SUPERVISED MODEL

- Supervised data
  - Supervised data
  - Class variables
- Supervised techniques
  - Classification used for categorical class variable
  - Regression used for numerical class variable

Data Processing transforms data into standard form so as to enable the model to become more applicable, scalable and high in performance.

DATA PREPROCESSING I

Data preprocessing is one of the important phase in the development of Machine Learning models. The key idea of this phase is to assess data on its quality to generate better machine learning models.

- The assessment of data is generally done on the following parameters:
  - **Incomplete:** Incompleteness refers to a situation where proportion of data is not available for some reasons. The unavailability can be in terms of attribute values or an observation.
  - **Noise:** It is a meaningless input/value of a feature(s).
  - **Inconsistency:** Inconsistency is different values taken by same variable at different places.

The high proportion of any of these parameters in the data is a sign of poor data quality.

DATA PREPROCESSING II

Example Illustration

Age	Name	Salary (in K)
34	Ravi	10.13
23	Raju	
67	Kavita	56.23
	Nargis	
	Jayanti	

Data set 1: Example of missing values in the data set

Age	Name	Salary (in K)
34	Ravi	10.13
23	=	89.90
777	Kavita	56.23
77	Nargis	235.6
78	Jayanti	67.78

Data set 2: Example of noise in the data set

Age	Name	Salary (in K)
34	Ravi	10.13
23	Raju	89.90
67	Kavita	56.23
38	Nargis	235.6
45	Jayanti	47.78

Data set 3: Example of Complete the data set

Age	Name	Salary (in K)
34	Ravi	178.13
23	Raju	89.90
67	Kavita	56.23
38	Nargis	235.6
65	Jayanti	47.78

Data set 4: Example of inconsistency w.r.t in the data set 3

Figure 15: Data quality on parameters of incompleteness, noise and inconsistency

DATA PREPROCESSING III

Dealing with Data Incompleteness

There are several methods that can help coping with the issue of Data incompleteness.

- The major methods are:
  - Ignore the data point(s)
  - Replace all incomplete values with "NA" or "Unknown"
  - Each continuous variable can be replaced with mean, and mode can be used for discrete variables

DATA PREPROCESSING III

EXAMPLE ILLUSTRATION

Gender	Age	Age
Male	20	20
Male	21	21
Male	22	22
Male	23	
Female	24	24

If we miss the data point "23" of 4th observation. To avoid the unknown fact here,we can replace the value of the missing data point with the mean of the dataset.

Mean = ⅕(20+21+22+24) = 17.4

DATA PREPROCESSING III

EXAMPLE ILLUSTRATION

1	Male
2	Male
3	Female
4	
5	Male

Example 1

1	Male
2	Male
3	Male
4	N.A
5	Male

Example 2

DATA PREPROCESSING IV

Key characteristics of Data Incompleteness techniques and how to choose among these techniques?

Technique	Key Characteristics and when to choose?
1. Replace with NA	1. Suitable only when fraction of data set is incomplete
2. Replace with Mean	2. May impact performance
3. Ignore data points (conditionally)	3. No special skills are required to apply these techniques

Table 7: Characteristics of Data Incompleteness Techniques

DATA PREPROCESSING V

Dealing with Noise

Noise in data is meaningless value with no significance. The most obvious causes of noise are: measurement and typo errors.

The presence of noise in data may cause hindrance developing appropriate machine learning model.

Following are popular data preprocessing methods for dealing with noise.

1. Binning
2. Machine learning algorithms especially based on proximity (distance) measures such as, K nearest neighbour. These algorithms helps in predicting the most suitable value for the feature value as "noise" in the data set.

DATA PREPROCESSING VI

Binning - Example

Consider we are given data on Age : 4, 8, 15, 21, 21, 24, 25, 28, 34. Binning first sort the data given and uses one of following three methods to smooths the variable Age from noise.

Partition into (equal-frequency) bins

Bin 1:: 4, 8, 15

Bin 2:: 21, 21, 24

Bin 3:: 25, 28, 34

Smoothing by bin means

Bin 1:: 9, 9, 9 $(4 + 8 + 15 = 27 \longrightarrow 27/3 = 9)$

Bin 2:: 22, 22, 22 $(21 + 21 + 24 = 66 \longrightarrow 66/3 = 22)$

Bin 3:: 29, 29, 29 $(25 + 28 + 34 = 87 \longrightarrow 87/3 = 29)$

Smoothing by bin boundaries

Bin 1:: 4, 4, 15	Minimum: 4	Maximum:15
Bin 2:: 21, 21, 24	Minimum: 21	Minimum: 24
Bin 3:: 25, 25, 34	Minimum: 25	Minimum: 34