



AI/ML Project Lifecycle

JARGONS

10 Stages of AI Project Lifecycle

- 1 EDA:** Exploratory data analysis. The process of exploring the data to identify the patterns and structures.
- 2 Feature Engineering:** The process of selecting the right attributes of the data & transforming it to train the model, this step is like selecting the right ingredients for making the required pizza.

Requirements and Scope of Work (SOW)

- 1 SOW:** Scope of Work. This is typically the first step in any project where the boundaries of the work are defined.
- 2 KANBAN:** One of the popular Agile techniques where the task is created as a board and moved across different stages of progress.

Data Collection

- 1 PII:** Personally Identifiable Information. PII is generally masked in a database to protect the identity of the person.
- 2 Web Scraping:** Process of extracting data from the website.
- 3 Beautiful Soup:** A Python library that is used to extract data from the website.

Data Preparation & Exploratory Data Analysis

- 1 EDA:** Exploratory Data Analysis is an approach to analyzing datasets to summarize their main characteristics, often using statistical graphics and other data visualization methods.
- 2 IDE:** Integrated Development Environment. Jupyter notebook is an IDE which can be used to develop a python program.
- 3 Apache Spark:** It is an open-source distributed computing system designed for big data processing and analytics.
- 4 Map and Reduce:** Map is the process used in distributed computing to assign work to different computers on cloud and Reduce is the process of collecting individual results and aggregating them to final answer.
- 5 Pyspark:** Python library for Apache Spark framework.
- 6 DataFrame:** A virtual table where you can store and organize data in rows and columns, just like how you might organize information in a chart or spreadsheet.

Feature Engineering

- 1 Feature:** A feature is a characteristic of an object or entity that helps you identify that object or entity. For example, pointy ears, round ears, and whiskers are the features of a cat.
- 2 Feature Engineering:** Feature Engineering is a process of transforming raw data into meaningful features (which can be new columns) such that these features help in improving the performance of the model being trained. Domain understanding and math/statistics can be used for doing feature engineering.
- 3 Scikit Learn:** A library used to perform feature learning in machine learning problems.
- 4 Scaling:** A technique used to adjust the data points to a scale that can easily interpreted such as a scale between 0 to 1.

Model Selection & Training

- 1 Train – Test – Split:** A popular technique used in ML projects where the original dataset is split into two parts namely train dataset and test dataset. The most popular split is 80% (test) and 20% (train).
- 2 Train Dataset:** This dataset is used exclusively for training the model and none of the data points used here are repeated in the test dataset.
- 3 Test Dataset:** This dataset is used to test the model after the model is trained.

Model Evaluation Metrics (Accuracy, Prediction, Recall & F1 Score)

- 1 Accuracy:** The percentage of correct predictions made with respect to overall predictions.
- 2 Precision:** The percentage of correct predictions made for a given class (e.g. class="dog") with respect to all predictions which resulted in the value of a given class (e.g. class="dog").
- 3 Recall:** The percentage of correct predictions made for a given class (e.g. class="dog") with respect to total number of instances of the class (e.g. class="dog").
- 4 F1-Score:** F1-Score is a harmonic mean of precision & recall, providing a single metric that balances both.

Model Evaluation Metrics: When to use which Metric?

- 1 K-Fold Cross-Validation:** K-Fold Cross-Validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The dataset is divided into k number of subsets (or folds), and the model is trained on $k-1$ folds and tested on the remaining fold. This process is repeated k times, with each fold used exactly once as the test set, providing a robust estimate of model performance.

Model Fine Tuning

- 1 Model Fine-Tuning:** The process of optimizing to get a combination of parameters that gives the best-performing model. Fine tuning also means using pre-trained generic models to train them on a small dataset which is domain-specific.
- 2 Hyperparameters:** Configurations that influence the learning rate, hidden layers in a neural network which can be optimized to improve the model performance and behavior.
- 3 Hyperparameter Tuning:** Process of figuring out the best configuration for the model which will result in the best model performance.
- 4 Grid Search CV:** Grid Search Cross Validation: A technique used in Hyperparameter tuning. Here a predefined set of values are passed into the model to get different combinations and cross-validated to evaluate each combination's performance in order to select the best possible combination.
- 5 Transfer Learning:** Transfer learning is a machine learning technique where a pre-trained model is used as the starting point for training a new model on a different but related task, often resulting in faster training and improved performance.

Model Deployment

- 1 API:** Application Program Interface. It is a set of rules and protocols that allows different software applications to communicate with each other, enabling data exchange and interaction between systems or services.
- 2 API End Points:** A specific URL that allows someone to fetch specific information from a service provider. Each endpoint is mapped to a specific service provided by the API.
- 3 ML OPS:** The process of streamlining / automating the end-to-end deployment and monitoring of a machine learning model.

Deployment & Monitoring Using ML Ops

- 1 Annotation:** Process of providing additional information to the key data points including labels and classes. This will help the machine to train better.
- 2 Version Control System:** A system like GitHub that tracks and stores the changes to model over a period.
- 3 Batch Inference:** A machine learning process where predictions are made on a set of data points simultaneously, often for efficiency in handling large datasets.
- 4 ML OPS:** The process of streamlining / automating the end-to-end deployment and monitoring of a machine learning model.
- 5 Near Real-Time Inference:** Process of making predictions on new data with minimal delay, typically within seconds or milliseconds of data arrival.
- 6 Kubernetes:** It is an open-source platform that provides tools to automate load balancing, resource allocation and deployment.
- 7 Serverless Architecture:** A model in which the server capacity is scaled up or down automatically based on the need which reduces the manual server load management time.
- 8 HITL:** Human-in-the-Loop. It is a machine learning approach that combines automated algorithms with human expertise to improve model accuracy and reliability by incorporating human feedback during the training or inference process.
- 9 Data Drift:** A phenomenon in which the statistical properties of the datapoints become outdated leading to model degradation. New and relevant dataset is added to the model to avoid this phenomenon.