

UNIT-3

Introduction to Machine Learning

Unit III: Introduction to Machine Learning	<p>3a.Explain the Installation process of Anaconda Jupyter Notebook</p> <p>3b.Explain the reading data procedure file(csv/excel) in python</p> <p>3c.Explain the machine learning workflow.</p> <p>3d.Explain how to evaluate performance of model using evaluation metrics.</p> <p>3e.Explain the significance of hyper parameter tuning to improve the accuracy of the model.</p> <p>3f. List various cross validation methods in machine learning</p>	<p>3.1 Definition of Machine Learning. Reading csv, excel files in python,</p> <p>3.2 Classification of Machine learning- Supervised and Unsupervised, and Reinforcement</p> <p>3.3 Introduction to Predictive modeling. Stages of predictive modeling(Machine learning Pipeline)- Problem definition, Hypothesis generation, Data Extraction/collection, Data exploration and transformation. Splitting dataset into training and test set, Model selection, Model deployment / implementation.</p> <p>3.4 Evaluation Metrics - Confusion matrix Accuracy, Precision and Recall, Sensitivity and Specificity.</p> <p>3.5 Validation-k-fold cross validation, Hyperparameter tuning.</p>
---	--	--

3.1 Definition of Machine learning:

Machine learning is an application of AI that enables systems to learn and improve from experience without being explicitly programmed.

OR

Machine learning focuses on developing computer programs that can access data and use it to learn for themselves.

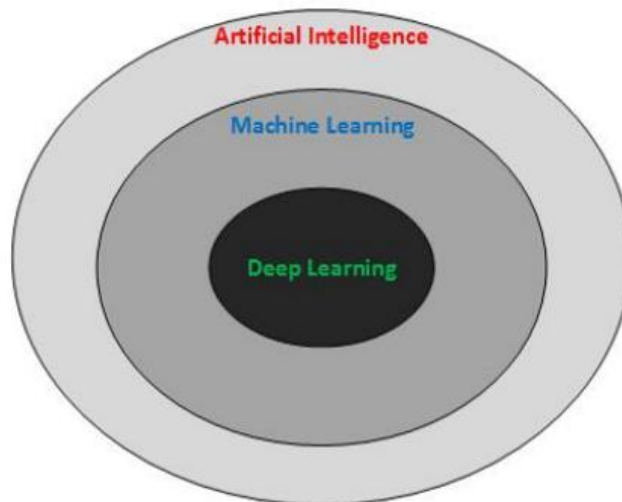
OR

Machine learning is a method of teaching machines to learn things and improve predictions/behaviour based on data on their own.

Applications of Machine Learning:

Social Media <ul style="list-style-type: none">• Recommendation Engine•	Banking <ul style="list-style-type: none">• Credit Scoring•	e-Commerce <ul style="list-style-type: none">• Discount Optimization• ...	Search Engines <ul style="list-style-type: none">• Search algorithm Optimization• ...	Insurance <ul style="list-style-type: none">• Agent Performance Model• ...	Sports <ul style="list-style-type: none">• Predicting the score of match• ...	Healthcare <ul style="list-style-type: none">• Healthcare Claims Fraud detection• ...	Media <ul style="list-style-type: none">• Recommendation•
Automobile <ul style="list-style-type: none">• Demand forecasting•	Manufacturing <ul style="list-style-type: none">• Predictive Asset Maintenance•	Market Place <ul style="list-style-type: none">• Predicting higher CTR• ...	Pharma <ul style="list-style-type: none">• Medicine research• ...	Network <ul style="list-style-type: none">• Optimizing network bandwidth• ...	Telecom <ul style="list-style-type: none">• Product Pricing• ...	Restaurant <ul style="list-style-type: none">• Customer Segmentation• ...	Real State <ul style="list-style-type: none">• Price optimization•
Medical <ul style="list-style-type: none">• Predicting patient health•	Electronics <ul style="list-style-type: none">• Price Optimization•	Retail <ul style="list-style-type: none">• Market Basket Analysis• ...	Airline <ul style="list-style-type: none">• Estimating Air traffic• ...	Transportation <ul style="list-style-type: none">• Finding the right route• ...	Education <ul style="list-style-type: none">• Skill Assessment• ...	Dairy <ul style="list-style-type: none">• Healthcare Claims Fraud detection• ...	Defence <ul style="list-style-type: none">• Identifying Flying Objects•

It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine algorithms are used in a wide variety of applications, such as medical diagnostic, email and computer vision, where it is difficult or infeasible to develop a conventional algorithm effectively performing the task.



Introduction to Pandas:

Pandas is a very powerful python data analysis toolkit for reading, filtering, manipulating, visualizing and exporting data. That means it is a python library which can perform all the functions required for processing data very efficiently.

List of Best Python IDEs for Machine Learning

- Scientific Python Development Environment (Spyder)
- Thonny
- JupyterLab/Jupyter notebook
- Google colab
- PyCharm
- Visual Code
- Atom

Functionalities of Pandas:

- Reading different varieties of data like csv, excel and json etc.
- Functions for filtering, selecting and manipulating data.
- Helps for visually exploring data i.e plotting data for visualization and exploration purpose.

Reading spreadsheet(csv and excel) file in pandas:

Different file that pandas can read are given in the following table.

Pandas can help you read data from different types of files		
Format Type	Data Description	Reader
text	CSV	read_csv
text	JSON	read_json
text	HTML	read_html
text	Local clipboard	read_clipboard
binary	MS Excel	read_excel
binary	HDF5 Format	read_hdf
binary	Feather Format	read_feather
binary	Msgpack	read_msgpack
binary	Stata	read_stata
binary	SAS	read_sas
binary	Python Pickle Format	read_pickle
SQL	SQL	read_sql
SQL	Google Big Query	read_gbq

We will focus on two types of files i.e csv and excel. **read_csv** function to read csv file and **read_excel** function to read excel file.

Steps to read csv and excel file in Jupyter notebook inside pandas.

- Data is present in two files, first is a csv file named data.csv and second is an excel file named data.excel.
- First import library first i.e. pandas.

```
In [1]: # importing pandas library
import pandas as pd
```

- Let us read first csv file. csv file can be read in pandas using read_csv function and file we are going to read is data.csv
- First of all we create a variable i.e df (which will store the data frame). We will call the read_csv function inside the pandas package.
- We need to supply the name of the file that we want to read with its full path where it is stored.

```
In [ ]: # reading the csv file
df=pd.read_csv("data")
```

- Now the file has been read. To confirm that particular file only read, we can check the first few rows using the **head** function.

```
In [3]: df.head()
```

Out[3]:

	ID	Clump_Thickness	Cell_Size_Uniformity	Cell_Shape_Uniformity	Marginal_Adhesion	Single_Epithelial_Cell_Size	Bare_Nuclei	Bland_Chromatin	Norm
0	1000025	5	1	1	1	2.0	1	3	
1	1002945	5	4	4	5	NaN	10	3	
2	1015425	3	1	1	1	2.0	2	3	
3	1016277	6	8	8	1	3.0	4	3	
4	1017023	4	1	1	3	2.0	1	3	

- This confirms that the data.csv file has been read properly.
- To read an excel file which is very much similar to read csv file. Only difference is instead of read_csv function, we have to use read_excel function.

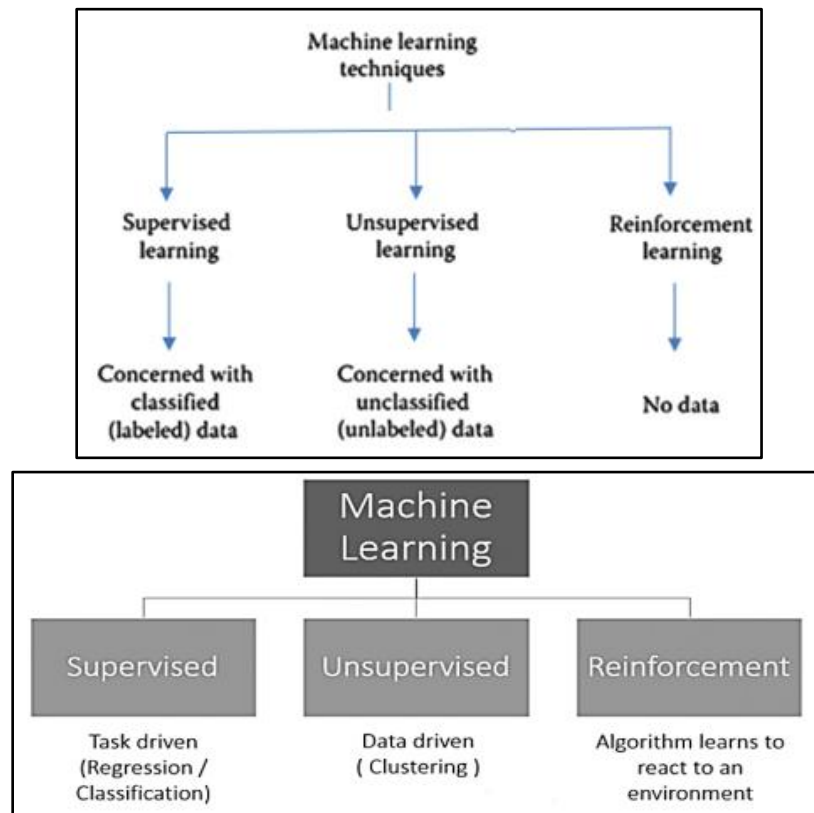
```
In [6]: # reading excel file
df1=pd.read_excel("data.xlsx")

In [7]: df1.head()
```

Out[7]:

	ID	Clump_Thickness	Cell_Size_Uniformity	Cell_Shape_Uniformity	Marginal_Adhesion	Single_Epithelial_Cell_Size	Bare_Nuclei	Bland_Chromatin	Norm
0	1000025	5	1	1	1	2.0	1	3	
1	1002945	5	4	4	5	NaN	10	3	
2	1015425	3	1	1	1	2.0	2	3	
3	1016277	6	8	8	1	3.0	4	3	
4	1017023	4	1	1	3	2.0	1	3	

3.2 Classification of machine learning:



SUPERVISED LEARNING:

- Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well **labeled**. Which means data is already **tagged with the correct answer**.
- After that, the machine is provided with a new set of examples (data) so that the supervised learning algorithm analyses the training data (set of training examples) and produces **a correct outcome** from labeled data.
- Supervised learning is classified into two categories of algorithms:
 - **Classification:** A classification problem is when the output variable is a categorical variable, such as “Red” or “blue”, “disease” or “no disease”.
 - **Regression:** A regression problem is when the output variable is a real value/continuous value, such as “dollars” or “weight”.

UNSUPERVISED LEARNING:

- Unsupervised learning is the training of a machine using information that is **neither classified nor labeled** and allowing the algorithm to act on that information without guidance.

- Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

REINFORCEMENT LEARNING:

- The machine is exposed to an environment where it gets trained by **trial and error method**.
- The machine learns from past experience and tries to capture the best possible knowledge to make accurate decisions based on the feedback received.

3.3 Introduction to Predictive modeling:

Predictive Modeling Does Two Functions:

- Making use of past data and other attributes
- Predict the future using this data

Example:

- Recommending movies to users based on attributes like gender, age, location, past movies.
- Identifying the factors of sales reduction of shampoo products last year.
- Viewing website's today traffic using google analytics
- Predicting stock prices will go up or down tomorrow.

Different stages of Predictive modeling/Machine learning Pipeline:

We can broadly divide the model building life cycle in six stages:



1. Problem definition:

- It is actually to define a problem you want to solve.
- It is identifying the right problem statement, ideally formulating the problem mathematically.
- A problem statement defines the gap between your desired goal and the current state of things.

2. Hypothesis generation:

- List down all possible variables which might influence the problem objective.
- Quality of the model is heavily dependent on the quality of the hypothesis.
- Let the problem statement be “**To predict the default rate of the customer.**”
- The factors that can impact the customer will default or not?
 - Income: Higher income means higher chance of financial stability which leads to a lesser default rate.
 - Job type: If the job is more stable that can lead to lower default rate.
 - Credit history: Your previous good repayment behaviour suggests that you know how to use credit products and that leads to lower default rate.
 - Education: Higher education is directly proportional to awareness about the use of credit products and that may impact the default rate also.
- Hypothesis Generation should be done before data collection to know all factors which might affect the problem without being biased. It stops time wastage in analyzing all available data.

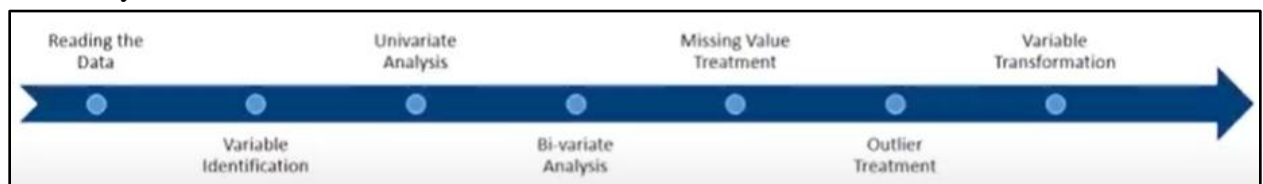
3. Data Extraction/Collection:

- After hypothesis generation, we will come to know which all factors affect the problem.
- We extract or collect from different sources and combine those for exploration and model building.

4. Data exploration and Transformation:

- Necessity of Data exploration: Data is available in tabular form and making sense out of it is very hard. We need some insights from this data for effective analysis. This separates a good analyst Vs Bad analyst.
- Good analyst-Knows his/her data very well. So he can modify data to choose the best one from available data.

Bad analyst: Relies on tools and libraries.

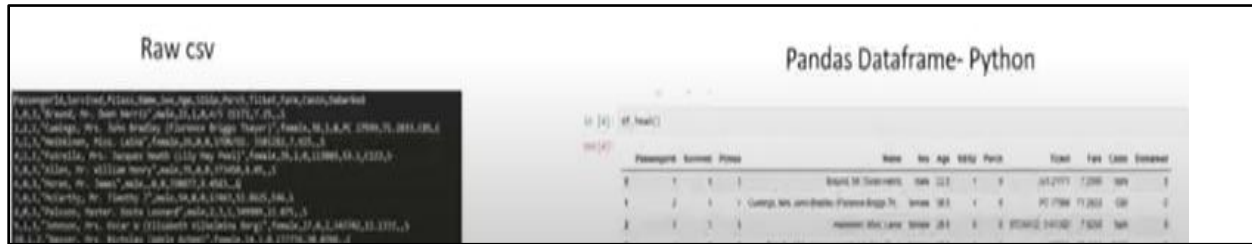


Steps of Data exploration:

- Reading data
- Variable identification
- Univariate Analysis
- Bivariate Analysis
- Missing value treatment
- Outlier treatment
- Variable transformation

i. Reading data:

We read raw data available into the analysis system like csv file into pandas.



ii. Variable identification:

We identify the predictor and target variable.

What is variable identification?

Variable identification is a process of identifying which variables are

- Independent and dependent variables.
- Continuous and categorical variables.

Independent and dependent variables:

- Dependent variables are the variables which we are trying to predict.
Ex: Survived variable is dependent variable in titanic dataset.
- Independent variables are the variables which help in predicting the dependent variables.
Ex: Age, Sex, Fare etc in titanic dataset.

Continuous and Categorical variables:

- Categorical variables are discrete in nature.
Ex: Survived, Sex in titanic dataset.
- Continuous variables can have infinite number of possible values
Ex: Fare, Age in titanic dataset.

How to identify categorical and continuous variables in Pandas.

- Categorical variables- Stored as Objects in Pandas.
- Continuous variables- Stored as int or float on Pandas.

Pandas provides a method called **dtypes** which tells where the variable is object type or int/float.

iii. Univariate Analysis:

- We analyze variables one by one using methods such as bar plots or histograms.
- **What is univariate analysis?**
 - Explore one variable at a time.
 - Summarize the variable
 - Make sense out of that summary to discover insights, anomalies/outliers etc.
- There are different types of methods for different variables.

- **Univariate analysis for continuous variables.**

- While working into continuous variables, we will see central tendency and dispersion such as **mean, median and standard deviation.**
- Distribution of variables: symmetric/right skewed/left skewed.
- It helps us find missing values.
- To find the presence of outliers.
- Two methods to analyse univariate analysis for continuous variables

Tabular method:

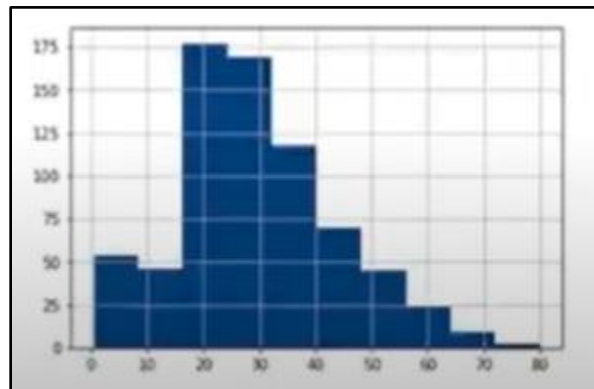
For analyzing mean, median, standard deviation and missing Values

describe() gives count, minimum value, mean, std div, 25%, 50%, 75% and maximum value of all continuous variables.

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	445.000000	0.383838	2.308642	29.599118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	445.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

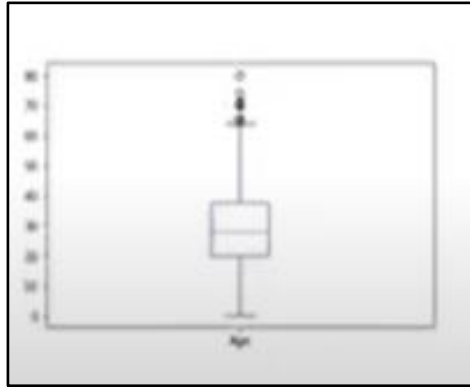
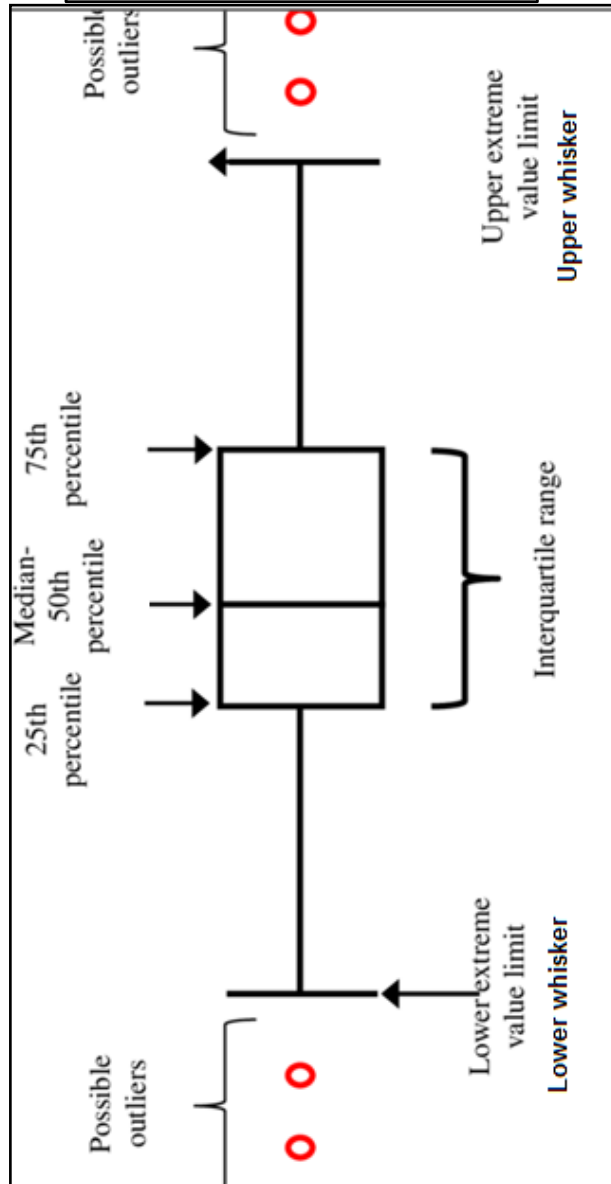
Graphical method: For checking distribution of variables, Presence of Outliers.

- **Histogram-**Gives the distribution of variable.



To generate histogram the function in pandas is `pd.dataframe.hist()`

- **Boxplot:** To detect outliers.



- **Univariate analysis for Categorical variables.**

- Count- absolute frequency of each category in a categorical variable
- count%-percentage of frequency of categorical variable rather than absolute frequency.
- Two methods to analyze univariate analysis for categorical variables

Tabular method:

Frequency tables

In pandas, frequency tables can be created by

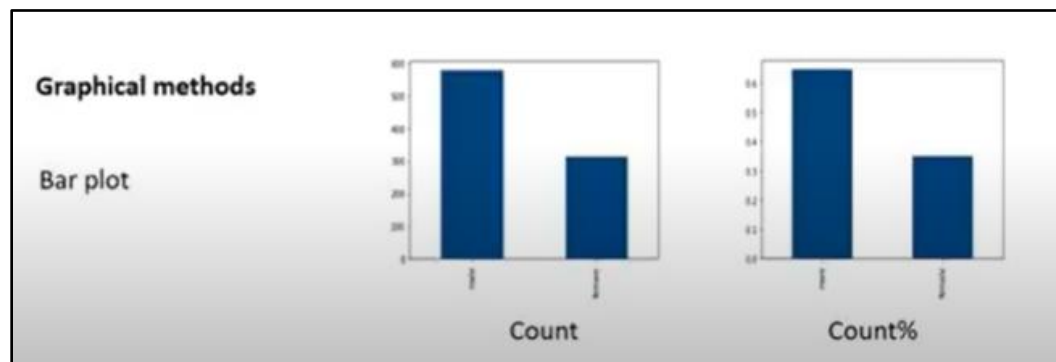
`df['Sex'].value_counts()`

Tabular methods	male	577	}	Count
	female	314		
Frequency Table	male	0.647587	}	Count%
	female	0.352413		

- **Graphical method:**

Bar plots

To visualize the frequency of categorical variables.



In Pandas ,bar plot can be drawn using the following function.

`df['Sex'].value_counts().plot.bar()`

iv. **Bivariate analysis:**

What is Bivariate analysis:

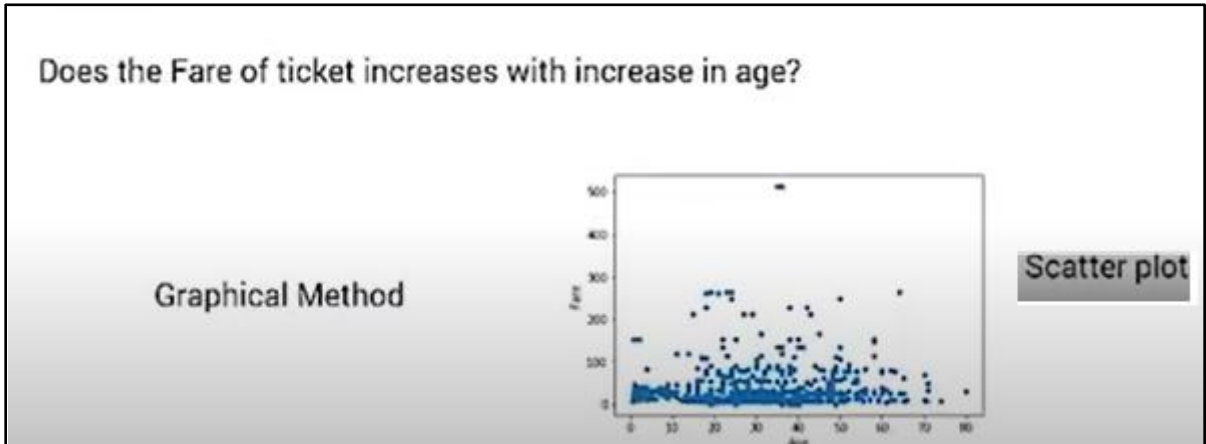
- When two variables are studied together for their empirical relationship.
 - When you want to see whether the two variables are associated with each other.
- Ex: Relation between age and height.

Why do we need bivariate analysis:

- It helps in prediction[when two variables are associated with other]
- It helps in detecting anomalies.

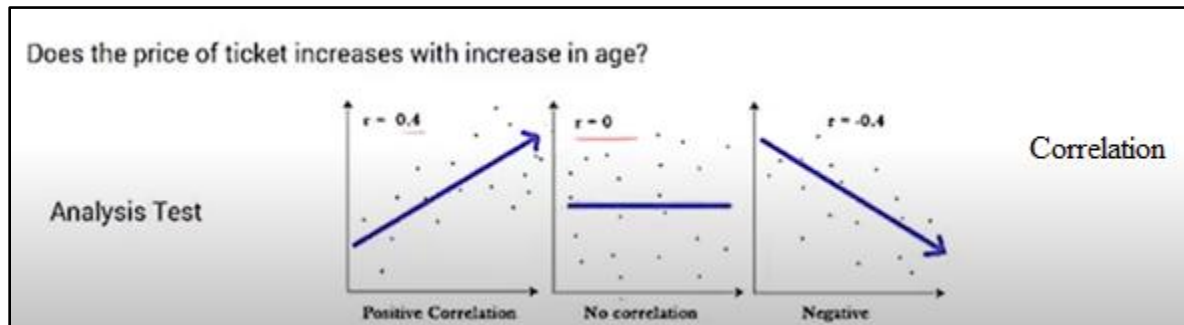
Types of Bivariate analysis.

- Continuous-Continuous Analysis



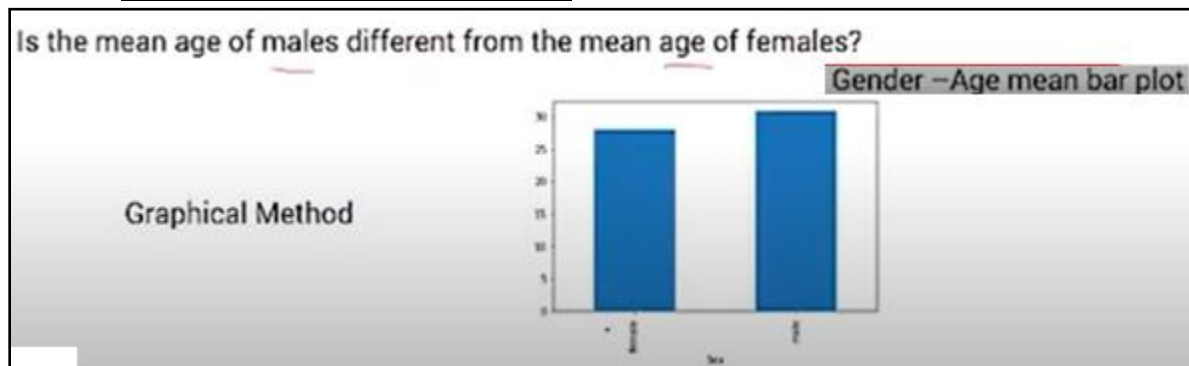
Fare and Age are continuous variables. To understand the relationship between these variables, we can visualize using scatterplot.

To quantify the relationship, between two continuous variables, we check co-relation between them.



- If r (correlation coefficient) is positive value, then positive correlation i.e, if one variable increases, other also increases.
- If r is 0, there is no linear relation between two variables.
- If r is negative. then positive correlation i.e, if one variable increases, other also decreases..

- Categorical-Continuous Analysis



Gender (male and female) is categorical variable and Age is continuous variable. Bar plot gives the relation between these two variables.

- **Categorical-Categorical Analysis**

Does gender have any effect on the survival rates?

Two way table

	Survived	0	1
Graphical Method	Sex		
	female	81	233
	male	468	109

v. **Missing value treatment**

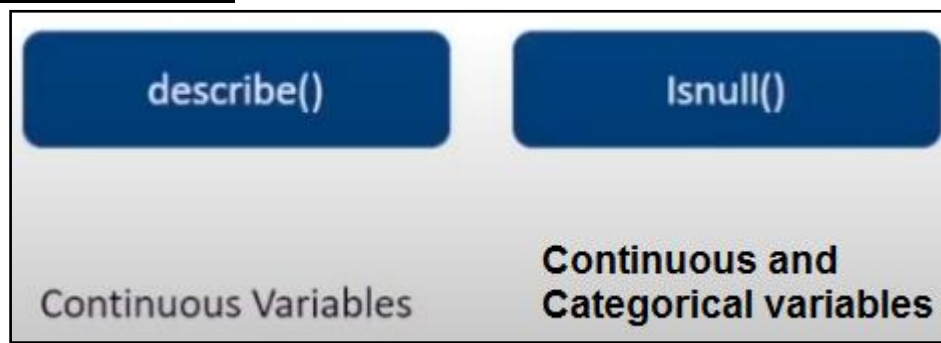
Reasons for Missing values in a dataset.

- Non-response- When we collect data on people's income and they may not answer.
- Error in Data collection
- Error in reading data

Types of Missing Values

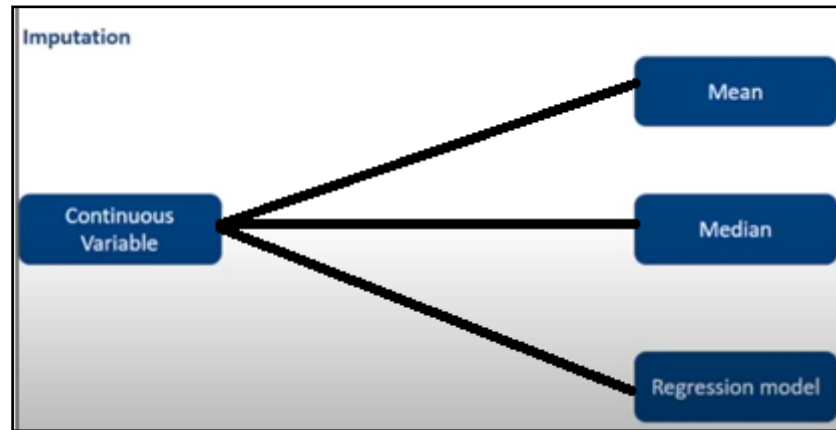
- Missing Completely at Random(MCR)-When missing values have no relation to the variable in which missing values exist and also no relation with any other variable in the dataset.
- Missing At Random(MAR)-Missing values have no relation in which missing values exist i.e., they have relation with other variables.
- Missing Not At Random(MNAR)- Missing values have relation in which missing values exist.

Identifying Missing values

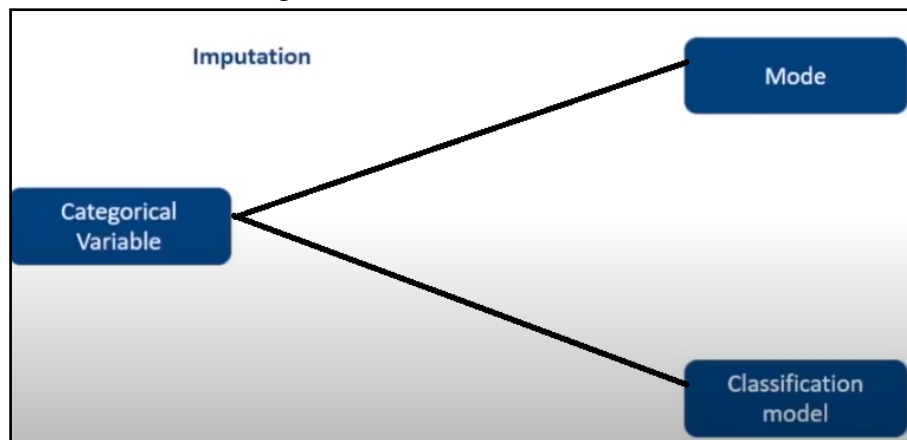


Different methods to Deal with Missing values

- **Imputation**
Imputation methods for Continuous variables:



Imputation methods for Categorical variables:



- **Deletion**

Row wise deletion

Age	Variable 1	Variable 2	Variable 3	Variable 4
10	77		47	
15	114	114	80	114
25	98		56	98
35	72	72	45	
40	119		74	119
45	76	76	56	
55	98	98	78	98
60	114	114	114	114
65	76	76	76	
70	78		78	
75	88	88	88	
80	112	112	112	112
85	89		89	
90	97		97	97

Row wise deletion

If we perform this deletion, Entire row will be deleted if even one missing value is there in any one row.

Age	Variable 1	Variable 2	Variable 3	Variable 4
10	77		47	
15	114	114	80	114
25	98		56	98
35	72	72	45	
40	119		74	119
45	76	76	56	
55	98	98	78	98
60	114	114	114	114
65	76	76	76	
70	78		78	
75	88	88	88	
80	112	112	112	112
85	89		89	
90	97		97	97

Column wise deletion

We can perform column wise deletion in which we lose entire column if a column has missing values.

Both of these methods result in loss of information. we generally stick to imputation method but not deletion method.

vi. Outlier Treatment

What are Outliers?

We all have heard of the idiom ‘**odd one out**’ which means something unusual in comparison to the others in a group.

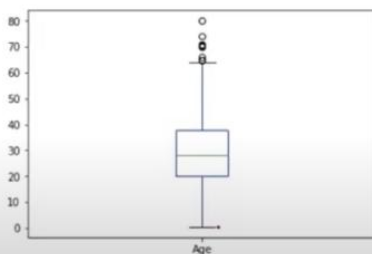
Similarly, an Outlier is an observation in a given dataset that lies far from the rest of the observations. That means an outlier is vastly larger or smaller than the remaining values in the set.

Reasons of Outliers

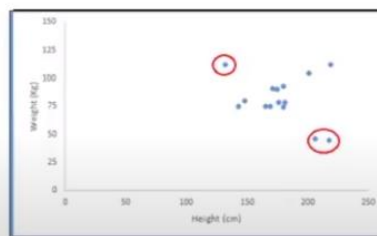
- Data Entry Error-An outlier may occur due to the variability in the data, or due to experimental error/human error.
- Measurement Error- experimental error or heavy skewness in the data(heavy-tailed distribution).Measuring some particular data in km instead of m.
- Processing Errors
- Change in the underlying population

Identifying Outliers

Graphical Methods



Boxplot



Scatter plot

Univariate Outliers

Bivariate Outliers

Treating Outliers-

- Deleting observations
- Transforming and Binning values
- Imputing outlier like missing values
- Treat them as separately

vii. Variable Transformation

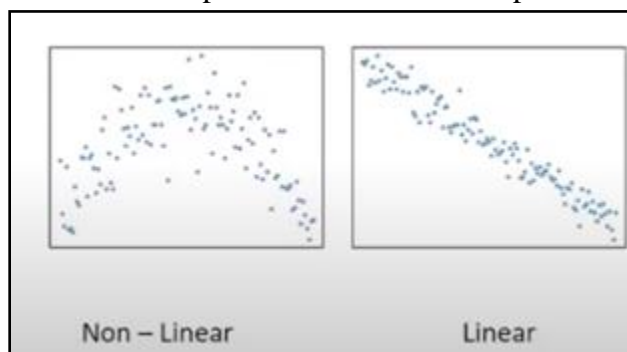
What is variable transformation

Variable transformation is the process by which –

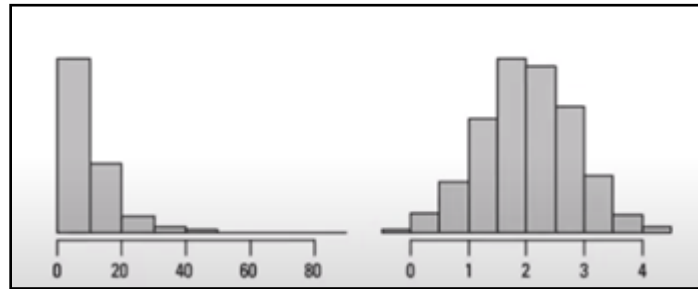
1. We replace a variable with some function of that variable. Example – replacing a variable x with its logarithm.
2. We change the distribution or relationship of a variable with others

Why do we use variable transformation

1. Change the scale of a variable
Ex: If 10 variables are measured in km and 1 in miles
2. Transforming non linear relationships into linear relationships

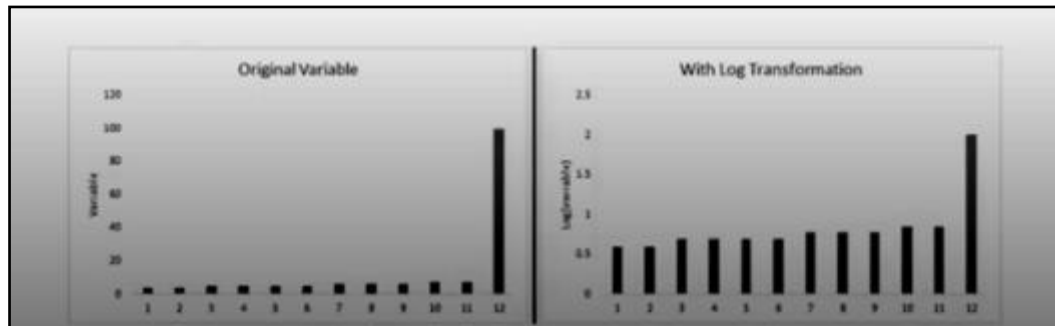


3. Create symmetric distributions from skewed distributions

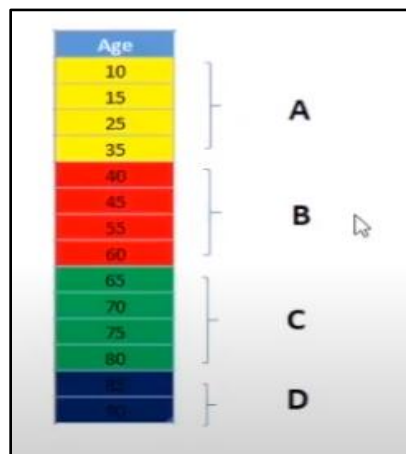


Methods of variable transformation-

- Logarithm-Taking log of the variable reduces right skewedness of the variable.



- Square root-Used for right skewed variable with positive values only.
- Cube root-Used for right skewed variable with positive or negative values.
- Binning-Used for converting continuous variables to categorical variables.



Splitting Dataset into training and test set



- The data should ideally be divided into 3 sets – namely, train, test, and holdout cross-validation or development (dev) set.
- **Train Set:**
The train set would contain the data which will be fed into the model. In simple terms, our model would learn from this data.
- **Validation set/Dev Set:**
The development set is used to validate the trained model. This is the most important setting as it will form the basis of our model evaluation. If the difference between error on the training set and error on the dev set is huge, it means the model has high variance and hence, a case of over-fitting.
- **Test Set:**
The test set contains the data on which we test the trained and validated model. It tells us how efficient our overall model is.
- The size of the train, dev, and test sets remains one of the vital topics of discussion. Though for general Machine Learning problems a train/dev/test set ratio of 80/20/20 is acceptable.

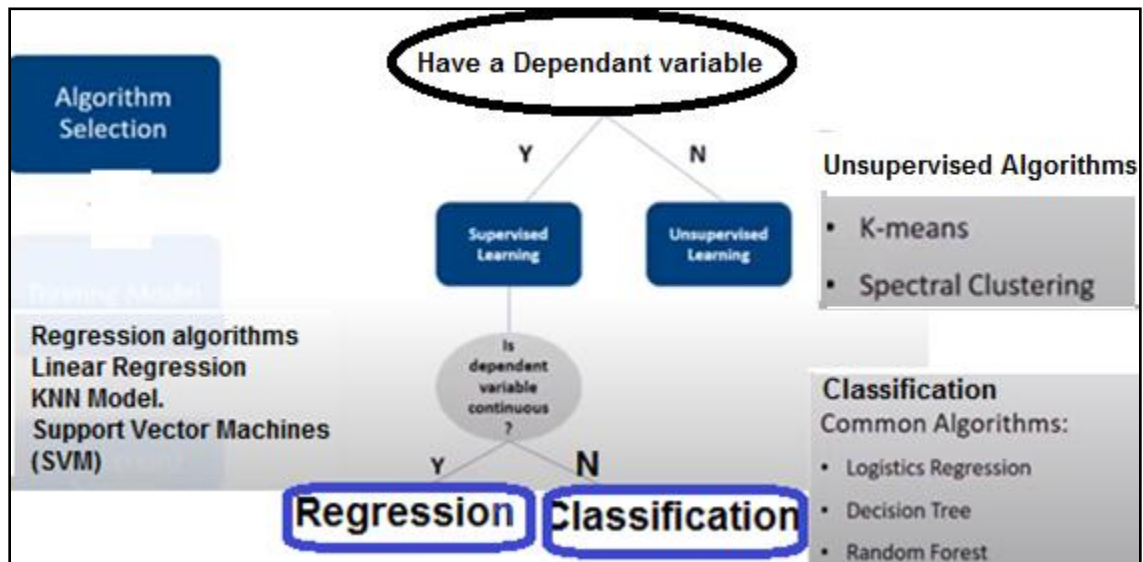
Train (80%)	Dev (20%)	Test (20%)
-------------	-----------	------------

5. Model Building/Predictive Modeling:

It is a process to create a mathematical model for estimating/predicting the future behavior based on the past data.

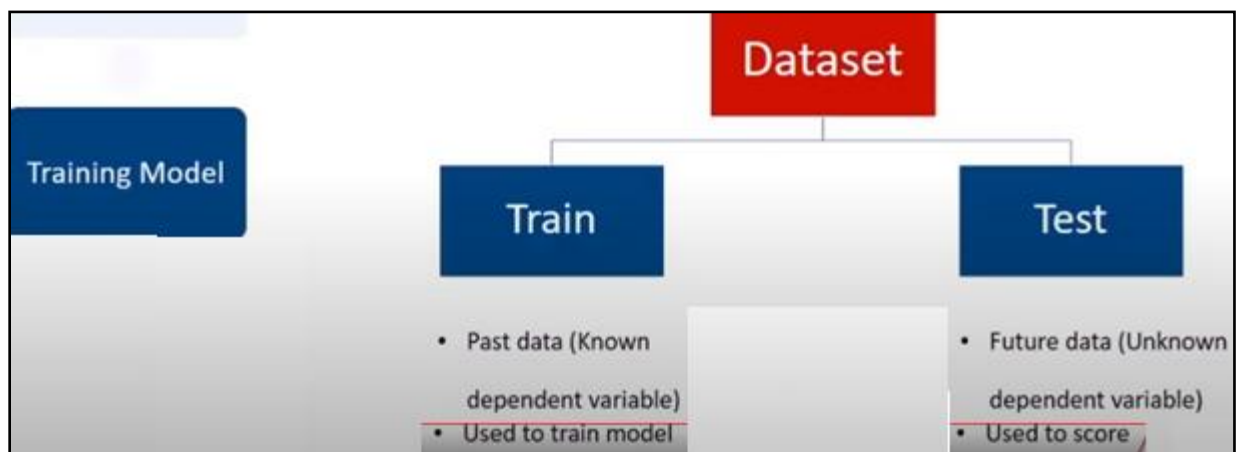


i. Algorithm Selection-



ii. Training Model-

It is a process to learn relationship/correlation between independent and dependent variables.



iii. Prediction/Scoring

It is a process to estimate/predict dependent variable of test data set by applying model rules.

6. Model Deployment:

- Machine learning model deployment is **the process of placing a finished machine learning model into a live environment where it can be used for its intended purpose.**
- Models are often integrated with apps through an **API(Application Programming Interface)** so they can be accessed by end users.

3.4 Evaluation Metrics:

i. Confusion matrix:

- A confusion matrix is an $N \times N$ matrix, where N is the number of classes being predicted. For the problem in hand, we have $N=2$, and hence we get a 2×2 matrix.
- It is a matrix of size 2×2 for binary classification with actual values on one axis and predicted on another.

		ACTUAL	
		Negative	Positive
PREDICTION	Negative	TRUE NEGATIVE	FALSE NEGATIVE
	Positive	FALSE POSITIVE	TRUE POSITIVE

Confusion Matrix

- The target variable has two values: Positive or Negative
- The columns represent the actual values of the target variable
- The rows represent the predicted values of the target variable

True Positive (TP)

- The predicted value matches the actual value
- The actual value was positive and the model predicted a positive value

True Negative (TN)

- The predicted value matches the actual value
- The actual value was negative and the model predicted a negative value

False Positive (FP) – Type 1 error

- The predicted value was falsely predicted
- The actual value was negative but the model predicted a positive value
- Also known as the Type 1 error

False Negative (FN) – Type 2 error

- The predicted value was falsely predicted
- The actual value was positive but the model predicted a negative value

ID	Actual Survived?	Predicted Survived?	Notation
ID1	1	0	FN
ID2	1	1	TP
ID3	1	0	FN
ID4	0	0	TN
ID5	1	1	TP
ID6	1	1	TP
ID7	0	1	FP
ID8	0	0	TN

		Prediction outcome	
		positive	negative
Actual value	positive	<i>TP</i> 3	<i>FN</i> 2
	negative	<i>FP</i> 1	<i>TN</i> 2

ii. **Accuracy:**

- Accuracy is of prime importance to select the best model. Accuracy is another metric for evaluating classification models.
- **Accuracy can be defined as the ratio of total number of correct predictions to the total number of predictions.**

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

iii. **Sensitivity/Recall:**

- Number of samples correctly identified as positive out of total true positives.
- Recall is the fraction of retrieved instances among all relevant instances. This means recall tells us the proportion of total number of positives which were identified correctly.
- Therefore, **recall is the ratio of true positives to the sum of true positives and false negatives.**

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

iv. **Precision:**

Out of all the positive predicted, what percentage is truly positive.

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

v. **Specificity:**

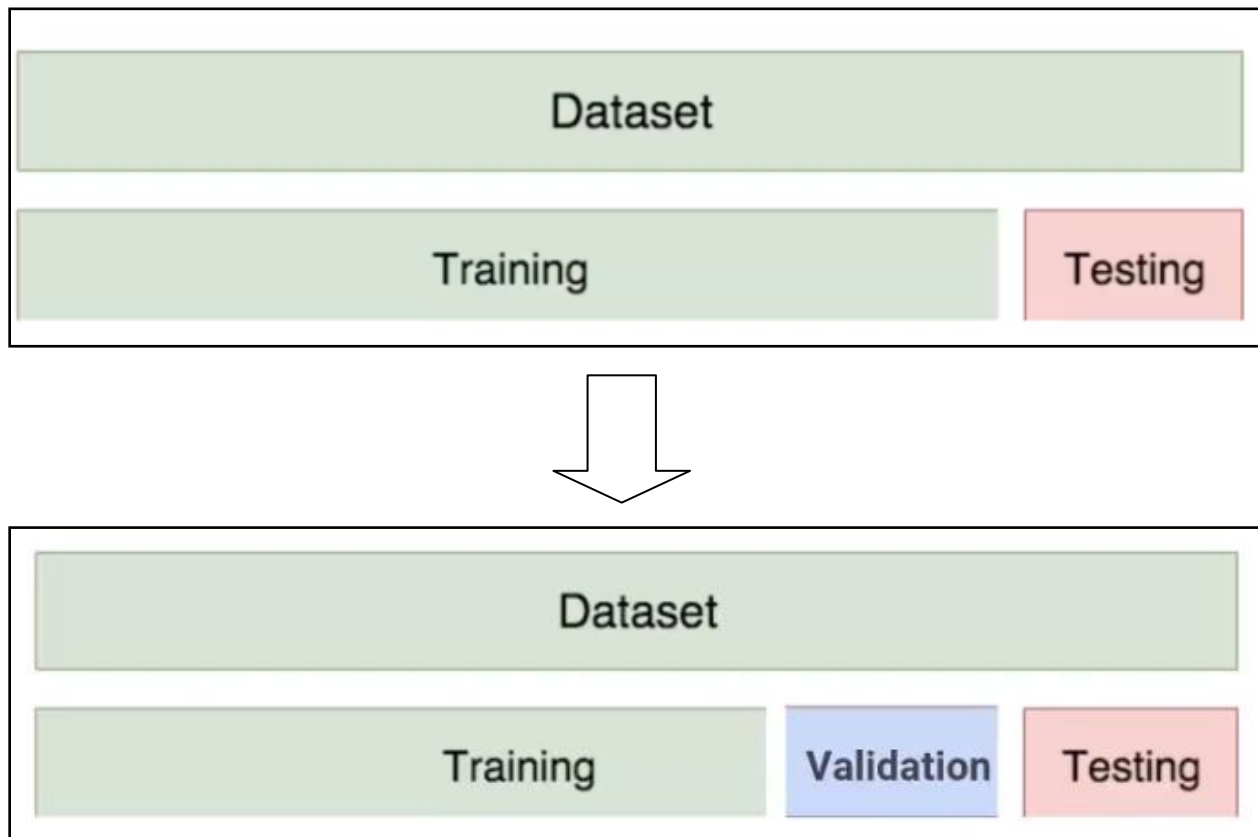
Percentage of true negative instances out of the overall actual negative instances present in the dataset.

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$$

3.5 **Validation:**

1. **Training Set:** Used to train the model.
2. **Validation Set:** Used to optimize model parameters.
3. **Test Set:** Used to get an unbiased estimate of the final model performance.

To optimize the model parameters, we use a partition of train data set as validation set.



Optimization is the process where we train the model iteratively that results in a maximum and minimum function evaluation. It is one of the most important phenomena in Machine Learning to get better results.


Different methods of validation techniques:

- Hold-out validation
- Stratified Hold-out validation
- Leave one out
- K-fold cross validation

K-fold cross validation

K Fold Cross Validation : Procedure

- Shuffle the dataset randomly.
- Split the dataset into k groups
 - pick a group as a hold out
 - Take the remaining groups as training and fit a model
 - Predict and evaluate on the hold out
- Repeat the above procedure with every group

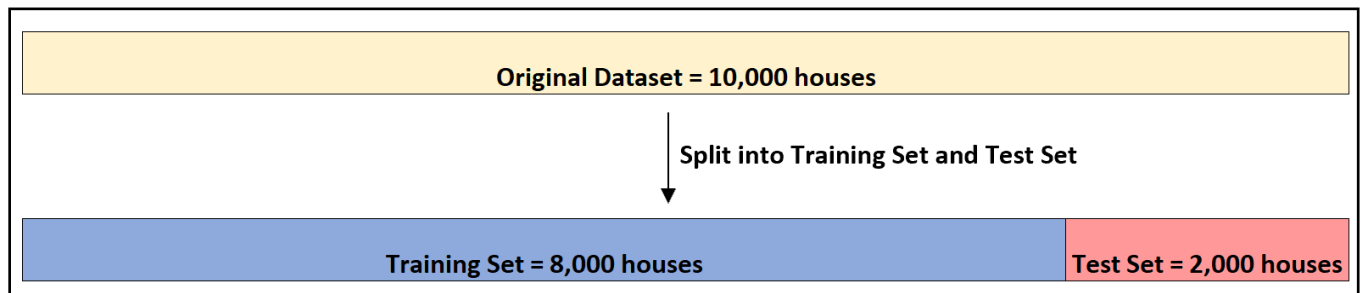


1. Randomly divide a dataset into k groups, or “folds”, of roughly equal size.
2. Choose one of the folds to be the holdout set. Fit the model on the remaining $k-1$ folds.
Calculate the test MSE (mean squared error) on the observations in the fold that was held out.
3. Repeat this process k times, using a different set each time as the holdout set.
4. Calculate the overall test MSE to be the average of the k test MSE's.

Example:

Suppose a real estate investor wants to use (1) number of bedrooms, (2) total square feet, and (3) number of bathrooms to predict the selling price of a given house.

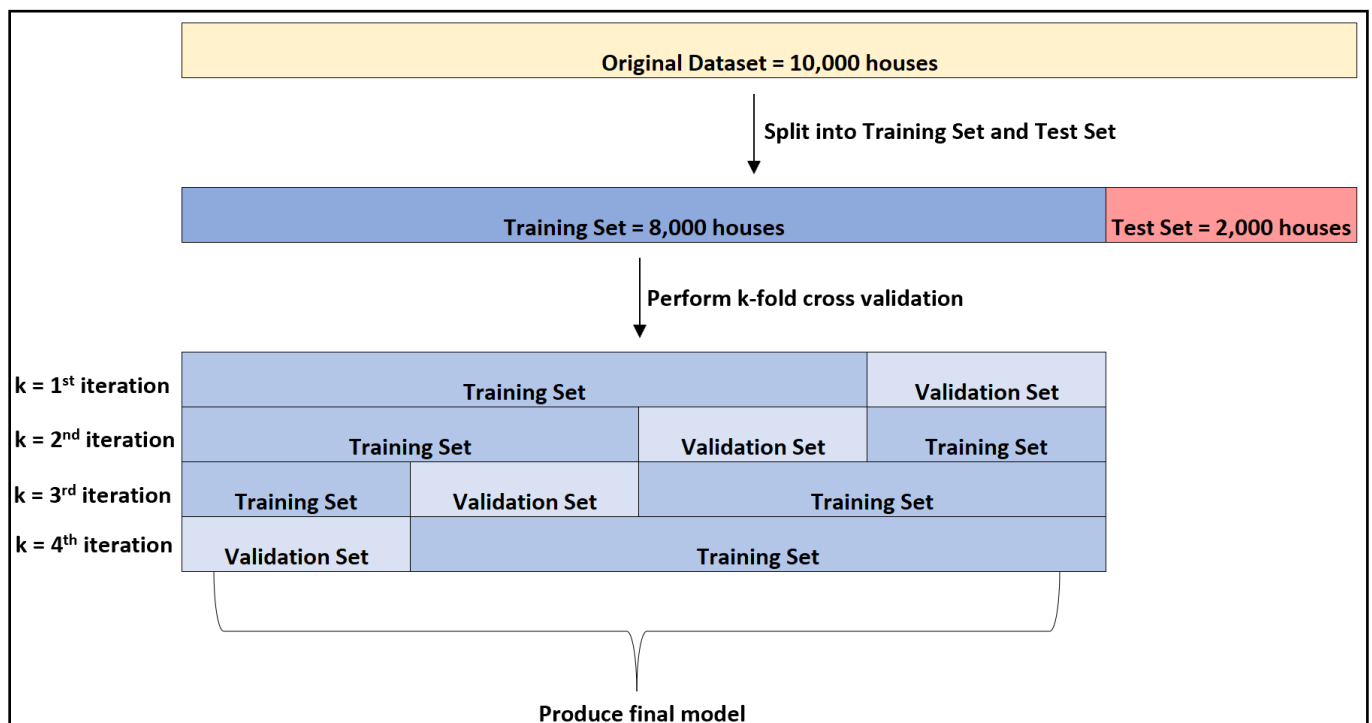
Suppose he has a dataset with this information on 10,000 houses. First, he'll split up the dataset into a training set of 8,000 houses and a test set of 2,000 houses:



Next, he'll fit a multiple linear regression model to the dataset four times. Each time he'll use 6,000 houses for the training set and 2,000 houses for the validation set.

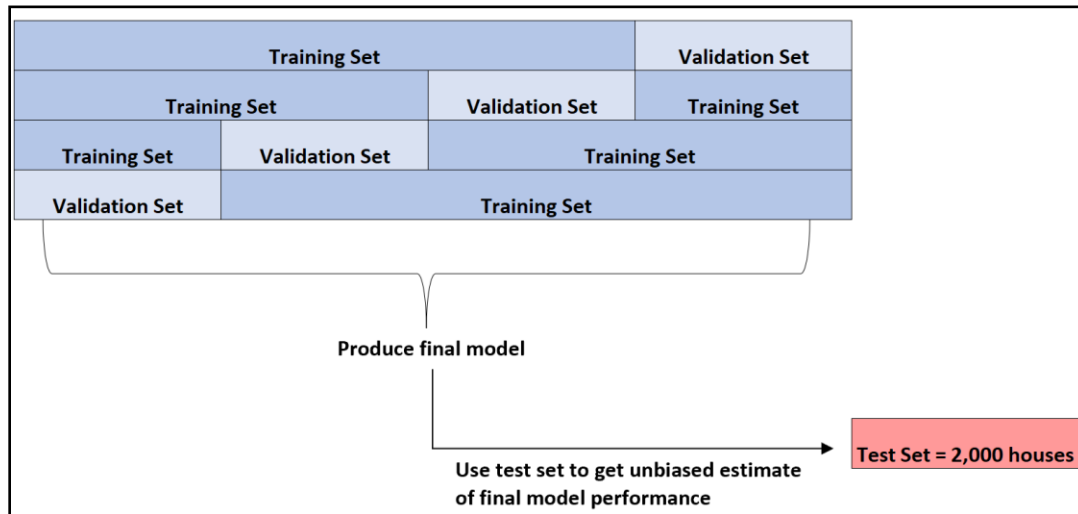
This is known as **k-fold cross validation**.

The training set is used to train the model and the validation set is used to assess the model performance. Each time he will use a different group of 2,000 houses for the validation set.



He may perform this k-fold cross validation on several different types of regression models to identify the model that has the lowest error (i.e. identify the model that fits the dataset best).

Only once he has identified the best model will he then use the test set of 2,000 houses that he held out at the beginning to get an unbiased estimate of the final model performance.



For example, he might identify a specific type of regression model that has a mean absolute error of **8,345**. That is, the mean absolute difference between the predicted house price and actual house price is \$8,345.

He may then fit this exact regression model to the test set of 2,000 houses that has not yet been used and find that the mean absolute error of the model is **8,847**.

Hyperparameter tuning:

HYPERPARAMETERS

- They are required for estimating the model parameters.
- They are estimated by hyperparameter tuning.
- The choice of hyperparameters decide how efficient the training is.
- Hyperparameter tuning (or hyperparameter optimization) is the process of determining the right combination of hyperparameters that maximizes the model performance.

Hyperparameter tuning methods

- Random Search
- Grid Search
- Bayesian Optimization
- Tree-structured Parzen estimators (TPE)

Significance of Hyperparameter tuning

- The outcome of hyperparameter tuning is the best hyperparameter setting. After evaluating a number of hyperparameter settings, the hyperparameter tuner outputs the setting that yields the best performing model.
- It can give you optimized values for hyperparameters, which maximizes your model's predictive accuracy.

- Hyperparameters are the knobs or settings that can be tuned before running a training job to control the behavior of an ML algorithm.
- They can have a big impact on model training as it relates to training time, infrastructure resource requirements (and as a result cost), model convergence and model accuracy.

2 Marks Questions

1. Define Machine learning.
2. List the steps of data exploration.
3. Define accuracy, precision w.r.t Machine learning model.
4. List the stages of predictive modeling.
5. Define Specificity, Sensitivity w.r.t Machine learning model.
6. List the methods of variable transformation.
7. List any 4 applications of ML.
8. List functions of Pandas.
9. Classify ML.
10. Classify supervised learning.
11. List functions of Predictive modeling.
12. Differentiate categorical and continuous variables.
13. List types of graphical methods of Univariate analysis of continuous variables.
14. Draw a labeled boxplot .
15. Define Univariate and Bivariate analysis.
16. List types of Bivariate analysis.
17. List reasons for missing value in a dataset.
18. List types of missing values.
19. List methods to identify missing values.
20. List reasons of outliers.
21. List methods to treat outliers.
22. List methods of variable transformation.
23. List the steps of predictive modeling.
24. Define model deployment.
25. List any 4 evaluation metrics.
26. Draw a labeled 2x2 confusion matrix.
27. Define TP, TN w.r.t confusion matrix.
28. Define FP, FN w.r.t confusion matrix.

29. List different methods of validation techniques.
30. What are hyperparameters.
31. List different hyperparameter tuning methods.

4 Marks Questions

1. Explain the confusion matrix with one example.
2. Explain Bivariate analysis.
3. Classify Machine learning and explain each type.
4. Explain the Steps to read csv and excel file in Jupyter notebook inside
5. pandas.
6. Explain the process of model building
7. Explain Univariate analysis.
8. Explain missing value treatment.
9. Explain outlier treatment.
10. Explain variable transformation.
11. Explain significance of hyperparameter tuning.

6 Marks Questions

1. Explain different stages of predictive modeling.
2. Explain different stages of Data exploration.
3. Explain evaluation metrics w.r.t. ML model.
4. Explain k-fold validation technique with example.
5. Explain how to evaluate performance of a model using evaluation metrics.