UNIT-3
PREPARING TO MODEL AND PREPROCESSING

Prepared by : Ritesh Prajapati

INTRODUCTION TO MACHINE LEARNING(4350702)

SEMESTER-5 DIPLOMA COMPUTER ENGINEERING

Topics

3.1.1 Machine Learning activities

- Preparing to Model
- Learning: Data Partition, Model Selection
- Performance Evaluation
- Performance Improvement

3.2.1 Types of Data

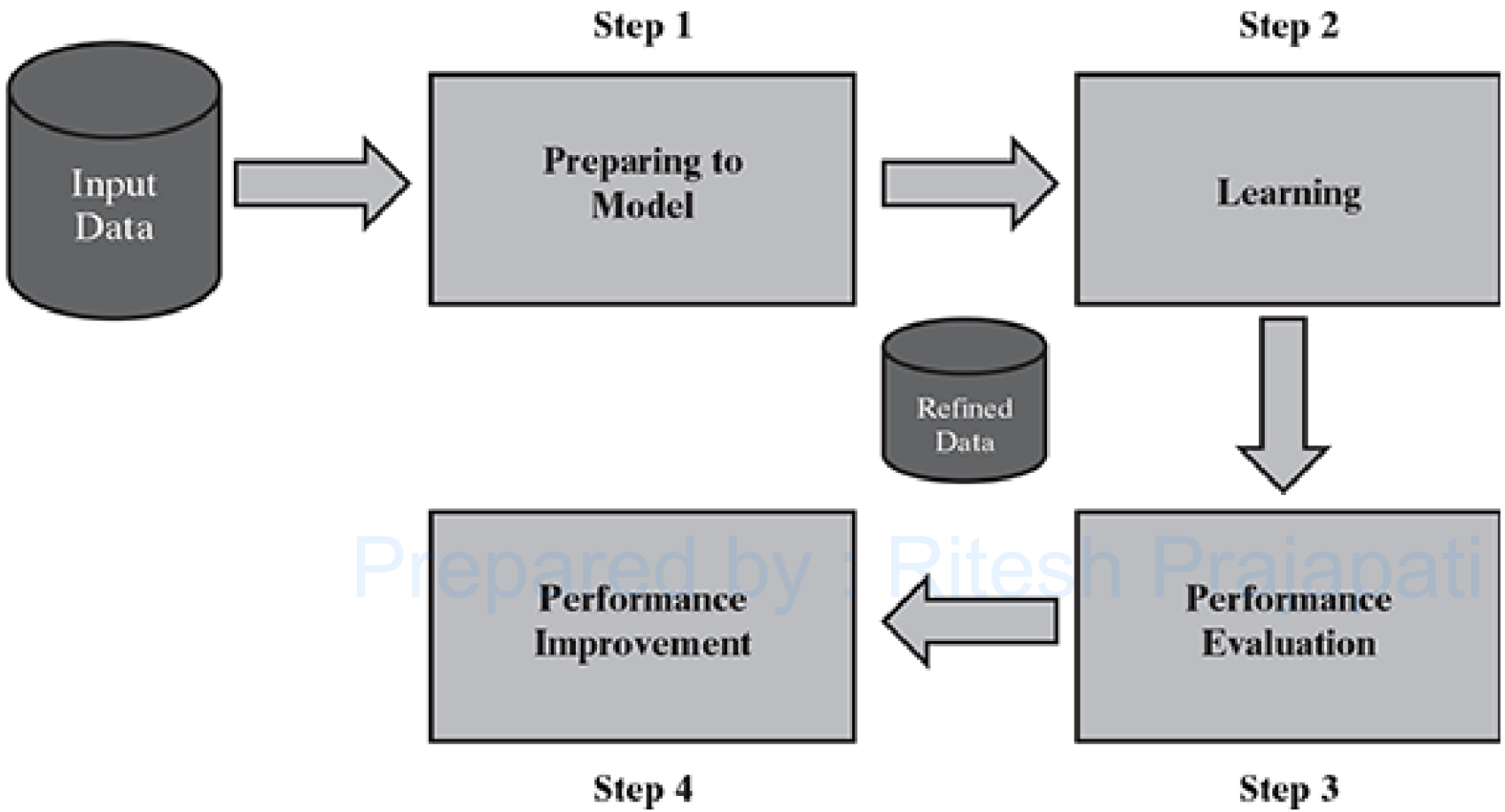
Prepared by : Ritesh Prajapati

3.2.2 Data quality and remediation

3.2.3 Data Pre-Processing

- Dimensionality reduction
- Feature subset selection

Machine Learning activities



Prepared by : Ritesh Prajapati

Machine Learning activities

Step #	Step Name	Activities Involved
Step 1	Preparing to Model	• Understand the type of data in the given input data set • Explore the data to understand data quality • Explore the relationships amongst the data elements, e.g. inter-feature relationship • Find potential issues in data • Remediate data, if needed • Apply following pre-processing steps, as necessary: ✓ Dimensionality reduction ✓ Feature subset selection
Step 2	Learning	• Data partitioning/holdout • Model selection • Cross-validation
Step 3	Performance evaluation	• Examine the model performance, e.g. confusion matrix in case of classification • Visualize performance trade-offs using ROC curves
Step 4	Performance improvement	• Tuning the model • Ensembling • Bagging • Boosting

Prepared by : Ritesh Prajapati

Preparing to Model

- Data Collection is the first step of the machine learning.
- Two types- Labelled and Unlabelled

- The quantity and quality of the collected data will determine the efficiency of the output.
- The more will be the data, the more accurate will be the prediction.

Prepared by : Ritesh Prajapati

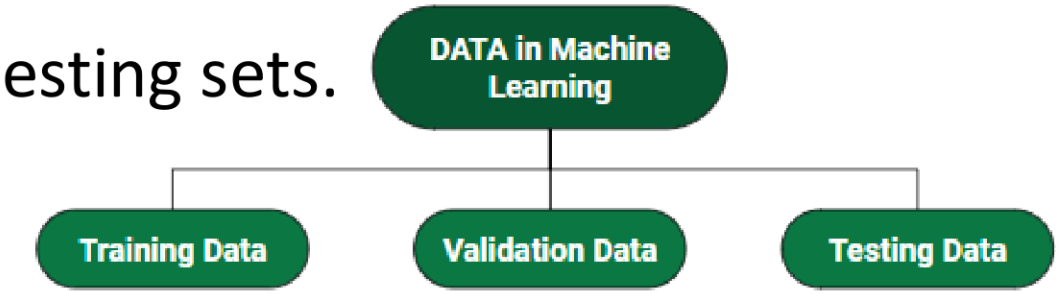
- Steps for data collection
  - Identify various data sources(files, database, internet, or mobile devices.)
  - Collect data
  - Integrate the data obtained from different sources
- Obtain set of data, also called as a **dataset**.

Preparing to Model - Activities

- Understand the type of data in the given input data set
- Explore the data to understand data quality
- Explore the relationships amongst the data elements, e.g. inter-feature relationship
- Find potential issues in data
- Remediate data, if needed
- Apply following pre-processing steps, as necessary:
  - Dimensionality reduction
  - Feature subset selection

Prepared by : Ritesh Prajapati

Learning -1. Data Partition



- Data is divided into training, validation and testing sets.
- **Training set :**
  - Used to train the model
  - It has input samples(features) and their matching values(labels)
- **Validation set :**
  - Used to do evaluation of the model
  - It helps in adjusting the model
- **Testing set :**
  - Used to give a fair assessment of how well the trained model performs
  - When we feed in the inputs of Testing data, our model will predict some values(without seeing actual output).
- Division depends on various factors –dataset size, problem complexity
- Popular division is 70 + 15 + 15

Prepared by : Ritesh Prajapati

Learning -2. Selection of Model

- The structured representation of raw input data to the meaningful pattern is called a model.
- Model selection is selecting the best algorithm and model for a specific problem or dataset.

Two types:

Prepared by : Ritesh Prajapati

- Models for supervised learning, which primarily focus on solving **predictive** problems
- Models for unsupervised learning, which solve **descriptive** problems.

Learning -2. Predictive Model

- Predictive models try to predict certain value using the values in an input data set.
- The learning model attempts to establish a relation between the target feature (the feature being predicted) and the predictor features.

- The models which are used for **predict the value of a category or class to which a data instance belongs to** are known as classification models.
  - Predicting win/loss in a cricket match
  - Predicting whether a transaction is fraud

Prepared by : Ritesh Prajapati

- The models which are used for **prediction of the numerical value of the target feature of a data instance** are known as regression models.
  - Prediction of revenue growth in the succeeding year
  - Prediction of rainfall amount in the coming monsoon

Learning -2. Descriptive Model

- Descriptive models are used to **describe a data set** or gain insight from a data set. there is no target feature.
- Based on the value of all features, **interesting patterns or insights** are derived about the data set.
- Models which group together similar data instances, i.e. **data instances having a similar value of the different features** are called clustering models.
  - Customer grouping or segmentation based on social, demographic, ethnic, etc. factors
  - Grouping of music based on different aspects like genre, language, time-period, etc.
- Models related to **pattern discovery** is used for market basket analysis of transactional data.
  - Possibility of purchasing one product based on the purchase of another product is determined.

Prepared by : Ritesh Prajapati

Performance Evaluation -1. Cross-Validation

- Cross validation is a technique to evaluate the performance of a model
- Steps :
  - Reserve a subset of the dataset as a validation set.
  - Provide the training to the model using the training dataset.
  - Evaluate model performance using the validation set. If the model performs well with the validation set, perform the further step, else check for the issues.
- Types:
  - Validation Set Approach
  - Leave-P-out cross-validation
  - Leave one out cross-validation
  - **K-fold cross-validation**
  - Stratified k-fold cross-validation

Prepared by : Ritesh Prajapati

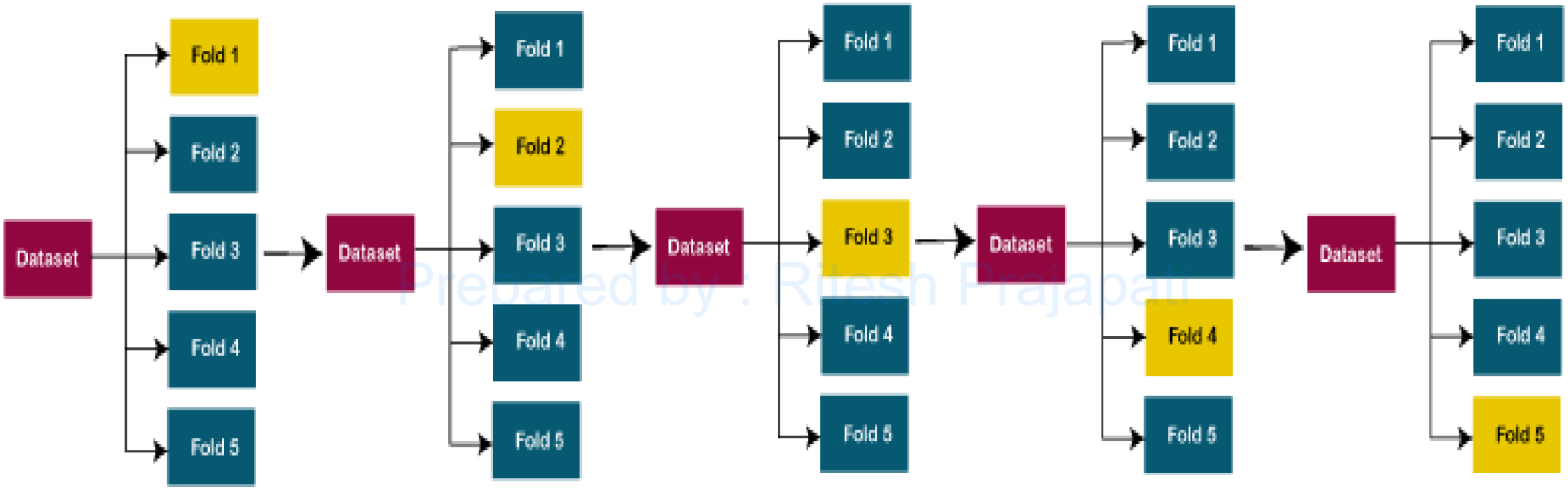
Performance Evaluation -1. Cross-Validation

- K-Fold Cross-Validation
- In this approach divide the input dataset into K groups of samples of equal sizes. These samples are called **folds**.
- Then using one of these folds as a validation set, and training the model on the remaining folds.
- This process is repeated multiple times, each time using a different fold as the validation set
- Finally, the results from each validation step are averaged to produce estimate of the model's performance.

Prepared by : Ritesh Prajapati

Performance Evaluation -1. Cross-Validation

5-Fold Cross-Validation



Prepared by : Ritesh Prajapati

Yellow- Validation set and Blue- Training set

Performance Evaluation- 2. Confusion Matrix

- Used for classification supervised machine learning
- **A matrix containing correct and incorrect predictions in the form of TPs, FPs, FNs and TNs is known as confusion matrix.**

		ACTUAL VALUES	
		Positive	Negative
PREDICTED VALUES	Positive	TP	FP
	Negative	FN	TN

Prepared by : Ritesh Prajapati

- Four possibilities with regards to the cricket match win/loss prediction:

1. model predicted win and the team won – TRUE POSTIVE
2. model predicted win and the team lost – FALSE POSITIVE (Type-1 Error)
3. model predicted loss and the team won – FALSE NEGATIVE (Type-2 Error)
4. model predicted loss and the team lost – TRUE NEGATIVE

Performance Evaluation- 2. Confusion Matrix

- **Model Accuracy** : how often the model predicts the correct output.

$$\text{Model accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

- **Error rate** : how often the model gives the wrong predictions.

$$\text{Error rate} = \frac{FP + FN}{TP + FP + FN + TN}$$

- **Precision**: proportion of positive predictions which are truly positive,

$$\text{Precision} = \frac{TP}{TP + FP}$$

Prepared by : Ritesh Prajapati

- **Recall** : proportion of TP cases over all actually positive cases.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F-measure** :It gives the right measure using which performance of different models can be compared.

$$F\text{-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Performance Evaluation- 2. Confusion Matrix

- Example

Index	1	2	3	4	5	6	7	8	9	10
Actual	Dog	Dog	Dog	Not Dog	Dog	Not Dog	Dog	Dog	Not Dog	Not Dog
Predicted	Dog	Not Dog	Dog	Not Dog	Dog	Dog	Dog	Dog	Not Dog	Not Dog
Result	TP	FN	TP	TN	TP	FP	TP	TP	TN	TN

- Confusion Matrix:

		Actual	
		Dog	**Not Dog**
Predicted	**Dog**	True Positive (TP =5)	False Positive (FP=1)
	Not Dog	False Negative (FN =1)	True Negative (TN=3)

Accuracy = (5+3)/(5+3+1+1) = 8/10 = 0.8

Precision = 5/(5+1) =5/6 = 0.8333

Recall = 5/(5+1) =5/6 = 0.8333

F-measure = (2* 0.8333* 0.8333)/

(0.8333+ 0.8333) = 0.8333

Prepared by : Ritesh Prajapati

Performance Improvement- 1. Model parameter tuning

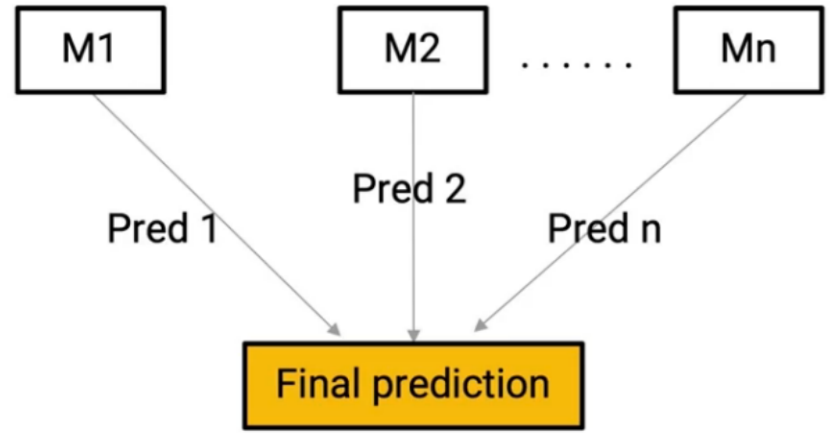
- One effective way to improve model performance is by tuning model parameter.
- Model parameter tuning is the process of adjusting the model fitting options.
- Examples:
  - In k-Nearest Neighbour (kNN), using different values of ‘k’ or the number of nearest neighbours to be considered, the model can be tuned.
  - a number of hidden layers can be adjusted to tune the performance in neural networks model

Prepared by : Ritesh Prajapati

- Most machine learning models have at least one parameter which can be tuned.

Performance Improvement- 2. Ensemble Model

- The approach of combining different models with diverse strengths is known as ensemble.
- Ensemble methods combine weaker learners to create stronger ones.

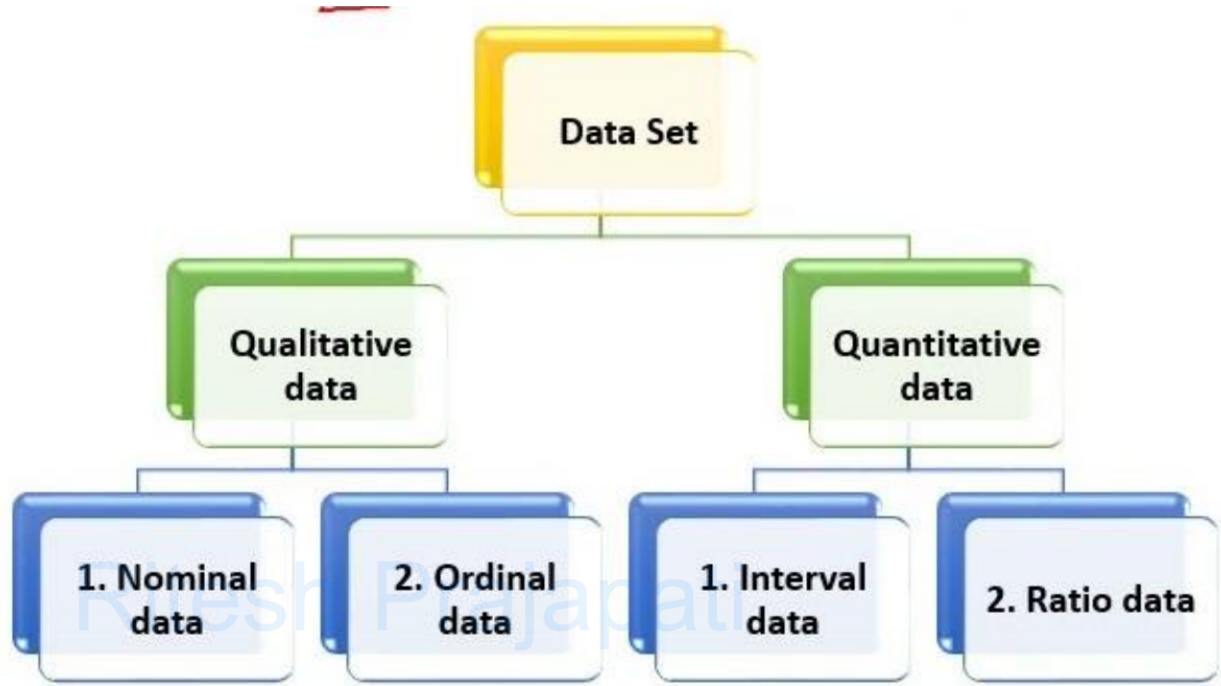


Prepared by : Ritesh Prajapati

- Bagging : Generates multiple training data sets. These training data sets are used to train a set of models using the same learning algorithm. Then the outcomes of the models are combined.
- Boosting : Weaker learning models are trained on resampled data and the outcomes are combined based on the performance of different models.

Types of Data

- Data Set
- Qualitative data
- 1. Nominal data
- 2. Ordinal data
- Quantitative data
- 1. Interval data
- 2. Ratio data



Prepared by : Ritesh Prajapati

Types of Data - Qualitative

- **Qualitative data** provides information about the quality of an object or information which cannot be measured.
  - quality of performance of students in terms of ‘Good’, ‘Average’, and ‘Poor’
  - name or roll number of students - cannot be measured
- **Nominal data** is one which has no numeric value, but a named value. It is used for assigning named values to attributes.
  - Blood group: A, B, O, AB, etc.
  - Gender: Male, Female, Other
- **Ordinal data** are like nominal data that can also be naturally ordered. It can be arranged in a sequence of increasing or decreasing value so that we can say whether a value is better than or greater than another value.
  - Customer satisfaction: ‘Very Happy’, ‘Happy’, ‘Unhappy’, etc.
  - Grades: A, B, C, etc.

Prepared by : Ritesh Prajapati

Types of Data - Qualitative

- **Examples of Nominal Data :**
- Hair color (blonde, gray, brown, black, etc.)
- Nationality (Kenyan, British, Chinese, etc.)
- Relationship status (married, cohabiting, single, etc.)
- Preferred mode of public transportation (bus, train, tram, etc.)
- Blood type (O negative, O positive, A negative, and so on)
- Political parties voted for (party X, party Y, party Z, etc.)
- Employment status (employed, unemployed, retired, etc.)
- **Examples of Ordinal Data :**
- Economic status (poor, middle income, wealthy)
- Income level in non-equally distributed ranges ($10K-$20K, $20K-$35K, $35K-$100K)
- Course grades (A+, A-, B+, B-, C)
- Education level (Elementary, High School, College, Graduate, Post-graduate)
- Likert scales (Very satisfied, satisfied, neutral, dissatisfied, very dissatisfied)
- Military ranks (Colonel, Brigadier General, Major General, Lieutenant General)
- Age (child, teenager, young adult, middle-aged, retiree)

Prepared by : Ritesh Prajapati

Types of Data - Quantitative

- **Quantitative data** provides information about the quantity of an object – hence it can be measured.
  - the attribute ‘marks’ - can be measured
- **Interval data** is numeric data for which the order is known, and the exact difference between values is also known.
- Addition, subtraction. mean, median, mode can also be calculated.
- Do not have a ‘true zero’ value.
  - Example - Celsius temperature date, time, etc.
- **Ratio data** represents numeric data for which exact value can be measured.
- Addition, subtraction. mean, median, mode can also be calculated.
- Absolute zero is available for ratio data.
  - Example - height, weight, age, salary, etc.

Prepared by : Ritesh Prajapati

Types of Data - Quantitative

- **<u>Examples of Interval Data :</u>**
- Temperature in Fahrenheit or Celsius (-20, -10, 0, +10, +20, etc.)
- Times of the day (1pm, 2pm, 3pm, 4pm, etc.)
- Income level on a continuous scale ($10K, $20K, $30K, $40K, and so on)
- IQ scores (100, 110, 120, 130, 140, etc.)
- pH (pH of 2, pH of 4, pH of 6, pH of 8, pH of 10, etc.)
- Credit ratings (20, 40, 60, 80, 100)
- Dates (1740, 1840, 1940, 2040, 2140, etc.)

Prepared by : Ritesh Prajapati

- **<u>Examples of Ratio Data :</u>**
- Temperature in Kelvin (0, +10, +20, +30, +40, etc.)
- Height (5ft. 8in., 5ft. 9in., 5ft. 10in., 5ft. 11in., 6ft. 0in. etc.)
- Price of goods ($0, $5, $10, $15, $20, $30, etc.)
- Age in years (from zero to 100+)
- Distance (from zero miles/km upwards)

Types of Data - Summary

	Nominal	Ordinal	Interval	Ratio
Categorizes and labels variables	✓	✓	✓	✓
Ranks categories in order		✓	✓	✓
Has known, equal intervals			✓	✓
Has a true or meaningful zero				✓

Prepared by : Ritesh Prajapati

Data quality

- Success of machine learning depends largely on the quality of data.
- Not realistic to expect that the data will be flawless.
- Two types of problems(data quality issues):
  - Certain data elements without a value or data with a missing value.
  - Data elements having value surprisingly different from the other elements (outliers).
- Factors which lead to these data quality issues.

Prepared by : Ritesh Prajapati

1. **Incorrect sample set selection:**
   - Data used for prediction/analysis does not represent desired population or time period
2. **Errors in data collection:**
   - resulting in outliers and missing values

Data Remediation

- The issues in data quality need to be remediated to achieve the right amount of efficiency in the learning activity.
- Remedial steps need for the second point(human errors happen)

1. Handling outliers

- Outliers are data elements with an abnormally high value which may impact prediction accuracy
- **Remove outliers**: If the number of records which are outliers is not many, a simple approach may be to **remove them.**
- **Imputation**: to **impute** the value with mean or median or mode. The value of the most similar data element may also be used for imputation.
- **Capping**: Limiting extreme values by **capping** the maximum and minimum values within a reasonable range.

Prepared by : Ritesh Prajapati

Data Remediation

2. Handling missing values

- **Eliminate records having a missing value of data elements**
  - Remove the records having data elements missing value .
- **Imputing missing values**
  - To assign a value to the data elements having missing values.
  - Mean/mode/median is most frequently assigned value.

Prepared by : Ritesh Prajapati

DATA PRE-PROCESSING

1. **Dimensionality reduction**

- Most of the machine learning algorithms perform better if the dimensionality of data set, i.e. **the number of features in the data set**, is reduced.
- Dimensionality reduction refers to the techniques of reducing the dimensionality of a data set by creating new attributes by combining the original attributes.
  - **Principal Component Analysis (PCA).**
  - PCA is a technique to convert a set of correlated variables into a set of transformed, uncorrelated variables called principal components.
  - The principal components are a linear combination of the original variables.
  - They are orthogonal to each other.
  - Since principal components are uncorrelated, they capture the maximum amount of variability in the data.

Prepared by: Ritesh Prajapati

DATA PRE-PROCESSING

2. **Feature subset selection**

- Feature subset selection is to find out **the optimal subset of the entire feature set.**
- It will reduce computational cost without any major impact on the learning accuracy.
- A feature is irrelevant if it **plays an insignificant role** in classifying a set of data instances.
- A feature is redundant when the information contributed by the feature is **more or less same as any other features**.

Prepared by : Ritesh Prajapati

THANK YOU

Prepared by : Ritesh Prajapati