be useful for tasks such as drug discovery and disease diagnosis.

Recommender Systems
Used to group users with similar interests or behavior to create more personalized recommendations for products or services.

Clustering Algorithm

K-means Clustering. is the most commonly used clustering algorithm. It's a centroid-based algorithm and the simplest unsupervised learning algorithm.

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters where K defines the number of pre-defined clusters that need to be created, if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

vi) Neural networks

Machine learning deep learning algorithm inspired by the human brain.

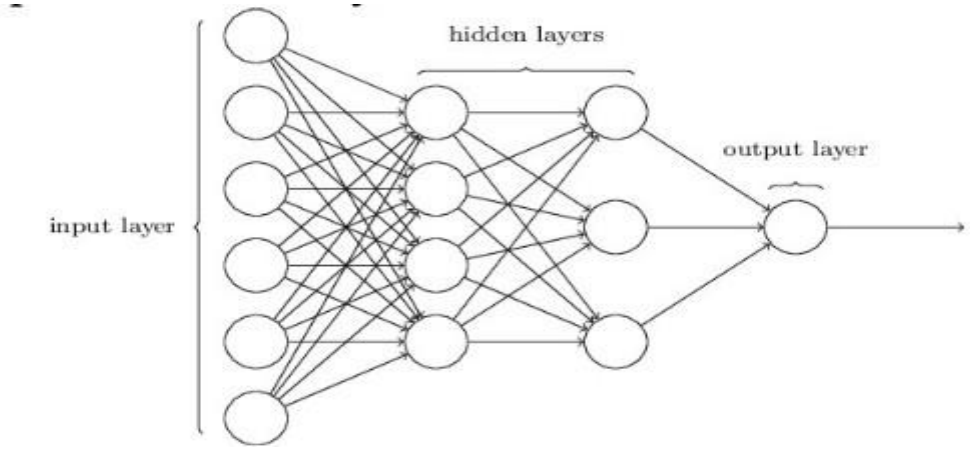
The neural networks consist of interconnected nodes or neurons that process and learn from data, enabling tasks such as pattern recognition and decision making in machine learning.

A neural network consists of multiple layers called the input layer, output layer, and hidden layers.

In each layer every node (neuron) is connected to all nodes (neurons) in the next layer with parameters called ‘weights’. .

- Input Layer: This layer receives the initial data or features. Each neuron in the input layer represents a feature or input variable.
- Hidden Layers: The hidden layers enable the network to learn complex patterns and representations in the data. Deep neural networks have multiple hidden layers, contributing to the term "deep learning."

- Output Layer: The layer produces the final result or prediction. The number of neurons in the output layer is determined by the nature of the task classification, regression, etc.



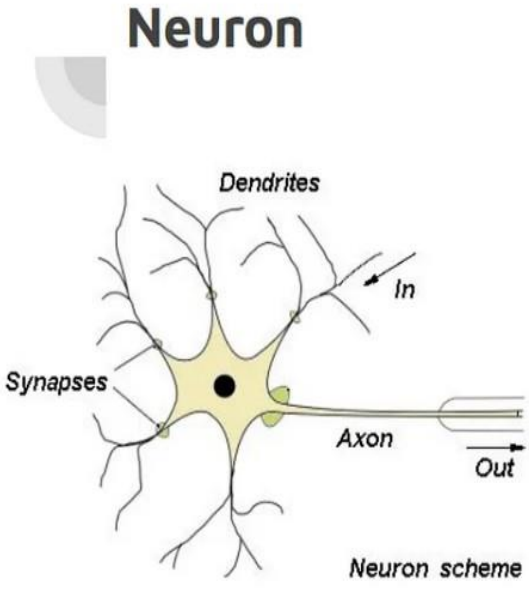
Basics of neural network

A neuron is the fundamental cellular unit of the nervous system.

Human brain consists of multiple connected neurons forming a neural network; similarly a network of artificial neurons called **perceptrons** form a Deep neural network.

That means,

- Multiple "**Neurons**" (human brain neurons) form a **Brain** of a neural network.
- Multiple "**Perceptrons"** (artificial neurons) from a **Deep neural network**.



Biological	Artificial
Dendrites	Inputs
Cell Nucleus	Nodes
Synapses	Weights
Axon	Outputs

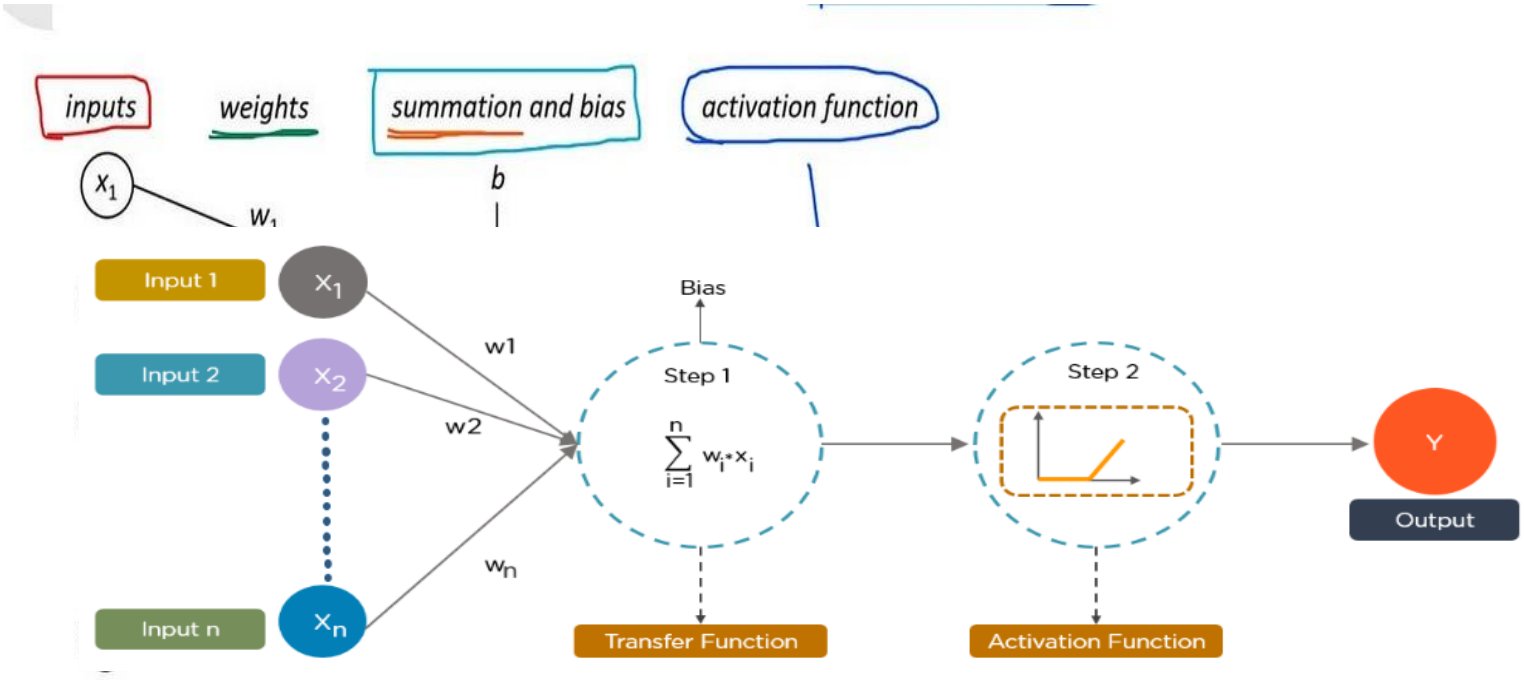
A Perceptron Architecture - network of artificial neurons

A neural network models a neuron with weighted inputs, summing and transforming them to produce an output.

It has two functions:

- Summation
- Transformation (Activation Function)

N.B: A perceptron consists of inputs, weights, biases, **and an** activation function**.**



Weight

The weight shows the **effectiveness of a particular input**. The more the weight of input, the more it will have an impact on the neural network.

Bias

Bias is an additional parameter in the Perceptron which is used to adjust the output along with the weighted sum of the inputs to the neuron which helps the model in a way that it can fit best for the given data.

Bias is the error that is introduced by the model's prediction and the actual data.

Bias = Predicted -Actual

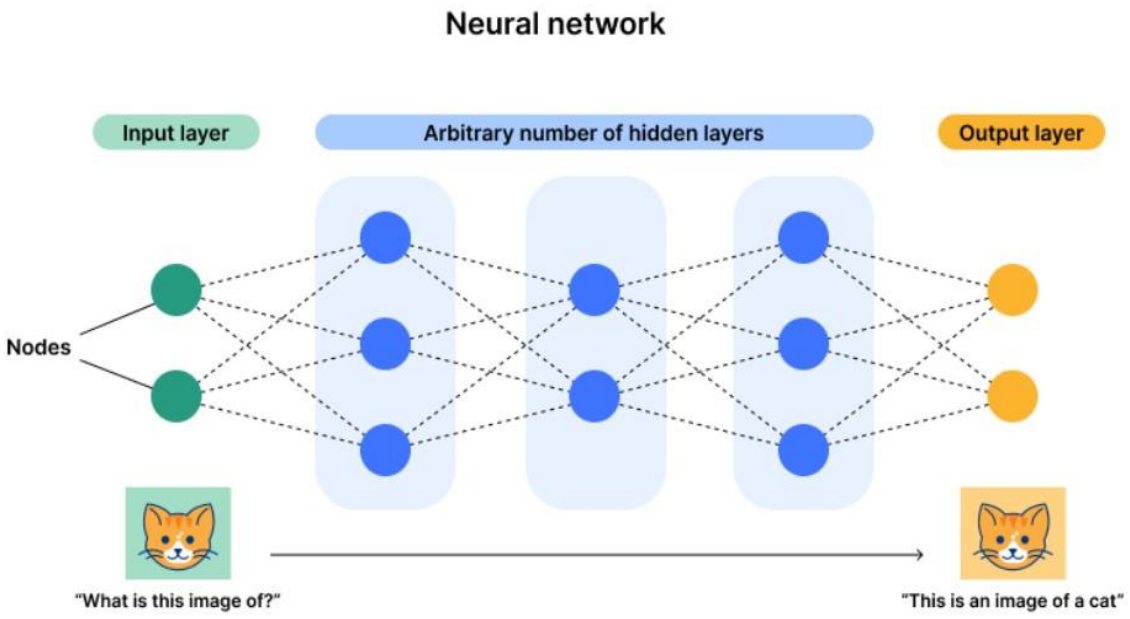
Activation Function

An activation function is a function that is added to an artificial neural network **in order to help the network learn complex patterns in the data**.

Activation functions **add non-linearity** to the model.

Neural networks are used in deep learning to draw conclusions from unlabeled data without human intervention.

Example a deep learning model built on a neural network could be able to identify items in a photo it has never seen before.



How neural network learns

Neural networks consist of interconnected layers of nodes, or "neurons", which transmit signals to each other.

Each connection between neurons has a weight, which determines the strength of the signal that is passed on.

The goal of training a neural network is to adjust these weights in a way that allows the network to accurately map inputs to outputs.

Neural networks learn by adjusting weights based on error minimization using two techniques

i. Forward propagation.
ii. Backward propagation

Forward Propagation

Forward propagation is the initial phase of data processing in a neural network.

The input data is fed into the network and passed through various layers. Each neuron in these layers processes the input and passes it to the next layer, ultimately leading to the output layer. This process is linear and straightforward, moving in one direction: from input to output.

Back Propagation

Back propagation is a learning phase where once the forward propagation is complete and an output is produced, the network compares the output to the desired outcome.

The difference, or error, is then used to adjust the network's weights and biases.

Bias = Predicted -Actual

This process is iterative and involves moving backward through the network, fine-tuning it to minimize the error.

Pseudo code / steps

The training process begins with the initialisation of the weights, which is often done randomly. The network is then presented with a set of training data, which includes both the inputs and the correct outputs.

Through forward propagation the network makes a prediction based on the inputs and the current weights, and the difference between this prediction and the correct output is calculated. This difference is known as the error (Bias).

The next step is to adjust the weights in order to minimise this error done using a technique called back propagation.

In **back propagation,** the error is propagated backwards through the network, from the output layer to the input layer.

The weights are adjusted proportionally to their contribution to the error. This process is repeated for each piece of training data, in a cycle known as an epoch.

The number of epochs and the rate at which the weights are adjusted, known as the learning rate, are key parameters in the training process.

Too few epochs or a low learning rate can result in under fitting, where the network fails to learn the underlying patterns in the data.

Simillary, too many epochs or a high learning rate can lead to over fitting, where the network becomes too specialised to the training data and performs poorly on new data.

Neural networks learn from data by iteratively adjusting their weights based on the error they make in predicting the correct output.

Underfitting,: When a model has not learned the patterns in the training data well and is unable to generalize well on the new data

A machine learning algorithm is said to have underfitting when a model is too simple to capture data complexities.

It represents the inability of the model to learn the training data effectively result in poor performance both on the training and testing data.

An underfit model has poor performance on the training data and will result in unreliable predictions.

To address underfitting problem of the model, we need to use more complex models, with enhanced feature representation, and less regularization.

Overfitting.

Overfitting occurs when the model cannot generalize and fits too closely to the training dataset instead.

N.B: An Over fit models don't generalize, which is the ability to apply knowledge to different situations.

Over fitting happens due to several reasons, such as:

- The training data size is too small and does not contain enough data samples to accurately represent all possible input data values.
- The training data contains large amounts of irrelevant information, called noisy data.
- The model trains for too long on a single sample set of data.
- The model complexity is high, so it learns the noise within the training data.

Case example 1

Consider a use case where a machine learning model has to analyze photos and identify the ones that contain dogs in them. If the machine learning model was trained on a data set that contained

majority photos showing dogs outside in parks , it may learn to use grass as a feature for classification, and may not recognize a dog inside a room.

Case example 2

Suppose the model learns the training dataset, like the Y student. They perform very well on the seen dataset but perform badly on unseen data or unknown instances. In such cases, the model is said to be Overfitting.

Case example 3

Consider a machine learning algorithm that predicts a university student's academic performance and graduation outcome by analyzing several factors like family income, past academic performance, and academic qualifications of parents. However, the test data only includes candidates from a specific gender or ethnic group. In this case, overfitting causes the algorithm's prediction accuracy to drop for candidates with gender or ethnicity outside of the test dataset

How to detect over fitting

Test the machine learning models on more data with comprehensive representation of possible input data values and types.

A high error rate in the testing data indicates overfitting. A method of testing for overfitting is

K fold cross-validation.

Cross-validation is a testing methods used in practice where data scientists divide the training set into K equally sized subsets or sample sets called folds. The training process consists of a series of iterations.

During each iteration, the steps are:

- Keep one subset as the validation data and train the machine learning model on the remaining K-1 subsets.
- Observe how the model performs on the validation sample.
- Score model performance based on output data quality.

Techniques to Reduce Overfitting

1. Improving the quality of training data reduces overfitting by focusing on meaningful patterns, mitigate the risk of fitting the noise or irrelevant features.

2. Increase the training data can improve the model’s ability to generalize to unseen data and reduce the likelihood of overfitting.

3. Reduce model complexity.

Bias and Variance in Machine Learning decision making models

Bias refers to the error due to overly simplistic assumptions in the learning algorithm.

These assumptions make the model easier to comprehend and learn but might not capture the underlying complexities of the data.

It is the error due to the model's inability to represent the true relationship between input and output accurately.

When a model has poor performance both on the training and testing data means high bias because of the simple model, indicating underfitting

Variance, is the error due to the model's sensitivity to fluctuations in the training data.

It's the variability of the model's predictions for different instances of training data.

High variance occurs when a model learns the training data's noise and random fluctuations rather than the underlying pattern.

As a result, the model performs well on the training data but poorly on the testing data, indicating overfitting.

Further research:- Research on Tensor flow

TensorFlow a free and open-source software library for machine learning and artificial intelligence has a particular focus on training and inference of deep neural networks

vii)Predictive analytics

Predictive analytics looks for past patterns to measure the likelihood that those patterns will reoccur.

It draws on a series of techniques to make these determinations, including (AI), data mining machine learning, modeling, and statistics.

Algorithms that make predictions about future outcomes using historical data combined with statistical modeling, data mining techniques and machine learning.

Models created to evaluate past data, uncover patterns, and analyze trends to forecast future trends.

Predictive Analytics models / Algorithms

Linear regression – A model is used to predict a continuous dependent variable based on one or more independent variables.
Decision trees – A model that use a tree-like structure to represent decisions and their possible consequences.
Random Forest – A model that combines multiple decision trees to improve accuracy and avoid overfitting.

Support Vector Machines (SVM) – A Supervised Learning algorithms used for Classification as well as Regression problems.

Neural networks – These models are inspired by the structure and function of the human brain and can be used for both classification and regression.

Naive Bayes Supervised machine learning algorithm that is used for classification tasks such as text classification. They use principles of probability to perform classification tasks.

Predictive Analytics Case applications

Bank industry. Credit scoring makes extensive use of predictive analytics.

When a consumer or business applies for credit, data on the applicant's credit history and the credit record of borrowers with similar characteristics are used to predict the risk that the applicant might fail to repay any new credit that is approved.

Manufacturing. Forecasting is essential in manufacturing to optimize the use of resources in a supply chain.

Insurance industry Underwriting

Insurance companies examine applications for new policies to determine the likelihood of having to pay out for a future claim.

The analysis is based on the current risk pool of similar policyholders as well as past events that have resulted in payouts.

Marketing

Marketing professionals planning a new campaign look at how consumers have reacted to the overall economy. They can use these shifts in demographics to determine if the current mix of products will entice consumers to make a purchase.

Stock Traders

Active traders look at a variety of historical metrics when deciding whether to buy a particular stock or other asset.

Fraud Detection

Financial services use predictive analytics to examine transactions for irregular trends and patterns. The irregularities pinpointed can then be examined as potential signs of fraudulent activity.

This may be done by analyzing activity between bank accounts or analyzing when certain transactions occur.

Predictive analytics in Agriculture

Prediction of crop and its yield, efficiently determines the crop to be sowed in a particular season and also predicts the amount of produce.

Predictive analytics can be used in many steps of the agricultural cycle, from crop selection to harvesting. The use of predictive modeling and analytics can:

- **Select the best crop for your field:** By using soil analysis data, historical weather, and other parameters farmers can make the best crop selection for any given condition.
- **Optimize irrigation** – analytics can aid in predicting crop stress periods, as well as optimal amounts of irrigation needed according to crop growth stages.
- **Optimize land preparation:** GPS-enabled field management maps can be correlated with productivity maps to optimize field operations.
- **Optimize crop protection:** Predictive analytics can help predict outbreaks of pests and crop disease using factors such as soil parameters and ongoing weather conditions.
- **Increase productivity and yields**: Using predictive analytics can build management zones, help optimize crop growth, track season progress, and take measures when needed.
- **Evade lower ROI** – Predictive analytics can Identify fields and subfields where ROI is repeatedly lower, and suggest if these fields should be let out of production.
- **Mitigate supply chain uncertainty:** Unpredictable weather, severe storms, drought, and changing insect behaviors due to weather are all environmental factors that impact the agribusiness supply chain. Using data can assist farmers to prepare farmers for these challenges and making decisions based on sound data.
- **Reduce detrimental environmental effects:** predictive analytics can help understand conditions where environmental pollution risks are high, relate actions to environmental footprint, and help evade them.

The Data Mining Process

The data mining process is usually broken into the following steps.

1: Understand the Business

Domain understanding - Entails identifying the key stakeholders in the research and looking for clarity and understanding of any useful knowledge that may be required. It is at this point that the goals are established.

Example: Understand the goals the company is trying to achieve by mining data? What is their current business situation? What are the findings of a SWOT analysis?

2: Understand the Data

Begins with data collection. This is followed by verifying the data for completeness, redundancy, and missing data. At this point data usefulness in terms of meeting the desired goal is also confirmed.

3: Prepare the Data

This step entails data cleaning and selecting the relevant feature subset. The goal in this step is to achieve a dataset that is suitable for selected methods of data mining
Data is cleaned, standardize and outliers, assessed for mistakes
An outlier can be defined as *a data point that deviates significantly from the normal pattern or behavior of the data*

4: Build the Model

Entails selecting the methods (classification, regression, or clustering) to be used for knowledge generation and applying those methods to the data. The generated data is also tested.

Algorithms used to build data mining models to search for relationships, trends, associations, or sequential patterns.

5: Evaluate the Results

Aassessing the findings of the data model or models. The outcomes from the analysis may be aggregated, interpreted, and presented to decision-makers.

6: Implement Change and Monitor

The data mining process concludes with management taking steps in response to the findings of the analysis. The company may decide the information was not strong enough or the findings were not relevant, or the company may strategically pivot based on findings.

Applications of Data Mining

Sales

Data mining encourages smarter, more efficient use of capital to drive revenue growth. Consider the point-of-sale register at a supermarket. For every sale, POS register collects the time a purchase was made and what products were sold. Using this information, the shop can strategically craft its product line.

Marketing

Businesses can use data mining to understand where its clients see ads, what demographics to target, where to place digital ads, and what marketing strategies most resonate with customers. This includes aligning marketing campaigns, promotional offers, cross-sell offers, and programs to the findings of data mining.

Manufacturing

For companies that produce their own goods, data mining plays an integral part in analyzing how much each raw material costs, what materials are being used most efficiently, how time is spent along the manufacturing process, and what bottlenecks negatively impact the process. Data mining helps ensure the flow of goods is uninterrupted.

Fraud Detection

The heart of data mining is finding patterns, trends, and correlations that link data points together.

Companies can use data mining to identify outliers or correlations that should not exist. For example, a company may analyze its cash flow and find a reoccurring transaction to an unknown account. If this is unexpected, the company may wish to investigate whether funds are being mismanaged.

Human Resources

Human resource departments often have a wide range of data available for processing including data on retention, promotions, salary ranges, company benefits, use of those benefits, and employee satisfaction surveys. Data mining can correlate this data to get a better understanding of why employees leave and what entices new hires.

Customer Service

Customer satisfaction may be caused (or destroyed) by many events or interactions. Imagine a company that ships goods. A customer may be dissatisfied with shipping times, shipping quality,

or communications. The same customer may be frustrated with long telephone wait times or slow e-mail responses. Data mining gathers operational information about customer interactions and summarizes the findings to pinpoint weak points and highlight what the company is doing right.

DSS DEVELOPMENT

Business face a various of problems and therefore, variety of decisions to be taken

Example,

i. Allocation problem,
ii. Classification problem,
iii. Prediction problem,
iv. Problems with huge amount of data (especially for present and future),
v. Network-related problems,
vi. Complex problems with uncertainty, .

Note: These problems can be solved using different techniques.

For example, for allocation problems, Linear Programming (LP) algorithms can be applied;

For prediction problem, regression of different types depending on the type of problem can be applied;

For classification problem, decision tree or some other techniques are applicable;

Problems with huge amount of data can be handled by data warehousing techniques;

For network-related problems, networking methods or methods like Petri net can be applied.

Thus, different kinds of problems demand different kinds of analytical techniques

DECISION SUPPORT TOOLS

If a tool supports decision-making then that tool proves to be a decision support tool.

Examples

- Decision Tree
- Linear programming
- Predicate Logic
- Fuzzy theory & Fuzzy logic
- Network tools
- Markov Chain
- Case based reasoning

- Simulation etc

i)Decision Tree

A decision tree is a graphical tool for decision-making. Here a manager uses graphics to study alternative solutions available.

A decision tree is a graphical representation of the various alternatives available to solve a given problem to determine the most effective course of action.

This technique can be used for many different project management cases.

For example: Should we upgrade the software that we are using in our organization? Should we build a prototype for our new project? Should we select the low-price contractor? Should we select the low-budget project?

A decision tree is a **mathematical model** used to help managers make decisions.

- A decision tree uses **estimates** and **probabilities** to calculate likely outcomes.
- A decision tree helps to decide whether the net gain from a decision is worthwhile.
- Used to calculate the expected monitory value(EMV)

Decision tree analysis implementation steps

1. List all the decisions and prepare a decision tree for a project management situation.
2. Assign the probability of occurrence for all the risks.
3. Assign the impact of a risk as a monetary value.
4. Calculate the Expected Monetary Value (EMV) for each decision path.

The EMV a three-step process:

1. Determine the probability (P) an outcome will occur.
2. Determine the monetary value or impact (I) of the outcome.
3. Multiply P x I to calculate the EMV.

If a scenario presents multiple potential outcomes, calculate the EMV for each potential outcome and add them together to get the overall EMV.

Example 1

As a project manager you need to decide which sub-contractor is appropriate for your projects critical path activities. While selecting a sub-contractor, you should take into consideration the costs and delivery dates.

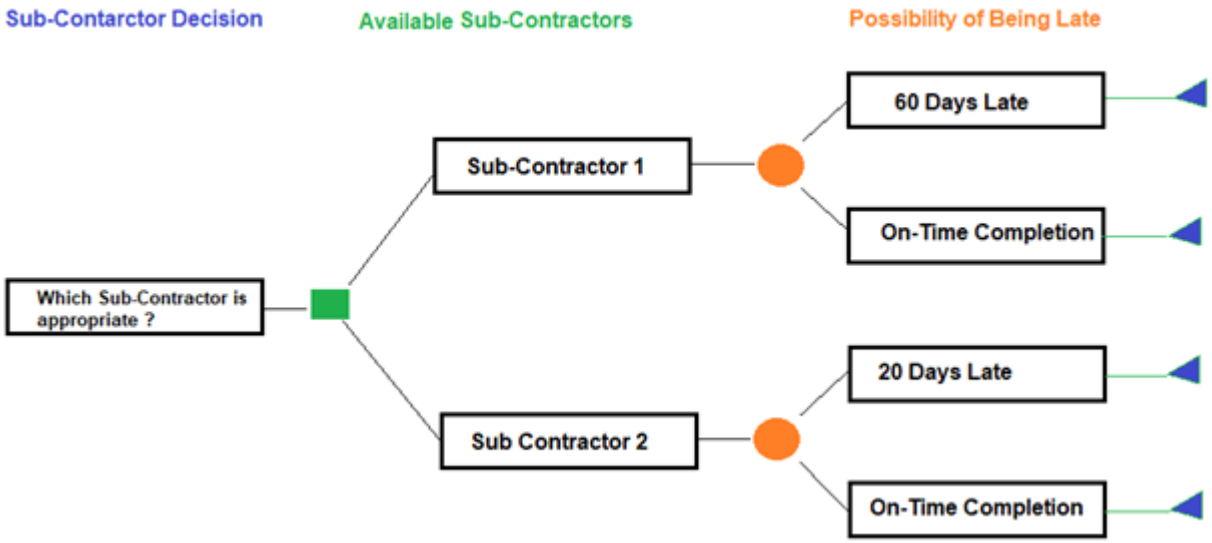
• Sub-contractor 1 bids $250,000. You estimate that there is a 30% possibility of completing 60 days late. As per your contract with the client, you must pay a delay penalty of $5,000 per calendar day for every day you deliver late.

• Sub-contractor 2 bids $320,000. You estimate that there is a 10% possibility of completing 20 days late. As per your contract with the client, you must pay a delay penalty of $5,000 per calendar day for every day you deliver late.

You need to determine which sub-contractor is appropriate for your projects critical path activities. Both sub-contractors promise successful delivery and high-quality work.

Solution

Step 1: List decisions and prepare a decision tree for a project management situation.



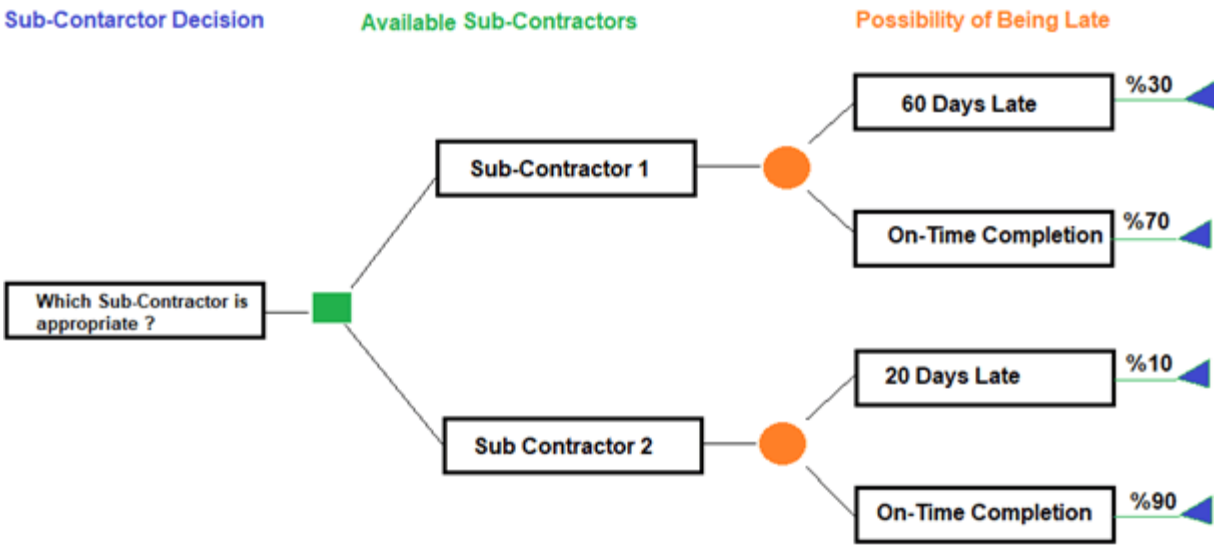
Step 2: Assign the probability of occurrence for the risks.

In this example, the possibility of being late for Sub-contractor 1 is 30% and for Sub-contractor 2 is 10 %. This means that the possibility of completing on-time for Sub-contractor 1 is 70% and for Sub-contractor 2 is 90 %.

In Figure 2 below the probability of occurrence for the risks are assigned.

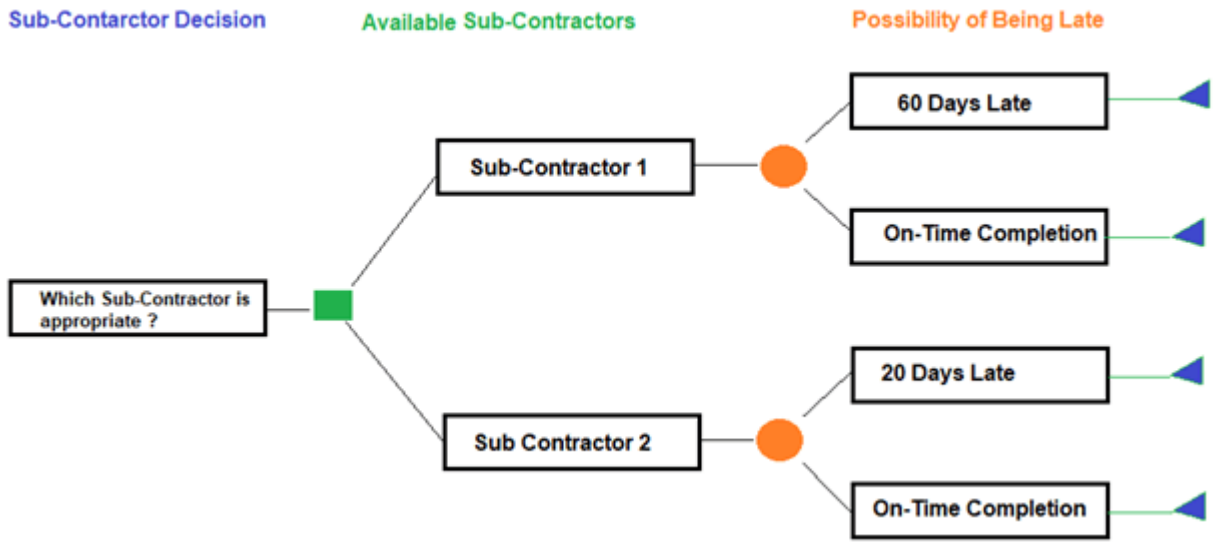
Step 2: Assign the probability of occurrence for the risks.

In this example, the possibility of being late for Sub-contractor 1 is 30% and for Sub-contractor 2 is 10 %. This means that the possibility of completing on-time for Sub-contractor 1 is 70% and for Sub-contractor 2 is 90 %. In Figure 2 below the probability of occurrence for the risks are assigned.



Step 3: Assign the impact of a risk as a monetary value.

In Step 3 we are calculating the value of the project for each path, beginning on the left-hand side with the first decision and cumulating the values to the final branch tip on the right side as if each of the decisions was taken and each case occurred. Figure 3 below shows the value of each path.



As shown in the figure, path values are calculated by the formulas given below.

Sub-Contractor 1
Path value of completing on-time = Bid Value = $ 250,000
Path value of being late = Bid Value + Penalty = $ 250,000 + 60 x $5,000 = $ 550,000

Sub-Contractor 2
Path value of completing on-time = Bid Value = $ 320,000
Path value of being late = Bid Value + Penalty = $ 320,000 + 20 x $5,000 = $ 420,000

Step 4: Calculate The Expected Monetary Value (EMV) for each decision path.

In Step 4, we are calculating the value of each node including both possibility nodes and decision nodes. We begin with the path values at the far right-hand end of the tree and then proceeding from the right to the left calculate the value of each node. This calculation is called "folding back" the tree.
The Expected Monetary Value (EMV) of each node will be calculated by multiplying Probability and Impact. Figure 4 below shows The Expected Monetary Value (EMV) of each path.

As shown in the figure, The Expected Monetary Value (EMV) of each path is below.

Sub-Contractor 1

EMV = %30 x $ 550,000 + %70 x $ 250,000 = $ 340,000

Sub-Contractor 2

EMV = %10 x $ 420,000 + %90 x $ 320,000 = $ 330,000
In this simple example Expected Monetary Values (EMV) are very close. Now we are selecting Contractor 2 because of low cost and low possibility of being late.

Linear Programming:

Linear programming, mathematical modeling technique in which a linear function is maximized or minimized when subjected to various constraints.

This technique has been useful for guiding quantitative decisions in business planning, in industrial engineering, and to a lesser extent in the social and physical sciences.

Linear programming is a method of optimising operations with some constraints. The main objective of linear programming is to maximize or minimize the numerical value

LP model. A model consisting of linear relationships representing a firm's objective and resource constraints

LP Applications

- Scheduling school buses to minimize total distance traveled
- Allocating police patrol units to high crime areas in order to minimize response time to 911 calls
- Scheduling tellers at banks so that needs are met during each hour of the day while

minimizing the total cost of labor

- Selecting the product mix in a factory to make best use of machine- and labor-hours available while maximizing the firm's profit
- Picking blends of raw materials in feed mills to produce finished feed combinations at minimum costs
- Determining the distribution system that will minimize total shipping cost
- Developing a production schedule that will satisfy future demands for a firm's product and at the same time minimize total production and inventory costs

Example: Two Mines

The Two Mines Company own two different mines that produce an ore which, after being crushed, is graded into three classes: high, medium and low-grade. The company has contracted to provide a smelting plant with 12 tons of high-grade, 8 tons of medium-grade and 24 tons of low-grade ore per week. The two mines have different operating characteristics as detailed below.

Mine	Cost per day (£'000)	Production (tons/day)		
		High	Medium	Low
X	180	6	3	4
Y	160	1	1	6

Consider that mines cannot be operated in the weekend. How many days per week should each mine be operated to fulfill the smelting plant contract?

Solution

What we have is a verbal description of the Two Mines problem.
What we need to do is to translate that verbal description into an *equivalent* mathematical description.
In dealing with problems of this kind we often do best to consider them in the order:

- Variables
- Constraints
- Objective

This process is often called *formulating* the problem (or more strictly formulating a mathematical representation of the problem).

Variables

These represent the "decisions that have to be made" or the "unknowns".
We have two decision variables in this problem:

x = number of days per week mine X is operated
y = number of days per week mine Y is operated
Note here that $x \geq 0$ and $y \geq 0$.

Constraints
It is best to first put each constraint into words and then express it in a mathematical form.
ore production constraints - balance the amount produced with the quantity required under the smelting plant contract

Ore

High $6x + 1y \geq 12$

Medium $3x + 1y \geq 8$

Low $4x + 6y \geq 24$

Days per week constraint - we cannot work more than a certain maximum number of days a week e.g. for a 5 day week we have
$x \leq 5$ $y \leq 5$

Objective
The objective is to minimize cost which is given by
180x + 160y

Complete mathematical representation of the problem:
Minimize $180x + 160y$

Subject to

$6x + y \geq 12$ $3x + y \geq 8$ $4x + 6y \geq 24$ $x \leq 5$ $y \leq 5$ $x, y \geq 0$

Predicate Logic

Formal language can be thought of as a type of language that has syntax to express the meanings

Predicate logic is a well understood formal language, with well-defined syntax, semantics and rules of inference.
Allows facts about the world to be represented as sentences formed from propositional symbols using Logic Symbols.

Logical connectives symbols used in logic are as follows.

- $\wedge$ And
- $\vee$ Or
- $\neg$ Not
- $\rightarrow$, = > Implies / Then
- $\leftrightarrow$, < = > Equivalent to

∃ there exist

∀ For all

- Implies: ⟹
- Therefore: ⊢

Application Example

Action Rules Approach to Solving Diagnostic Problems in Clinical Medicine

The beginning of the twenty first century have brought considerable advances in the field of computer-based medical systems. It resulted from noticeable improvements in medical care, from ease of storage and access of digital imaging through gathering of computerized medical data to accessing on-line literature, patient monitoring, therapy planning and computer support for medical diagnosis. Similarly to other domains, decision-support systems have proven to be valuable tools that help practitioners in facing challenging medical problems, such as diagnosis and therapy. Given that stakes happen to be extremely high, support of decision-making plays a particularly important role in the field of medicine. As an example, in the domain of hepatology, inexperienced clinicians have been found to make a correct diagnosis in jaundiced patients in less than 45% of cases. Incorrect diagnosis leads to suboptimal treatment that means waste of resources, time and sometimes of human life. Automating those processes that can be captured by focus on important signs and symptoms, eliminates human error and leads to overall improvements in the quality of medical care.