Variable Rationalization

- The data set may have a large number of attributes. But some of those attributes can be irrelevant or redundant. The goal of Variable Rationalization is to improve the Data Processing in an optimal way through attribute subset selection.
- This process is to find a minimum set of attributes such that dropping of those irrelevant attributes does not much affect the utility of data and the cost of data analysis could be reduced.
- Mining on a reduced data set also makes the discovered pattern easier to understand. As part of Data processing, we use the below methods of Attribute subset selection

1. Stepwise Forward Selection
2. Stepwise Backward Elimination
3. Combination of Forward Selection and Backward Elimination
4. Decision Tree Induction.

All the above methods are greedy approaches for attribute subset selection.

1. **Stepwise Forward Selection**: This procedure starts with an empty set of attributes as the minimal set. The most relevant attributes are chosen (having minimum p-value) and are added to the minimal set. In each iteration, one attribute is added to a reduced set.

- **Stepwise Backward Elimination**: Here all the attributes are considered in the initial set of attributes. In each iteration, one attribute is eliminated from the set of attributes whose p-value is higher than significance level.
- **Combination of Forward Selection and Backward Elimination:** The stepwise forward selection and backward elimination are combined so as to select the relevant attributes most efficiently. This is the most common technique which is generally used for attribute selection.
- **Decision Tree Induction**: This approach uses decision tree for attribute selection. It constructs a flow chart like structure having nodes denoting a test on an attribute. Each branch corresponds to the outcome of test and leaf

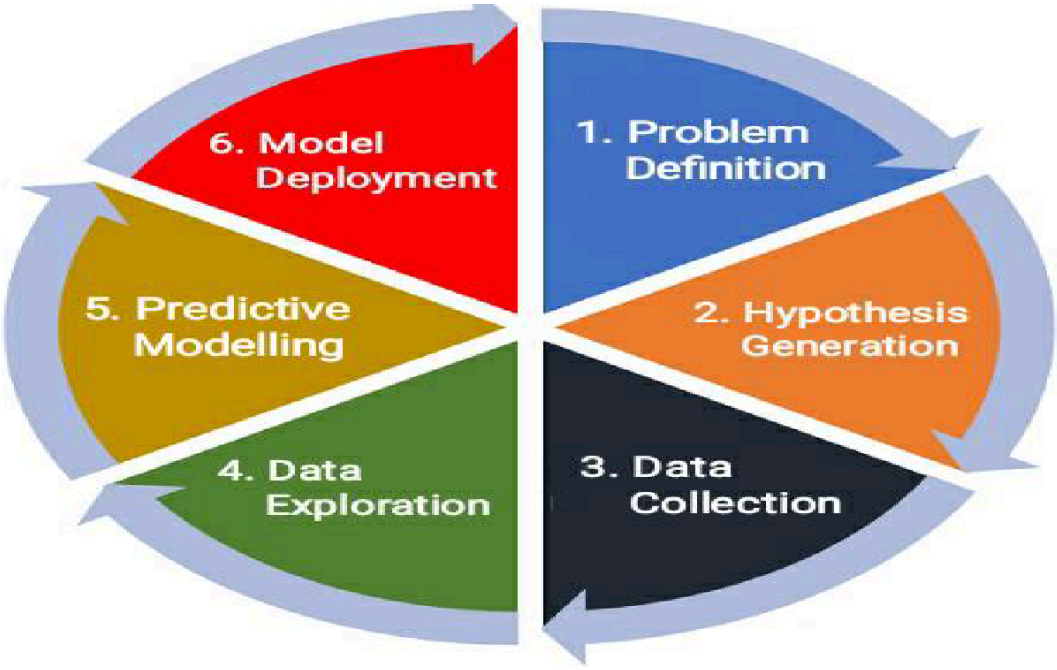
nodes is a class prediction. The attribute that is not the part of tree is considered irrelevant and hence discarded.

Model Building Life Cycle in Data Analytics:

When we come across a business analytical problem, without acknowledging the stumbling blocks, we proceed towards the execution. Before realizing the misfortunes, we try to implement and predict the outcomes. The problem-solving steps involved in the data science model-building life cycle.

Let's understand every model building step in-depth**,**

The data science model-building life cycle includes some important steps to follow. The following are the steps to follow to build a Data Model



1. Problem Definition

2. Hypothesis Generation

3. Data Collection

4. Data Exploration/Transformation

5. Predictive Modelling

6. Model Deployment

1. Problem Definition

The first step in constructing a model is to

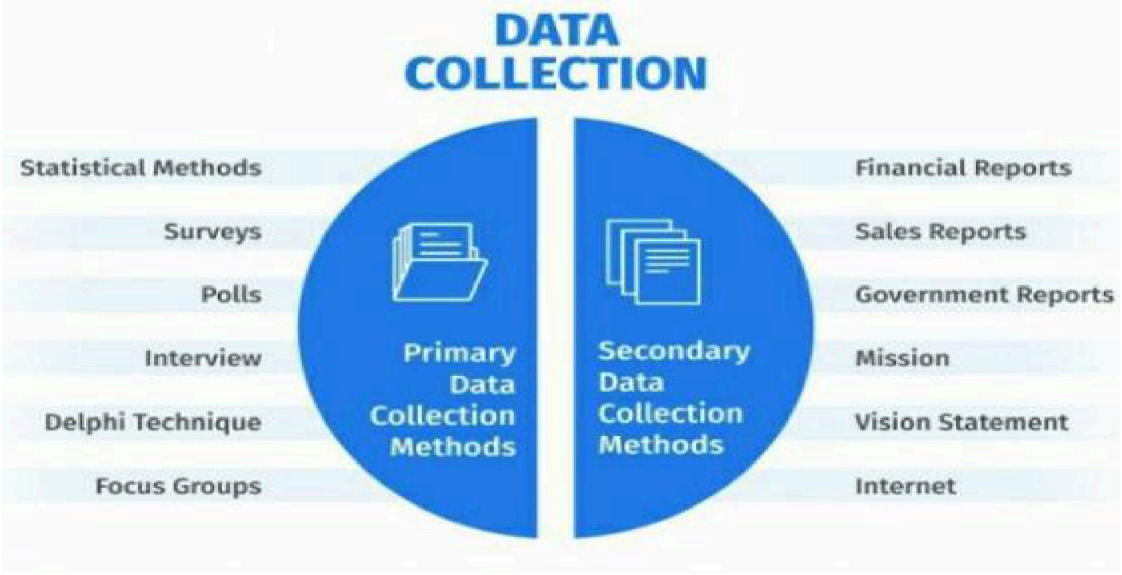
- Understand the industrial problem in a more comprehensive way. To identify the purpose of the problem and the prediction target, we must define the project objectives appropriately.
- Therefore, to proceed with an analytical approach, we have to recognize the obstacles first. Remember, excellent results always depend on a better understanding of the problem.

2. Hypothesis Generation

- Hypothesis generation is the guessing approach through which we derive some essential data parameters that have a significant correlation with the prediction target.
- Your hypothesis research must be in-depth, looking for every perceptive of all stakeholders into account. We search for every suitable factor that can influence the outcome.
- Hypothesis generation focuses on what you can create rather than what is available in the dataset.

3. Data Collection

- Data collection is gathering data from relevant sources regarding the analytical problem, then we extract meaningful insights from the data for prediction.



The data gathered must have:

- Proficiency in answer hypothesis questions.
- Capacity to elaborate on every data parameter.
- Effectiveness to justify your research.
- Competency to predict outcomes accurately.

4. Data Exploration/Transformation

- The data you collected may be in unfamiliar shapes and sizes. It may contain unnecessary features, null values, unanticipated small values, or immense values. So, before applying any algorithmic model to data, we have to explore it first.
- By inspecting the data, we get to understand the explicit and hidden trends in data. We find the relation between data features and the target variable.
- Usually, a data scientist invests his 60–70% of project time dealing with data exploration

only.

Feature Identification:

- You need to analyze which data features are available and which ones are not.
- Identify independent and target variables.
- Identify data types and categories of these variables.

Univariate Analysis:

- We inspect each variable one by one. This kind of analysis depends on the variable type whether it is categorical and continuous.
- Continuous variable: We mainly look for statistical trends like mean, median, standard deviation, skewness, and many more in the dataset.
- Categorical variable: We use a frequency table to understand the spread of data for each category. We can measure the counts and frequency of occurrence of values.

Multi-variate Analysis:

- The bi-variate analysis helps to discover the relation between two or more variables.

- We can find the correlation in case of continuous variables and the case of categorical, we look for association and dissociation between them.

Filling Null Values:

- Usually, the dataset contains null values which lead to lower the potential of the model. With a continuous variable, we fill these null values using the mean or mode of that specific column. For the null values present in the categorical column, we replace them with the most frequently occurred categorical value. Remember, don't delete that rows because you may lose the information.

5. Predictive Modeling

- Predictive modeling is a mathematical approach to create a statistical model to forecast future behavior based on input test data.

  Steps involved in predictive modeling:

- Algorithm Selection:

When we have the structured dataset, and we want to estimate the continuous or categorical outcome then we use supervised machine learning methodologies like regression and classification techniques. When we have unstructured data and want to predict the clusters of items to which a particular input test sample belongs, we use unsupervised algorithms. An actual data scientist applies multiple algorithms to get a more accurate model.

- Train Model:

After assigning the algorithm and getting the data handy, we train our model using the input data applying the preferred algorithm. It is an action to determine the correspondence between independent variables, and the prediction targets.

- Model Prediction:

We make predictions by giving the input test data to the trained model. We measure the accuracy by using a cross-validation strategy or ROC curve which performs well to derive model output for test data.

6. Model Deployment

- There is nothing better than deploying the model in a real-time environment. It helps us to gain analytical insights into the decision-making procedure. You constantly need to update the model with additional features for customer satisfaction.

- To predict business decisions, plan market strategies, and create personalized customer interests, we integrate the machine learning model into the existing production domain.
- When you go through the Amazon website and notice the product recommendations completely based on your curiosities. You can experience the increase in the involvement of the customers utilizing these services. That's how a deployed model changes the mindset of the customer and convinces him to purchase the product.

SUMMARY OF DA MODEL LIFE CYCLE:

- Understand the purpose of the business analytical problem.
- Generate hypotheses before looking at data.
- Collect reliable data from well-known resources.
- Invest most of the time in data exploration to extract meaningful insights from the data.
- Choose the signature algorithm to train the model and use test data to evaluate.
- Deploy the model into the production environment so it will be available to users and strategize to make business decisions effectively.