SNS COLLEGE OF TECHNOLOGY

(An Autonomous Institution)

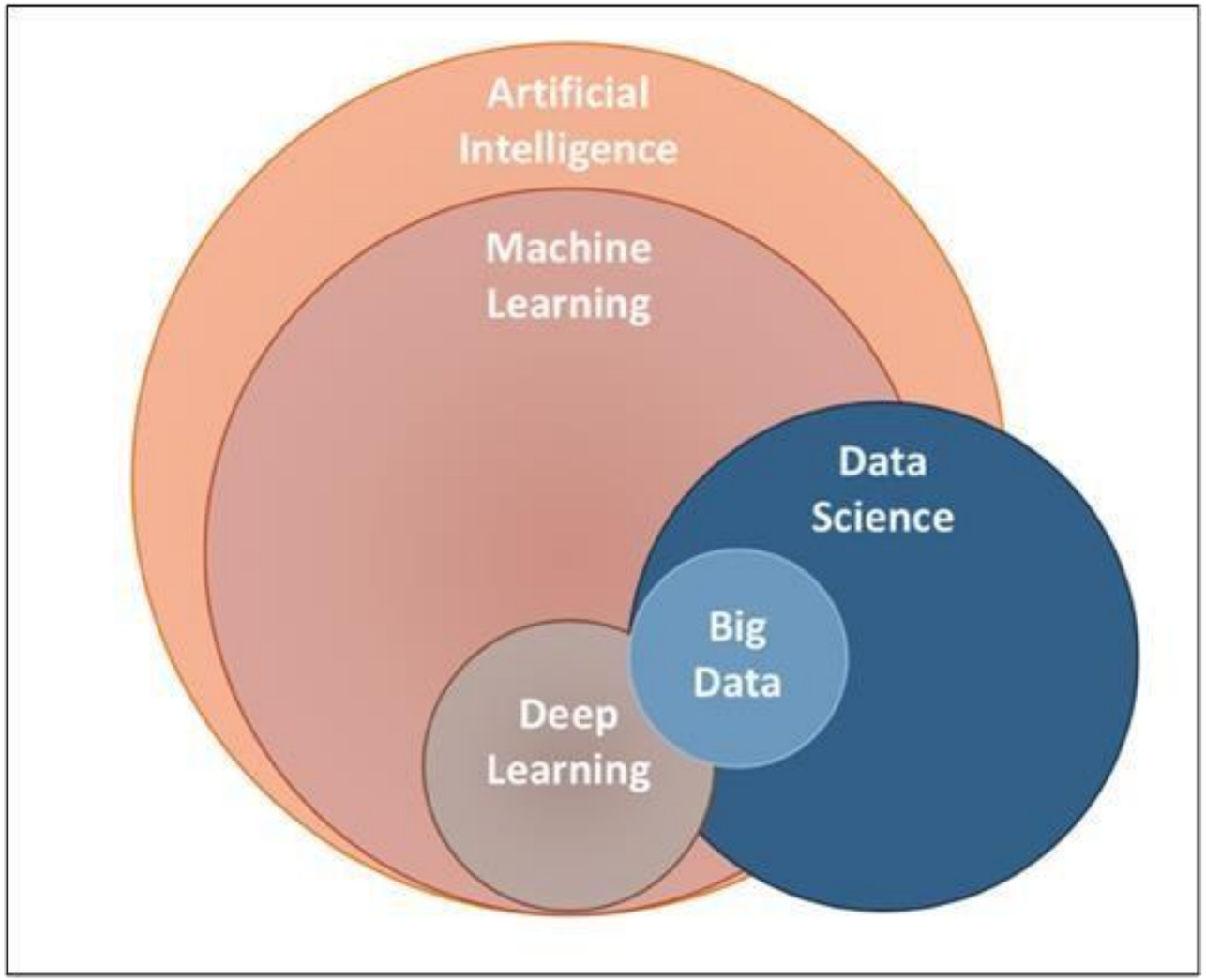
Approved by AICTE, New Delhi, Affiliated to Anna University, Chennai
Accredited by NAAC-UGC with 'A++' Grade (Cycle III) & Accredited by NBA (B.E - CSE, EEE, ECE, Mech & B.Tech.IT)
COIMBATORE-641 035, TAMIL NADU

DEPARTMENT OF COMPUTER APPLICATIONS

19CAE716 – DATA SCIENCE

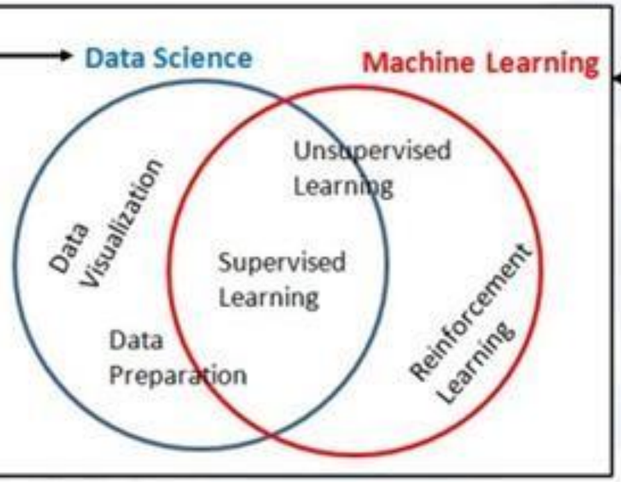
UNIT – I: INTRODUCTION TO DATA SCIENCE

TOPIC: Modeling Process


Artificial Intelligence
Machine Learning
Data Science
Big Data
Deep Learning

How Data Science and ML are related?

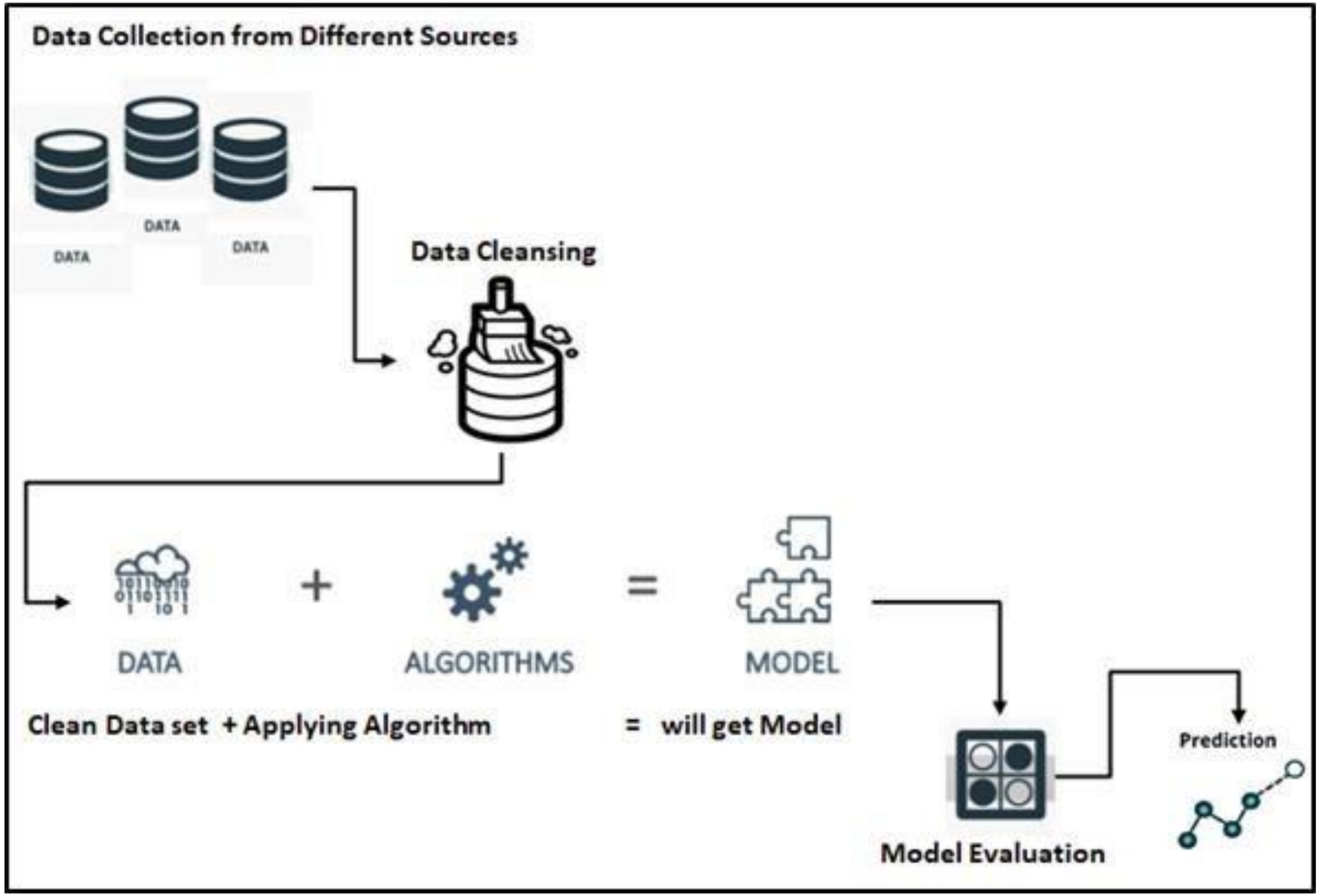
- Data Science is a broader concept and multidisciplinary.
- Data science is general process and method that analyse and manipulate data.
- Data science enables to find meaning and appropriate information from give data.
- Data Science creating an opportunity to use data for making key decisions in different business domains and technology.
- Data science provides vast and robust way of visualization techniques to under the data insights.



The main difference between the two is that data science as a broader term not only focuses on algorithms and statistics but also takes care of the entire data processing methodology.

- Machine learning fits within data science.
- Machine learning uses various techniques and algorithm
- Machine learning is highly iterative process
- Machine Learning algorithms are trained over instances.
- Machine Model are learn from past experiences and also analyse the historical data
- Machine Model able to identify patterns in order to make predictions about the future of the give data

Machine Learning Process



Machine Learning process

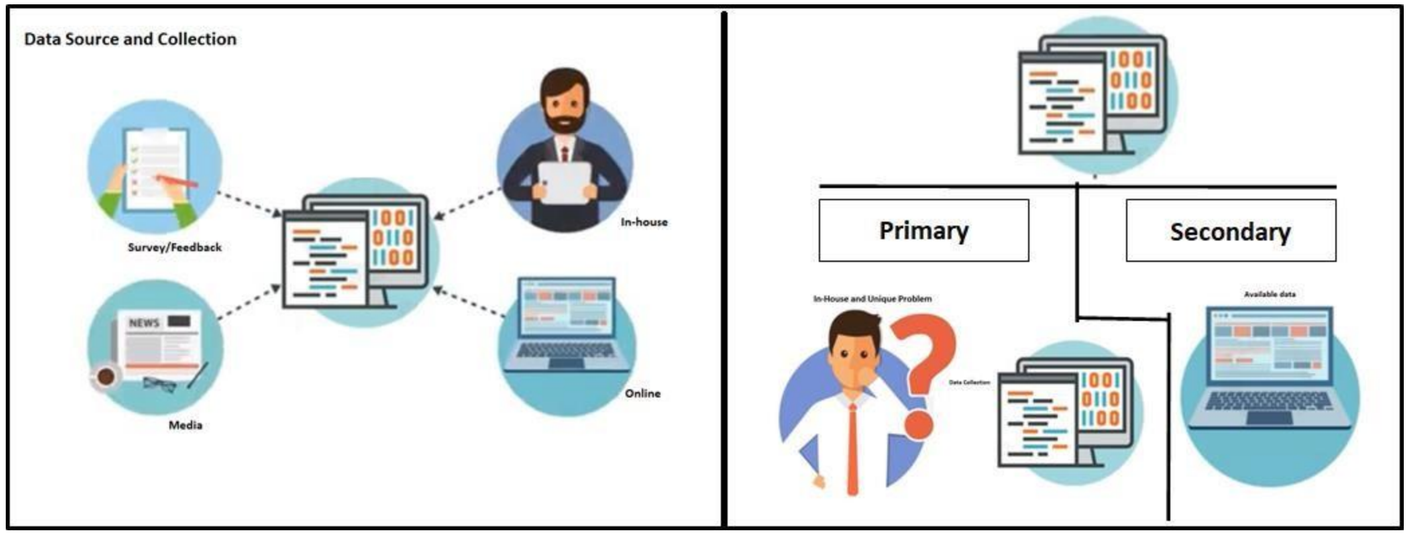
- Collection of Data
- Data Wrangling
- Model Building
- Model Evaluation
- Model Deployment

Machine Learning process

Collection of Data

Data collection from different sources could be internal and/or external to satisfy the business requirements/problems. Data could be in any format. CSV, XML.JSON, etc.,

here Big Data is playing a vital role to make sure the right data is in the expected format and structure.



Data Processing (EDA)

- Understanding the given dataset and helping clean up the given dataset.
- It gives you a better understanding of the features and the relationships between them
- Extracting essential variables and leaving behind/removing non-essential variables.
- Handling Missing values or human error.
- Identifying outliers.
- The EDA process would be maximizing insights of a dataset.

Feature engineering

1.Handling missing values in the variables

2.Convert categorical into numerical since most algorithms need numerical features.

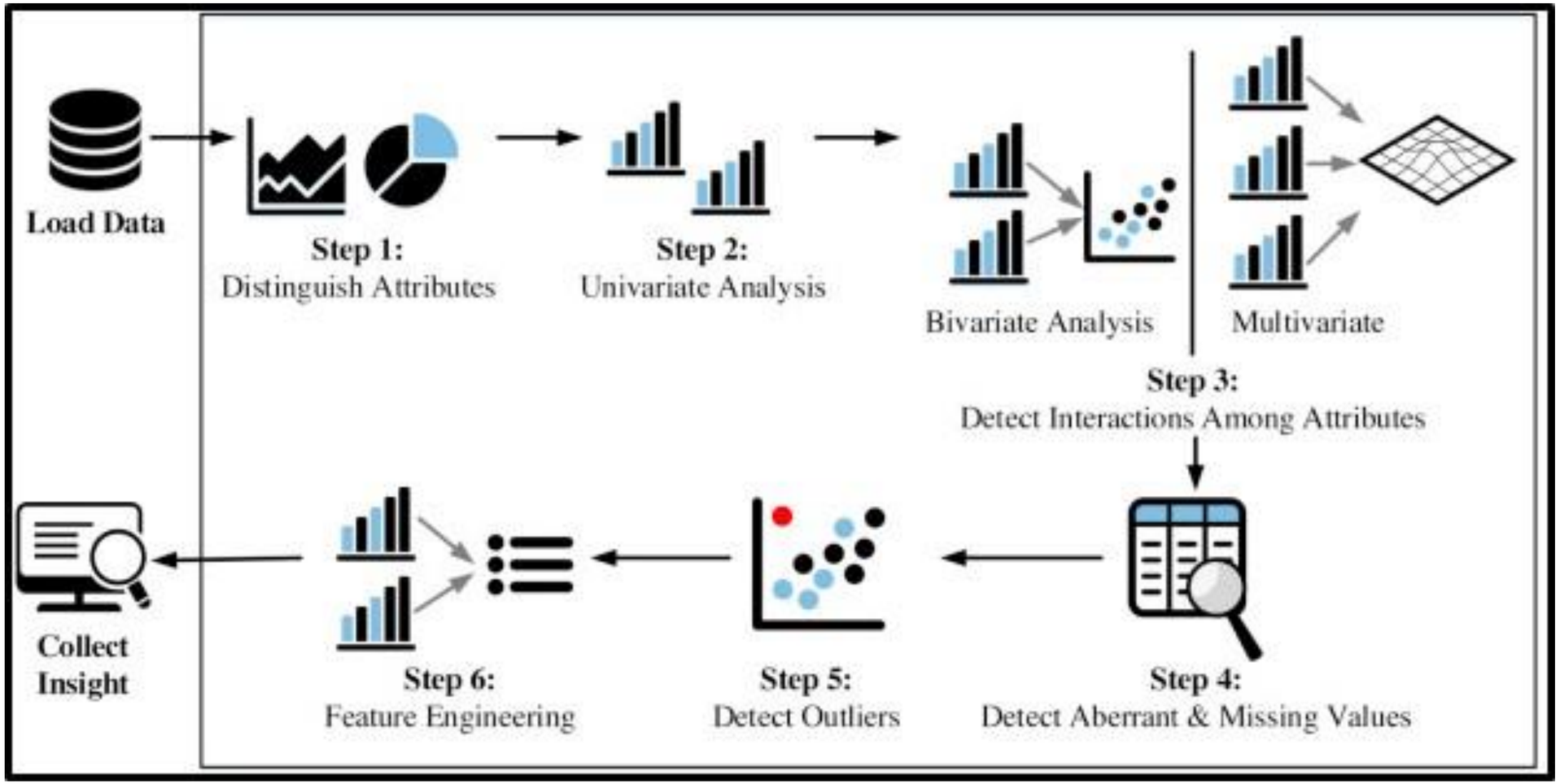
3.Need to correct not Gaussian(normal). linear models assume the variables have Gaussian distribution.

4.Finding Outliers are present in the data, so we either truncate the data above a threshold or transform the data using log transformation.

5.Scale the features. This is required to give equal importance to all the features, and not more to the one whose value is larger.
6.Feature engineering is an expensive and time-consuming process.

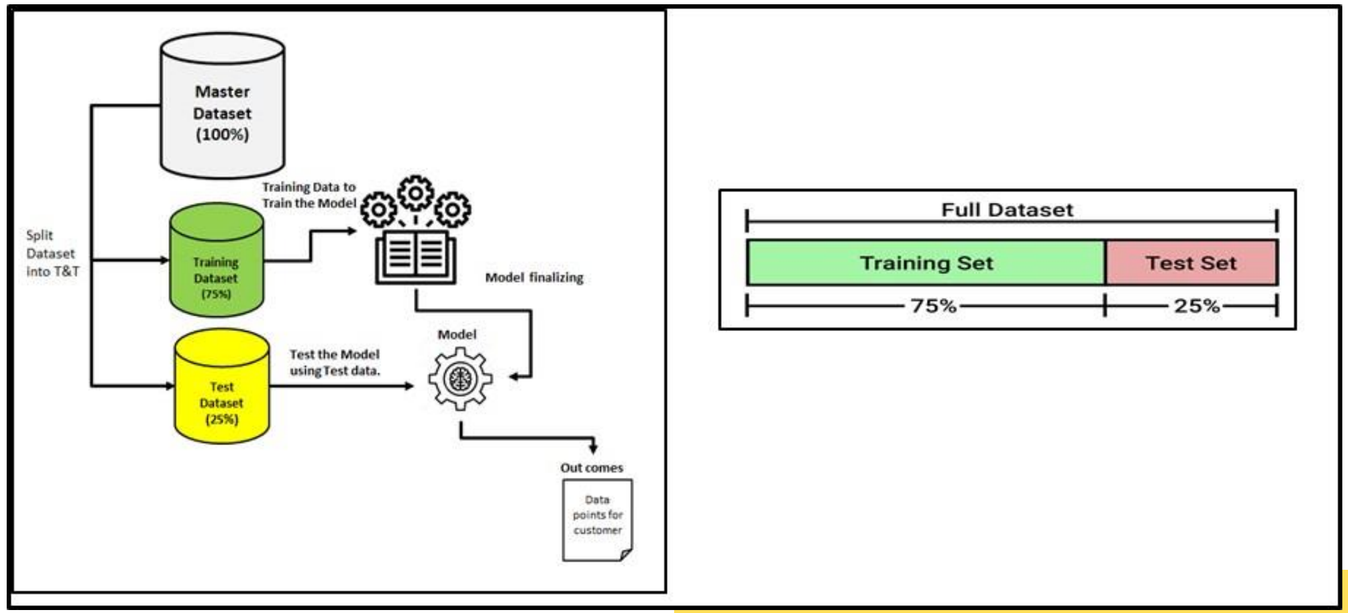
7.Feature engineering can be a manual process, it can be automated

Feature engineering



Training and Testing

- The training data is used to make sure the machine recognizes patterns of the data, cross-validation of data is used to ensure better accuracy and the efficiency of the algorithm which is used to train the machine.
- Test data is used to see how well the machine can predict new answers based on its training.
- The train-test split procedure is used to estimate the ML performance of algorithms when they are used to make predictions on data that is not used to train the model.



Training & Testing

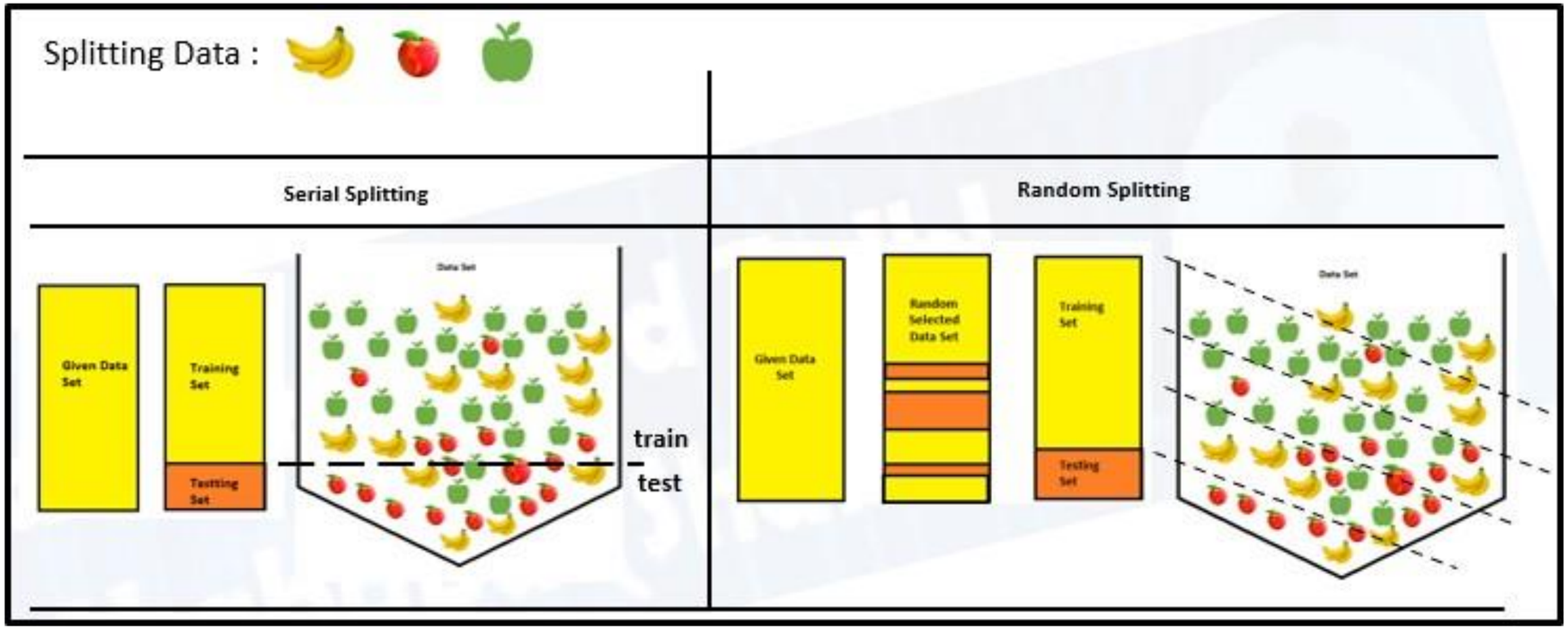
1. Training data is the data set on which you train the model.
2. Train data from which the model has learned the experiences.
3. Training sets are used to fit and tune your models.

1. Test data is the data which is used to check if the model has learnt good enough from the experiences it got in the train data set.
2. Test sets are “unseen” data to evaluate your models.

Train data: It trains our machine learning algorithm
Test data: After the training the model, test data is used to test its efficiency and performance of the model

purpose of the random state in train test split

- **Random state** ensures that the **splits** that you generate are reproducible. The **random state** that you provide is used as a seed to the random number generator. This ensures that the **random numbers** are generated in the same order.



Data Split into Training/Testing Set

1. We used to split a dataset into training data and test data in the machine learning space.
2. The split range is usually 20%-80% between testing and training stages from the given data set.
3. A major amount of data would be spent on to train your model
4. The rest of the amount can be spent to evaluate your test model.
5. But you cannot mix/reuse the same data for both Train and Test purposes
6. If you evaluate your model on the same data you used to train it, your model could be very overfitted. Then there is a question of whether models can predict new data.
7. Therefore, you should have separate training and test subsets of your dataset.

MODEL EVALUATION

Each model has its own model evaluation mythology, some of the best evaluations are here.

Evaluating the Regression Model.

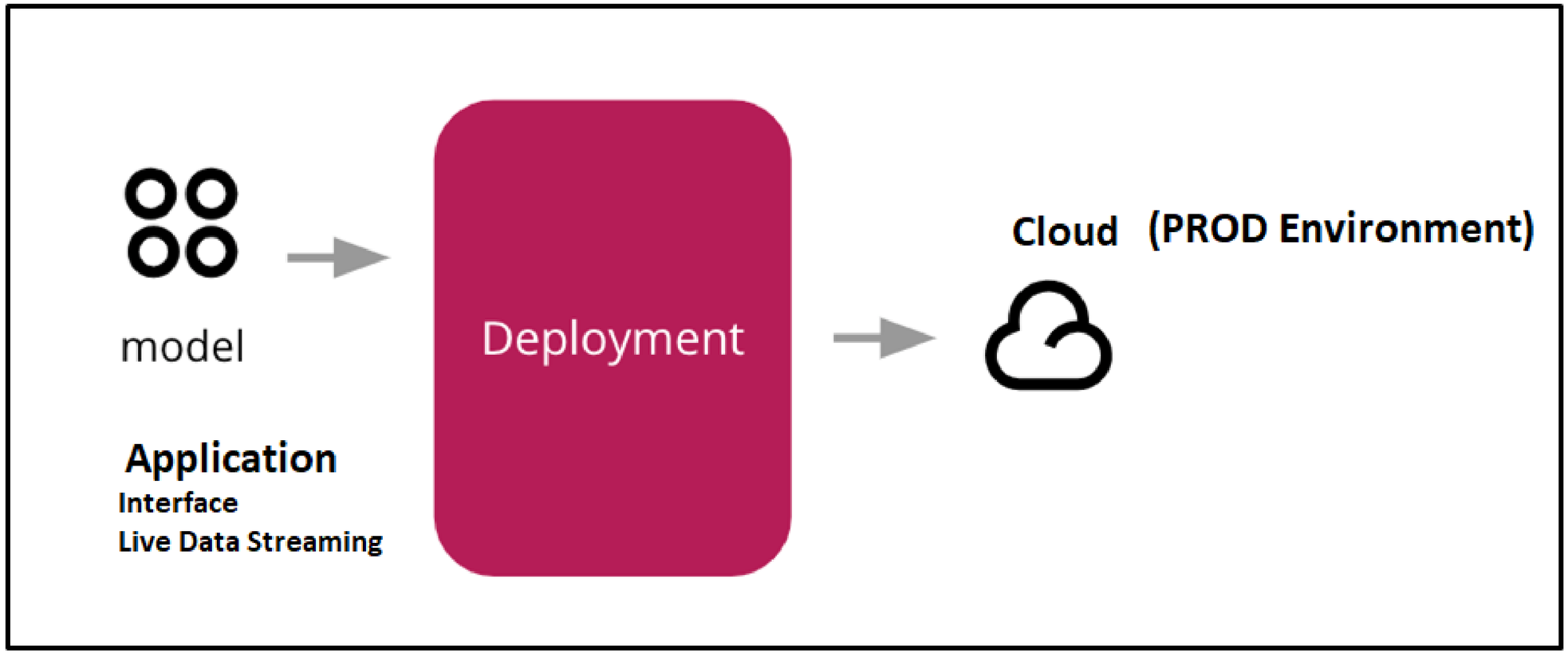
- Sum of Squared Error (SSE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Coefficient of Determination (R2)
- Adjusted R2

Evaluating Classification Model.

- Confusion Matrix.
- Accuracy Score.
- AUC and ROC.|

Deployment of an ML-model

Simply means the integration of the finalized model into a production environment and getting results to make business decisions.



Assessment

A model of language consists of the categories which does not include _____.

A. System Unit
B. structural units.
C. data units
D. empirical units

Different learning methods do not include?
Memorization
Analogy
Introduction
Deduction

Which of the following are ML methods?

A. based on human supervision
B. supervised Learning
C. semi-reinforcement Learning
D. All of the above

REFERENCES

1. Tom M. Mitchell, “Machine Learning”, McGraw-Hill Education (India) Private Limited, 2013.
2. Trevor Hastie, Robert Tibshirani, Jerome Friedman, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”, Springer; Second Edition, 2009.

THANK YOU