



Department of Data Science

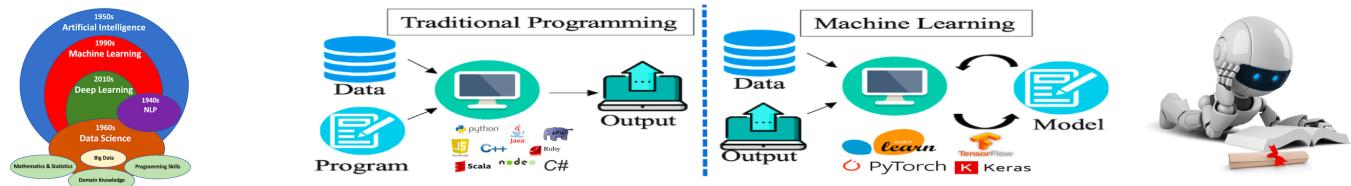
Course: Tools and Techniques for Data Science

Instructor: Muhammad Arif Butt, Ph.D.

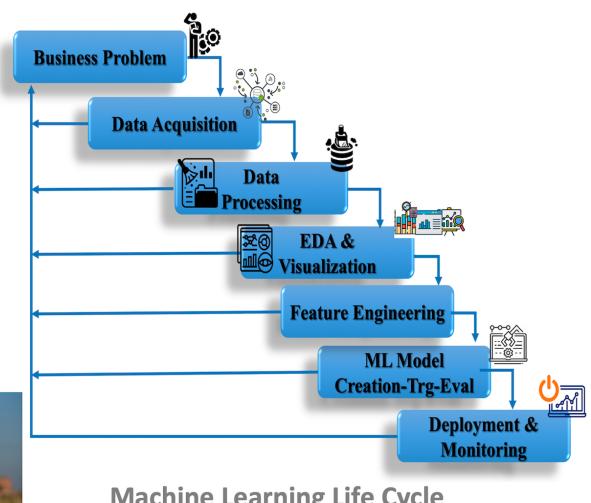
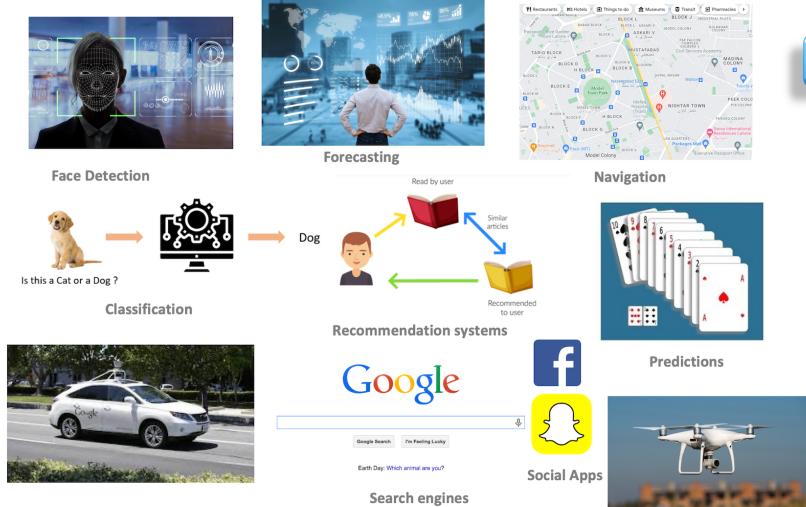
Lecture 6.1 (Overview of Machine Learning)

Open in Colab

([https://colab.research.google.com/github/arifpucit/data-science/blob/master/Section-4-Mathematics-for-Data-Science/Lec-4.1\(Descriptive-Statistics\).ipynb](https://colab.research.google.com/github/arifpucit/data-science/blob/master/Section-4-Mathematics-for-Data-Science/Lec-4.1(Descriptive-Statistics).ipynb))



ML is the application of AI that gives machines the ability to learn without being explicitly programmed

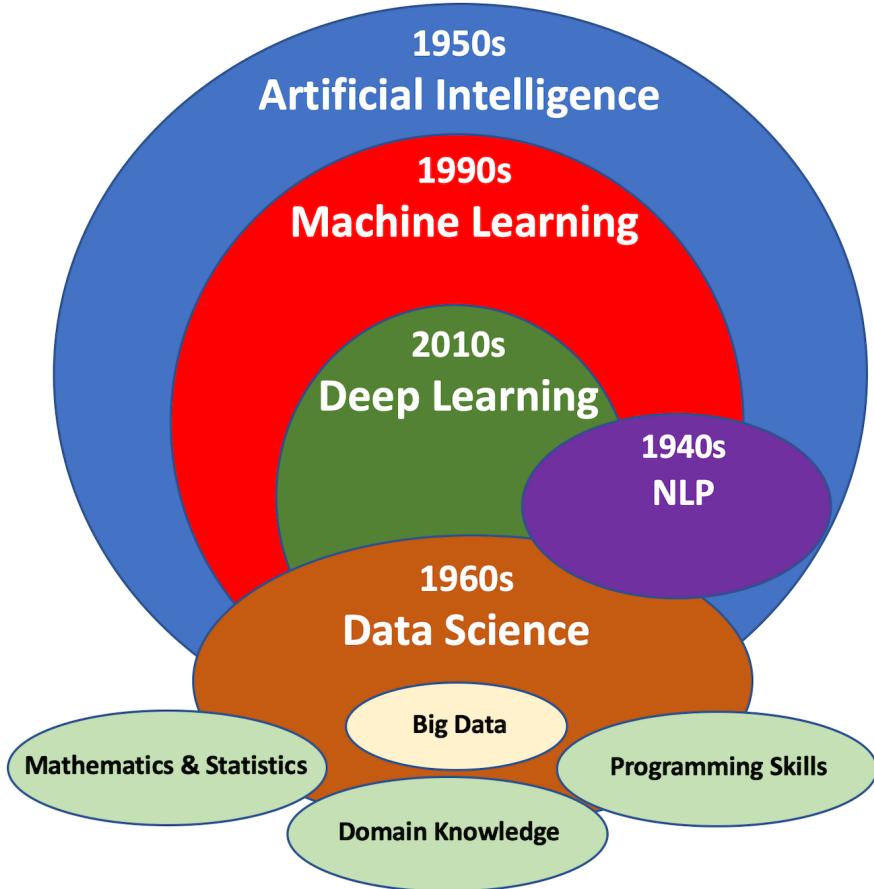


Learning agenda of this notebook

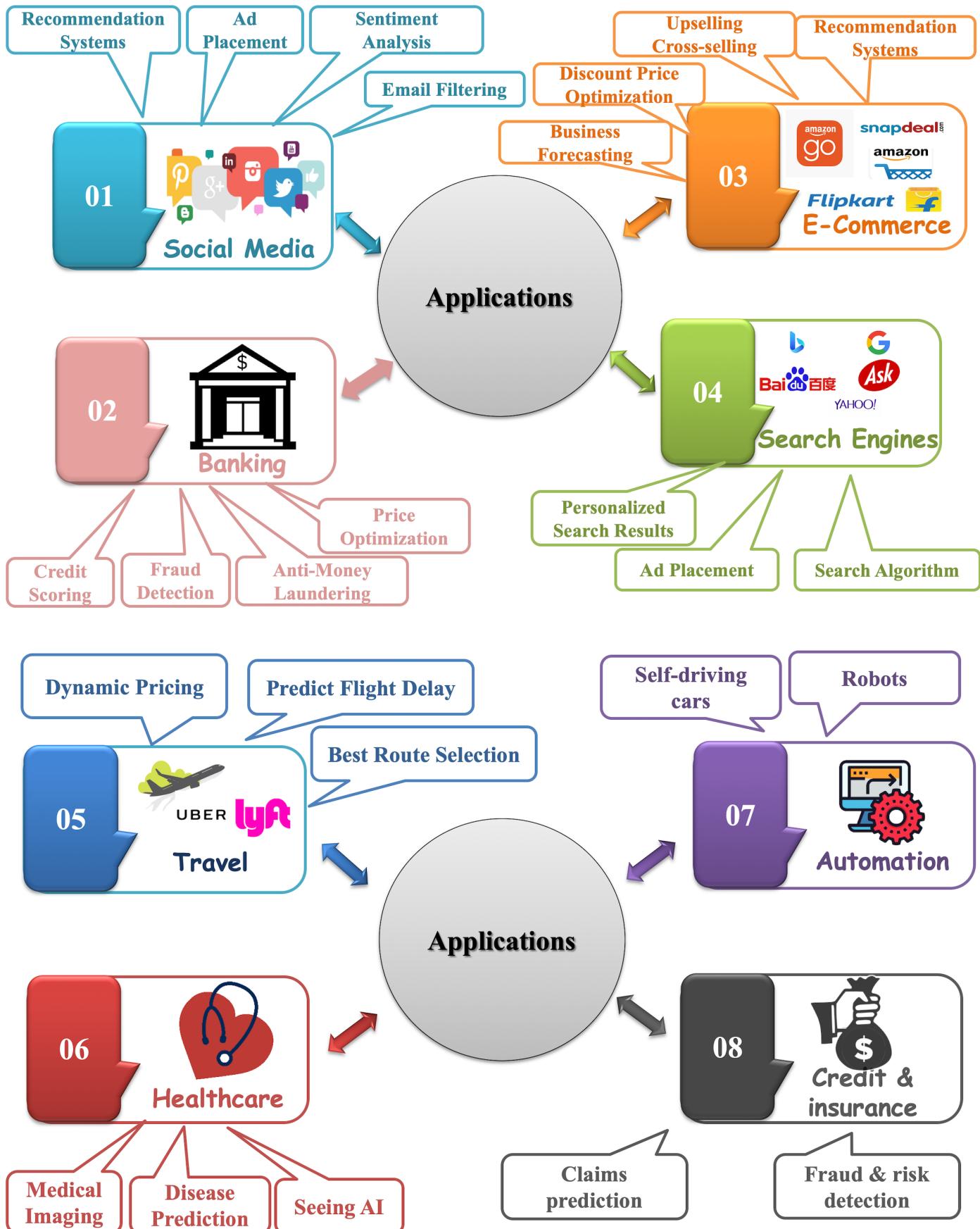
1. The Big Picture
2. Why to do Machine Learning? (Applications of ML)
3. What is Machine Learning and its Types?
4. How to do Machine Learning? (ML algorithms)
5. Machine Learning Development Life Cycle

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.



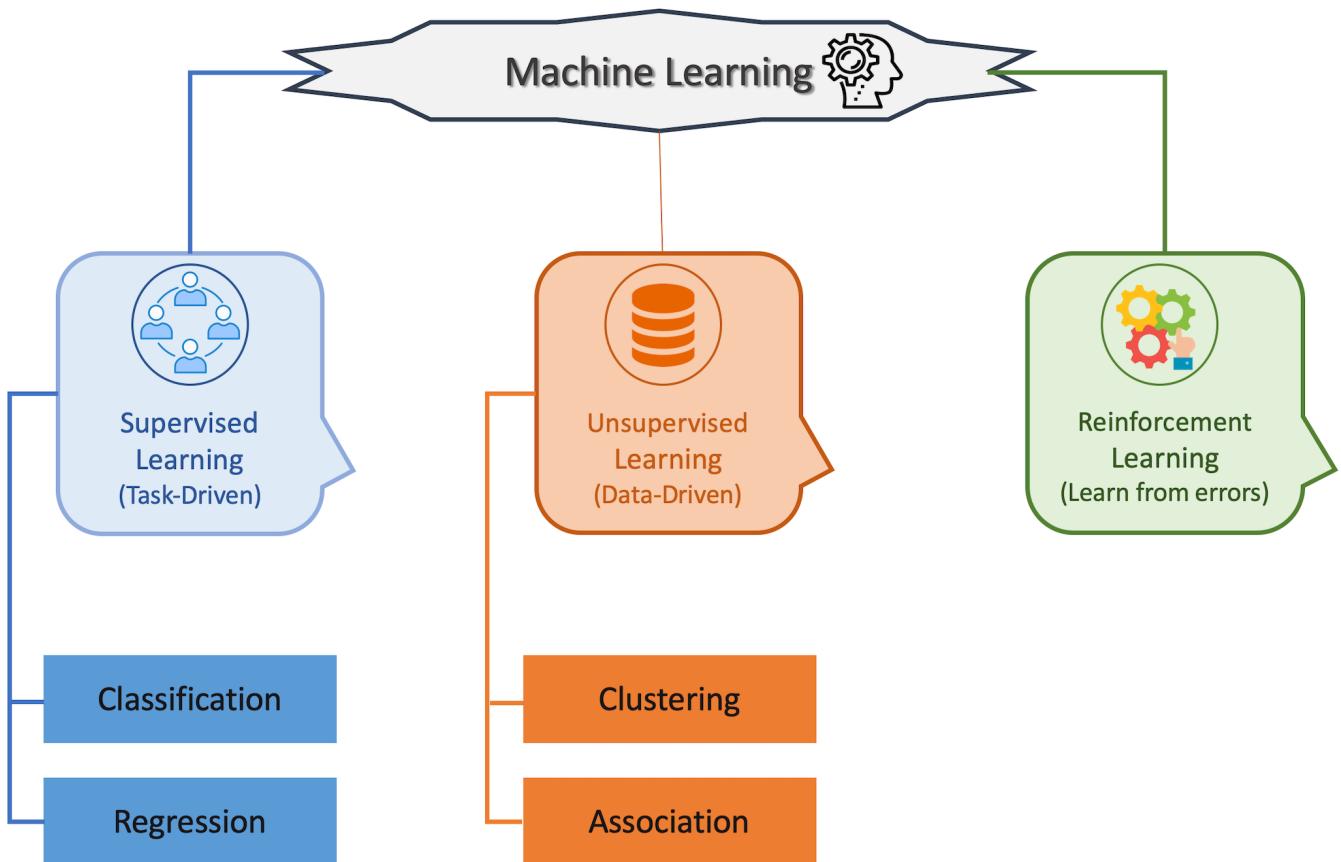
2. Why to do Machine Learning? (ML Applications)



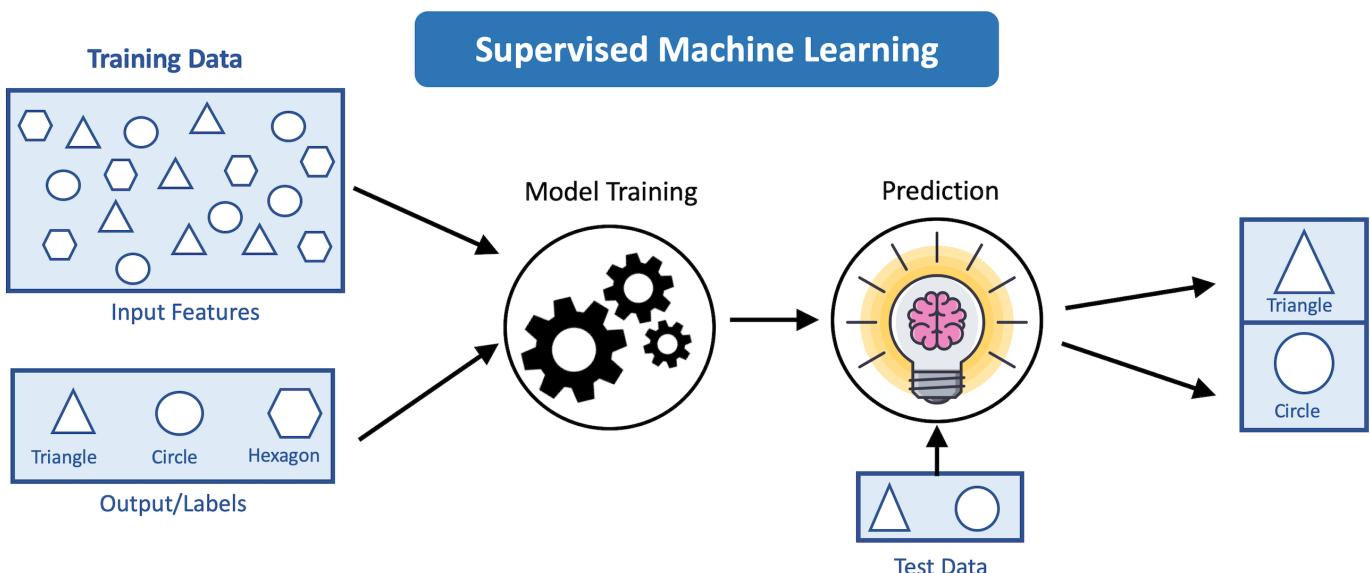
3. What is Machine Learning?

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

a. Types of Machine Learning



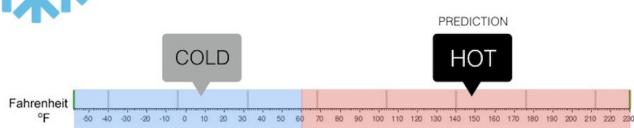
b. Supervised Machine Learning (Classification vs Regression)





Classification

Will it be Cold or Hot tomorrow?

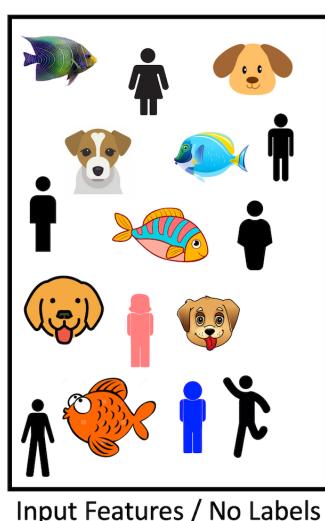


Regression

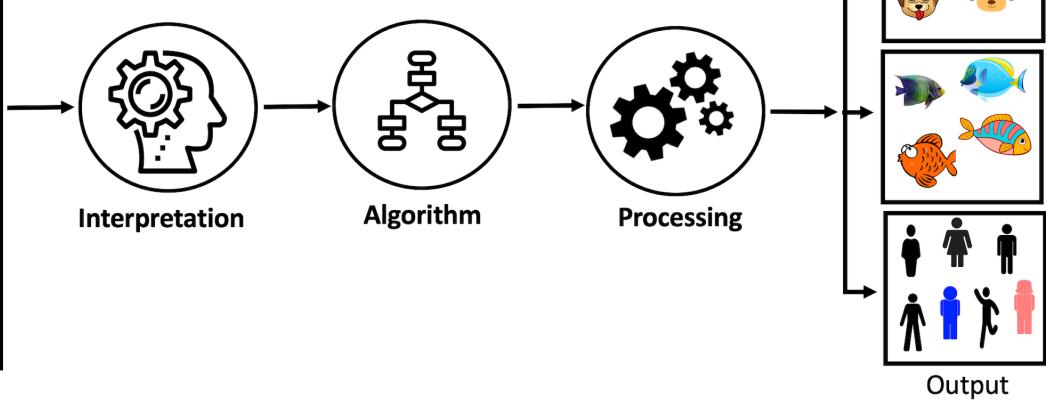
What is the temperature going to be tomorrow?



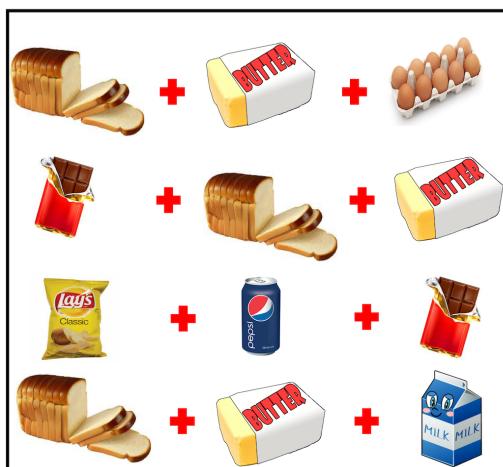
c. Unsupervised Machine Learning (Clustering)



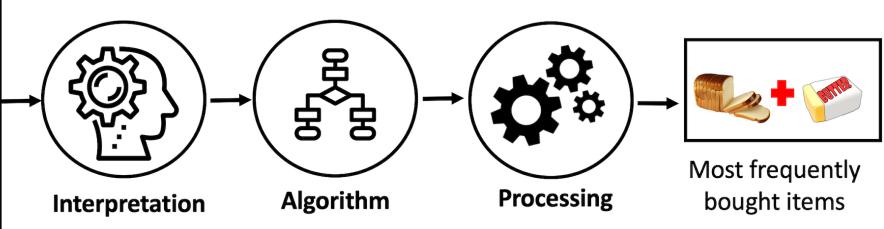
Unsupervised Machine Learning (Clustering)



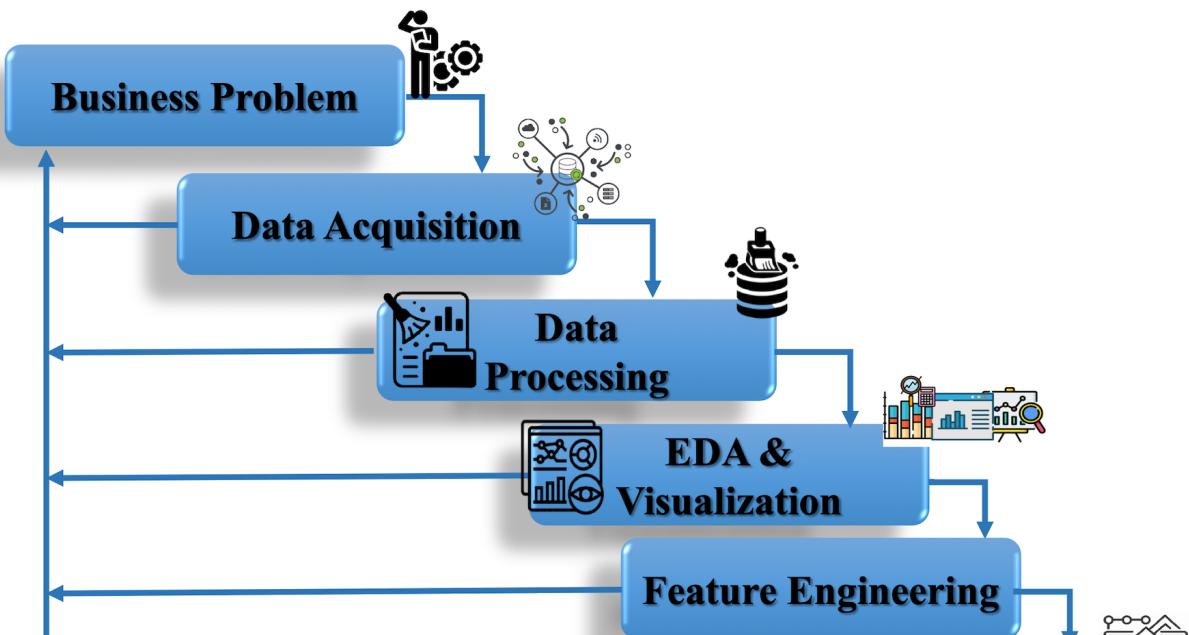
d. Unsupervised Machine Learning (Association)



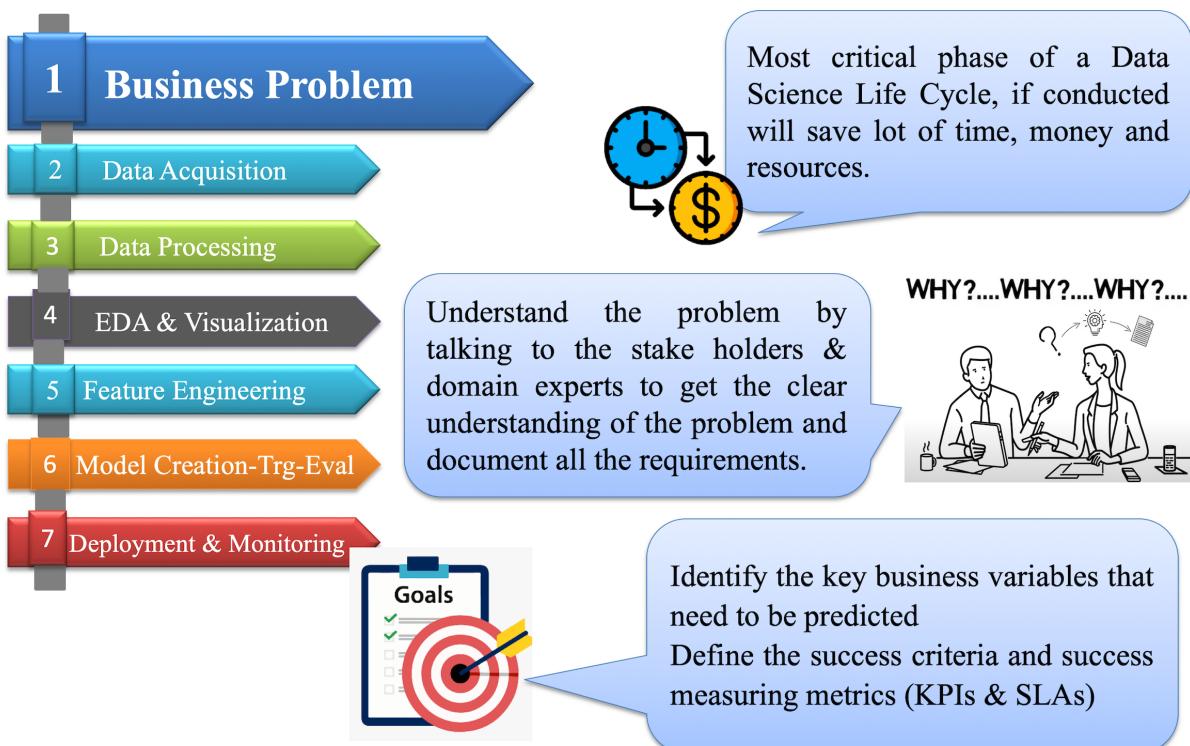
Unsupervised Machine Learning (Association)



5. Machine Learning Development Life Cycle

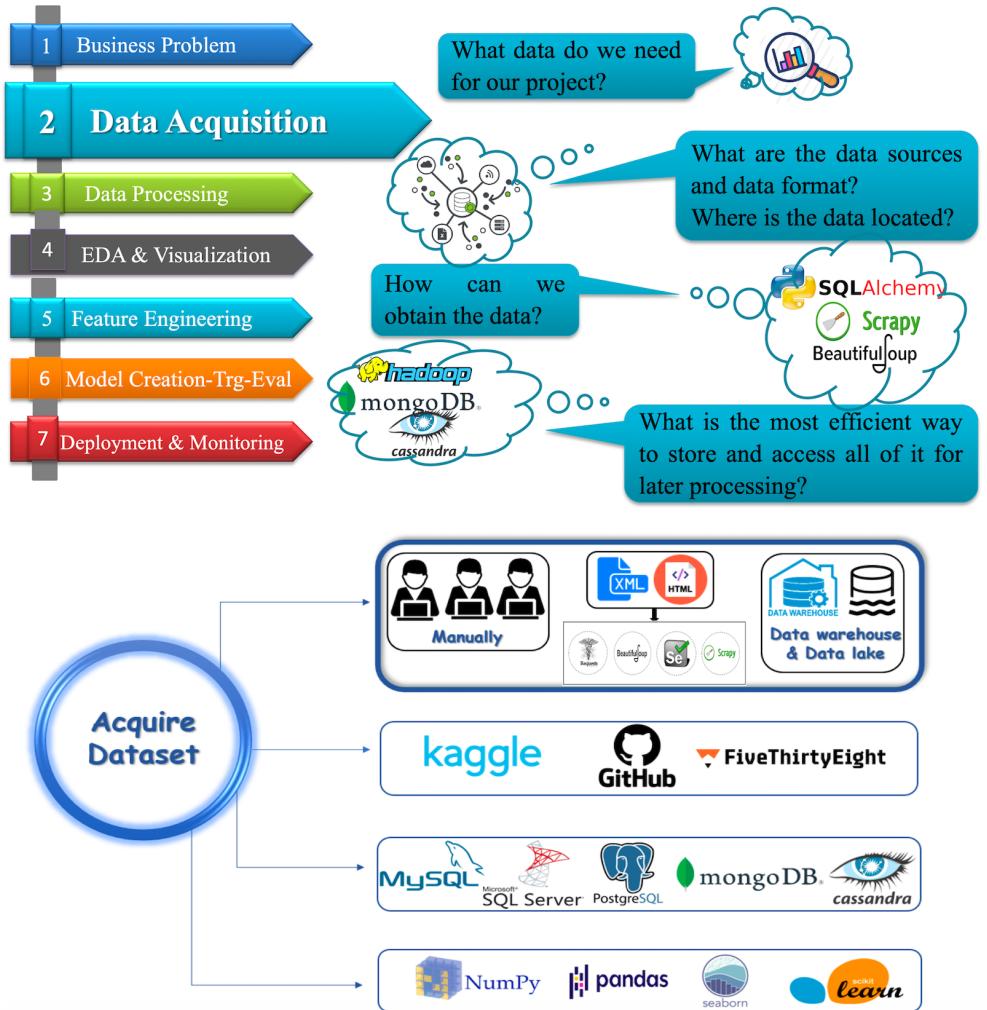


(i) Understand the Business Problem



(ii) Data Acquisition

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

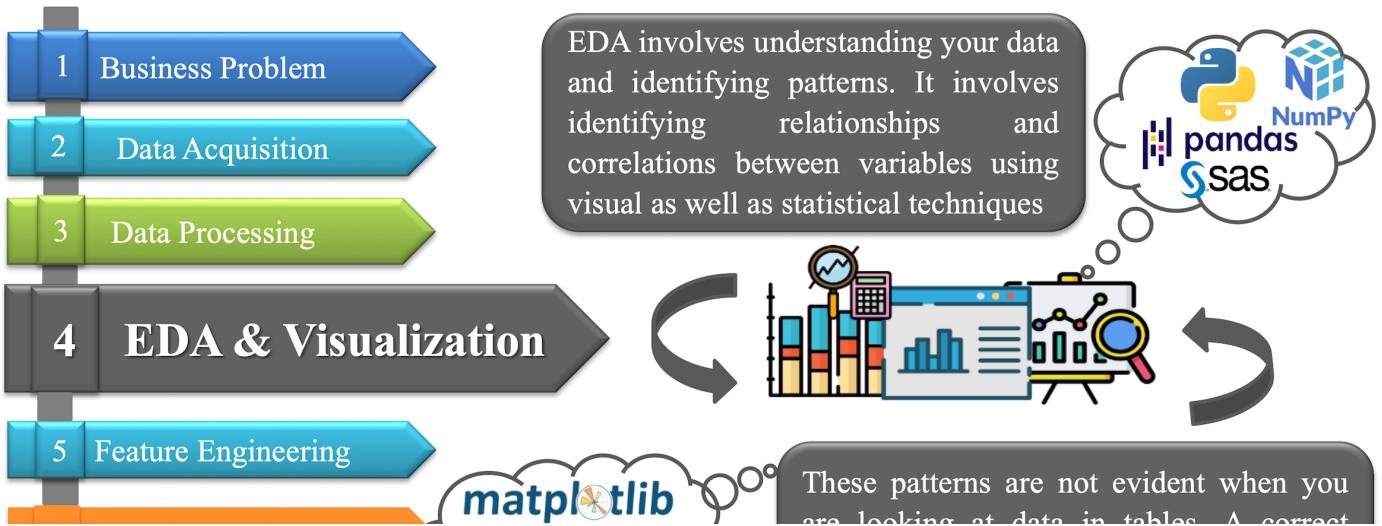
- Use Web scraping or Web API
- IoT Devices
- Crowd Sourcing (Amazon Mechanical Turk, Lionbridge AI)
- Data Augmentation

•

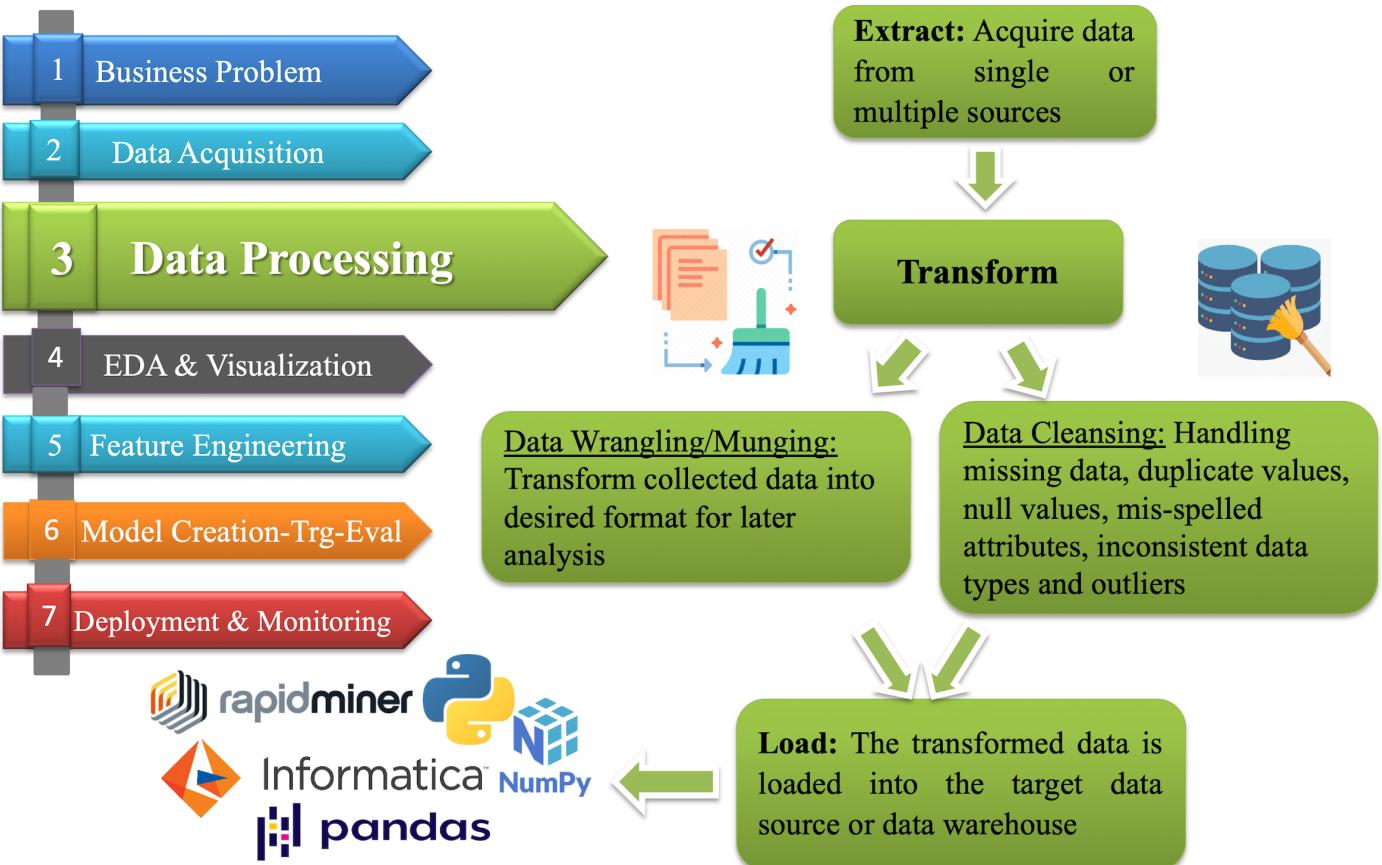
Techniques of Labeling your Datasets

- Data Acquisition using web scraping: <https://www.youtube.com/watch?v>

(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 |

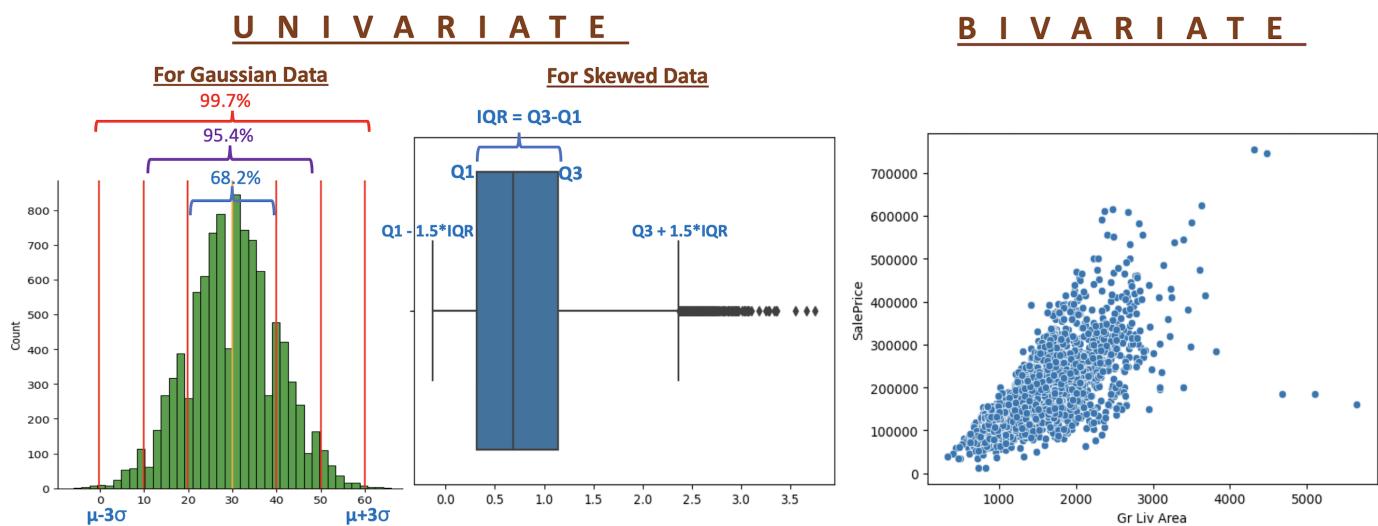
a. Detecting and Handling Outliers

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?

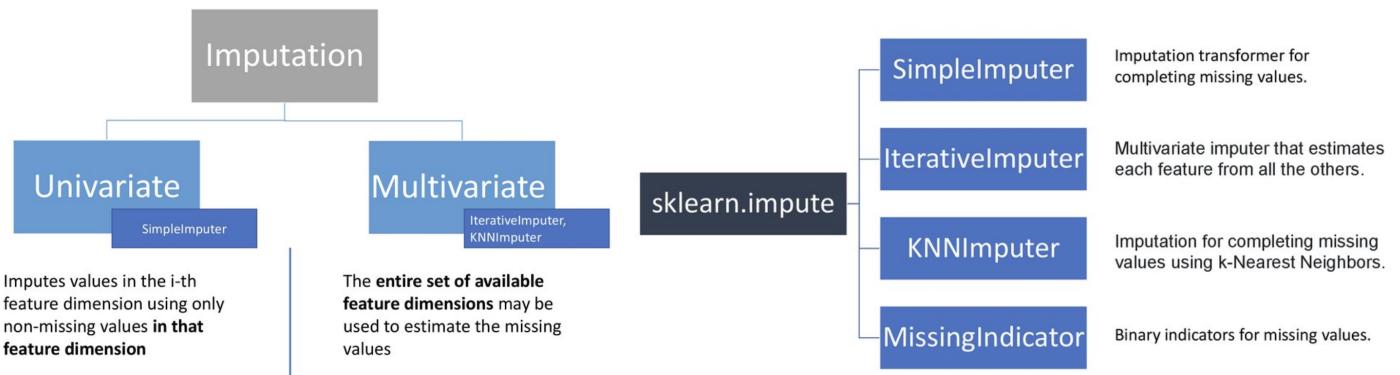
- Trimming (Remove)
- Capping/Winsorization (Replace with upper value)

- Discretization

b. Detecting and Handling Missing Values

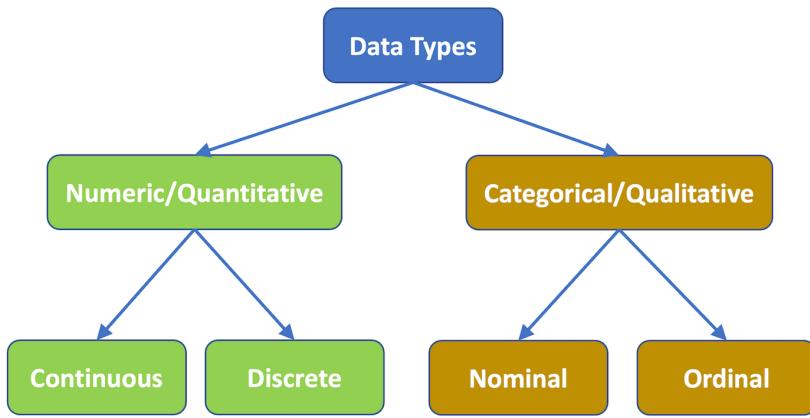
Missing data, or missing values, occur when no data value is stored for the variable in an observation

- **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



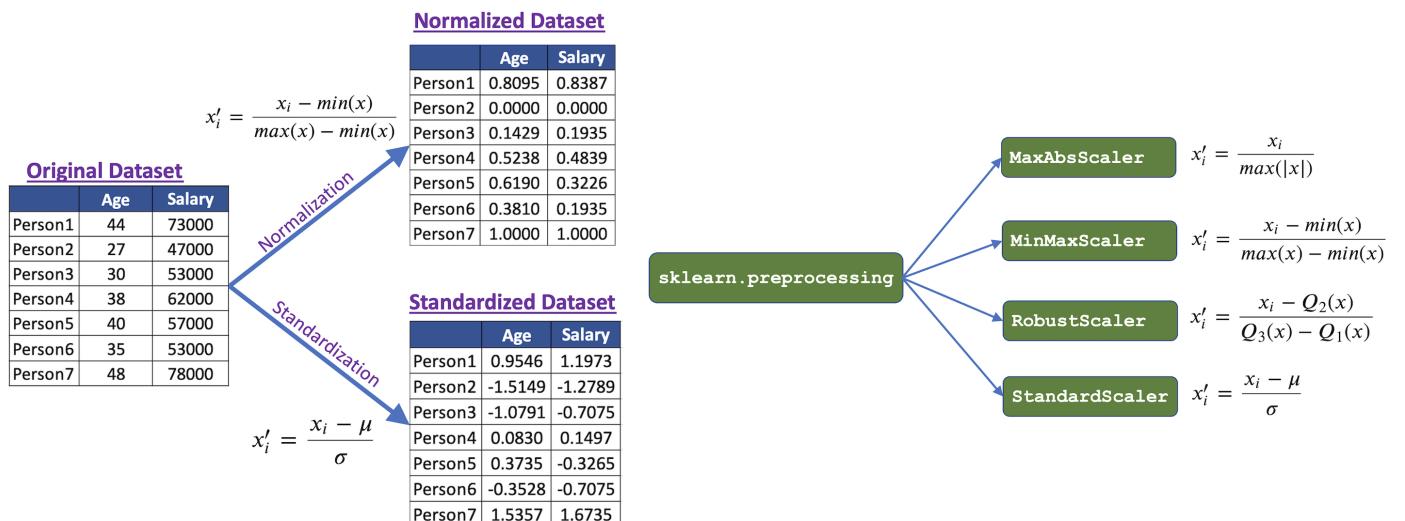
c. Encoding Categorical Features

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



- **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 Regression, Neural Network) require feature scaling.
- All **Tree based algorithms** like Decision Tree, Random Forest 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

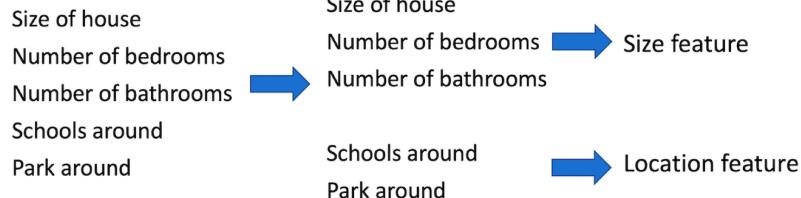
| City | Size | Covered Area | No of bedrooms | Trees near by | No of bathrooms | Schools near by | Construction Date | Price |
|-------------|------|--------------|----------------|---------------|-----------------|-----------------|-------------------|--------|
| Lahore | 2000 | 3500 | 3 | 1 | 3 | 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 |

5 Feature Engineering

6 Model Creation-Trg-Eval

7 Deployment & Monitoring

Merge the Features



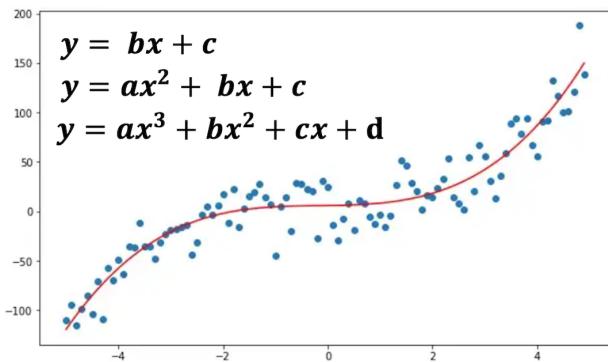
Principal Component Analysis: https://www.youtube.com/watch?v=Z2WDUUHkof8&list=PL7B2bn3G_wfAs3C49i12i_rblzvuU1dFN&index=50
[\(https://www.youtube.com/watch?v=Z2WDUUHkof8&list=PL7B2bn3G_wfAs3C49i12i_rblzvuU1dFN&index=50\)](https://www.youtube.com/watch?v=Z2WDUUHkof8&list=PL7B2bn3G_wfAs3C49i12i_rblzvuU1dFN&index=50)

f. Add-Ons

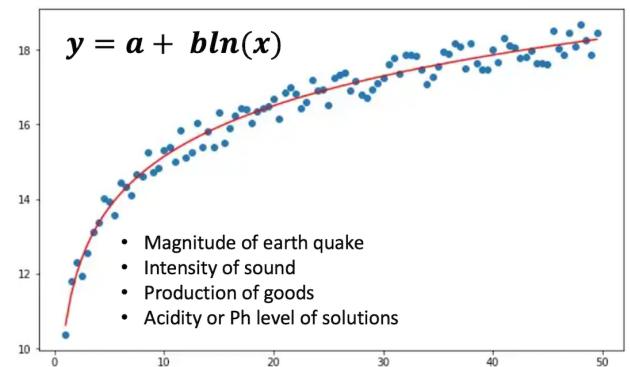
Polynomial Regression

It is a form of linear regression, which estimates the relationship between X and y as an n^{th} degree polynomial in X.

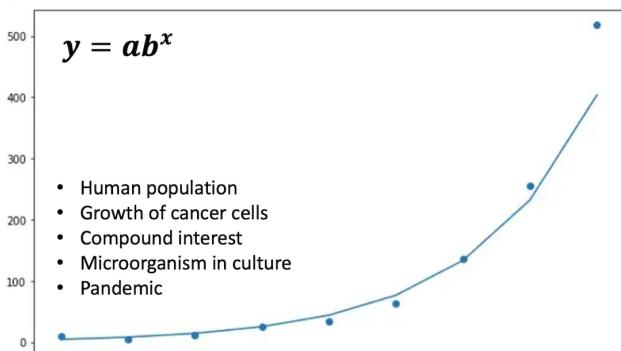
Polynomial Regression



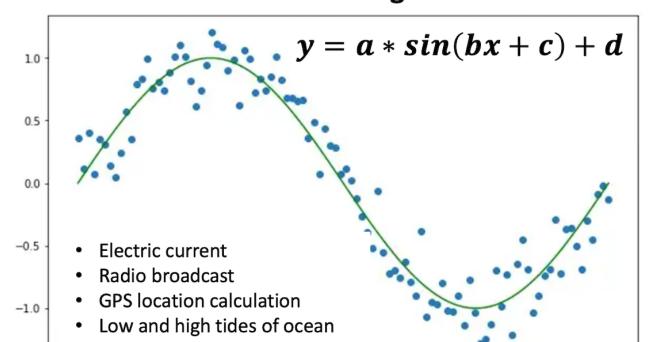
Logarithmic Regression



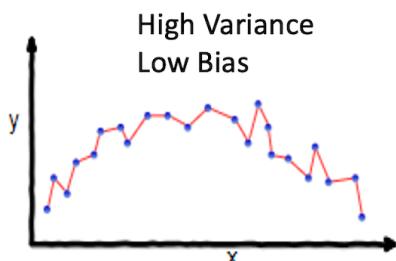
Exponential Regression



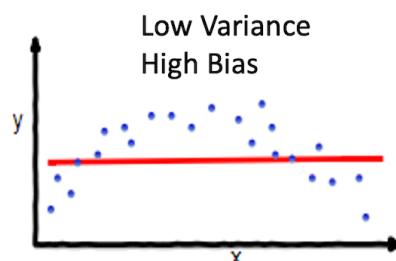
Sinusoidal Regression



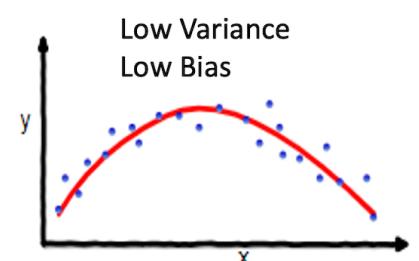
Bias-Variance Tradeoff and Overfitting



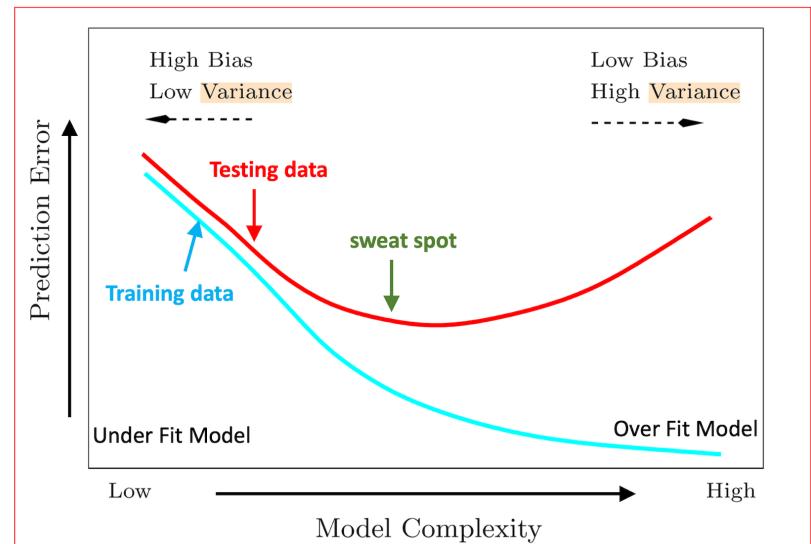
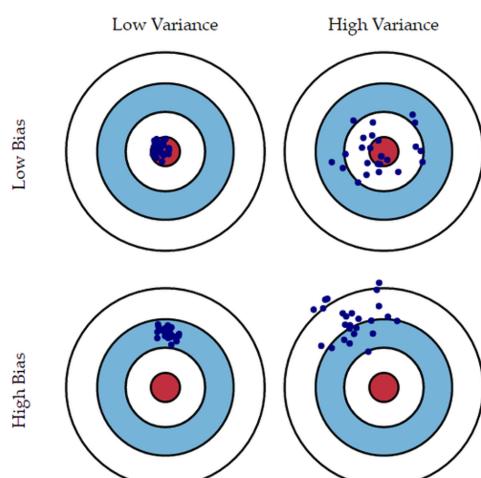
Over Fit Model



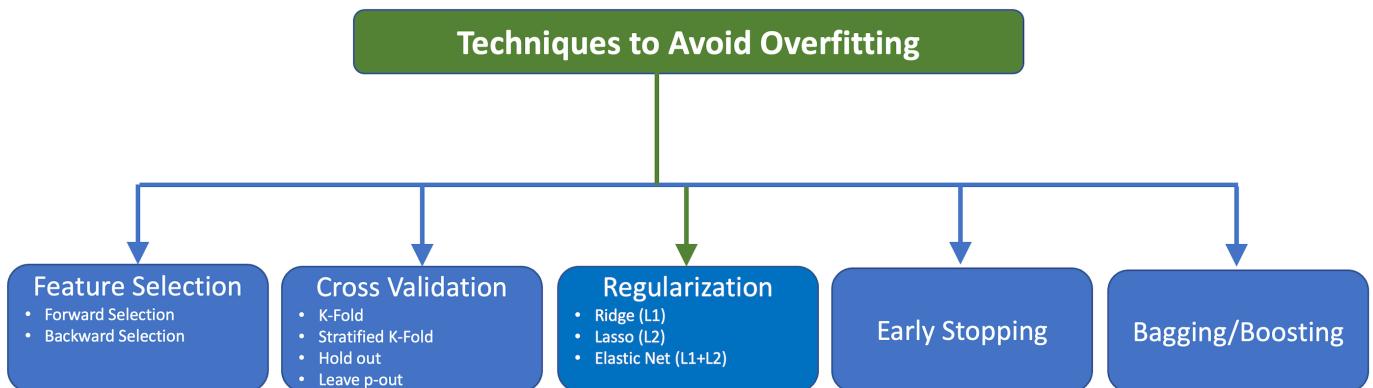
Under Fit Model



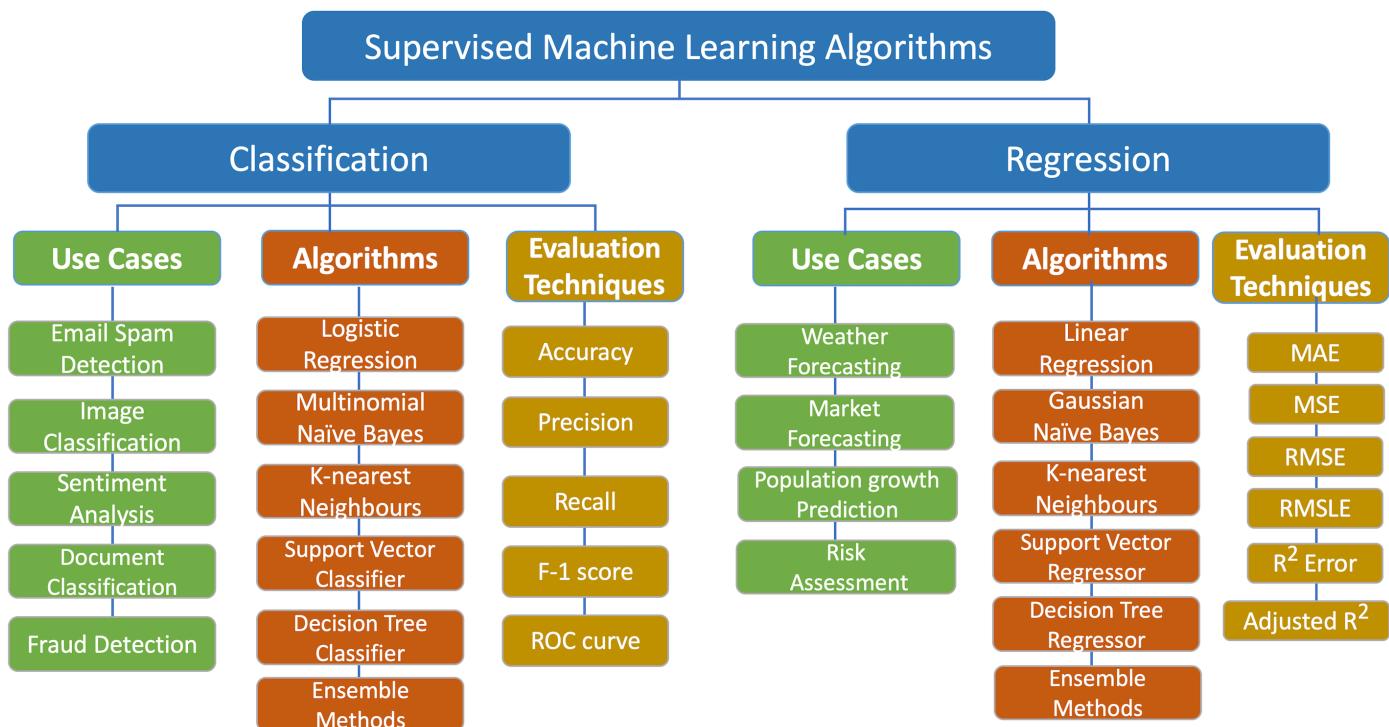
Good Fit Model

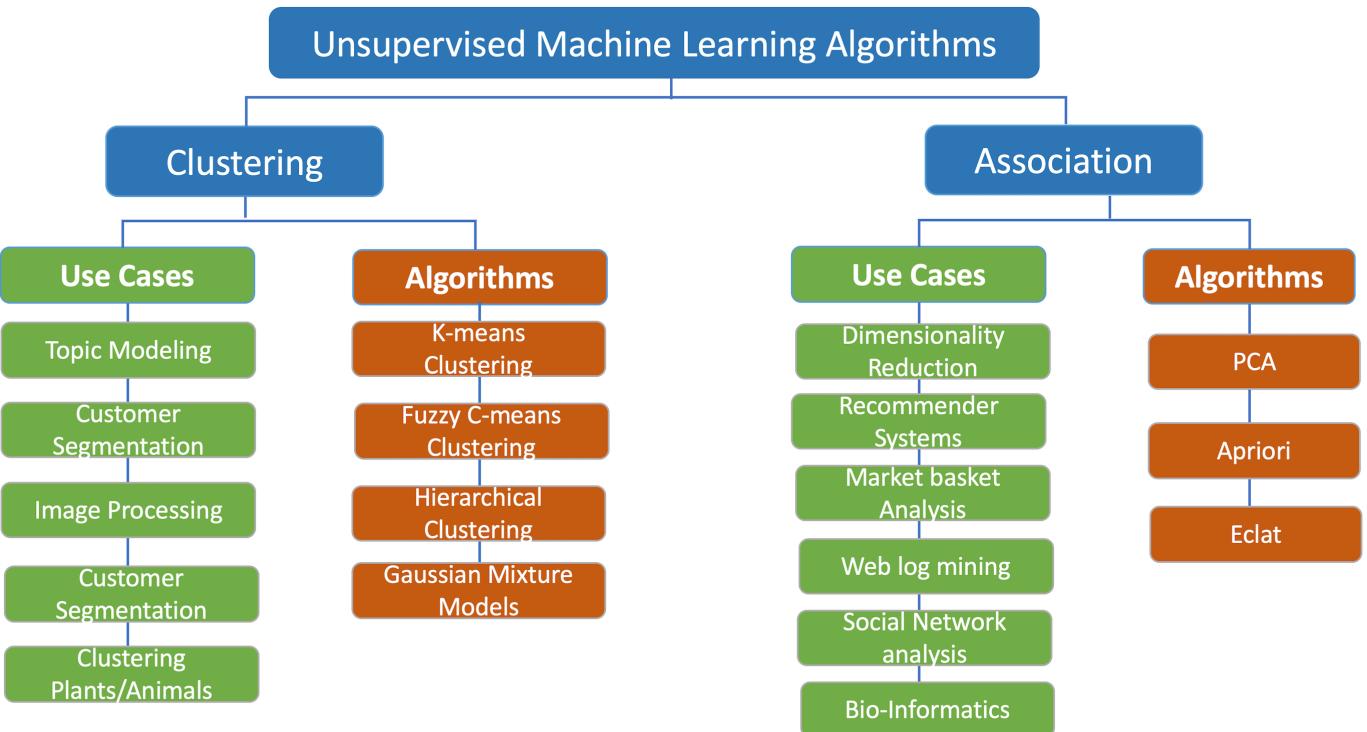


How to prevent overfitting?



(v) Machine Learning Models





Supervised Machine Learning Model Creation-Training-Evaluation



Classification

| # | Height (inches) | Weight (kgs) | B.P. Sys | B.P. Dia | Label |
|---|-----------------|--------------|----------|----------|-------|
| 1 | 62 | 70 | 120 | 80 | No |
| 2 | 72 | 90 | 110 | 70 | No |
| 3 | 74 | 80 | 130 | 70 | No |
| 4 | 65 | 120 | 150 | 90 | Yes |
| 5 | 67 | 100 | 140 | 85 | Yes |
| 6 | 64 | 110 | 130 | 90 | No |
| 7 | 69 | 150 | 170 | 100 | Yes |
| 8 | 66 | 125 | 145 | 90 | ? |
| 9 | 74 | 67 | 110 | 60 | ? |

Regression

| # | Height (inches) | Weight (kgs) | B.P. Sys | B.P. Dia | Label |
|---|-----------------|--------------|----------|----------|--------|
| 1 | 62 | 70 | 120 | 80 | 150.35 |
| 2 | 72 | 90 | 110 | 70 | 160.46 |
| 3 | 74 | 80 | 130 | 70 | 130.05 |
| 4 | 65 | 120 | 150 | 90 | 200.19 |
| 5 | 67 | 100 | 140 | 85 | 190.26 |
| 6 | 64 | 110 | 130 | 90 | 130.33 |
| 7 | 69 | 150 | 170 | 100 | 250.10 |
| 8 | 66 | 125 | 145 | 90 | ? |
| 9 | 74 | 67 | 110 | 60 | ? |

Annotations:

- A red arrow points from the "Features" column header to the first four columns of the Classification table.
- A red arrow points from the "Label" column header to the last column of the Classification table.
- A red arrow points from the "Features" column header to the first four columns of the Regression table.
- A red arrow points from the "Label" column header to the last column of the Regression table.
- A red arrow points from the "4-D Feature Vector (x_i)" label to the row where the fourth feature is 70 in the Classification table.
- A red arrow points from the "Training Data (X and Y)" label to the last two columns of the Classification table.
- A red arrow points from the "Test Data" label to the last two columns of the Regression table.

$$D = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), (\vec{x}_3, y_3), \dots (\vec{x}_n, y_n)\} \subseteq X \times Y$$

Add-Ons

Hyperparameter Tuning

Ser Hyperparameters

- 1 Hyperparameters are set manually by ML engineer/practitioner prior to the start of the model's training.

Model Parameters

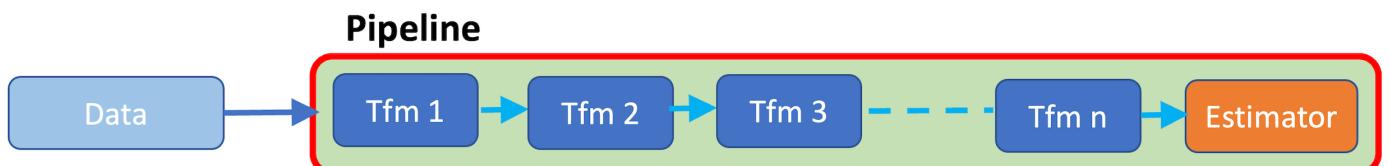
Model parameters are learnt by the learning algorithm during the training phase.

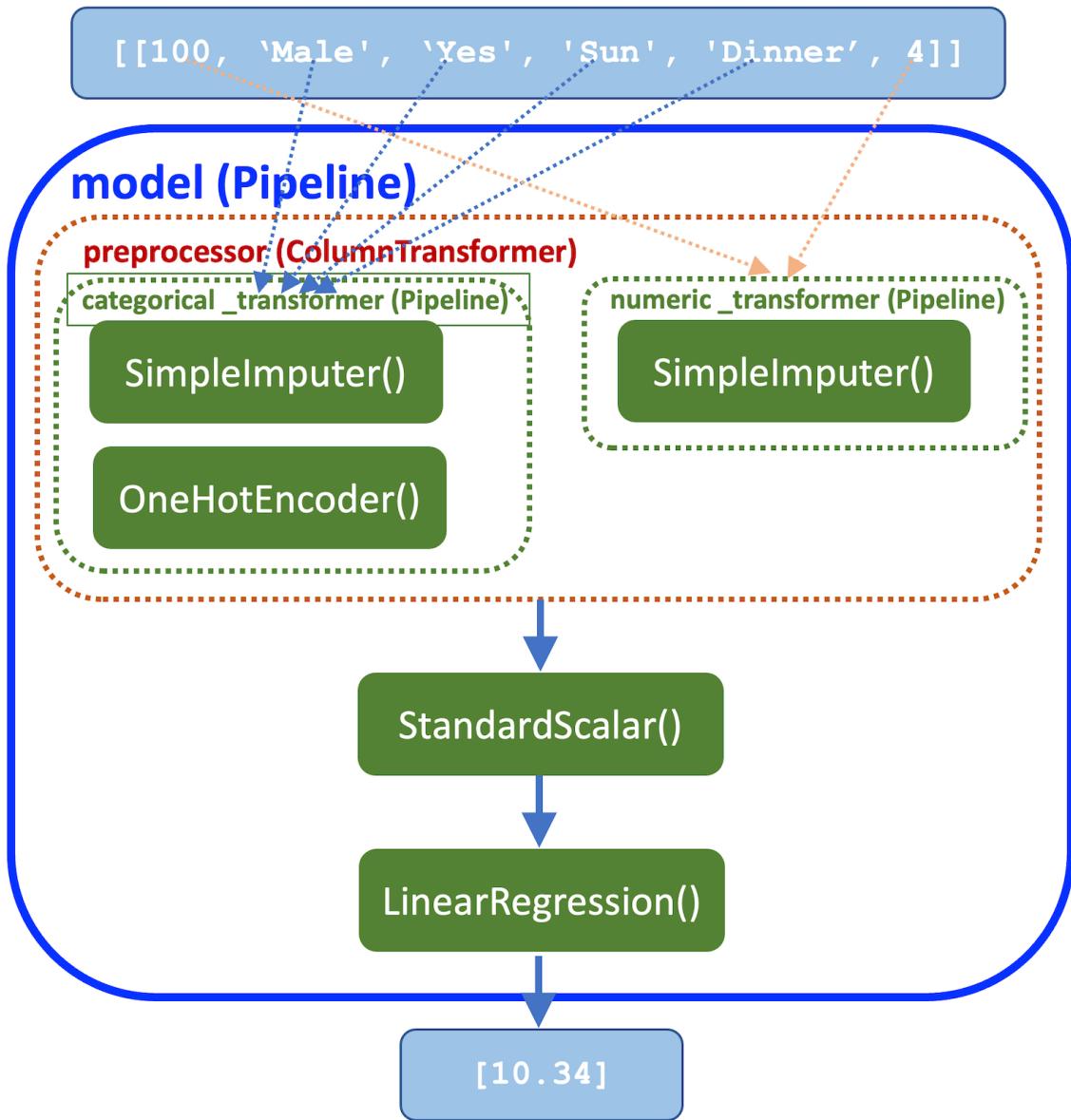
| Ser | Hyperparameters | Model Parameters |
|-----|--|--|
| 2 | Hyperparameters are used to optimize machine learning model. | Model's parameters are later used for prediction. |
| 3 | They are internal to the model. | They are external to the model. |
| 4 | Examples: Value of K in KNN, learning rate 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!

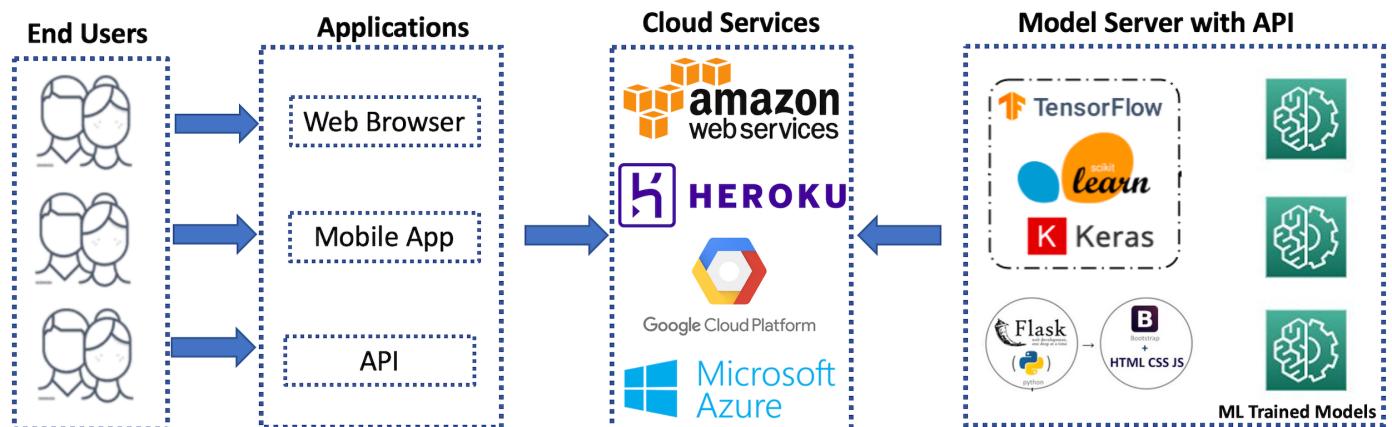
Machine Learning Pipelines

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

In []:

| | |
|---|--|
| 1 | |
|---|--|