EMBA Program
MB-511

Data Science for Managerial Decisions (MB 511)

Instructor
Anant Prakash Awasthi

Data Science for Managerial Decisions (MB 511)

EMBA Program
MB-511

References/Literature

- Jojo Moolayil, “Smarter Decisions The Intersection of IoT and Data Science”, PACKT, 2016.
- Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.
- David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”, EMC 2013
- Raj, Pethuru, “Handbook of Research on Cloud Infrastructures for Big Data Analytics”, IGI Global
- Management Information System, W.S Jawadekar, Tata Mc Graw Hill Publication.
- Management Information System, David Kroenke, Tata Mc Graw Hill Publication.
- MIS Management Perspective, D.P. Goyal, Macmillan Business Books.

Data Science for Managerial Decisions (MB 511)

EMBA Program
MB-511

Online Resources

Software Resources

ANACONDA.

Data Science for Managerial Decisions (MB 511)

EMBA Program
MB-511

Program Overview

- Introduction to Data Science
- Information Technology An Overview
- Applications of Data Science in various fields
- MIS and Control Systems
- Data Collection and Data Pre-Processing
- Building Information Systems
- Support Systems for Management Decisions

MIS and Control Systems

- Introduction to MIS and Control Systems
- Design and Implementation of MIS
- Control Systems in Action
- Challenges and Future Trends

EMBA Program
MB-511

Building Data Science Solutions

Introduction to Data Science terminologies

1. Data Science
2. Algorithm
3. Features/Variables
4. Artificial Intelligence
5. Machine Learning
6. Deep Learning
7. Natural Language Processing
8. Prediction
9. Regression
10. Classification
11. Supervised Machine Learning
12. Un-Supervised Machine Learning
13. Clustering
14. Neural Networks
15. Overfitting/Underfitting
16. Encoding
17. Embeddings
18. Sentiment Analysis
19. Large Language Models
20. A/B Testing

Building Data Science Solutions

Data Science

Data science for business involves using data analysis techniques and technologies to help businesses make better decisions, solve problems, and achieve their goals. It involves collecting, processing, analyzing, and interpreting large volumes of data to uncover insights that can drive strategic and operational decisions. These insights can range from understanding customer behavior and preferences to optimizing operations, identifying market trends, predicting future outcomes, and much more. Ultimately, data science for business aims to leverage data as a valuable asset to improve efficiency, effectiveness, and competitiveness in the marketplace.

An algorithm is a step-by-step set of instructions or a procedure designed to solve a particular problem or perform a specific task.

In the context of business and management, algorithms are often used in various software applications and systems to automate processes, analyze data, and make decisions. They can range from simple rules-based procedures to complex mathematical models, all aimed at improving efficiency, optimizing resources, and achieving organizational objectives. Essentially, algorithms help streamline operations, enhance decision-making processes, and drive innovation within a company.

Features

In machine learning, features (also known as input variables or independent variables) are the individual measurable properties or characteristics of the data that are used to make predictions or decisions.

Features are the inputs to a machine learning model and play a crucial role in determining its performance.

Features are fundamental components of machine learning models, and careful selection, preprocessing, and engineering of features are essential for building accurate and effective predictive models.

Building Data Science Solutions

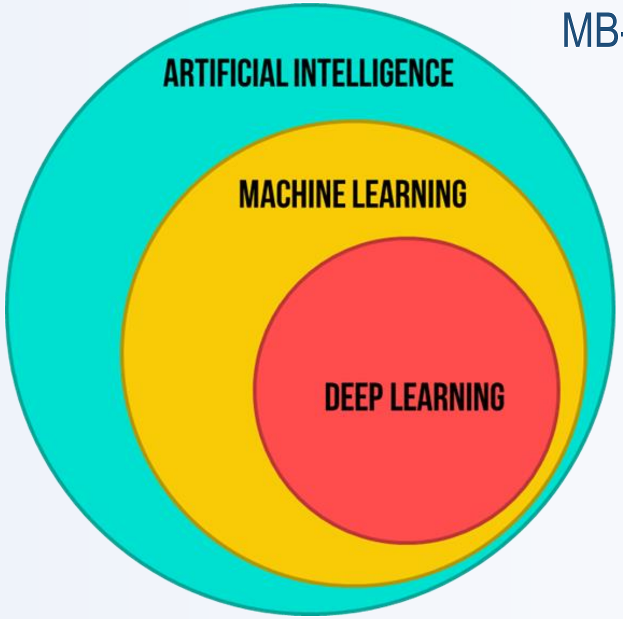
Artificial Intelligence (AI)

EMBA Program
MB-511

From a management perspective, Artificial Intelligence (AI) can be defined as the integration of advanced computing technologies that enable machines to perform tasks that typically require human intelligence. This includes tasks such as understanding natural language, recognizing patterns in data, making decisions, and even learning from experience.

AI represents a transformative force that can drive business growth, improve decision-making, and enhance competitiveness in today's rapidly evolving digital landscape.

Effective management of AI initiatives involves leveraging its capabilities to achieve strategic objectives while addressing ethical and societal considerations.

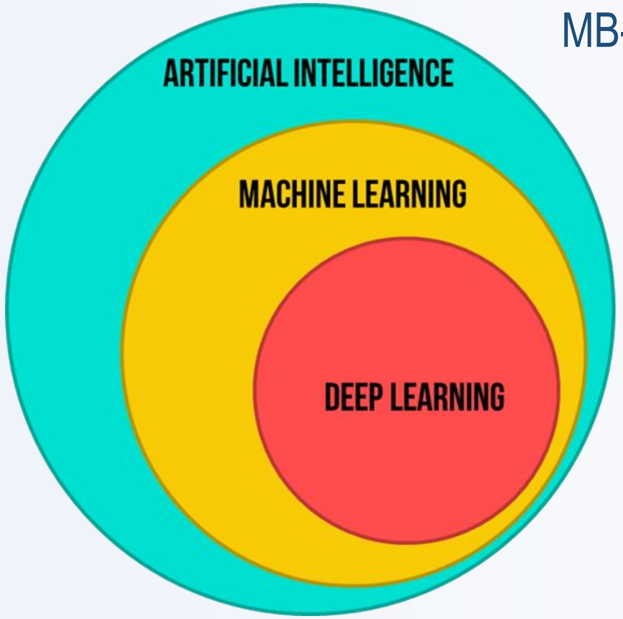


Building Data Science Solutions

Machine Learning (ML)

Machine learning can be defined as a powerful technology that enables computers to learn from data and make predictions or decisions without being explicitly programmed to do so. In simpler terms, it's like having a smart assistant that can analyze large amounts of data to identify patterns, trends, and insights that can help businesses make better decisions, improve processes, and achieve their goals.

Machine learning empowers businesses to leverage their data assets more effectively, driving innovation, efficiency, and competitiveness in today's data-driven economy. By harnessing the power of machine learning, businesses can unlock new opportunities, solve complex problems, and achieve their strategic objectives.

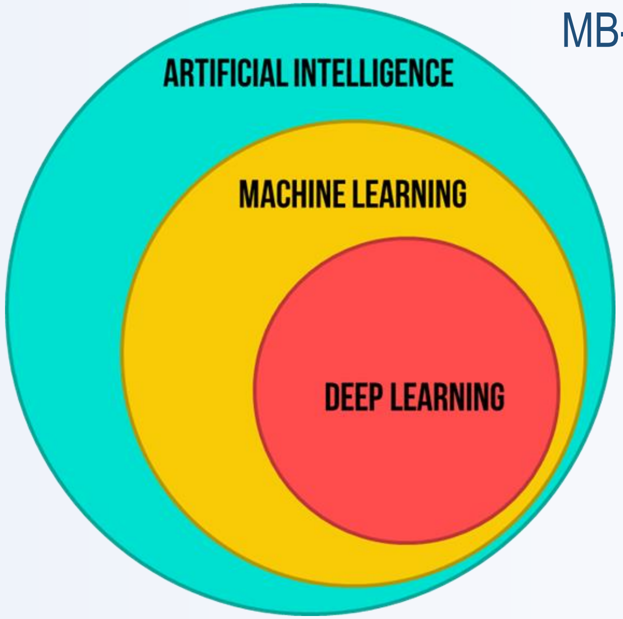


Building Data Science Solutions

Deep Learning (DL)

EMBA Program
MB-511

Deep learning can be defined as an advanced subset of machine learning that mimics the workings of the human brain to process and understand complex data. Deep learning algorithms are designed to automatically learn representations of data through multiple layers of abstraction, allowing them to extract intricate patterns and make highly accurate predictions.



Building Data Science Solutions

AI, ML and DL in a nutshell

EMBA Program
MB-511

Aspect	Artificial Intelligence (AI)	Machine Learning (ML)	Deep Learning (DL)
Definition	A broad field of computer science that aims to create intelligent machines capable of performing tasks that typically require human intelligence.	A subset of AI that involves developing algorithms that allow computers to learn from data and make predictions or decisions without being explicitly programmed.	A subset of machine learning that uses neural networks with multiple layers (deep architectures) to learn complex patterns and representations from data.
Approach	May involve various techniques including rule-based systems, expert systems, and symbolic reasoning.	Utilizes algorithms that learn from data and iteratively improve performance over time without being explicitly programmed.	Utilizes artificial neural networks with multiple layers to automatically learn representations of data through abstraction.
Scope	Encompasses a wide range of applications including natural language processing, robotics, computer vision, and more.	Widely used in applications such as predictive analytics, pattern recognition, classification, and clustering.	Primarily used in tasks such as image recognition, speech recognition, natural language processing, and more, where large amounts of data are available.
Model Complexity	Can range from simple rule-based systems to complex cognitive architectures.	Can range from simple linear regression models to complex ensemble methods and deep neural networks.	Typically involves complex neural network architectures with multiple hidden layers, requiring significant computational resources.

Building Data Science Solutions

AI, ML and DL in a nutshell

Aspect	Artificial Intelligence (AI)	Machine Learning (ML)	Deep Learning (DL)
Data Requirements	May require structured and unstructured data from various sources depending on the application.	Requires labeled or unlabeled training data for algorithm training and validation.	Requires large amounts of labeled or unlabeled data for training, which can be a limiting factor in some applications.
Training Process	May involve manual programming, knowledge engineering, or learning from experience.	Involves training algorithms on historical data to learn patterns and make predictions or decisions.	Involves training deep neural networks on large datasets using techniques such as backpropagation and stochastic gradient descent.
Interpretability	Outputs may not always be explainable or interpretable by humans.	Outputs can often be interpreted and understood based on the model's underlying logic or decision rules.	Outputs may be less interpretable due to the complex, hierarchical nature of deep neural networks.
Applications	Common applications include virtual assistants, autonomous vehicles, recommendation systems, and more.	Widely used in industries such as finance, healthcare, retail, and marketing for tasks such as fraud detection, customer segmentation, and demand forecasting.	Common applications include image and speech recognition, natural language processing, autonomous driving, and more.
Examples	IBM Watson, Siri, self-driving cars, facial recognition systems.	Logistic regression, decision trees, random forests, support vector machines.	Convolutional neural networks (CNNs), recurrent neural networks (RNNs), generative adversarial networks (GANs).

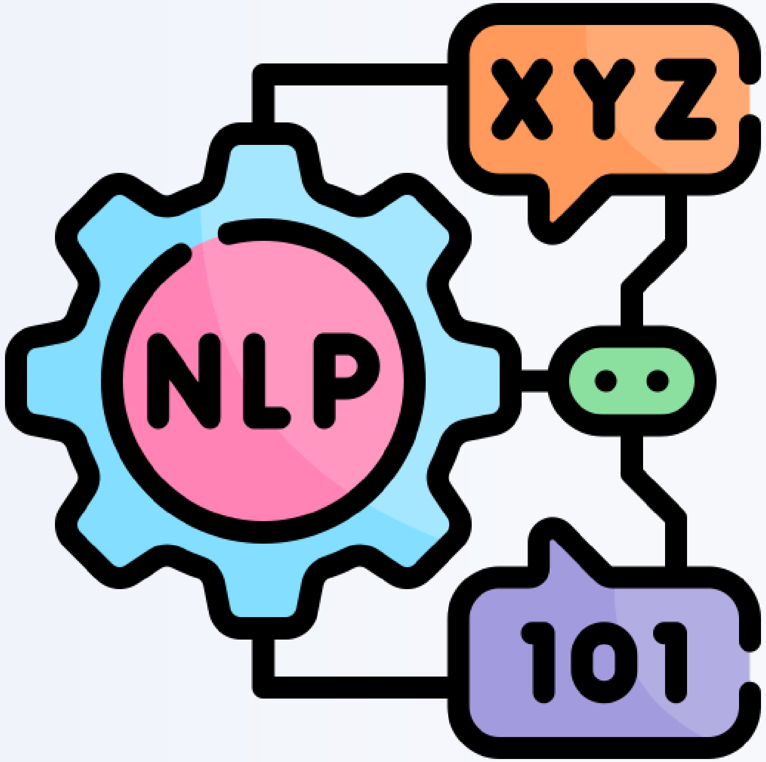
Building Data Science Solutions

Natural Language Processing (NLP)

EMBA Program
MB-511

Natural Language Processing (NLP) is a technology that enables computers to understand, interpret, and generate human language in a way that's similar to how people communicate.

Imagine you have a lot of text-based information, like emails, social media posts, or customer reviews. NLP helps businesses make sense of all this text by allowing computers to understand the meaning behind the words. NLP helps businesses unlock valuable insights from text-based data, improve customer interactions, automate tasks, and expand their global reach.



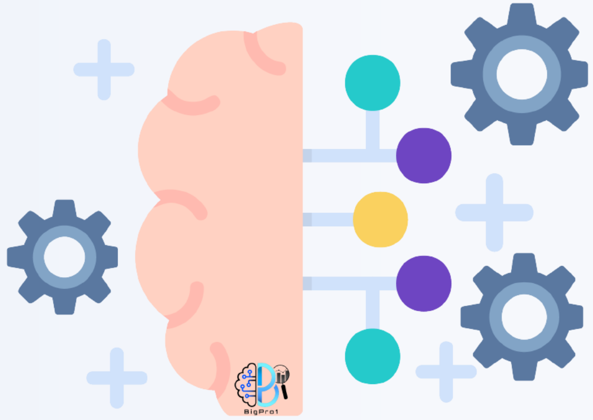
Building Data Science Solutions

Supervised & Unsupervised Machine Learning

EMBA Program
MB-511

Supervised machine learning is a type of machine learning where the algorithm learns from labeled data, meaning the input data is paired with the corresponding correct output. The algorithm learns to make predictions or decisions based on this labeled dataset.

Unsupervised machine learning is a type of machine learning where the algorithm learns from unlabeled data, meaning the input data does not have corresponding output labels. The goal of unsupervised learning is to find patterns, relationships, or structures within the data without explicit guidance.



Building Data Science Solutions

Supervised & Unsupervised Machine Learning in nutshell

Aspect	Supervised Machine Learning	Unsupervised Machine Learning
Input Data	Labeled data, consisting of input features and corresponding output labels.	Unlabeled data, consisting only of input features without corresponding output labels.
Learning Approach	Learns from labeled examples to predict or classify new instances.	Learns from unlabeled data to discover patterns, relationships, or structures within the data.
Goal	To make predictions or decisions based on input features.	To find patterns, groupings, or anomalies within the data.
Examples	Classification, Regression.	Clustering, Dimensionality Reduction, Anomaly Detection.
Evaluation Metrics	Accuracy, Precision, Recall (for classification).	No clear output labels, evaluation often subjective or based on domain knowledge.
Model Complexity	Typically more complex models with higher capacity to learn intricate patterns.	Simpler models, often relying on clustering or dimensionality reduction techniques.
Feedback Loop	Requires labeled data for training, feedback loop involves comparing predictions to actual labels.	Does not require labeled data, feedback loop may involve adjusting model parameters based on unsupervised learning techniques.
Use Cases	Predictive modeling, Sentiment Analysis, Image Recognition.	Customer Segmentation, Anomaly Detection, Data Exploration.
Interpretability	Often more interpretable, as predictions are based on labeled examples and can be traced back to input features.	May be less interpretable, as patterns and groupings are discovered without explicit guidance.
Data Requirements	Requires labeled data, which can be expensive and time-consuming to obtain.	Can work with unlabeled data, potentially reducing data collection and labeling efforts.

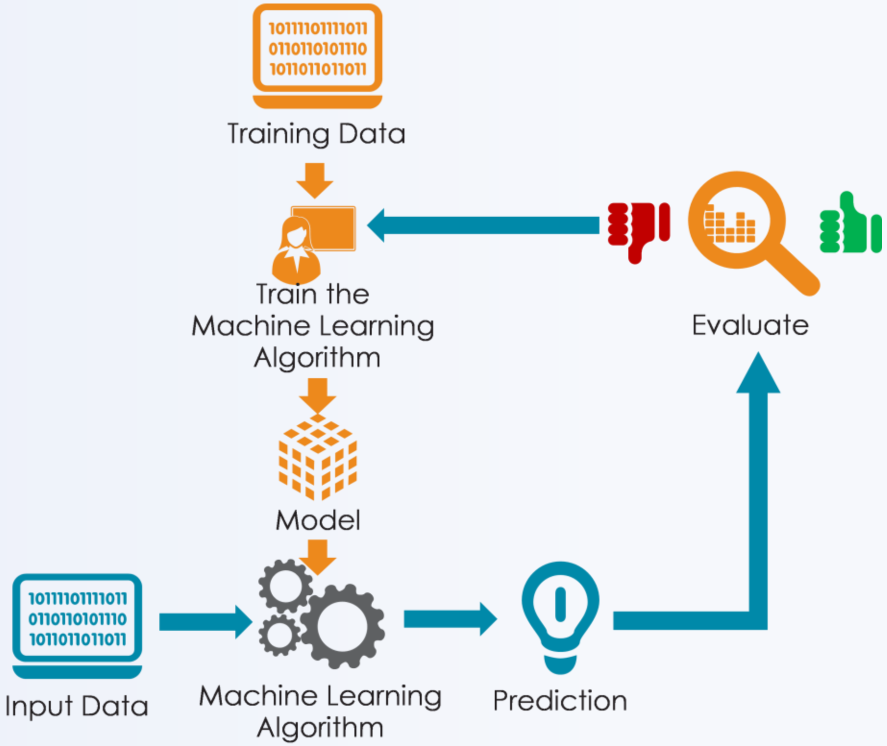
Building Data Science Solutions

Prediction

EMBA Program
MB-511

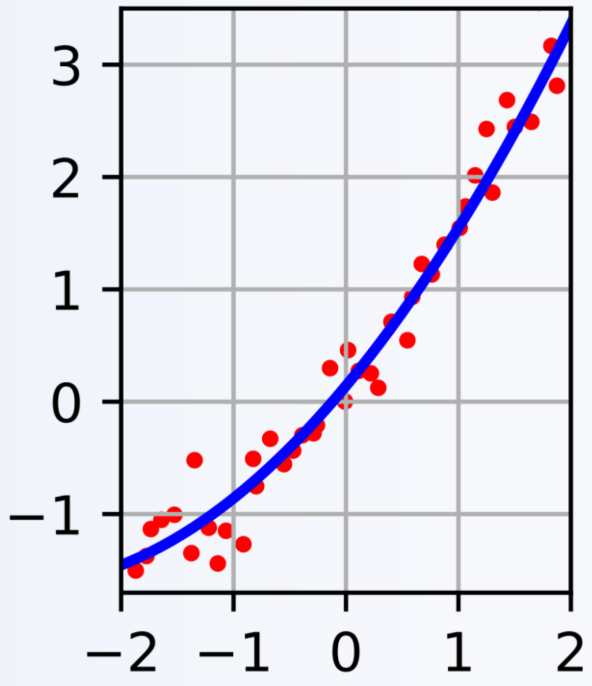
Imagine you're a business owner of an online retail store. You have a lot of historical data about your customers, such as their purchase history, demographics, and website behavior. Machine learning prediction can help you forecast/predict what your customers are likely to do in the future based on this data.

Machine learning prediction empowers businesses to anticipate future trends, behaviors, and outcomes based on historical data. By leveraging these predictions, businesses can make more informed decisions, improve customer experiences, and drive growth and profitability.



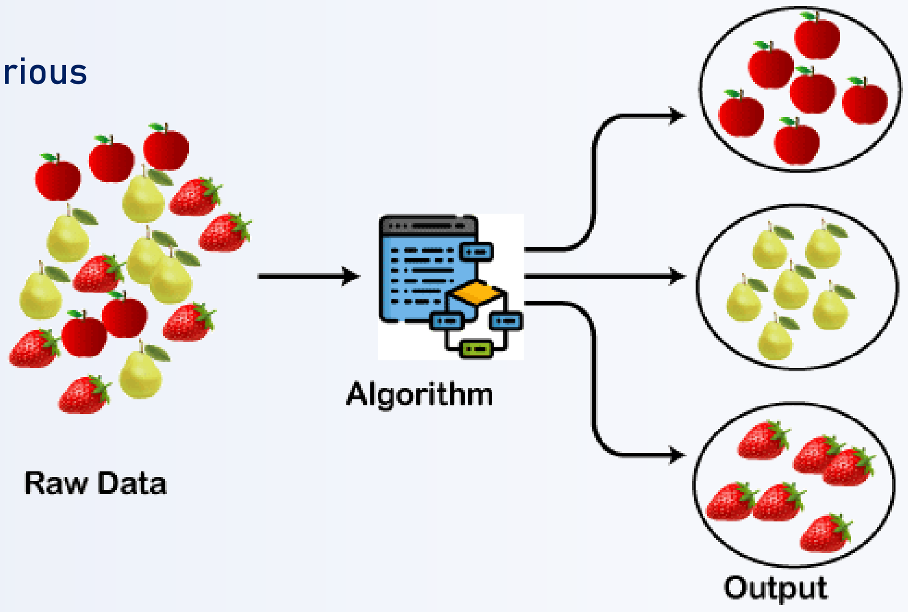
Regression problems are a type of supervised learning problem in machine learning where the goal is to predict a continuous numerical value based on input variables. In other words, regression models are used when the target or dependent variable is a real-valued number.

Regression problems involve predicting a continuous numerical value based on input variables, and regression analysis is widely used in various fields for prediction and forecasting tasks.



Classification problems are a type of supervised learning problem in machine learning where the goal is to predict the category or class label of a new observation based on input variables. In other words, classification models are used when the target or dependent variable is categorical.

Classification problems involve predicting the category or class label of a new observation based on input variables, and classification models are widely used in various fields for tasks such as sentiment analysis, image recognition, and fraud detection.



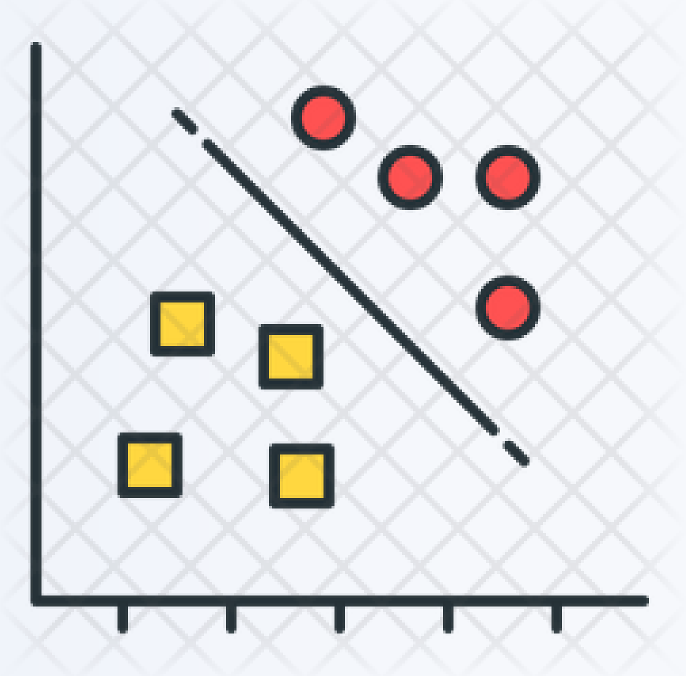
Building Data Science Solutions

Clustering

EMBA Program
MB-511

Clustering in machine learning is like organizing your business's customer base into different groups based on similarities. It's a way to group together customers who share common characteristics or behaviors, even if you didn't know about these groups beforehand.

Clustering in machine learning is a powerful tool for businesses to understand their customers and tailor their strategies accordingly. It helps you uncover hidden patterns, group customers into meaningful segments, and make data-driven decisions that drive growth and profitability.



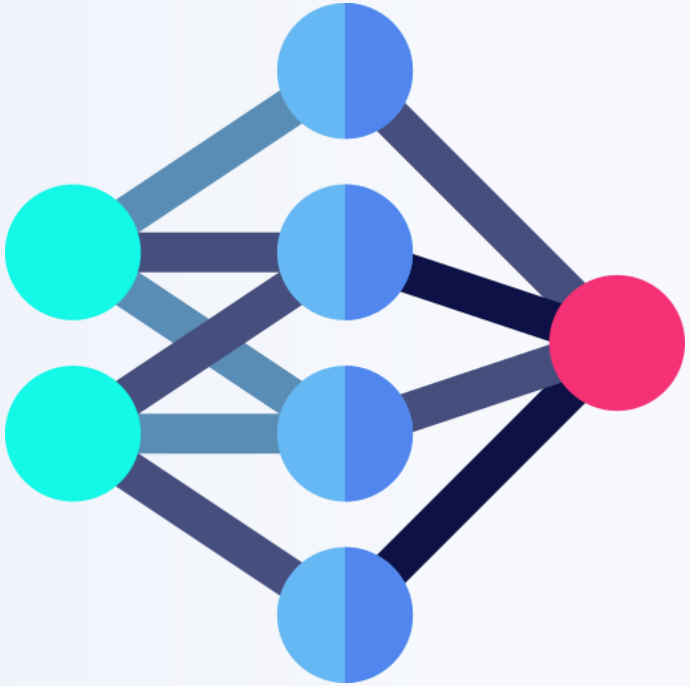
Building Data Science Solutions

Artificial Neural networks (ANN)

EMBA Program
MB-511

Neural networks are like virtual brains for computers. Just as our brains have interconnected neurons that process information, neural networks are made up of artificial neurons called nodes, organized into layers.

Neural networks are computer systems inspired by the human brain. They learn from data, recognize patterns, and make predictions, making them valuable tools for solving complex problems and driving innovation in various fields.



Building Data Science Solutions

Supervised Machine Learning use cases

EMBA Program
MB-511

Algorithm	Description	Use Cases in Business
Linear Regression	Predicts a continuous numerical value based on input features by fitting a linear equation to the data.	Sales forecasting, price optimization, demand prediction.
Logistic Regression	Predicts the probability of a binary outcome based on input features using a logistic function.	Customer churn prediction, credit risk assessment, fraud detection.
Decision Trees	Builds a tree-like structure to classify or regress data by recursively splitting based on feature values.	Customer segmentation, product recommendation, risk analysis.
Random Forest	Ensemble learning method that combines multiple decision trees to improve prediction accuracy.	Marketing campaign targeting, credit scoring, anomaly detection.
Support Vector Machines	Finds the optimal hyperplane that best separates data points into different classes in a high-dimensional space.	Image classification, sentiment analysis, text categorization.
Naive Bayes	Probabilistic classifier based on Bayes' theorem that assumes independence between features.	Email spam detection, document classification, sentiment analysis.
K-Nearest Neighbors (KNN)	Classifies data points based on the majority class among their k nearest neighbors in feature space.	Customer segmentation, recommendation systems, anomaly detection.
Neural Networks	Deep learning models composed of interconnected layers of nodes that learn complex patterns from data.	Image recognition, speech recognition, natural language processing.

Building Data Science Solutions

Un-Supervised Machine Learning use cases

EMBA Program
MB-511

Algorithm	Description	Use Cases in Business
K-Means Clustering	Divides data into 'k' clusters based on similarity, with each cluster represented by its centroid.	Customer segmentation, Market basket analysis, Anomaly detection, Product recommendation.
Hierarchical Clustering	Builds a hierarchy of clusters by either merging or splitting existing clusters based on similarity.	Customer segmentation, Taxonomy creation, Fraud detection, Image segmentation.
DBSCAN	Groups together data points that are closely packed based on density, identifying noise as outliers.	Anomaly detection, Customer segmentation, Geographic clustering, Fraud detection.
Principal Component Analysis (PCA)	Reduces the dimensionality of the data while preserving most of its variance, by transforming it into a new set of orthogonal variables.	Dimensionality reduction, Feature extraction, Data visualization, Anomaly detection.
t-Distributed Stochastic Neighbor Embedding (t-SNE)	Reduces high-dimensional data to a lower-dimensional space while preserving local structure.	Data visualization, Clustering validation, Exploratory data analysis, Feature engineering.
Gaussian Mixture Models (GMM)	Represents a mixture of several Gaussian distributions to model complex data distributions.	Clustering, Density estimation, Anomaly detection, Image segmentation.

Building Data Science Solutions

Artificial Neural Network (ANN) use cases

Neural Network Architecture	Description	Use Cases
Feedforward Neural Network (FNN)	Consists of input, hidden, and output layers where information flows in one direction from input to output. Each neuron in a layer is connected to all neurons in the subsequent layer.	Classification, Regression, Pattern recognition, Credit scoring.
Convolutional Neural Network (CNN)	Designed for processing grid-like data such as images. Utilizes convolutional layers to extract features hierarchically and pooling layers to reduce dimensionality.	Image classification, Object detection, Facial recognition, Medical image analysis.
Recurrent Neural Network (RNN)	Contains recurrent connections allowing information to persist over time. Suitable for sequence data processing where the output depends on the previous inputs.	Natural language processing, Speech recognition, Time series prediction, Sentiment analysis.
Long Short-Term Memory (LSTM)	A type of RNN architecture with memory cells and gating mechanisms to capture long-range dependencies and mitigate vanishing gradient problem.	Text generation, Machine translation, Stock price prediction, Video analysis.
Gated Recurrent Unit (GRU)	Similar to LSTM but with a simpler architecture. It combines the forget and input gates into a single update gate.	Speech recognition, Language modeling, Gesture recognition, Handwriting recognition.
Autoencoder	Consists of an encoder and a decoder. Encoder compresses input data into a latent representation, and decoder reconstructs the original input from the latent representation.	Dimensionality reduction, Feature learning, Anomaly detection, Denoising data.
Generative Adversarial Network (GAN)	Comprises two neural networks, a generator and a discriminator, trained simultaneously in a competitive manner. Generator creates realistic data samples, and discriminator distinguishes between real and fake samples.	Image generation, Data augmentation, Style transfer, Super-resolution.

Overfitting and Underfitting

In machine learning, overfitting happens when a model learns too much from the training data and captures noise or random fluctuations in the data as if they were real patterns. As a result, the model performs very well on the training data but fails to generalize to new, unseen data.

In machine learning, underfitting happens when a model is too simple to capture the underlying structure of the data. The model fails to learn from the training data and performs poorly both on the training data and on new, unseen data.

Building Data Science Solutions

Encoding

In machine learning, "encodings" refer to the process of representing categorical data (data that consists of categories or labels) in a numerical format that machine learning algorithms can understand and process effectively. Encodings are necessary because most machine learning algorithms require numerical input data.

These encoding techniques allow machine learning algorithms to effectively process categorical data, enabling them to learn from and make predictions on datasets that include categorical variables. Choosing the appropriate encoding technique depends on the nature of the categorical variable and the specific requirements of the machine learning task.

Building Data Science Solutions

Encoding - Examples

Encoding Method	Description	Use Cases
Label Encoding	Assigns a unique numerical label to each category in a categorical variable.	Ordinal categorical variables, such as "Low," "Medium," "High" levels of a feature.
One-Hot Encoding	Creates binary dummy variables for each category, where each category is represented by a binary vector.	Nominal categorical variables, such as colors ("Red," "Blue," "Green").
Binary Encoding	Represents each category as a binary number, with each digit of the binary number becoming a feature.	Large categorical variables, where one-hot encoding would lead to high dimensionality.
Target Encoding	Replaces each category with the mean (or other statistic) of the target variable for that category.	Variables with high cardinality, such as geographical regions or product categories.
Frequency Encoding	Replaces each category with the frequency (count) of occurrences of that category in the dataset.	Rare event detection, anomaly detection, or when the frequency of occurrence is relevant information.
Ordinal Encoding	Similar to label encoding but assigns numerical labels based on the ordinality (order) of categories.	Ordinal categorical variables, where the order of categories is meaningful, such as "Low," "Medium," "High" levels of a feature.

In simple terms, word embeddings in natural language processing (NLP) are numerical representations of words that capture their meaning and context in a way that computers can understand.

Imagine you're teaching a computer to understand language. Instead of treating each word as a standalone entity, word embeddings allow the computer to understand relationships between words based on their usage in context.

Word embeddings provide a way to represent words as numerical vectors that capture their meaning and context in a high-dimensional space. They are a fundamental component of many NLP models, enabling computers to understand and work with human language more intelligently.

Building Data Science Solutions

Words Embeddings - Examples

EMBA Program
MB-511

Word Embedding Method	Description	Use Cases
Word2Vec	A shallow neural network model that learns word embeddings by predicting words in context.	Natural language processing tasks such as sentiment analysis, document classification, and machine translation.
GloVe (Global Vectors for Word Representation)	Uses co-occurrence statistics to learn word embeddings, emphasizing global context.	Similar use cases as Word2Vec, often preferred for tasks where semantic relationships between words are crucial, such as word analogy tasks.
FastText	An extension of Word2Vec that represents words as bags of character n-grams, allowing for better handling of out-of-vocabulary words and morphologically rich languages.	Text classification, Language modeling, Named entity recognition, Text similarity tasks.
ELMo (Embeddings from Language Models)	Generates word embeddings by combining deep bidirectional language model representations.	Natural language understanding tasks such as question answering, sentiment analysis, and named entity recognition.
BERT (Bidirectional Encoder Representations from Transformers)	Utilizes transformer architecture to pre-train a deep bidirectional language model on large text corpora.	Cutting-edge natural language processing tasks such as text generation, question answering, and language understanding.
GPT (Generative Pre-trained Transformer)	A transformer-based model trained on a diverse range of internet text to generate coherent and contextually relevant text.	Text generation, Conversational agents, Content creation, Language understanding.

Large Language Models (LLMs) are powerful artificial intelligence models trained on massive amounts of data. These models, such as OpenAI's GPT (Generative Pre-trained Transformer) series or Google's BERT (Bidirectional Encoder Representations from Transformers), are capable of understanding and generating human-like text at scale.

Large Language Models offer tremendous potential for businesses across industries to enhance their operations, improve customer experiences, drive innovation, and gain a competitive edge in today's data-driven and digitally connected world. However, it's essential for businesses to consider ethical and privacy implications when deploying LLMs and ensure responsible and transparent use of AI technologies.

Building Data Science Solutions

Large Language Models (LLMs) – Use Cases

Large Language Model	Developer	Use Cases
LaMDA	Google AI	• Chatbots and virtual assistants • Dialogue generation • Summarization of factual topics
GPT-3	OpenAI	• Creative text formats (poems, code, scripts) • Machine translation • Question answering • Content generation
Megatron-Turing NLG	NVIDIA	• Scientific research paper generation • Summarization of complex information • Extractive question answering
WuDao 2.0	BAAI (China)	• Machine translation (especially Chinese languages) • Dialogue generation • Question answering
Jurassic-1 Jumbo	AI21 Labs	• Scientific writing assistance • Different creative text formats • Summarization • Question answering
LLaMA	Meta AI	• Generating different creative text formats • Machine translation (with focus on cultural context) • Question answering

A/B Testing

A/B testing, also known as split testing, is a method used to compare two versions of a process, webpage or app to determine which one performs better. It is commonly used in marketing, user experience design, and product development to optimize various aspects of a product or service.

- Website Design and Layout Optimization
- Email Marketing Campaigns
- Product Features and User Interface (UI) Enhancements
- Pricing and Promotions Optimization
- Mobile App Optimization

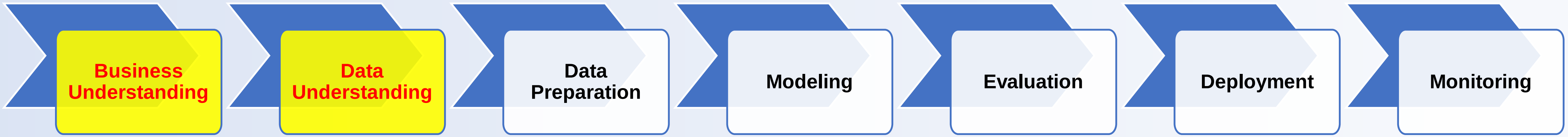
Building Data Science Solutions

Data Management in Information System

EMBA Program
MB-511

Scope:

- Role of Data Management in Information System
- Terminologies
- Hands on Sessions
    - Data Import
    - Data Export
    - Sub-setting data (selecting rows and columns)
    - Data types conversion in Pandas
    - Operators and New Variable Creation
    - Functions with Pandas
    - User defined functions in Pandas
    - Control structure in Pandas
    - Sorting data
    - Merging data
    - Reshaping data
    - Data Aggregation and Pivoting



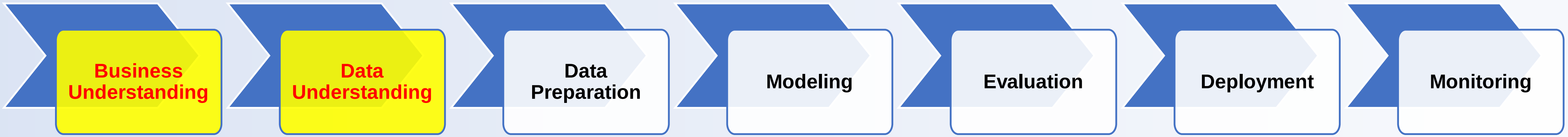
Building Data Science Solutions

Data Import/Export

EMBA Program
MB-511

Data import, involves taking data from an external source, such as a file or another system, and bringing it into a target system or application. This process may include mapping data fields, validating data integrity, and transforming the data into a format that is compatible with the target system.

Once imported, the data becomes available for use within the target system for analysis, reporting, or other purposes.



Building Data Science Solutions

Data Import/Export

EMBA Program
MB-511

Data export refers to the process of extracting data from a source system or application and transferring it to another location or format. This can involve exporting data from databases, spreadsheets, CRM systems, or any other data storage platform.

Data export typically involves selecting specific datasets, defining the export format, and initiating the transfer process. The exported data may be saved as files (such as CSV, Excel, or JSON), transferred to another database, or integrated into another application for further analysis or processing.

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment → Monitoring

Building Data Science Solutions

Data Sub-setting

EMBA Program
MB-511

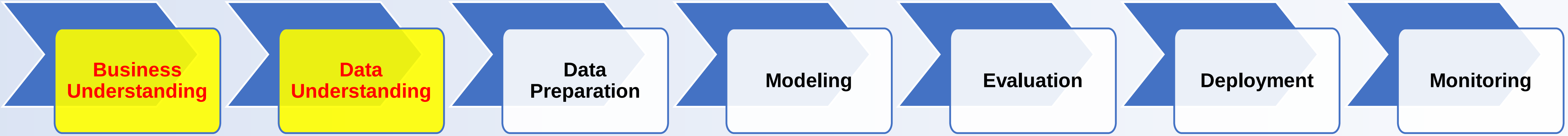
In the context of data management, data sub-setting refers to the process of selecting and extracting a subset or portion of a larger dataset based on specific criteria or requirements. This subset typically contains a smaller, more focused set of data that meets the needs of a particular use case or analysis.

Purpose of Data Sub-setting:

- Reducing Data Volume
- Improving Data Quality
- Enhancing Performance
- Enhancing Security and Compliance

Methods of Data Sub-setting:

- Sampling
- Querying
- Selecting or dropping Rows
- Selecting or dropping Columns



Building Data Science Solutions

Data Types conversion

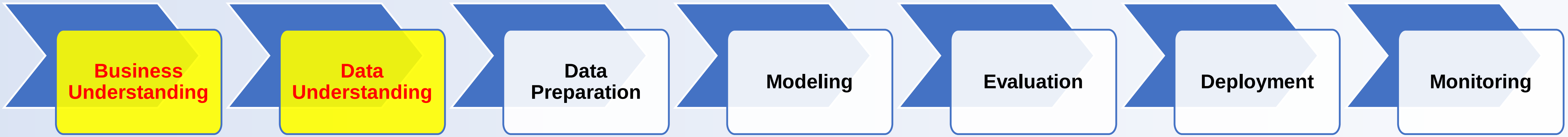
EMBA Program
MB-511

Data type conversion plays a crucial role in data management for ensuring data compatibility, consistency, and usability across different systems, applications, and processes. Here are some key roles of data type conversion in data management:

- Interoperability
- Data Integration
- Data Transformation
- Data Validation and Cleansing

Major Data Conversion Examples:

- Text >> Number
- Text >> Date
- Number >> Text
- Text or Number >> Boolean
- Date Format Interchange
- Date >> Text



Building Data Science Solutions

Operators

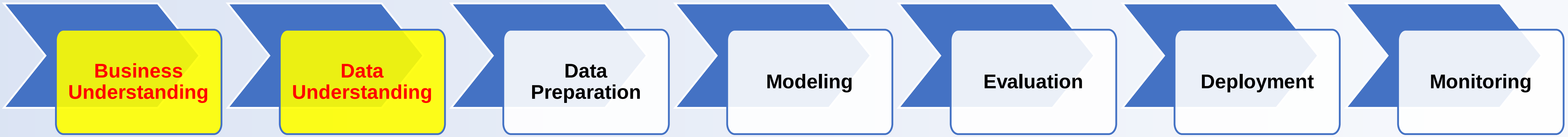
EMBA Program
MB-511

Arithmetic operators are mathematical symbols or functions used to perform basic arithmetic operations on numerical values. These operators are commonly used in programming languages and mathematical expressions to perform calculations.

- Addition (+)
- Subtraction (-)
- Multiplication (*)
- Division (/)
- Exponentiation () or Power (^)**
- Modulo (%)

Logical operators are used to perform logical operations on boolean values (true or false) or expressions. These operators are commonly used in programming languages and conditional statements to make decisions based on the truth or falsehood of conditions.

- AND (&)
- OR (|)
- NOT (!)

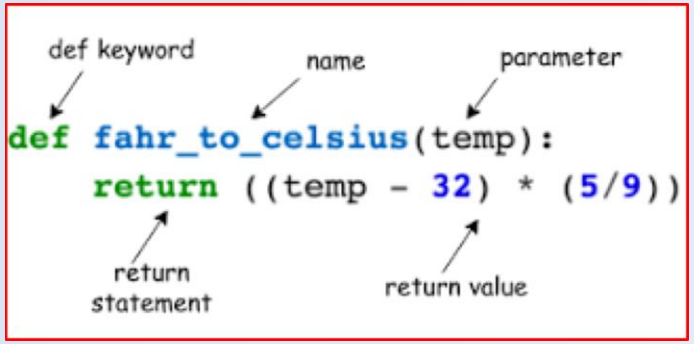


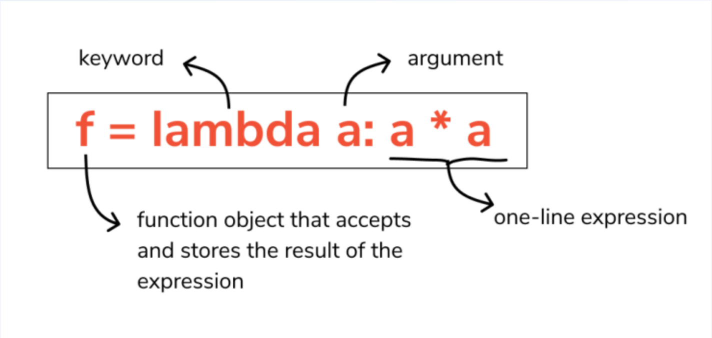
Building Data Science Solutions

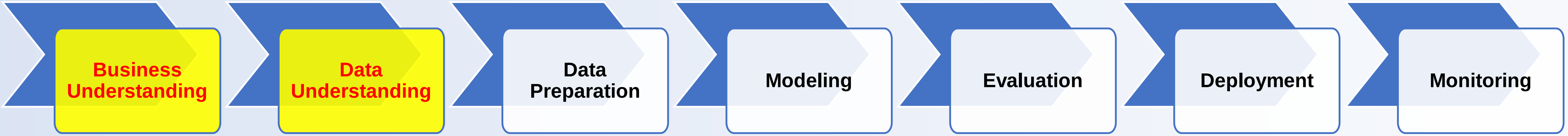
Functions and Anonymous Functions

In programming languages, a function is a self-contained block of code that performs a specific task or computation. Functions are used to organize code into reusable modules, promote code reuse, and improve code readability and maintainability.

A lambda function is a small anonymous function that can have any number of parameters but can only have one expression. Lambda functions are often used as inline functions where a full function definition is not necessary or convenient.







Building Data Science Solutions

Control Statements

control statements are used to control the flow of execution in a program. They allow you to make decisions, loop over sequences of data, and perform different actions based on conditions. The main types of control statements in Python are:

Conditional Statements (if, elif, else):

- The if statement is used to execute a block of code if a specified condition is true.
- The elif (else if) statement allows you to check additional conditions if the previous conditions are false.
- The else statement is used to execute a block of code if none of the previous conditions are true.

Building Data Science Solutions

Control Statements

EMBA Program
MB-511

control statements are used to control the flow of execution in a program. They allow you to make decisions, loop over sequences of data, and perform different actions based on conditions. The main types of control statements in Python are:

- Conditional Statements (if, elif, else)
- Looping Statements (for, while)
- Control Flow Statements (break, continue)
- Exception Handling (try, except, finally)

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment → Monitoring

Building Data Science Solutions

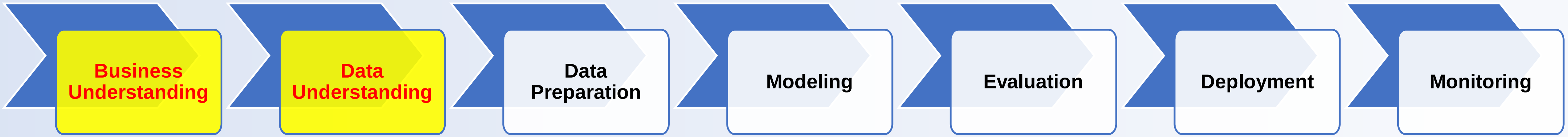
Conditional Statements (if, elif, else)

EMBA Program
MB-511

```python
x = 10
if x > 0:
    print("Positive")
elif x < 0:
    print("Negative")
else:
    print("Zero")
```

Conditional Statements (if, elif, else):

- The if statement is used to execute a block of code if a specified condition is true.
- The elif (else if) statement allows you to check additional conditions if the previous conditions are false.
- The else statement is used to execute a block of code if none of the previous conditions are true.



Building Data Science Solutions

Looping Statements (for, while)

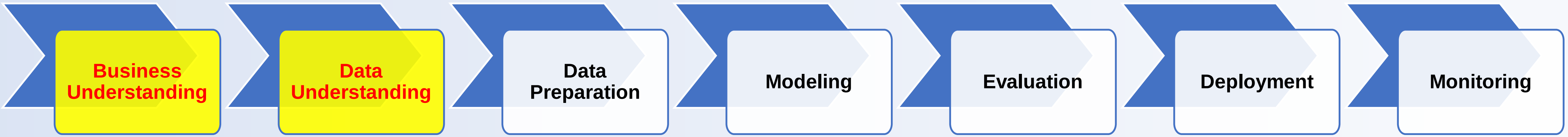
EMBA Program
MB-511

```python
fruits = ["apple", "banana", "cherry"]
for fruit in fruits:
    print(fruit)

num = 1
while num <= 5:
    print(num)
    num += 1
```

Looping Statements (for, while):

- The for loop is used to iterate over a sequence (such as a list, tuple, or string) or an iterable object.
- The while loop is used to repeatedly execute a block of code as long as a specified condition is true.



Building Data Science Solutions

Control Flow Statements (break, continue)

```python
for x in range(10):
    if x == 5:
        break
    print(x)

for x in range(10):
    if x == 5:
        continue
    print(x)
m += 1
```

EMBA Program
MB-511

Control Flow Statements (break, continue):

- The break statement is used to exit a loop prematurely, regardless of whether the loop condition is true or false.
- The continue statement is used to skip the rest of the code inside a loop for the current iteration and proceed to the next iteration.

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment → Monitoring

Building Data Science Solutions

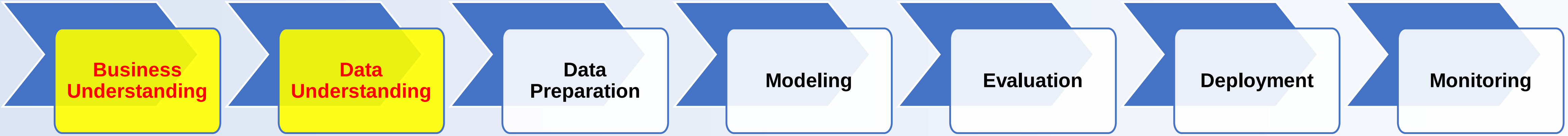
Exception Handling (try, except, finally)

EMBA Program
MB-511

```
try:
    result = 10 / 0
except ZeroDivisionError:
    print("Error: Division by zero")
finally:
    print("Execution completed")
)
m += 1
```

Exception Handling (try, except, finally):

- The try statement is used to enclose code that may raise exceptions.
- The except statement is used to handle specific exceptions that occur within the try block.
- The finally statement is used to execute code regardless of whether an exception occurred.



Building Data Science Solutions

Sorting & Merging

EMBA Program
MB-511

In pandas, sorting refers to arranging the rows of a DataFrame or Series in a specific order based on the values of one or more columns. Sorting can be done in ascending or descending order.

sort_values(): This method is used to sort the DataFrame by the values of one or more columns. You can specify the column(s) to sort by and the sorting order (ascending or descending). By default, the method sorts in ascending order.

In pandas, merging refers to combining two or more DataFrames based on common columns or indices. Pandas provides several merging options to perform different types of merges.

- Concatenation
- Merge
- Join

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment → Monitoring

Building Data Science Solutions

Data Reshaping & Pivoting

EMBA Program
MB-511

In pandas, data reshaping and pivoting refer to the process of transforming a DataFrame into a different shape or structure to better suit the analysis or presentation needs. This involves rearranging the layout of the data by altering the arrangement of rows and columns.

Methods in Pandas:

- Reshaping with stack() and unstack()
- Pivoting with pivot()
- Pivoting with pivot_table()
- Melting with melt()

Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment → Monitoring

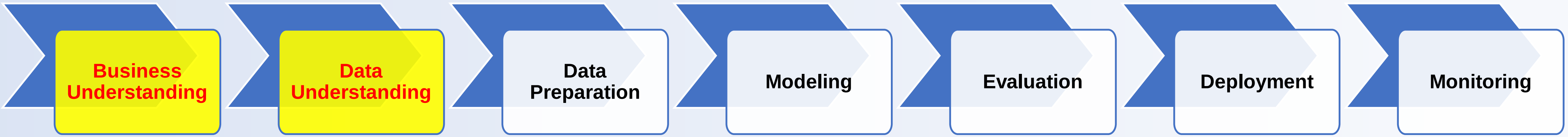
Building Data Science Solutions

Data Aggregation

data aggregation refers to the process of combining multiple rows of data into a single value based on some grouping criteria. The groupby() function is used to group data in a DataFrame based on one or more columns, and aggregation functions can then be applied to each group to summarize the data.

Methods in Pandas:

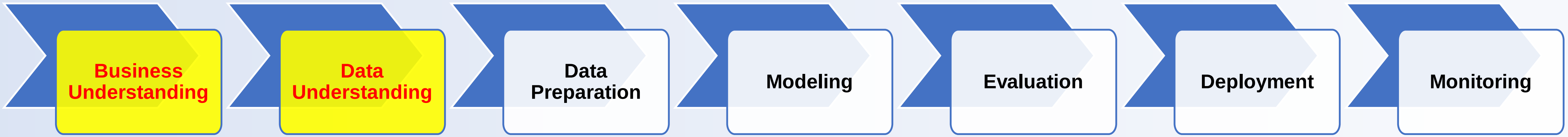
- Grouping with groupby() >> Applying Aggregation Functions
- Multiple & custom Aggregations
- Grouping by Multiple Columns



Quiz and Assignment 3

EMBA Program
MB-511

- Quiz 3 : April 04, 2024
- Assignment 3 : April 11, 2024



Have a question?

Feel Free to Reach out at

- +91-88846-52929 (WhatsApp)
- anant.awasthi@outlook.com (E-Mail)