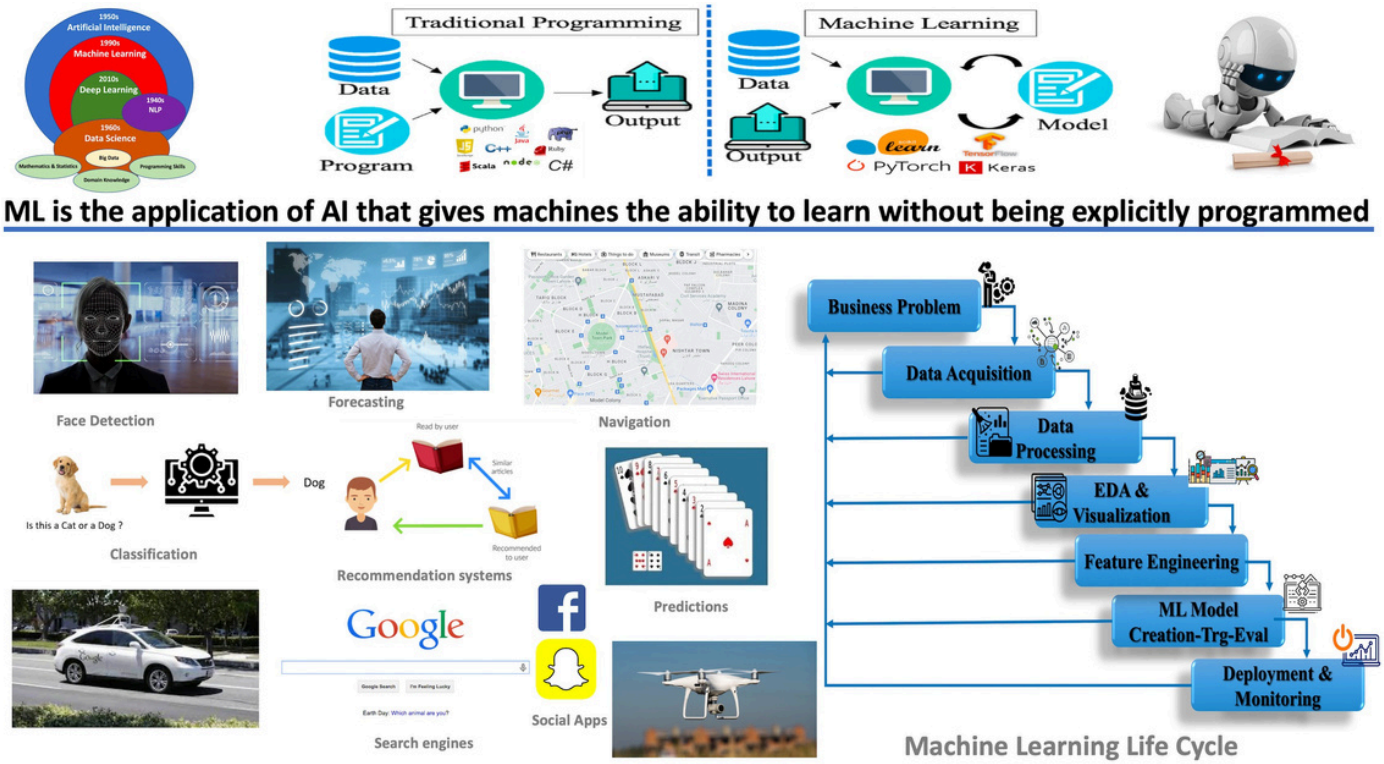
Introduction to Machine Learning

Instructor: Ehtisham Sadiq

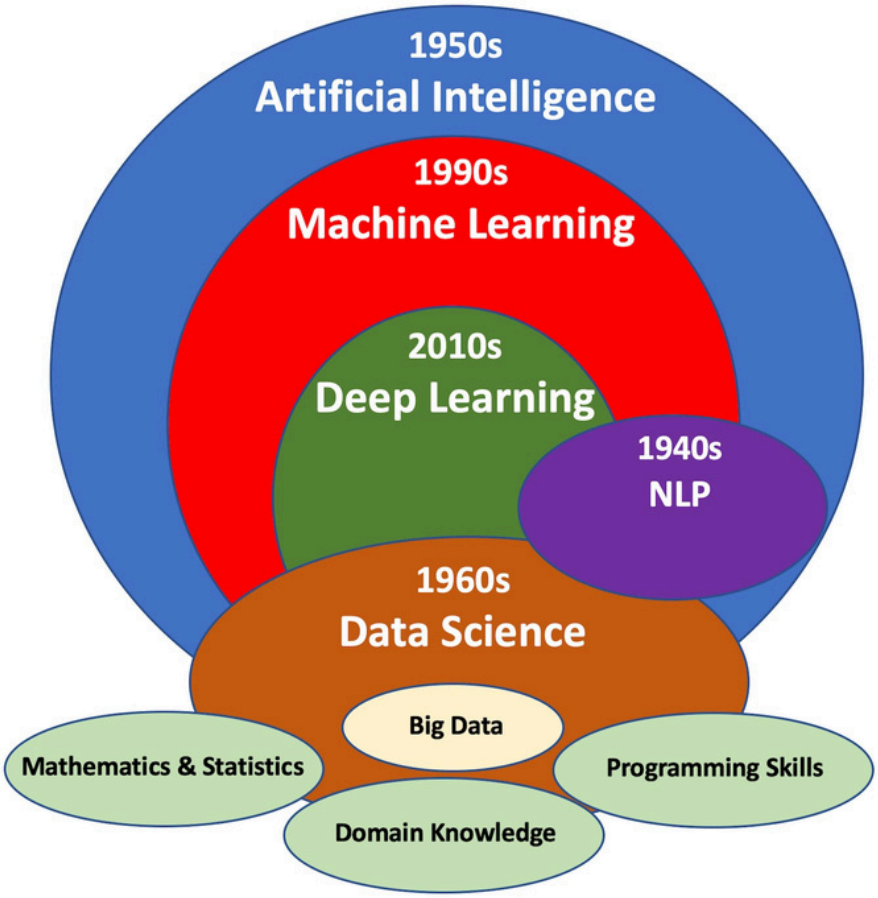


Learning agenda of this notebook

1. The Big Picture
2. Why to do Machine Learning? (Applications of ML)
3. What is Machine Learning and its Types?
4. How to do Machine Learning? (ML algorithms)
5. Machine Learning Development Life Cycle

1. The Big Picture

- **Artificial Intelligence:** AI is the intelligence demonstrated by machines to sense, act, reason, and adapt like humans.
  - Narrow A.I. make machines do one thing really well.
  - General A.I. make machines having multiple abilities like humans.



- **Machine Learning:** ML is the application of AI that gives machines the ability to learn without being explicitly programmed. A study of statistical computer algorithms that improve automatically through experience/data.

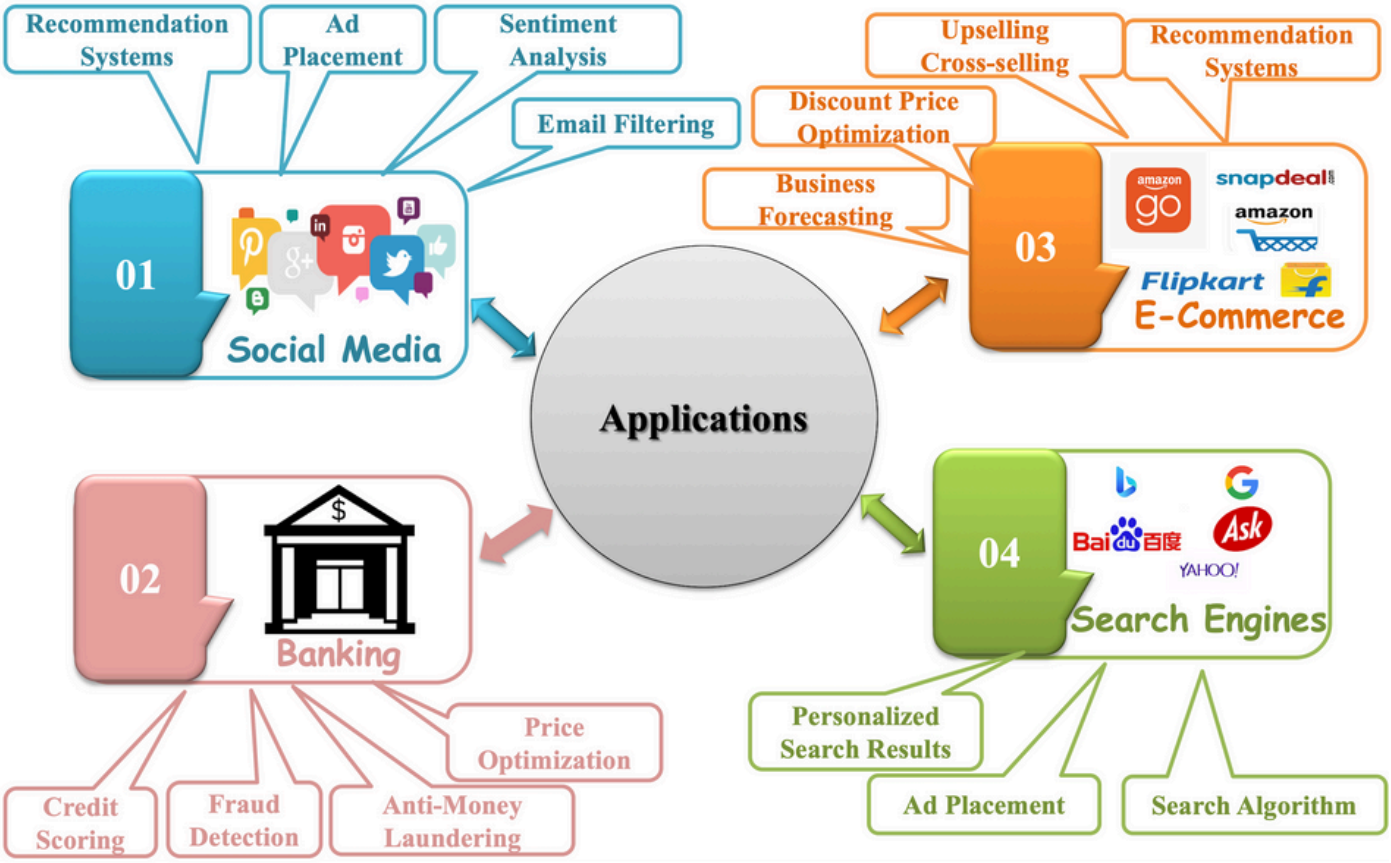
- **Deep Learning:** DL is a subfield of machine learning that attempts to mimic the human brain and use the concept of neurons or perceptrons.

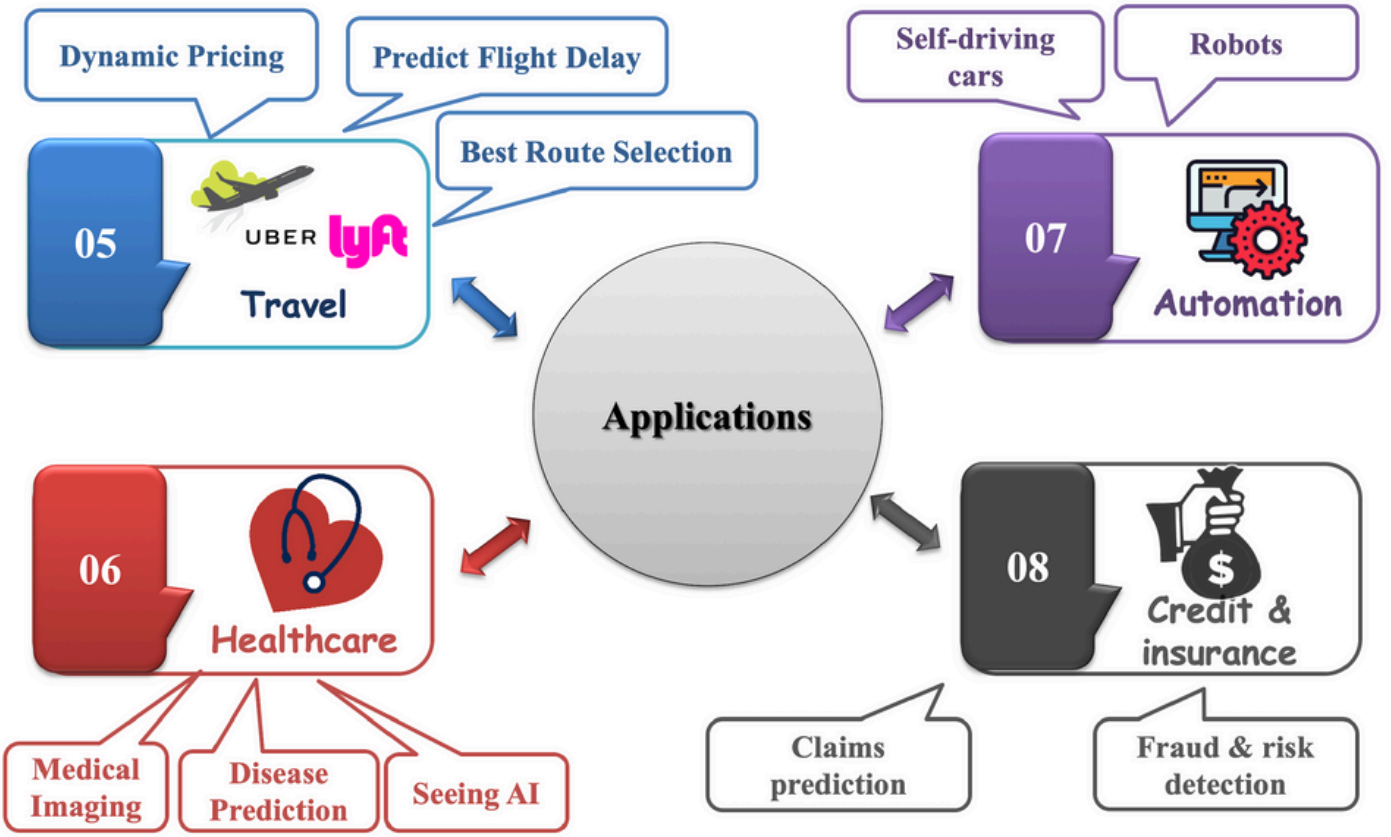
- **Natural Language Processing:** NLP enables machines to understand and respond to human language(s) in written, verbal and sign form.

- **Data Science:** DS is not exactly the subset of ML, but uses ML and DL to gain insights from both structured and unstructured data. DS is an inter disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data related to a business problem.

- **Big Data:** Big Data is a combination of structured, semistructured and unstructured data that is huge in volume, yet growing exponentially with time. It can be mined for information and used in machine learning projects, predictive modeling and other advanced analytics applications.

2. Why to do Machine Learning? (ML Applications)



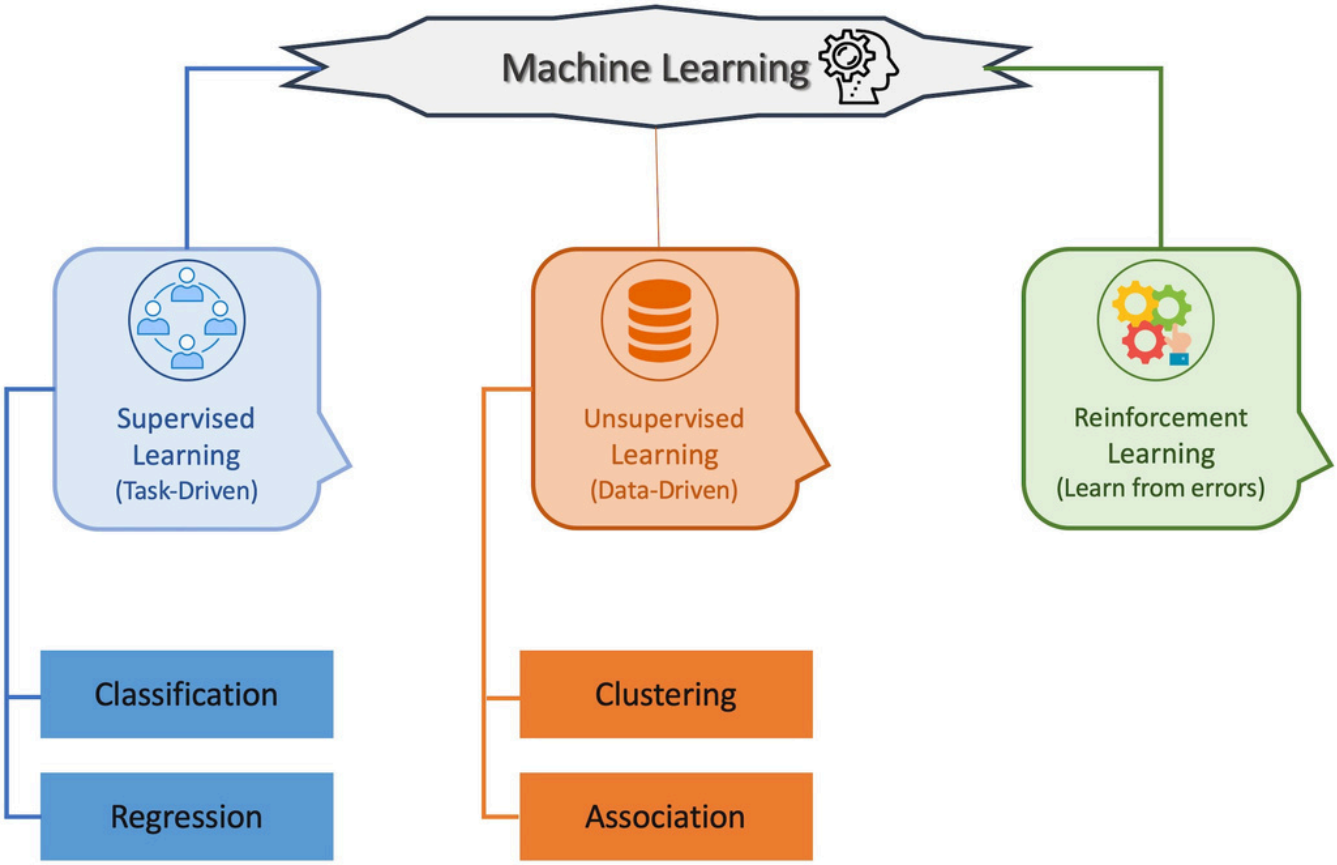


3. What is Machine Learning?

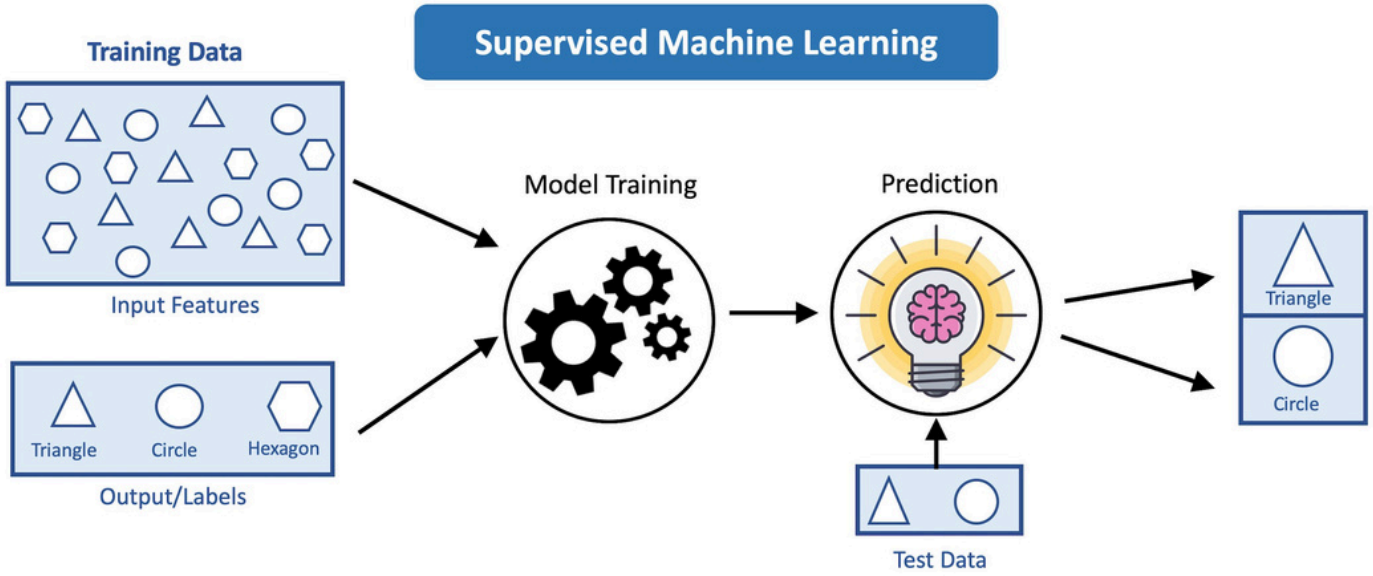
> An application of AI that gives machines the ability to learn without

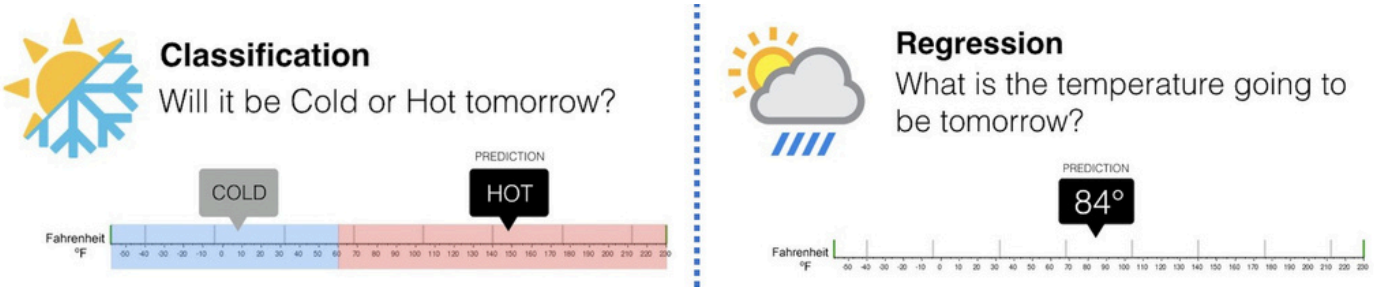
being explicitly programmed

a. Types of Machine Learning

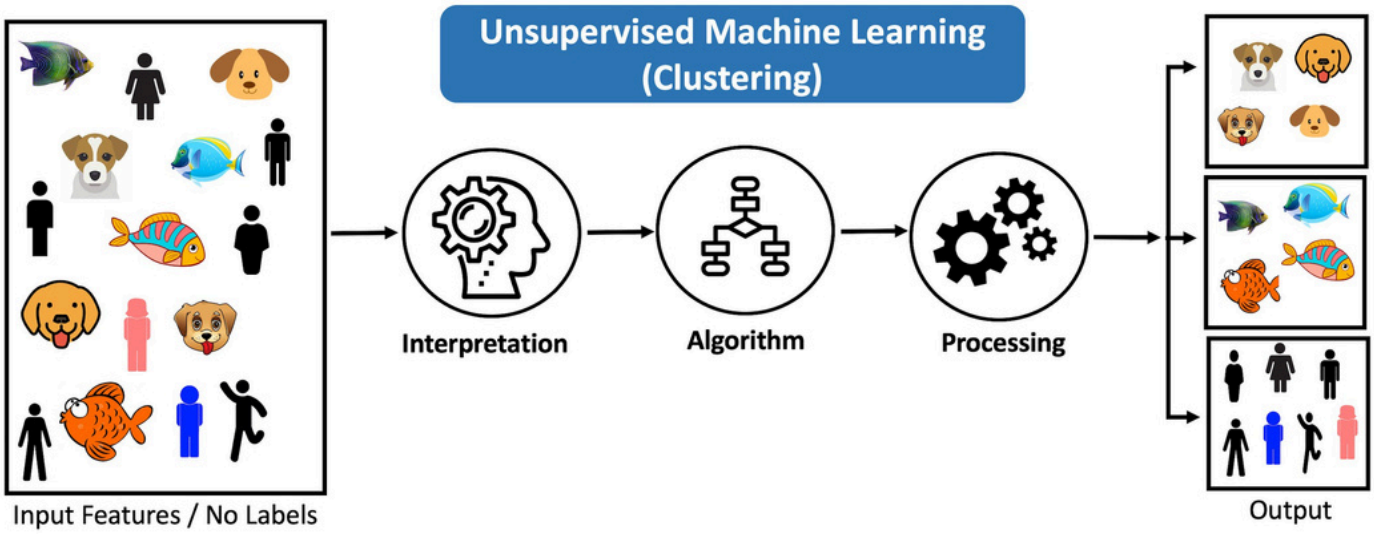


b. Supervised Machine Learning (Classification vs Regression)

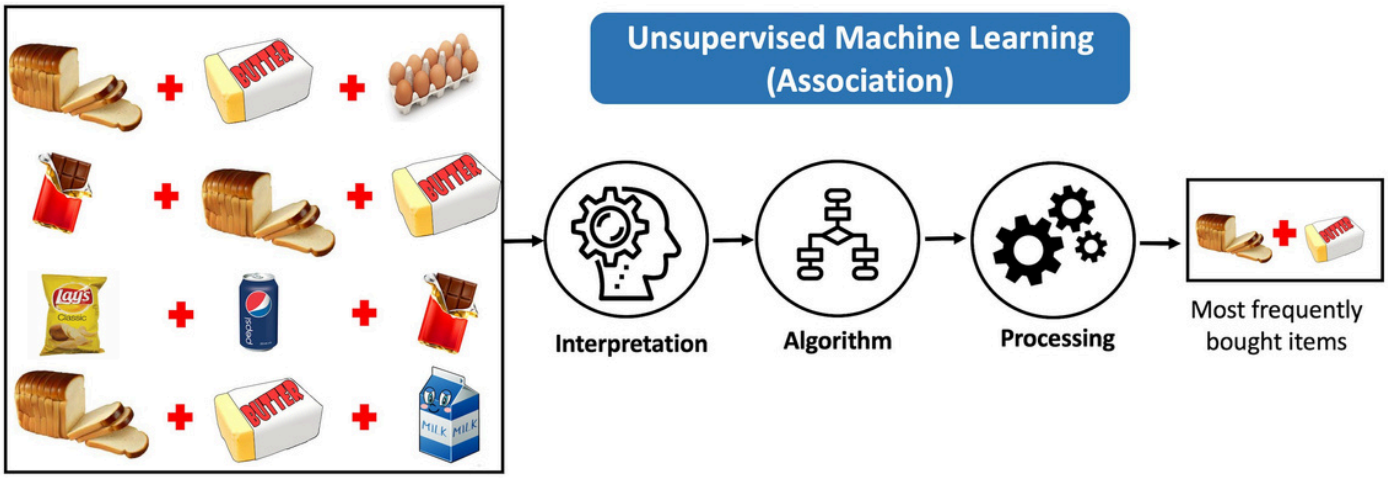




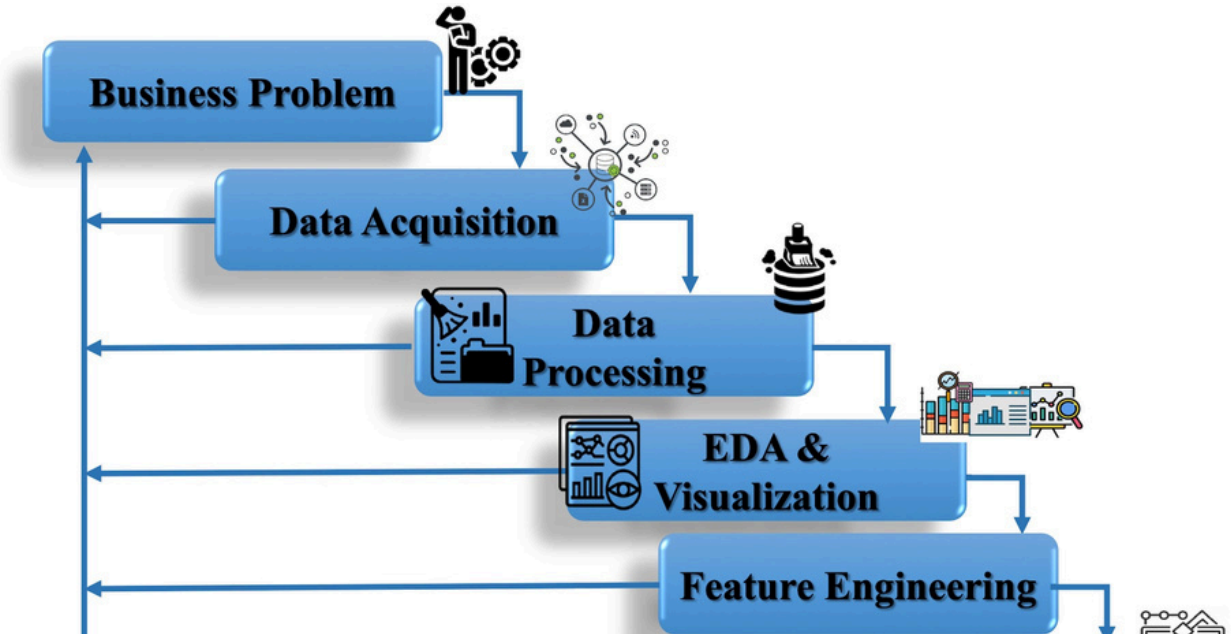
c. Unsupervised Machine Learning (Clustering)



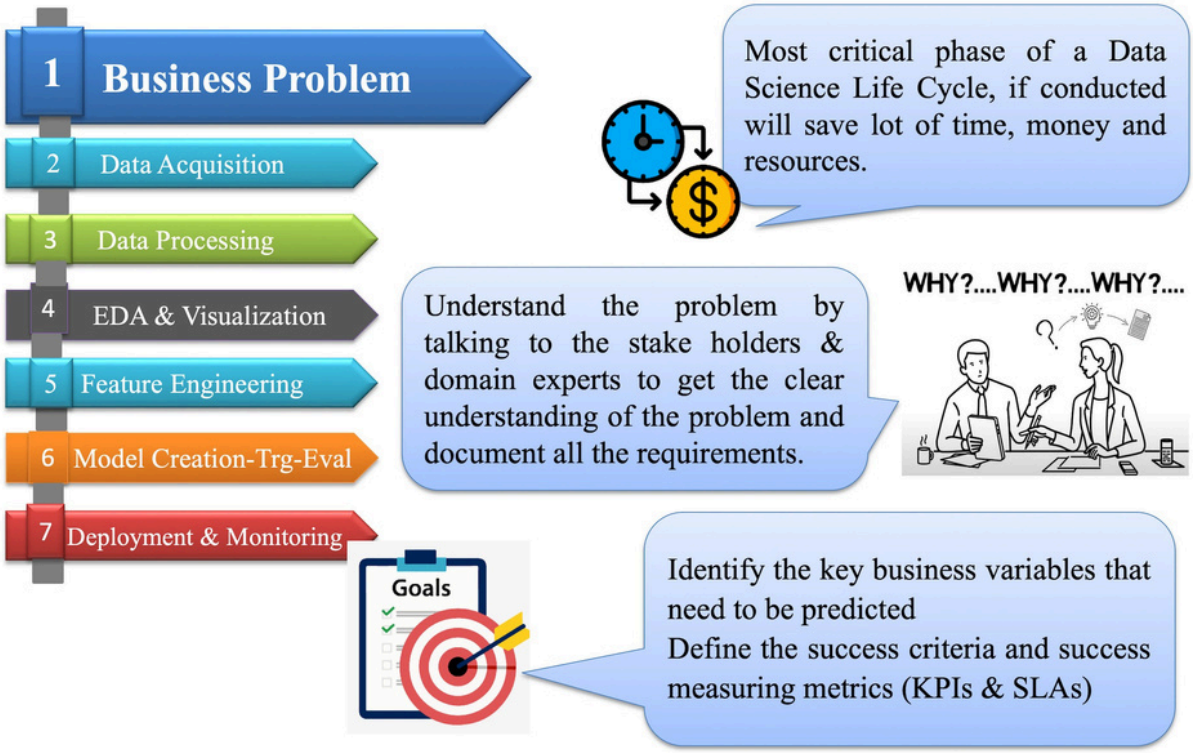
d. Unsupervised Machine Learning (Association)



5. Machine Learning Development Life Cycle

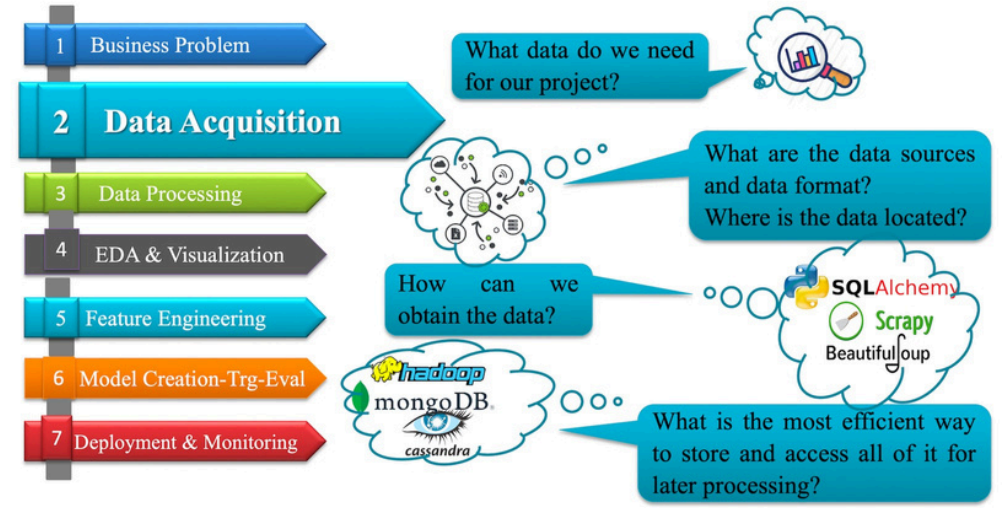


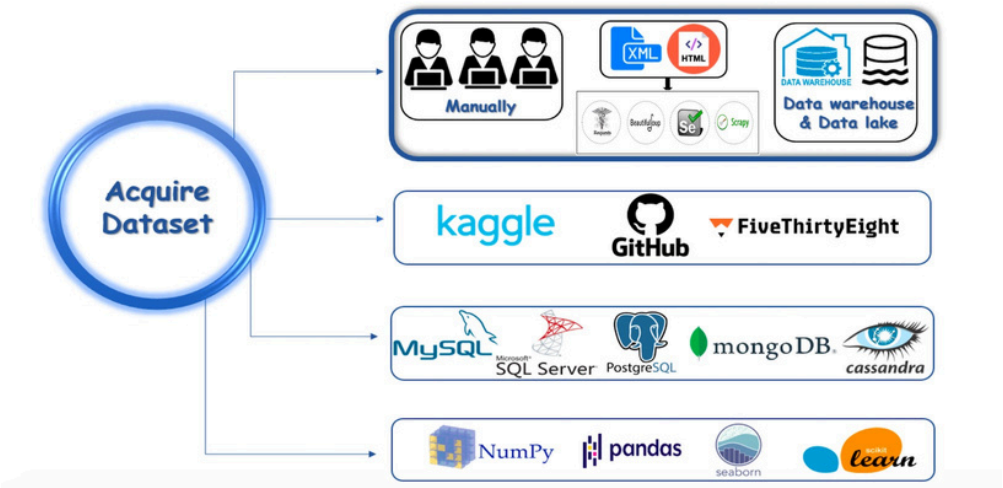
(i) Understand the Business Problem



(ii) Data Acquisition

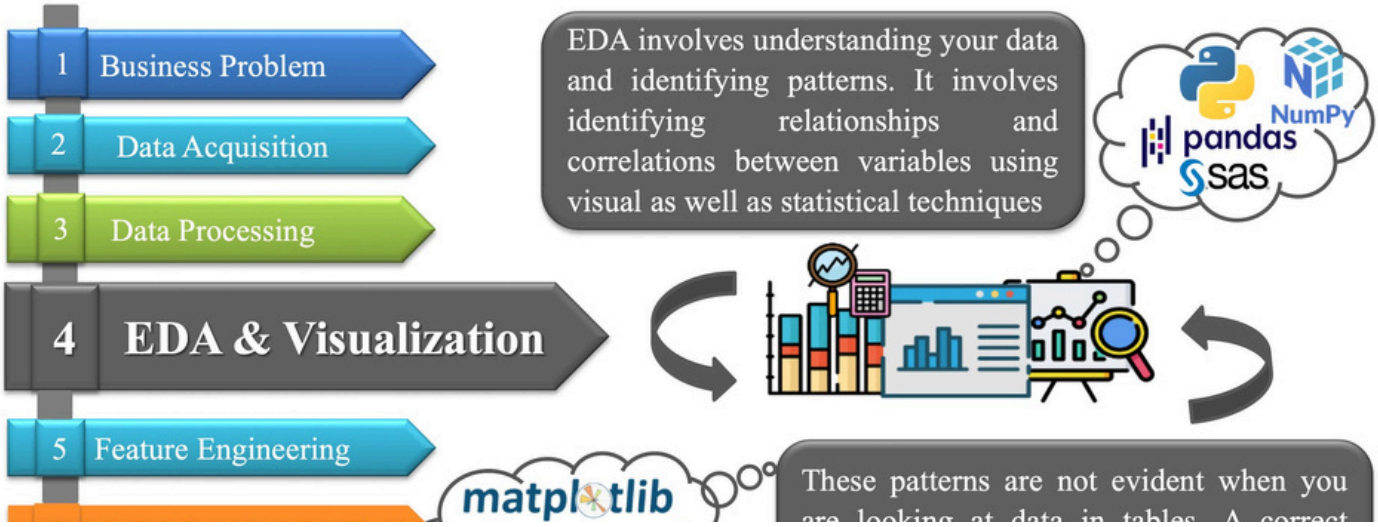
- **Use Libraries Built-in Datasets:**
  - Seaborn: (iris, titanic, tips, flights, panguins, car_crashes)
  - Scikit-learn: (iris, digits, diabetes, Bostan housing)
  - NLTK: (movie-reviews, product_reviews, twitter_samples, gutenberg, genesis, timeit, voice, wordnet, sentiword)
- **Use Public Dataset Repositories:**
  - ~~https://www.kaggle.com/ (https://www.kaggle.com/)~~
  - ~~https://data.gov/ (https://data.gov/)~~



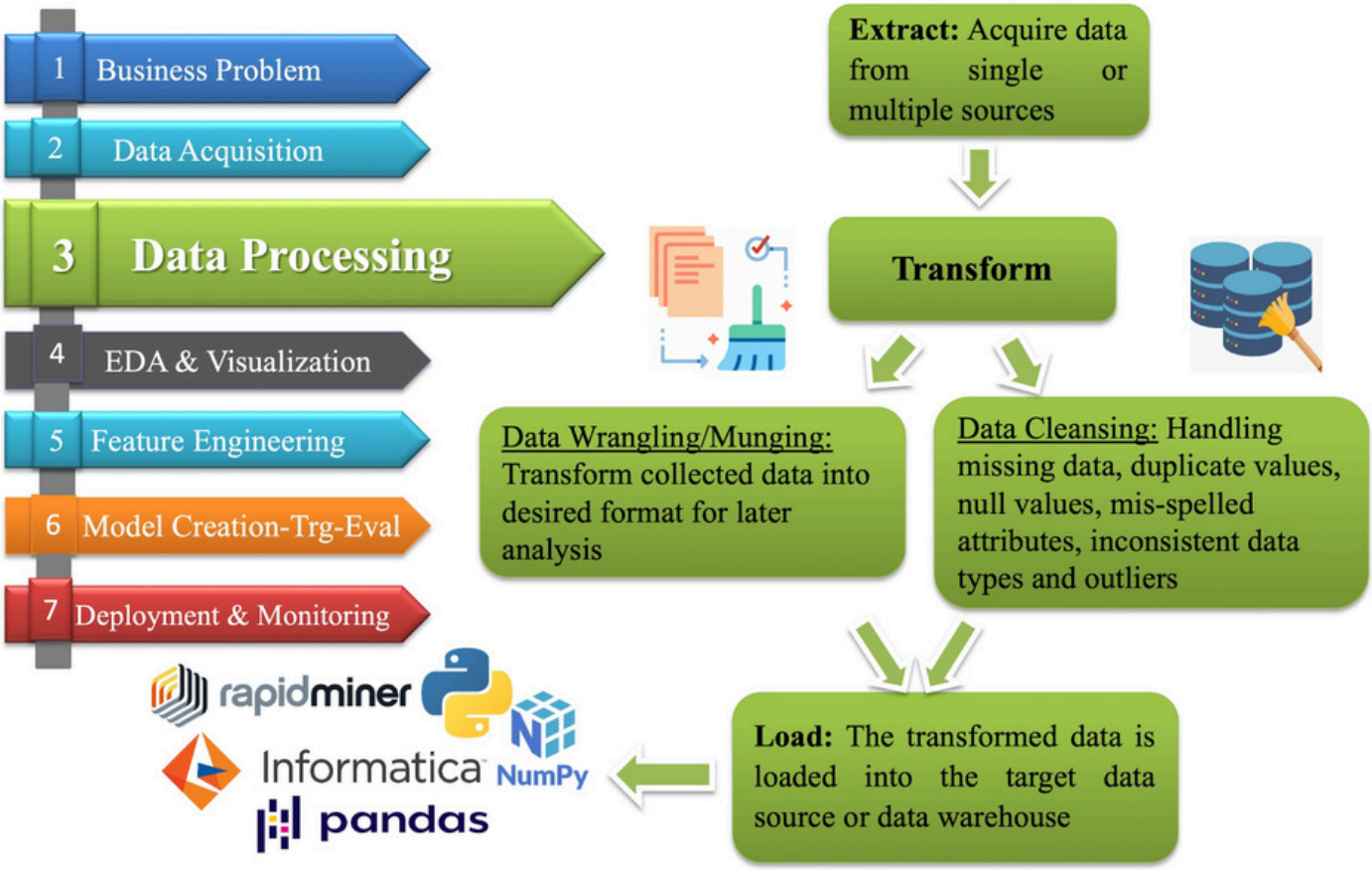


  - https://archive.ics.uci.edu/ml/index.php (https://archive.ics.uci.edu/ml/index.php)
  - https://github.com/ (https://github.com/)
- **Use Company's Database:** (SQL, NOSQL, Data warehouse, Data lake)
- **Generate your own Datasets:**
  - Use Web scraping or Web API
- IoT Devices
- Crowd Sourcing (Amazon Mechanical Turk, Lionbridge AI)
- Data Augmentation
-

(iii) Exploratory Data Analysis and Visualization



(iv) Data Pre-Processing

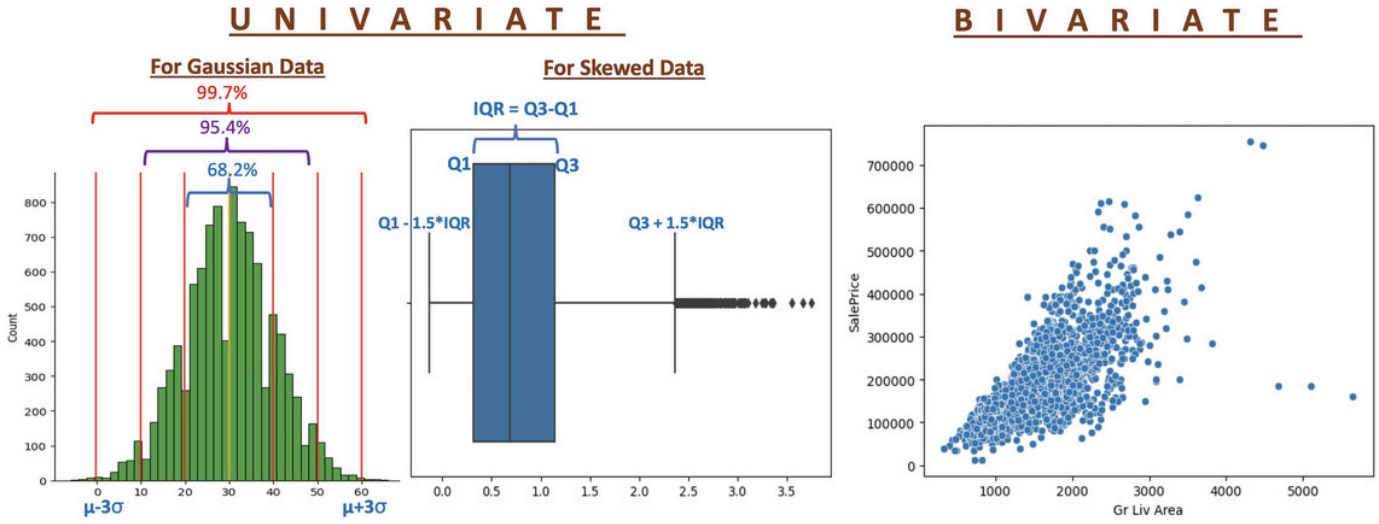


City	Size	Covered Area	No of bedrooms	Trees near by	No of bathrooms	Schools near by	Construction Date	Price
Lahore	2000	3500	3	1	3	1	25/10/2001	20.5 M
Karachi	2600	3000	2	0	4	1	16/05/1990	18 M
Islamabad	1800	2000	3	1	3	2	25/11/1995	20 M
Shaikhupura	1600	2600	1	2	NaN	0	08/06/2020	5 M
Lahore	2600	2000	3	3	1	1	03/09/2016	4 M
Karachi	3000	1000	2	2	1	NaN	19/01/1980	6 M
Islamabad	2000	3600	44	4	3	3	21/07/1999	30 M
Lahore	1000	2000	3	NaN	1	2	12/04/2015	10 M

a. Detecting and Handling Outliers

An outlier is a data point that differs significantly from other observations

- An outlier can be a result of a mistake during data collection in which case it must be handled appropriately.
- If your machine learning model is going to detect anomalies in a banking transaction, then you must NOT remove the outliers because this is what the model will be looking for later.
- All distance based models (KNN, K-Means, SVM) are sensitive to outliers, while tree based algorithms (Decision Tree, Random Forrest and Gradient Boosting) are insensitive to outliers.
- **How to identify outliers?**
  - Z Score Method (For normally distributed data)
  - IQR Method (For skewed distribution)
  - Percentiles Method (For other distributions)
  - Multivariate Analysis using Scatter Plot



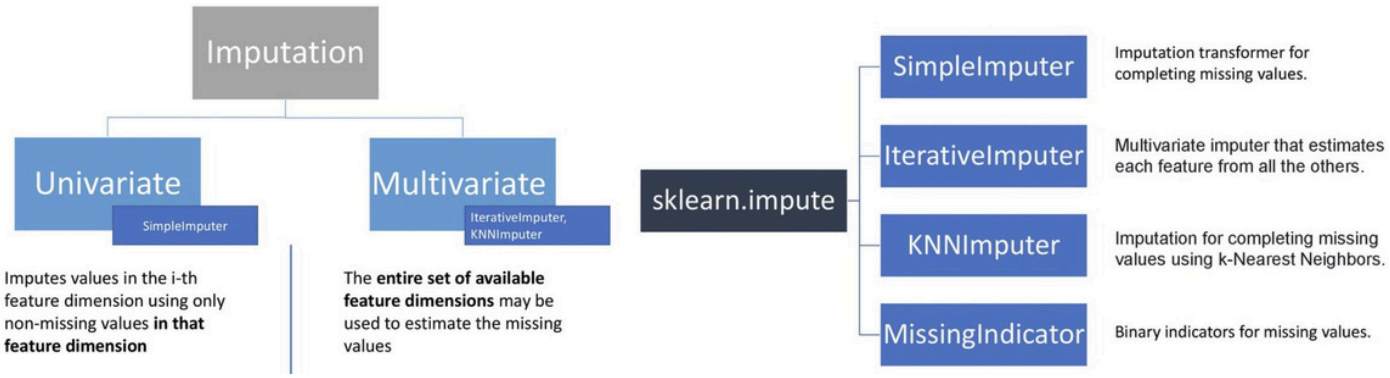
- **How to treat outliers?**
  - Trimming (Remove)
  - Capping/Winsorization (Replace with upper value)

- Discretization

b. Detecting and Handling Missing Values

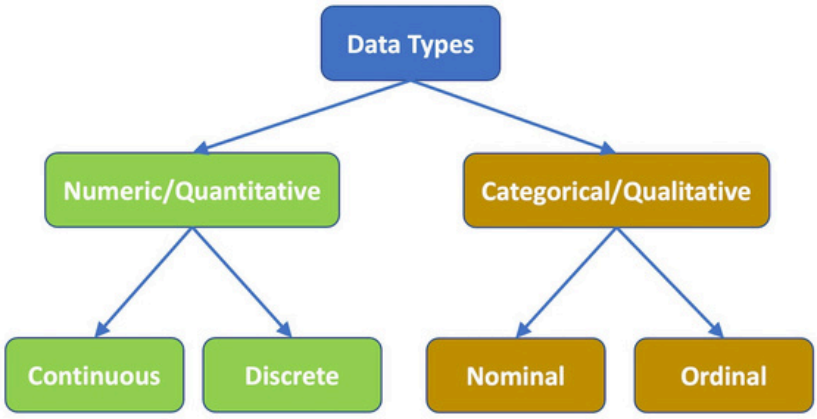
Missing data, or missing values, occur when no data value is stored for the variable in an observation

- **Deleting the Missing values:** Drop rows (List-wise deletion) having missing values or drop the entire column
- **Impute/Replace the Missing Values:**
  - **Univariate Imputation:**
    - Handling Missing Values using Panda's fillna() method
    - Handling Missing Values using sklearn's SimpleImputer() transformer
    - `Use of ColumnTransformer`
  - **Multivariate Imputation:**
    - Handling Missing Values using sklearn's IterativeImputer() transformer
    - Handling Missing Values using sklearn's KNNImputer() transformer



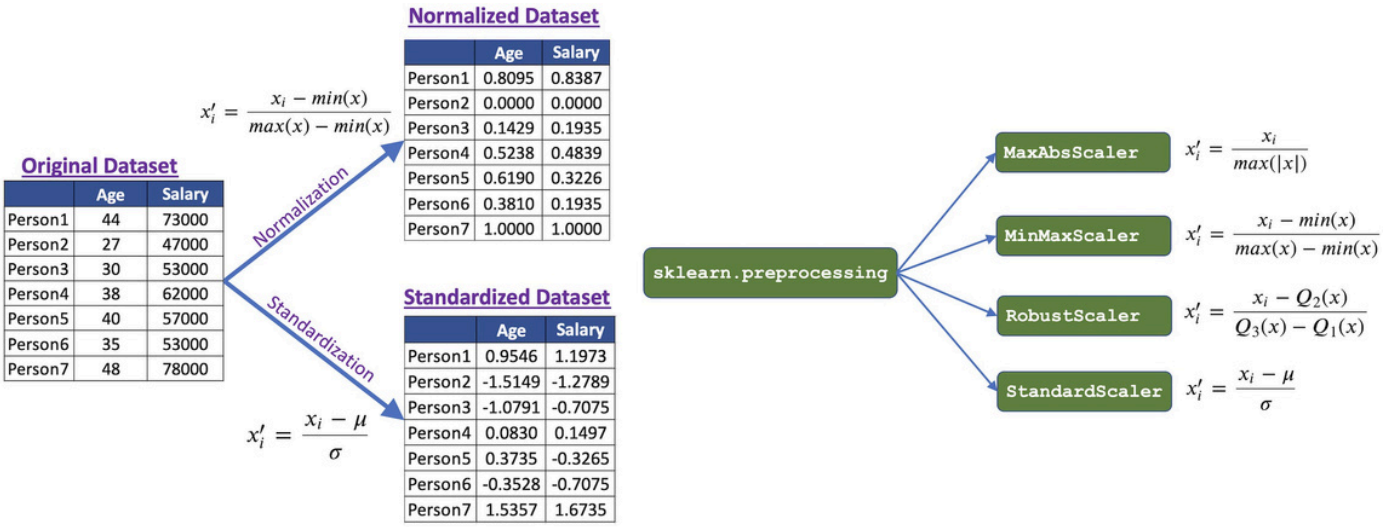
c. Encoding Categorical Features

Encoding categorical data is a process of converting it into numerical values, so that it could be fed to machine learning models



d. Feature Scaling

Many machine learning algorithms perform better or converge faster, when features are on a relatively similar scale and close to normally distributed



	Age	Salary
Person1	44	73000
Person2	27	47000
Person3	30	53000
Person4	38	62000
Person5	40	57000
Person6	35	53000
Person7	48	78000

	Age	Salary
Person1	0.8095	0.8387
Person2	0.0000	0.0000
Person3	0.1429	0.1935
Person4	0.5238	0.4839
Person5	0.6190	0.3226
Person6	0.3810	0.1935
Person7	1.0000	1.0000

	Age	Salary
Person1	0.9546	1.1973
Person2	-1.5149	-1.2789
Person3	-1.0791	-0.7075
Person4	0.0830	0.1497
Person5	0.3735	-0.3265
Person6	-0.3528	-0.7075
Person7	1.5357	1.6735

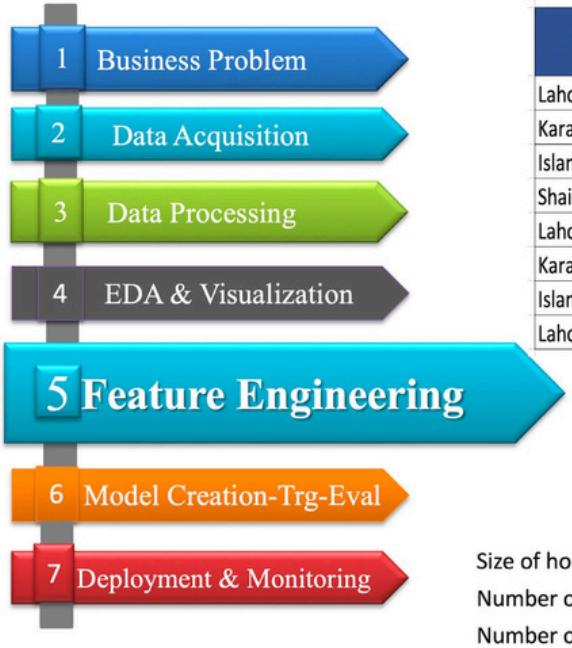
- **Normalization** is rescaling of the data from original range, so that all values are within the new range of 0 and 1
- **Standardization** is rescaling of the data from original range, so that all values are centered around mean of zero with a standard deviation of 1.

> - All **Distance based ML algorithms** (KNN, K-Means, SVM) require feature scaling, and **Gradient Descent based ML algorithms** (Linear Regression, Logistic Regresion, Neural Network) require feature scaling.
> - All **Tree based algorithms** like Decision Tree, Random Forrest and Gradient Boosting, as well as Naive Bayes and Linear Discriminant Analysis DO NOT require feature scaling.

e. Feature Engineering

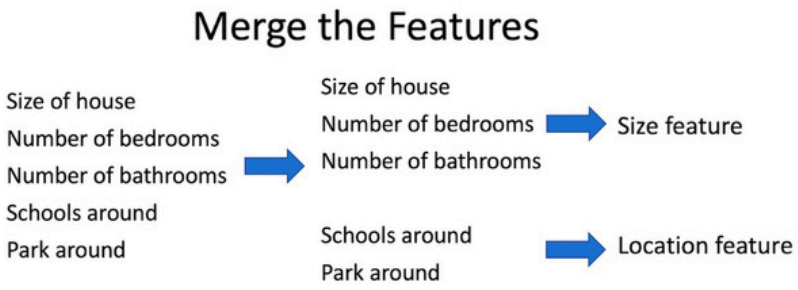
Feature Engineering is the process of using domain knowledge to extract features from raw data via data mining techniques

- **Extracting Information**
- **Combining Information**
- **Transforming Information**



Housing Data Set

City	Size	Covered Area	No of bedrooms	Trees near by	No of bathrooms	Schools near by	Construction Date	Price
Lahore	2000	3500	3	1	3	1	25/10/2001	20.5 M
Karachi	2600	3000	2	0	4	1	16/05/1990	18 M
Islamabad	1800	2000	3	1	3	2	25/11/1995	20 M
Shaikhupura	1600	2600	1	2	0	0	08/06/2020	5 M
Lahore	2600	2000	3	3	1	1	03/09/2016	4 M
Karachi	3000	1000	2	2	1	0	19/01/1980	6 M
Islamabad	2000	3600	4	4	3	3	21/07/1999	30 M
Lahore	1000	2000	3	0	1	2	12/04/2015	10 M

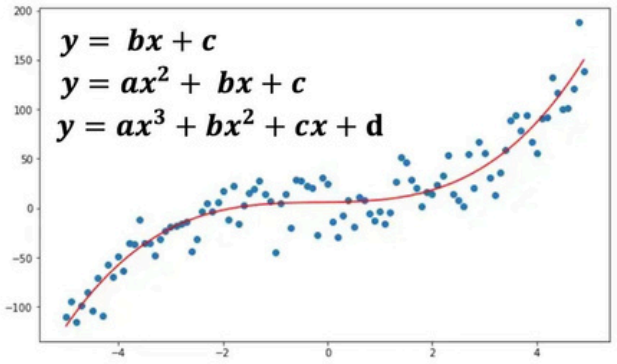


f. Add-Ons

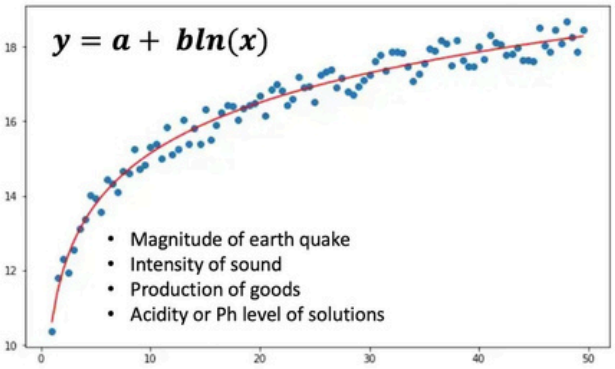
Polynomial Regression

It is a form of linear regression, which estimates the relationship between X and y as an thn degree polynomial in X.

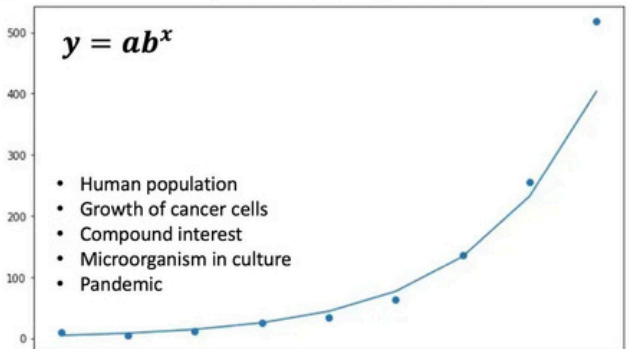
Polynomial Regression



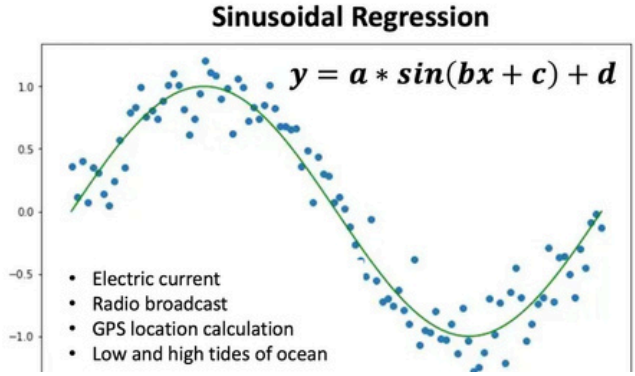
Logarithmic Regression



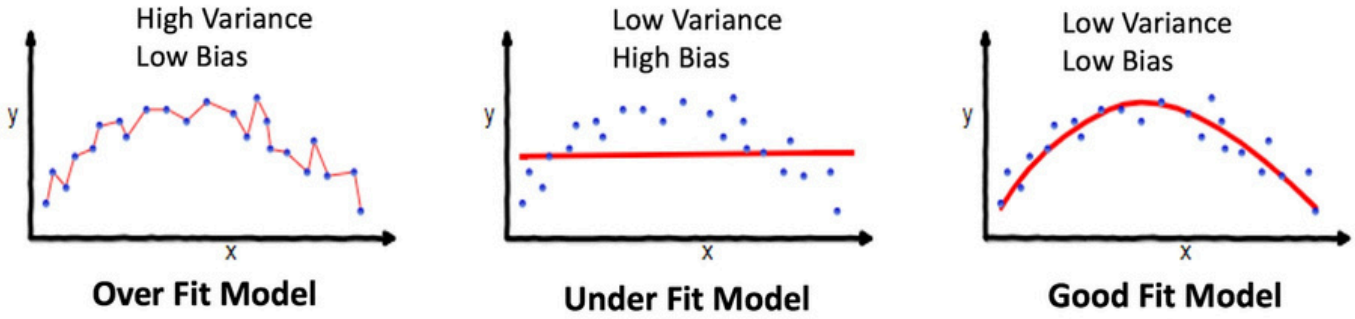
Exponential Regression

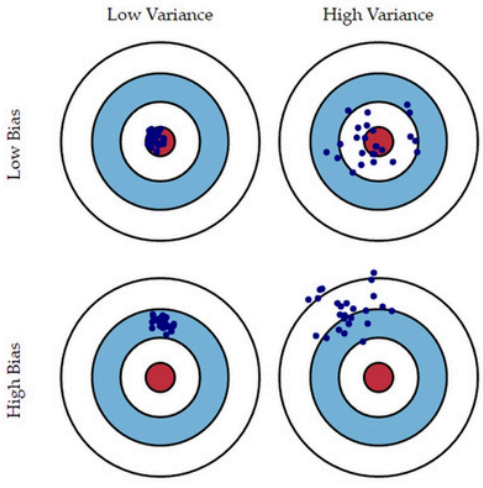


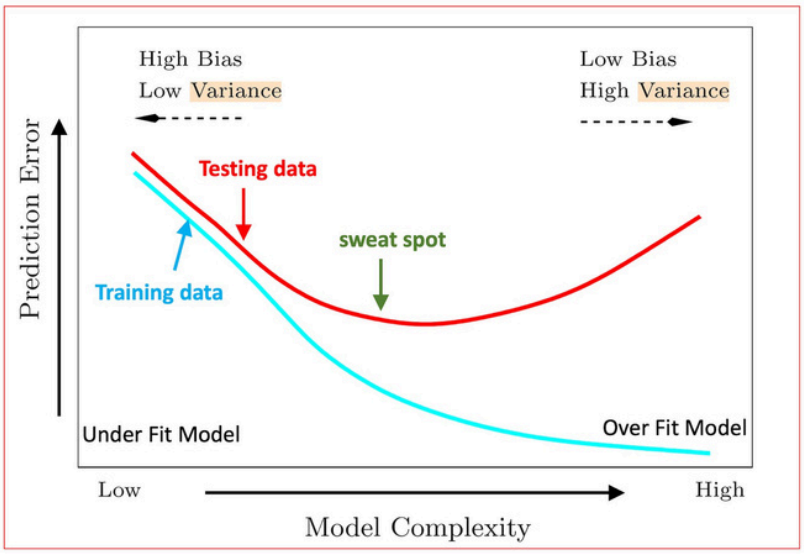
Sinusoidal Regression



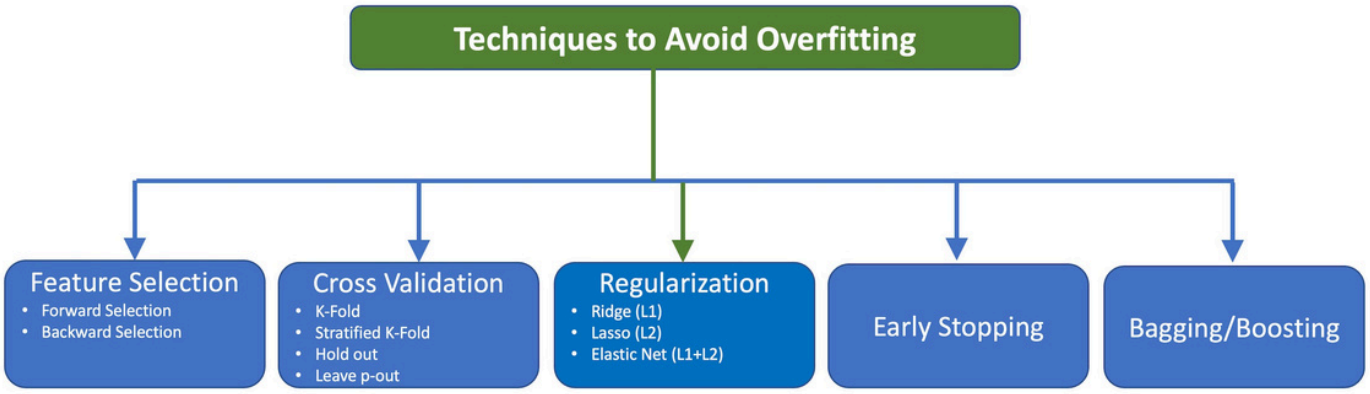
Bias-Variance Tradeoff and Overfitting



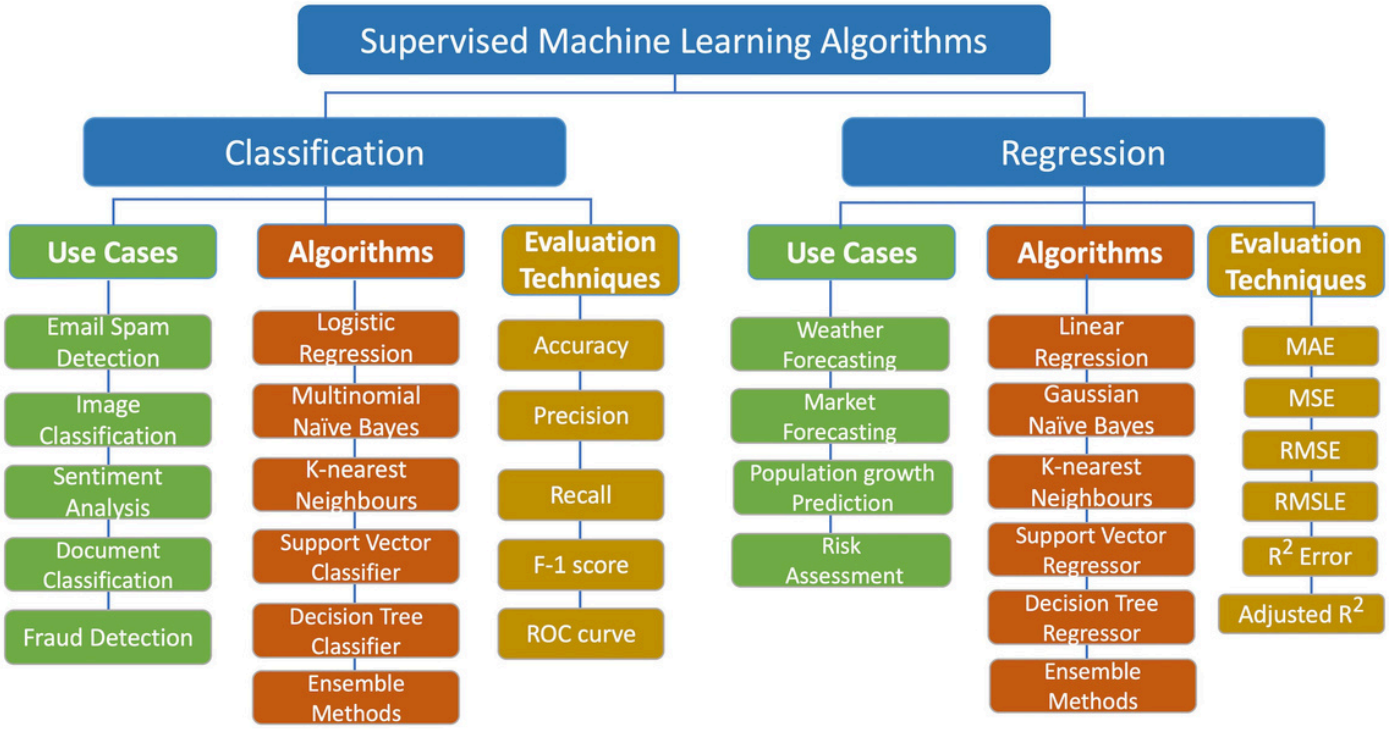


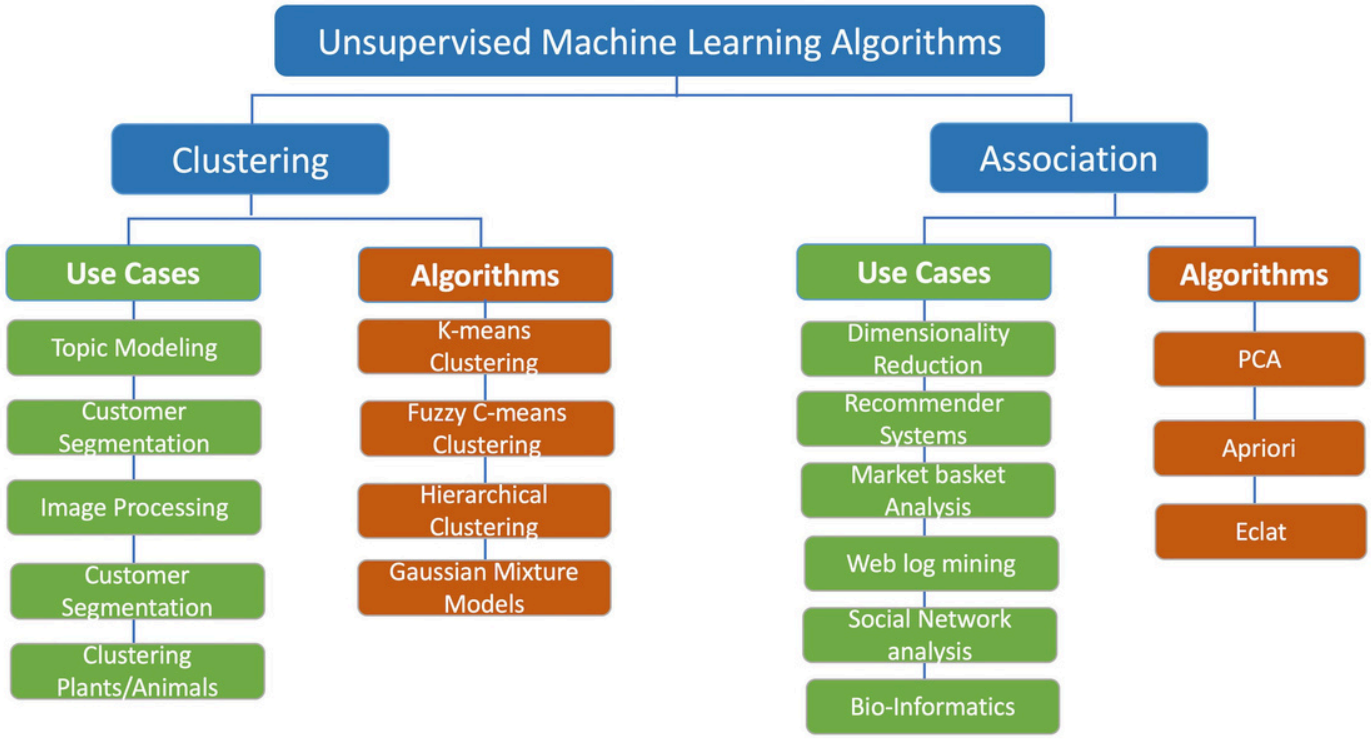


How to prevent overfitting?

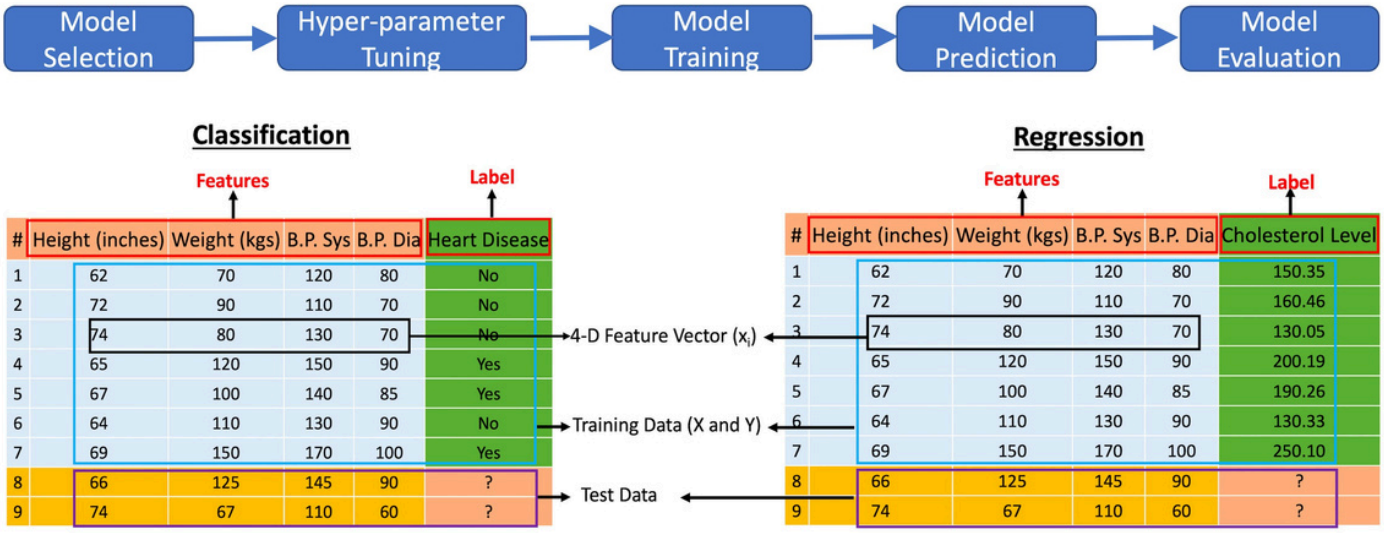


(v) Machine Learning Models





Supervised Machine Learning Model Creation-Training-Evaluation



$$D = \{(\vec{x_1}, y_1), (\vec{x_2}, y_2), (\vec{x_3}, y_3), \cdots (\vec{x_n}, y_n)\} \subseteq X \times Y$$

Add-Ons

Hyperparameter Tuning

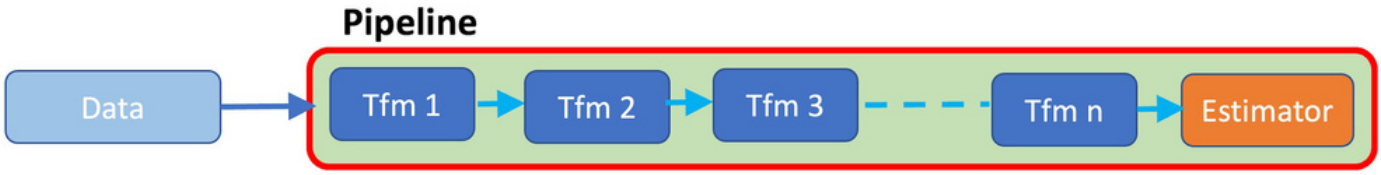
Ser	Hyperparameters	Model Parameters
1	Hyperparameters are set manually by ML engineer/practitioner prior to the start of the model's training.	Model parameters are learnt by the learning algorithm during the training phase.

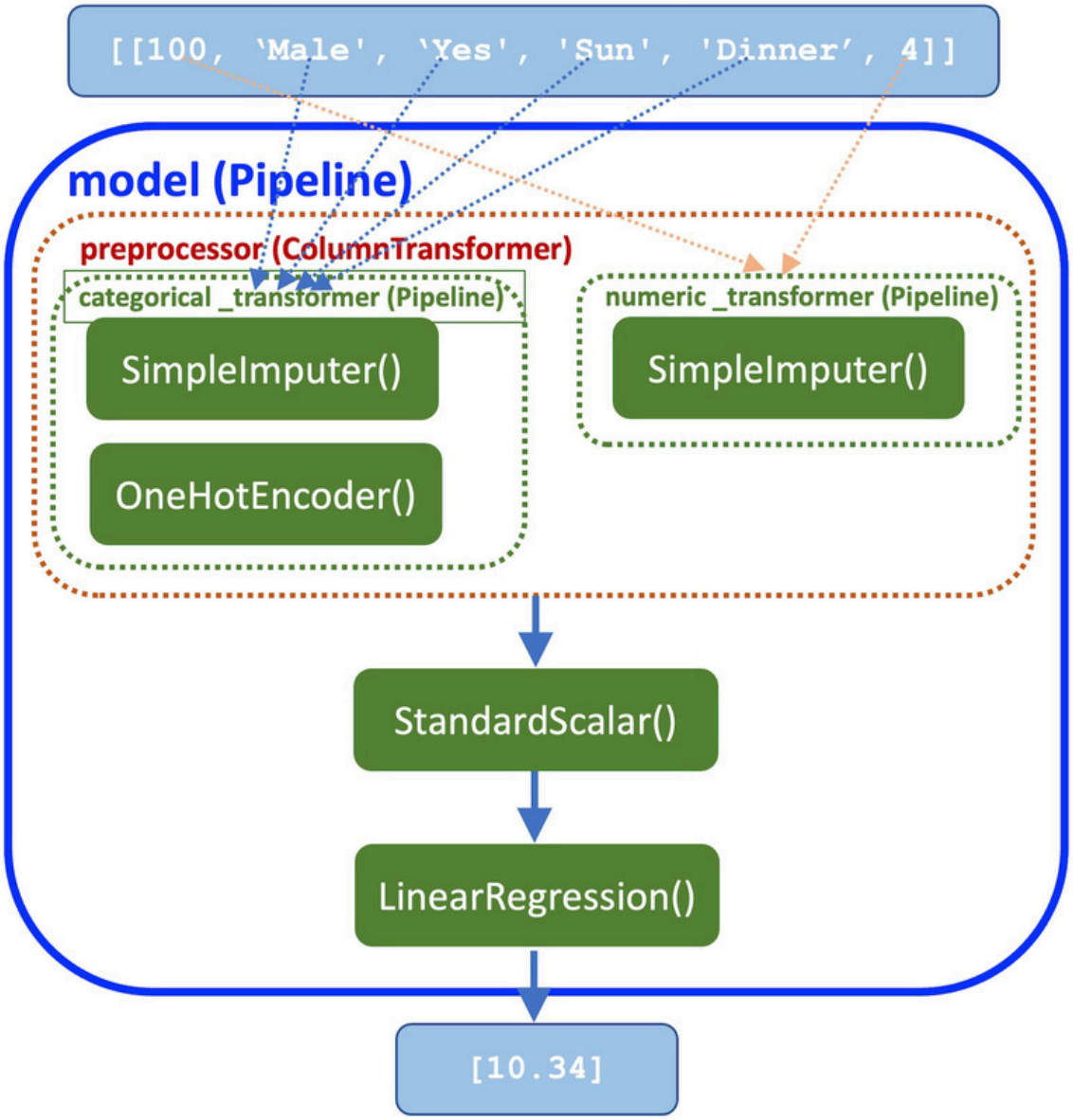
Ser	Hyperparameters	Model Parameters
2	Hyperparameters are used to optimize machine learning model.	Model's parameters are later used for prediction.
3	They are internal to the model.	Thy are external to the model.
4	Examples: Value of K in KNN, learning rate in gradient descent, number of iterations in gradient descent, number of layers in neural network, number of trees in RandomForrest, number of clusters in K-Mean Clustering.	Examples: Coefficients in a Linear or Logistic regression, support vectors in a support vector machine, and weights in an artificial neural network.

- By Hand: Select the hyperparameters values based on intuition/experience/guessing, train the model with the hyperparameters, and score on the validation data. Repeat process until you run out of patience or are satisfied with the results.
- GridSearchCV: Set up a grid of hyperparameter values and for each combination, train a model and score on the validation data. In this approach, every single combination of hyperparameters values is tried which can be very inefficient!

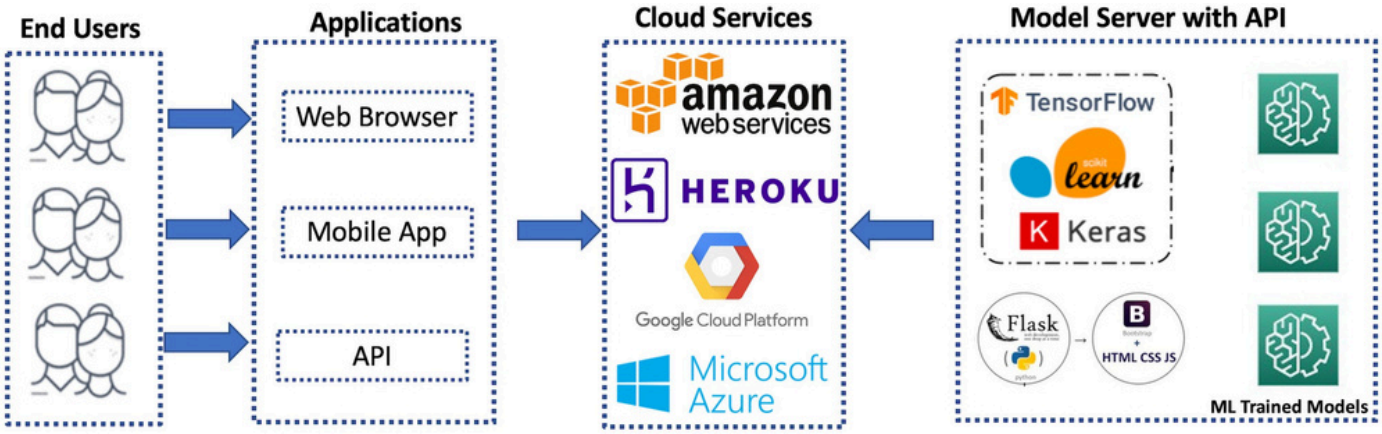
Machine Learning Pipelines

Pipeline is a mechanism that chains together multiple steps together so that the o/p of each step is used as i/p to the next step





(vii) Deployment and Monitoring



Batch Machine Learning vs Online Machine Learning