Course: Artificial Intelligence and Machine Learning Code: 20CS51I

WEEK- 2 Machine Learning

- **Pipeline**
  - **Data engineering**
  - **Machine Learning**
  - **Deployment**
- **What is Data Science?**
- **How Data Science works?**
- **Data Science uses**

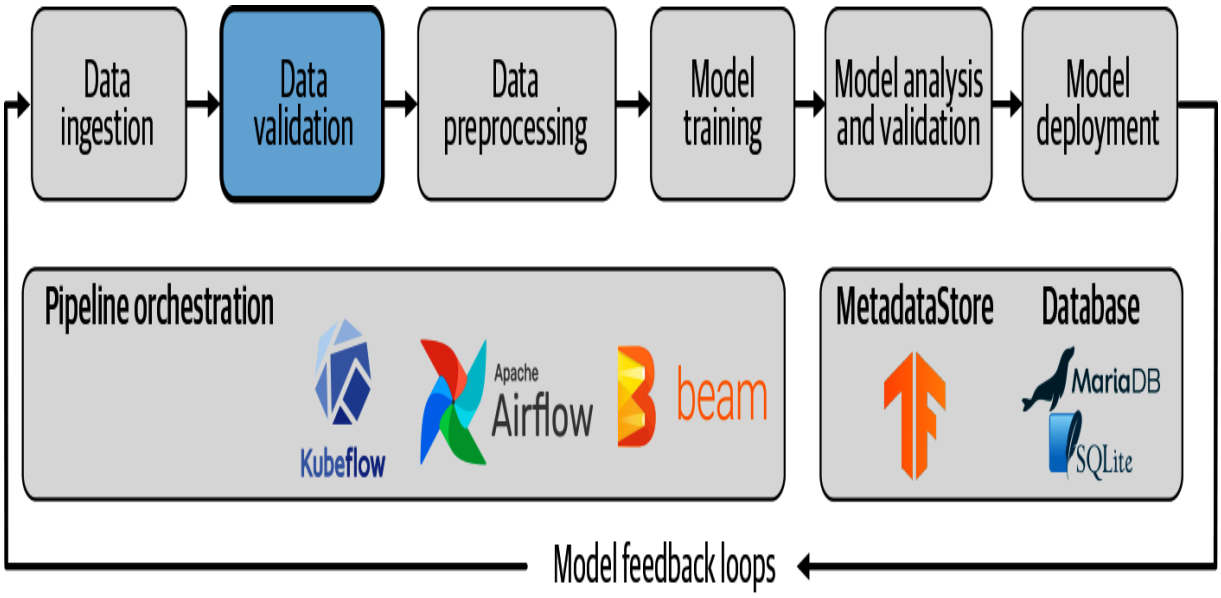
Session No. 2

Pipeline

Machine Learning

A machine learning pipeline is a way to control and automate the workflow it takes to produce a machine learning model. Machine learning pipelines consist of multiple sequential steps that do everything from data extraction and preprocessing to model training and deployment.

Machine learning pipelines are iterative as every step is repeated to continuously improve the accuracy of the model and achieve the end goal.



An example of ML Pipeline O'Reilly

The term **Pipeline** is used generally to describe the independent sequence of steps that are arranged together to achieve a task. This task could be machine learning or not. Machine Learning Pipelines are very common but that is not the only type of pipeline that exists. Data Orchestration Pipelines are another example.

According to **Microsoft docs**, there are three scenarios:

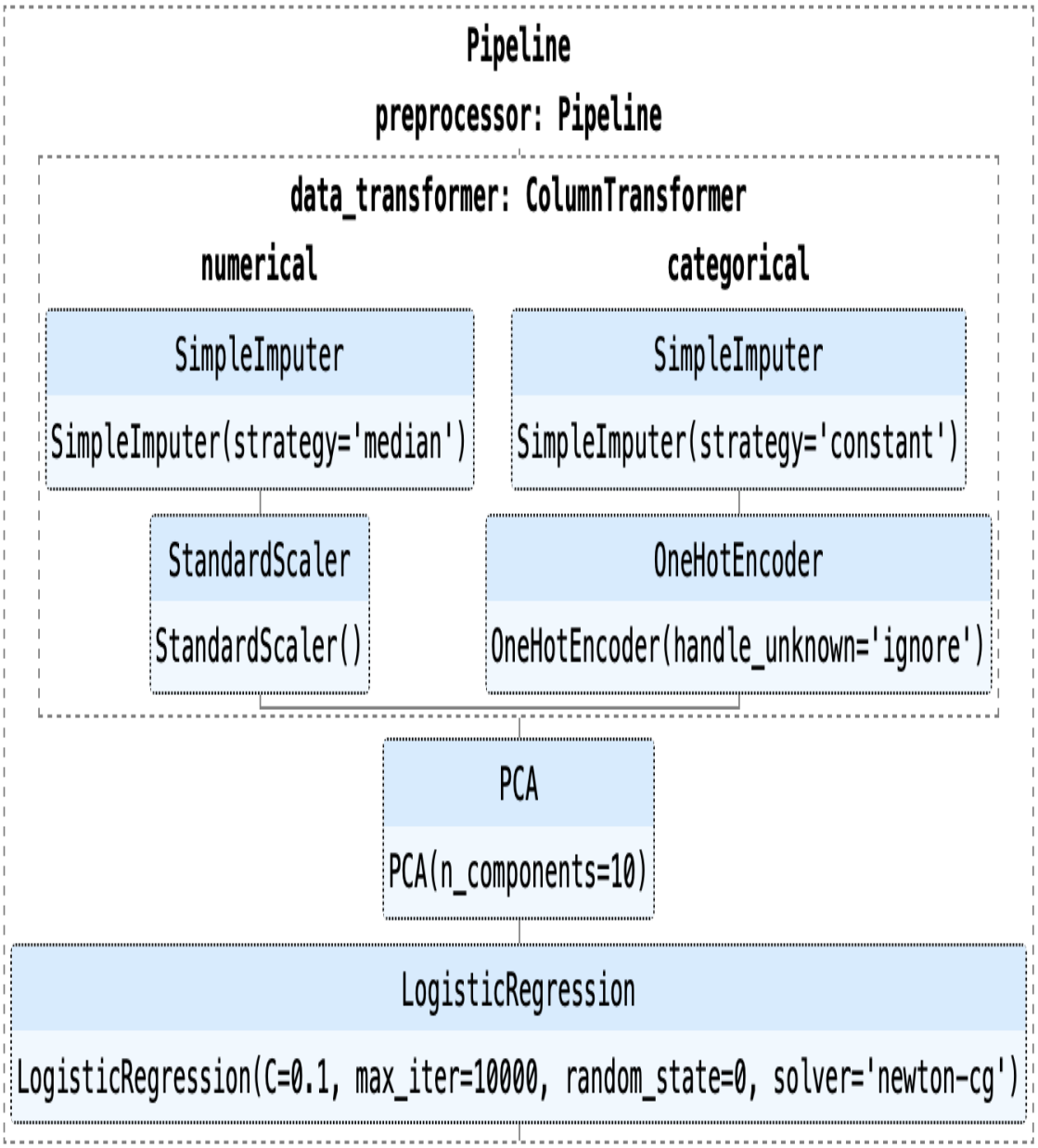
Scenario	Primary persona	Azure offering	OSS offering	Canonical pipe	Strengths
Model orchestration (Machine learning)	Data scientist	Azure Machine Learning Pipelines	Kubeflow Pipelines	Data -> Model	Distribution, caching, code-first, reuse
Data orchestration (Data prep)	Data engineer	Azure Data Factory pipelines	Apache Airflow	Data -> Data	Strongly typed movement, data-centric activities
Code & app orchestration (CI/CD)	App Developer / Ops	Azure Pipelines	Jenkins	Code + Model -> App/Service	Most open and flexible activity support, approval queues, phases with gating

Deployment

The deployment of machine learning models (or pipelines) is the process of making models available in production where web applications, enterprise software (ERPs) and APIs can consume the trained model by providing new data points, and get the predictions.

In short, Deployment in Machine Learning is the method by which you integrate a machine learning model into an existing production environment to make practical business decisions based on data. It is the last stage in the machine learning lifecycle.

Normally the term Machine Learning Model Deployment is used to describe deployment of the entire Machine Learning Pipeline, in which the model itself is only one component of the Pipeline.

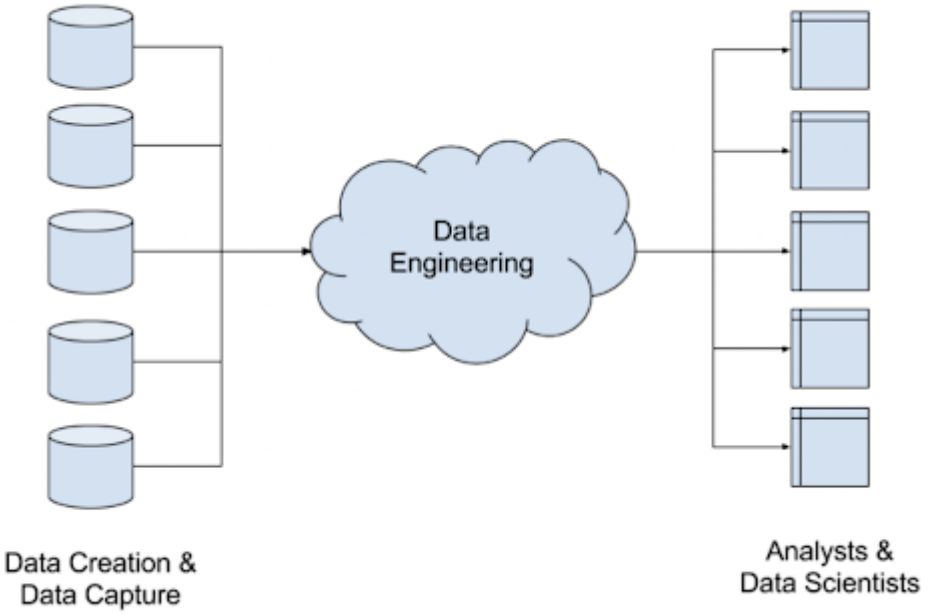


An example of a machine learning pipeline built using sklearn

As you can see in the above example, this pipeline consists of a Logistic Regression model. There are several steps in the pipeline that have to be executed first before training can begin, such as Imputation of missing values, One-Hot encoding, Scaling, and Principal Component Analysis (PCA).

Data engineering □

Data engineering is the process of designing and building systems that let people collect and analyze raw data from multiple sources and formats. These systems empower people to find practical applications of the data, which businesses can use to thrive.



What Do Data Engineers Do?

Data engineering is a skill that is in increasing demand. Data engineers are the people who design the system that unifies data and can help you navigate it. Data engineers perform many different tasks including:

Acquisition: Finding all the different data sets around the business

Cleansing: Finding and cleaning any errors in the data

Conversion: Giving all the data a common format

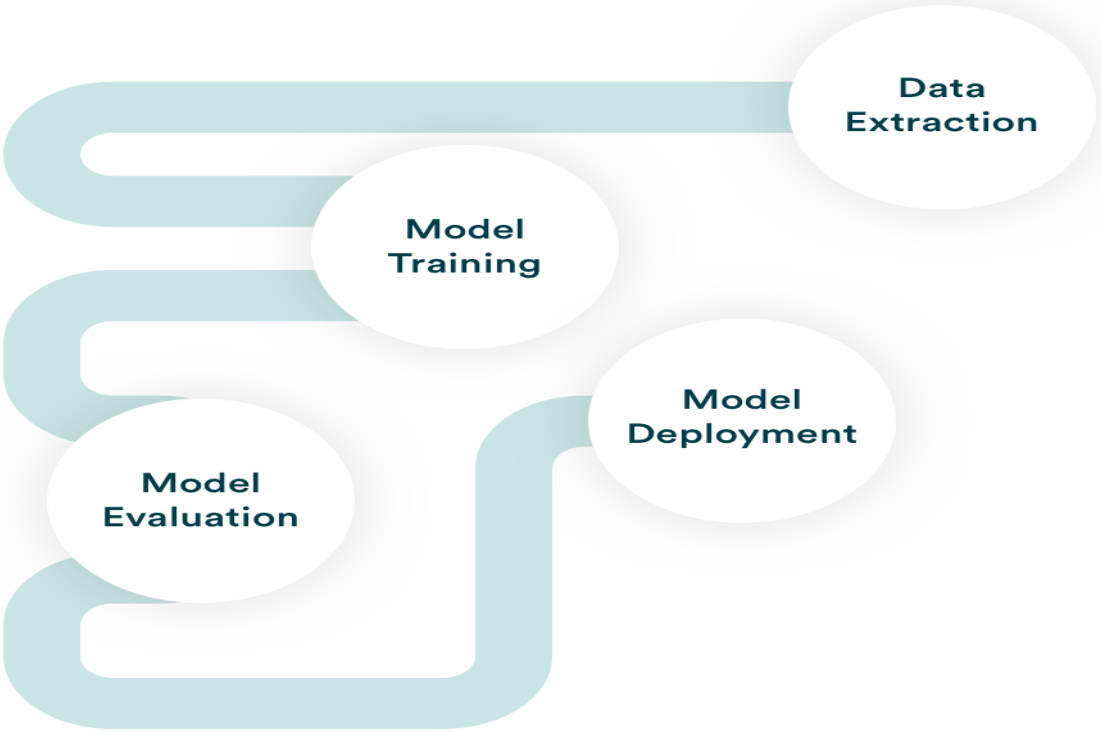
Disambiguation: Interpreting data that could be interpreted in multiple ways

Deduplication: Removing duplicate copies of data

Once this is done, data may be stored in a central repository such as a data lake or data lakehouse. Data engineers may also copy and move subsets of data into a data warehouse.

Machine Learning pipeline

A machine learning pipeline is a way to codify and automate the workflow it takes to produce a machine learning model. Machine learning pipelines consist of multiple sequential steps that do everything from data extraction and preprocessing to model training and deployment.



What is Data Science?

Data science is **the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions**. For example, finance companies can use a customer's banking and bill-paying history to assess creditworthiness and loan risk.

How Data Science works ?

Data science uses techniques such as **machine learning and artificial intelligence to extract meaningful information and to predict future patterns and behaviours**. Advances in technology, the internet, social media, and the use of technology have all increased access to big data.

Data Science uses

data science is used in **marketing, finance, and human resources, healthcare, government programmes, and any other industry that generates data**. Marketing departments use data science to determine which product is most likely to sell.