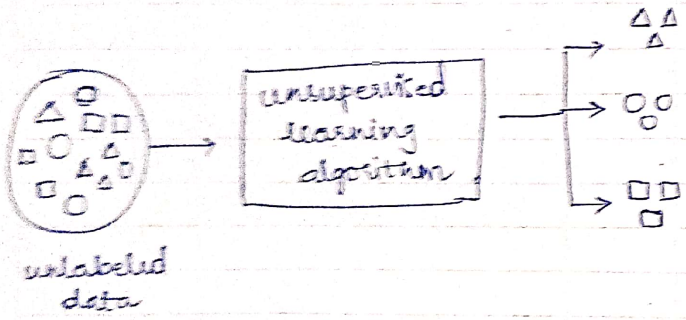
Applications of Regression:

- Demand forecasting in retails.
- Sales prediction for manage.
- Price prediction in real estates.
- Weather forecast
- Skill demand forecast on Job market.



Unsupervised learning:-

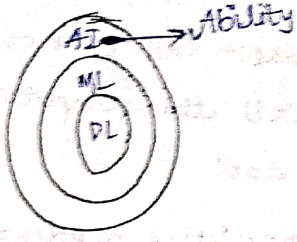
→ Unsupervised learning is a type of machine learning where algorithm learn patterns and insights from unlabeled data without explicit guidance.

→ Unsupervised learning aims to discover hidden structures and relationships within the data itself.

→ The process of unsupervised learning is referred as, pattern discovery (or) knowledge discovery.

Essentials of Machine Learning

Introduction:-



Applications of Machine learning:-

- Social Media.
- Face Recognition
- Speech.
- navigation
- Spam filter.
- Suggestions of Products.
- Fraud Detection
- Stock & weather forecast.
- Healthcare
- chatbots
- Autopilot cars.

** What is Machine Learning?

A computer program is said to learn from experience E with respect to some Task T. and performance measure P; If its performance at Task T as measured by P, improves with experience E.

Tom M. mitchell
Professor of machine learning.

Task(T):- This is the goal or activity the computer is trying to perform.

Experience :- This refers to the data or past activities the computer is trained on.

Performance Measure:- This is how we evaluate how well the computer is performing the task

ex:- education: predicting student perform

T(Task):- Predict whether a student will pass or fail a course.

E(Experience):- Historical student data - attendence, previous grades, time spent or learning performs.

P(Performance Measure):- Prediction accuracy or F_1 – score.

* Why do we learn Machine learning?

→ Solves complex problems that traditional programming can't.

→ Handles large volumes of Data quickly

→ Automates Repetitive Tasks with accuracy

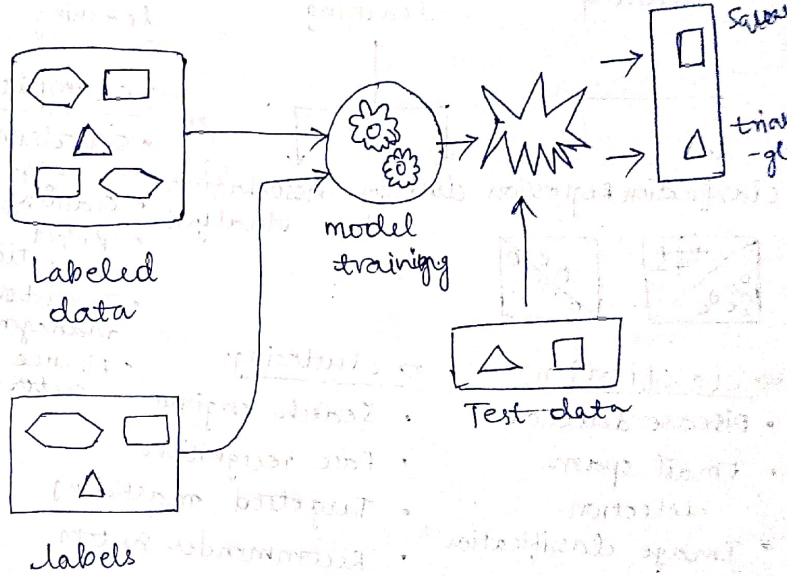
→ Personalized User experience.

* Types of Machine learning:-

1. Supervised learning:-

→ Also called predictive learning.

→ A machine learning predicts the class of unknown objects based on prior class related information of similar objects.



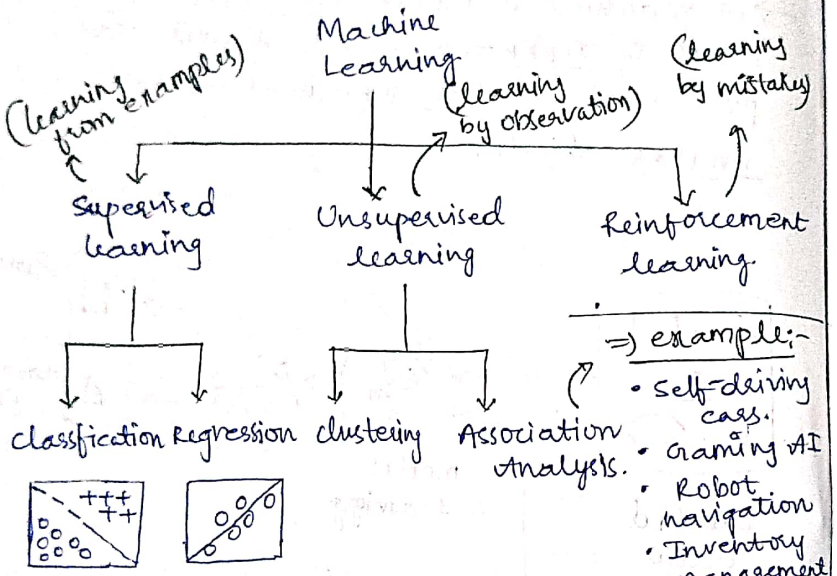
2. Unsupervised learning:-

→ also called descriptive learning.

→ A machine finds patterns of unknown objects by grouping similar objects together.

3. Reinforcement learning:-

→ A machine learns to act on its own to achieve the given goals.



⇒ example:-
- Self-driving cars.
- Gaming AI
- Robot navigation
- Inventory management
- Finance sector.

⇒ classification:-
- Disease detection
- Email spam detection.
- Image classification
- Bank loan prediction

⇒ clustering:-
- Search engines
- Face recognition
- Targetted marketing
- Recommender system.

⇒ Regression:-
- Hour price prediction
- Stock price prediction.

⇒ Association Rule mining:-
- Market Basket analysis
- Medical diagonsis
- census data.

* Supervised learning:-

→ Supervised learning is a machine learning approach where an algorithm learns from a labeled dataset to make predictions or classifications.

→ Essentially, it's like having a teacher guiding the algorithm to learn the relationship b/w inputs & outputs.

→ The algorithm uses this labeled data to build a model that can then predict outcomes for new unseen data.

→ labeled data is the type of data which contains both the features(input) and the target(output) is known as labeled data.

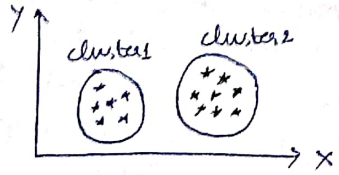
Some examples of supervised learning:-
- predicting of the result of a game.
- Predicting whether a tumour is malignant or benign.
- Predicting the price of domains like real estate, stocks, etc.
- classifying the test such as classifying a set of emails or spam or non-spam.

Applications of classification:-
- Image classification.
- Prediction of disease.
- win-loss prediction of games.
- Prediction of natural calamity like earthquake, flood, etc.
- Recognition of handwriting.

next Applications of regression
1st page:

7/7

1. clustering:- is an unsupervised learning technique where data points are grouped into clusters based on their similarity.



2. Association analysis:-

→ It identifies the association between data elements.

ex:- Market Basket Analysis:-

TID	Items Bought
1	(Butter, Bread)
2	(Diaper, Bread, milk, Beer)
3	(Milk, chicken, Beer, Diaper)
4	(Bread, Diaper, chicken, Beer)
5.	(Diaper, Beer, cookies, icecream)

Market Basket transactions.

Frequent itemset → (Diaper, Beer)

Possible associations: Diaper → Beer.

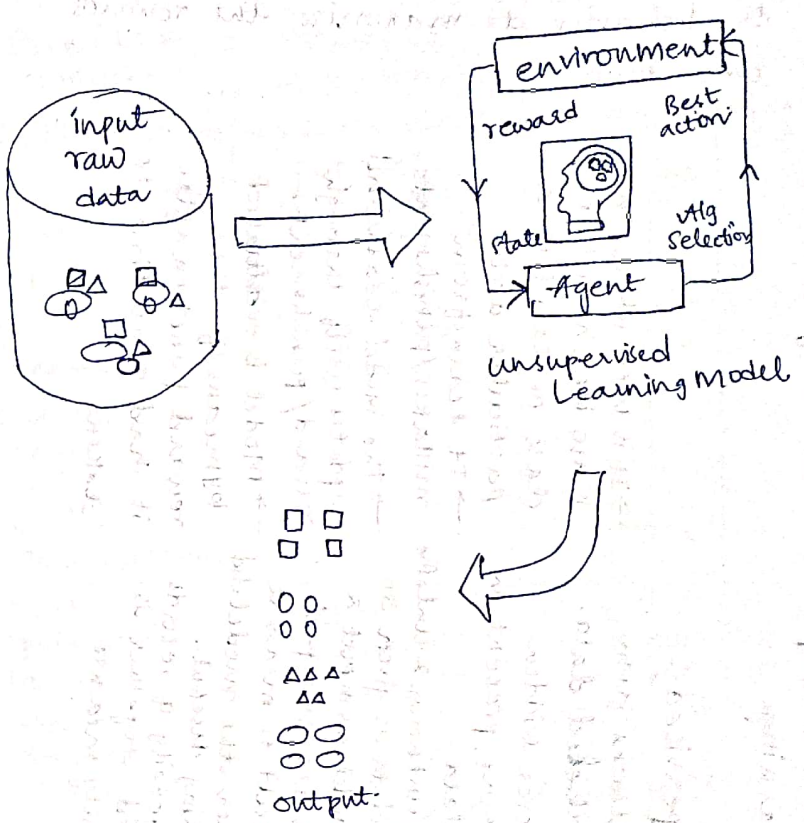
* Applications of Unsupervised learning:-

1. customer segmentation.
2. Anomaly Detection
3. Recommendation systems.
4. Social Network Analysis.
5. Text categorization.
6. Language Translation.

Differences between supervised and unsupervised learning:-

Aspect	Supervised	Unsupervised
Definition	learning from labeled data	learning from unlabeled data.
Goal	Predicts outputs from inputs.	Discover hidden patterns or structures
Data	Requires input-output pairs.	only input data is provided.
Examples	classification, Regression.	clustering, Dimensionality Reduction.
Algorithms	Linear Regression, SVM,	

3. Reinforcement Learning:-



Reinforcement learning (RL):-

→ Reinforcement learning (RL) is a machine learning paradigm where an agent learns to make decisions by interacting with an environment to maximize a cumulative reward.

→ It's a trial-and-error process where the agent learns through feedback, similar to how humans learn by doing.

comparison - supervised, Unsupervised & reinforcement learning

Supervised	Unsupervised	
→ It learns from labelled data. → Predicts output from inputs. → Model is built based on training data. → The model performance can evaluated based on how many mis-classification have been done based on a comparision b/w predicted & actual.	→ It learns from unlabeled data. → Discover hidden patterns present in the data. → Any unknown & unlabe-led data is given to the model as input & records are grouped. → Difficult to measure whether the model did something useful. Homogeneity of records grouped together is the only measure.	→It is used when there is no idea about the class or label of a particular data. → It learns from mistakes / punishments. → The model learns and updates itself through reward / punishment. → Model is evaluated by means of the reward function after it had some time to taken.

→ The agent receives rewards or penalties for its actions, and it adjust its behavior to maximize the rewards over time.

Supervised learning Algorithm	Unsupervised learning Algorithm	Reinforcement learning.
• Standard Algorithms:- 1. Naive Bayes 2. K-nearest neighbour(KNN) 3. Decisiontree 4. Linear Regression. 5. logistic regression 6. Support vector machine(SVM)	• standard Algorithms:- 1. K-means 2. principal component Analysis(PCA) 3. Apriori algorithm. 4. DBSCAN etc	• standard Algorithms:- 1. Q-learning 2. Sarsa.
• Practical Applications include;- 1. Hand writing recognition. 2. Stock market prediction. 3. Disease prediction 4. Fraud detection	• Practical Applications include;- 1. Market Basket Analysis. 2. Recommendation Systems 3, customer Segmentation etc	• Practical Applications include;- 1. Self-driving Cars. 2. Intelligent robots 3. AlphaGo Zero (the latest version of Deepminds AI System playing GO).
Simple to understand.	more difficult to understand & implement than supervised learning.	Most complex to understand and apply.

* State-of-the-art languages and Tools in Machine Learning:-

⇒ Languages:-

- Python:- Most widely used due to its simplicity and vast ML libraries.
- R:- Preferred in statistics-heavy applications.
- MATLAB:- widely used for numerical computing, data analysis, and algorithm development
- SAS:- (Statistical Analysis System) Offers several robust tools for building and deploying models.
- Julia:- Known for high-performance numerical computing.

⇒ Tools & Frameworks:-

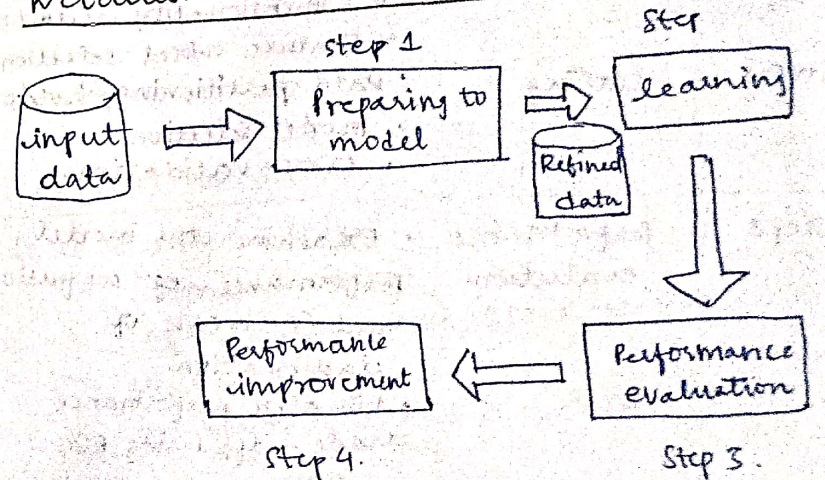
- TensorFlow:- Open-source library developed by Google for deep learning.
- PyTorch:- preferred for research and development due to its flexibility
- Scikit-learn:- A Python library for classical machine learning algorithms.
- Keras:- High-level neural networks API, runs on top of TensorFlow.
- XGBoost:- Libraries for efficient gradient boosting.
- Jupyter Notebook:- Interactive development environment widely used for ML experiments.

* ML in Healthcare & Pharma:-

→ proactive health monitoring and alerts.
→ Disease Identification / Diagnosis.
→ Personalized treatment.
→ Disease / epidemic outbreak prediction.
→ clinical trial Research

9/7

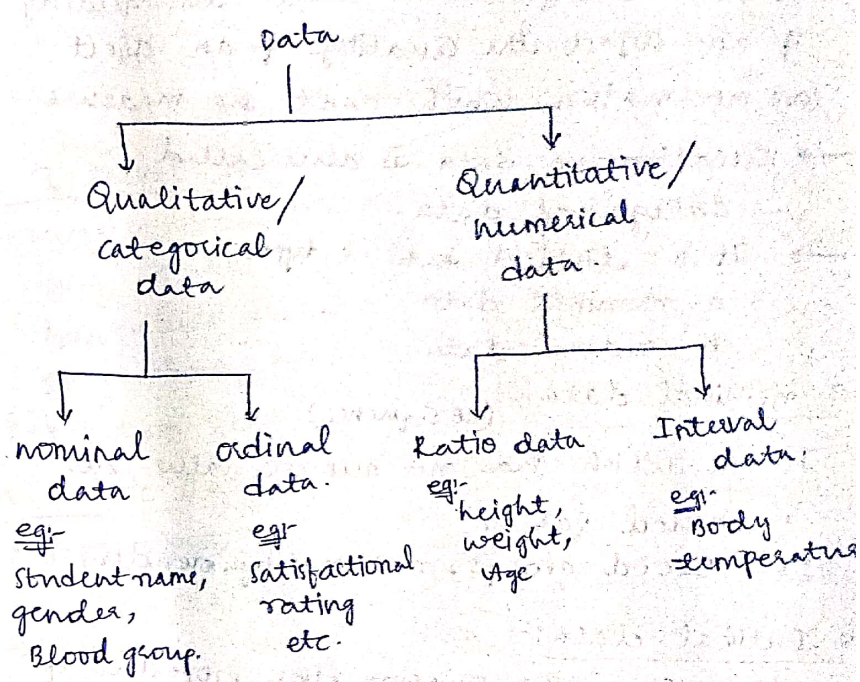
Detailed Machine learning process:-



Machine learning Operations:

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 eg:- inter-feature relationship • Find potential issues in data : Remediate data, if needed. : Apply following pre-processi -ng steps, as necessary:- ✓ Dimensionality reduction. ✓ Feature subset selection.
Step 2.	Hosting	• Data partitioning / holdout. • Model selection. • cross-validation.
Step 3	Performance evalution.	• Examine the model performance. eg:- confusion matrix in case of classification. • Visualize performance trade-offs using ROC curves.
Step 4	Performance improvement.	• Tuning the model. • Ensembling • Bagging • Boosting.

Types of data:-



1. Qualitative data:-

It is divided into:-

a. nominal data.
   eg:- student name, blood group.
b. ordinal data.
   eg:- grade, satisfactional level.

2. Quantitative data:-

It is ◎ interval data.

a. Ratio data.
   eg:- height, age.
b. Interval data.
   eg:- Body temperature.

Qualitative data:-

→ It provides information about the quality of an object the quality of an object (or) information which can't be measure.

→ Qualitative data is also called categorical data.

→ It is divided into 2 types:-
  a. nominal data
  b. Ordinal data.

a. nominal data:- (no sequence)

Is one which has no numeric value but a named value.

ex:- Blood groups, nationality, gender.

b. Ordinal data:-

→ In addition to possesing the properties of nominal data can also be naturally ordered.

→ Ordinal data also assigns named values to attributes and they can be arranged in a sequence of increasing or decreasing value.

ex:- 1. customer satisfaction - very happy, happy, unhappy.
  2. Grades - A, B, C, etc.
  3. hardness of metal - very hard, hard, soft etc.

Quantitative data:-

Two types of Quantitive data:-
  1. Ratio data
  2. Interval data.

b. Interval data:-

→ numeric data for which not only the order is known, but the exact difference b/w values is also.

ex:- 1. Body temperature - celsius temp.

→ no true zero.

a. Ratio data:-

→ represents a numeric data for which value can be measured.

→ absolute zero is available for ratio data.

→ can be added, subtracted, multiplied or divided.

ex:- 1. height, weight, age, salary etc.

* Data Exploration:-

→ Understand the central tendency-
- Mean
- Median
- Mode.

→ Understand data spread
- Standard deviation, variance

→ Understand data value position.
- Box plot.
- Histogram

→ To Understand the nature of numeric variables we can apply measure of central tendency of data.

Mean Vs Median for Auto MPG.

	mpg	cylinders	displace-ment	horse-power	weight	acceration
Median	23	4	148.5	?	2604	15.5
Mean	23.51	5.455	193.4	?	2970	15.57
Deviation	2.17%	26.67%	23.22%		5.59%	0.45%
	Low	high	High		low	low.

model year, orgin

.
.
.

Understanding data spread:-

→ To understand data spread:-
1. Dispersion of data that is spread in the data.
2. position of the different data values.
3. To measure the how much different values of data are spread out variance of the data is used.

→ consider the data values of two attributes
- Attribute 1 values - 44, 46, 48, 45, 47

$$\text{mean} = \frac{44+46+45+45+47}{5} = 46$$

44, 45, 46, 47, 48

median = 46

- Attribute 2 values - 34, 46, 59, 39, 52

$$\text{mean} = \frac{34+46+59+39+52}{5} = 46.$$

median = 34, 39, 46, 52, 59 → 46.

→ Both the set of values have a mean and median of 46.

→ First set of values is more connected (or) clustered around the mean/median value.

$$\text{variance}(x) = \sum_{i=1}^{n} \frac{x_i}{n} - \left(\sum_{i=0}^{n} \frac{x_i}{n}\right)^2$$

$$\text{standard deviation}(x) = \sqrt{\text{variance}}$$

$x \rightarrow$ variable

$n \rightarrow$ no. of observations

→ calculate variance for attribute 1

44, 46, 48, 45, 47

$$\text{Variance} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

$$= \frac{44^2+46^2+48^2+45^2+47^2}{5} - \left(\frac{44+46+48+45+47}{5}\right)^2$$

$$= \frac{1936+2116+2304+2025+2209}{5} - \left(\frac{230}{5}\right)^2$$

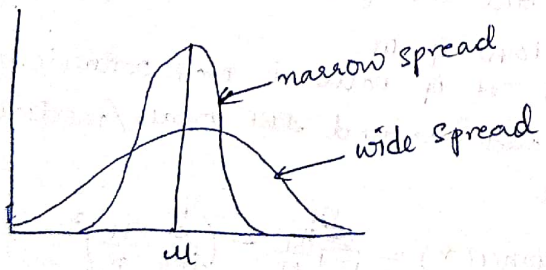
$$= \frac{10590}{5} - (46)^2 = 2.$$

For attribute 2,

$$= \frac{34^2+46^2+59^2+39^2+52^2}{5} - \left(\frac{34+46+59+39+52}{5}\right)^2$$

$$= 79.6.$$

large value of variance (or) standard deviation indications more dispersion in the data and viceversa.



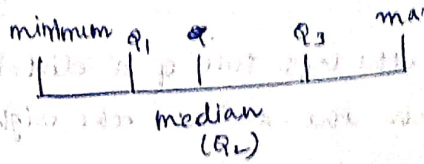
Same mean and different dispersion.

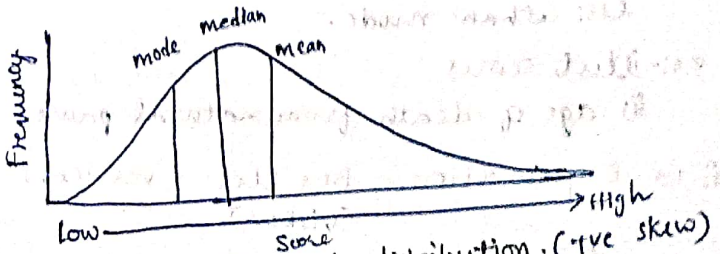
⇒ If variance is less → there is narrow spread.
variance is more → wide spread.

Data value position :→ (Boxplot is also called as Box & Whisker plot)

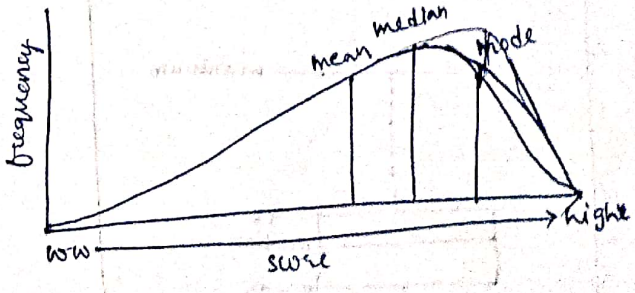
Any data set attribute has five values

- Minimum
- First quartile (Q_1)
- Median (Q_2)
- Third Quartile (Q_3)
- Maximum





(a) Right-skewed distribution. (+ve skew)



(b) Left-skewed distribution.

Tail on the right side is ... the tail on the left side

Mean is greater than median. median is greater than mode.

ex:- income distribution.

Left Skew:-

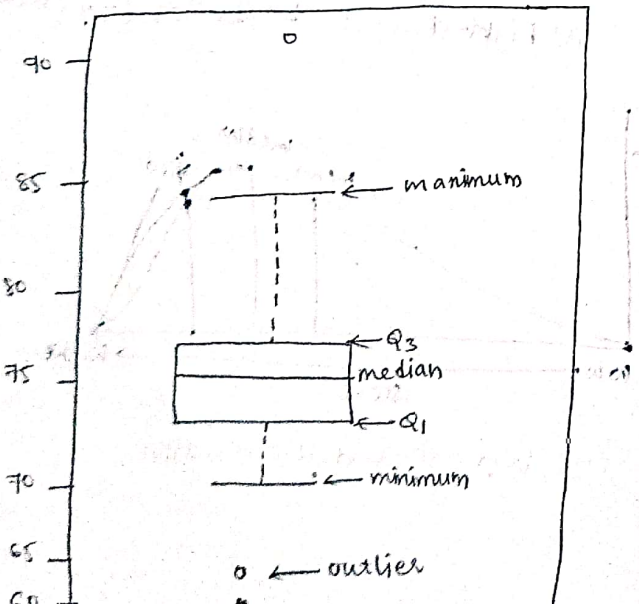
where tail on the left side of a distribution is longer than the tail on the right.

mean is less than median - median is less than mode.

ex:- i) test scores.

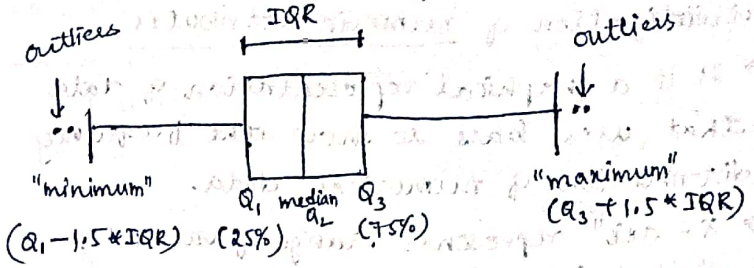
ii) age of death from natural pause.

Data Exploration – Box plot. (vertical) (viskers)



Horizontal plot:-

(Interquartile Range)

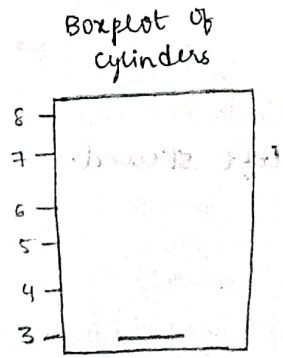


$IQR = Q_3 - Q_1$

$min = Q_1 - 1.5 * IQR$

$max = Q_3 + 1.5 * IQR$

Box plot of Auto-MPG Attributes:-



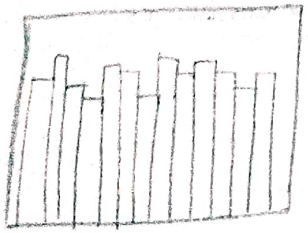
Histogram :-

→ Histogram which helps in effective visualization of numeric attributes.

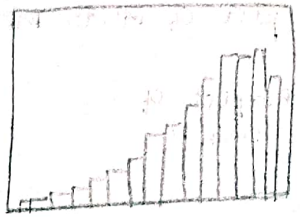
→ It is a graphical representation of data that uses bars to show the frequency distribution of numerical data.

→ "x-axis" represents "range of values", divided into intervals (bins).

→ "y-axis" represents "frequency count of data points" within each bin.



Symmetric, Uniform.



Left skewed.

Right skewed.

Exploring Categorical data:

→ Statistical data:-

⇒ mode is applicable on categorical attributes, and can also apply to numeric data.

⇒ like mean and median, mode is also a statistical measure of central tendency of a data.

⇒ mode of a data is the data value which appears more frequently in a dataset.

⇒ mean, median cannot be applied for categorical variables.

eg:- names:

Rama
Sita
Rama
Latha
Rama

```
df['Names'].mode()[0]
```

o/p:- Rama.

→ An attribute may have one or more modes.

→ Single mode is called unimodal, two modes are called Bimodal and multiple modes are called multimode.

Exploring relationship between variables:

→ There are multiple plots to explore the relationship between variables

→ The basic and most commonly use plot is scatter plot

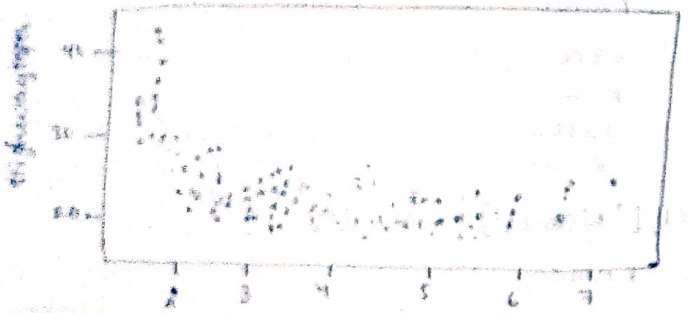
→ A scatter plot helps in visualizing bivariate relationships, i.e relationship between two variables.

→ It is a two-dimensional plot in which points or dots are drawn on coordinates provided by values of the attributes.

→ example, in a dataset there are two attributes - att_1 & att_2.

Scatter plot:

Engine Displacement Vs Highway MPG



engine displacement (Litres)

Pair Plot:

* Preprocessing steps:-

1. Dimensionality reduction:-

→ High-dimensional data sets need a high amount of computational space and time.

→ not all features are useful - they degrade the performance of machine learning algorithms.

→ Most of the machine learning algorithms perform better if the dimensionality of dataset, i.e the number of features in the data set, is reduced.

→ Dimensionality reduction helps in reducing irrelevance and redundancy in features.

→ Also, it is easier to understand a model if the number of features involved in the learning activity is less.

Other Pre-processing steps:-

→ Dimensionality reduction
- Principal component analysis (PCA)
- Singular value Decomposition (SVD)
- Linear Discriminant Analysis (LDA).

→ Feature subset selection achieved by
- Removing irrelevant features.
- Selecting a subset of potentially redundant features.

Selecting a Model:-

* → Input variables can be denoted by x, while individual input variables are represented as $x_1, x_2, x_3 \ldots x_n$ and

* → Output variable by symbol Y.

→ The relationship b/w x and Y is represented in the general form:

$$Y = f(x) + e$$

'f' → target function.
'e' → random error term.

Cost Function:-

→ A cost function (also called error function) helps to measure the extent to which the model is going wrong in estimating the relationship between x and Y.

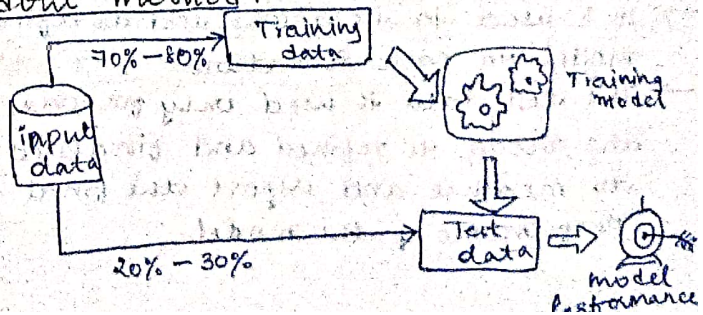
→ Cost function can tell how bad the model is performing.

Loss function:-

→ Loss function is almost synonymous to cost function - loss function is usually a function defined on a data point cost function is for the entire training data set.

Training a model (for supervised learning)

- Holdout method:-



Holdout method :-

→ A model is trained using the labelled input data.

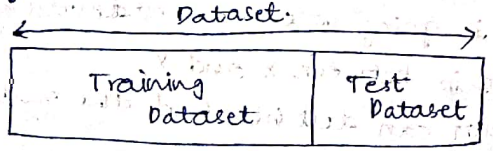
→ A part of the input data is held back (that is how the name holdout orginates) for evaluation of the model.

→ This subset of the input data is used as the test data for evaluating the performance of a trained model.

~~→ The remaining 2~~

→ In general 70% - 80% of the input data is used for model training.

→ The remaining 20% - 30% is used as test data for validation of the performance of the model.



→ In certain cases, the input data is partitioned into three portions :-
- a training data.
- a test data.
- validation data.

→ The validation data is used in place of test data, for measuring the model performance.

→ It is used in iterations and to refine the model in each iteration.

→ The test data is used only for, once, after the model is refined and finalized, to measure and report the final performance of the model.

K-fold Cross-Validation method

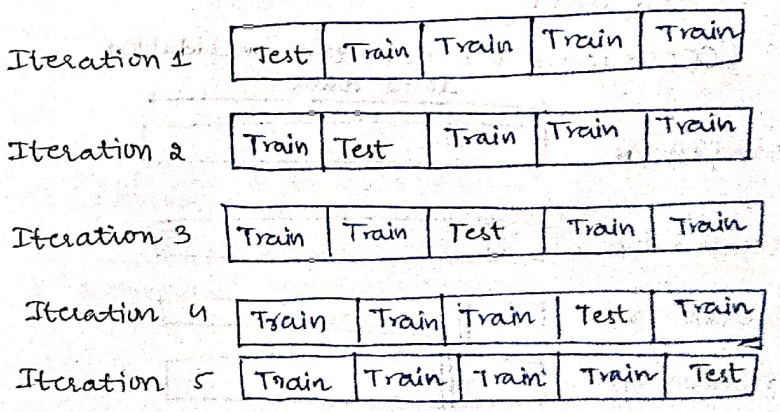
→ K-fold cross-validation is a technique used to evaluate the performance of a machine learning model by dividing the dataset into K subsets (or folds)

K-Fold cross validation :-

1. Divides the dataset into K equal-sized folds.
2. For each of the K iterations :
   - Use K-1 folds to train the model.
   - Use the remaining 1 fold to test the model.
3. Average the results across all K trails to get a reliable performance estimate.

10 or 5 fold cross validation are commonly used.

K-Fold cross validation



Benefits:-

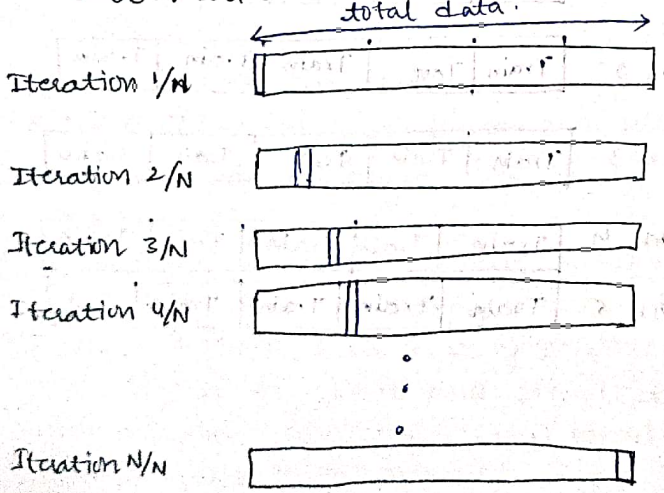
→ More reliable than the holdout method.
→ Every sample is used for both training and testing.
→ Reduced bias and variance
→ Prevents over fitting.

LOOCV (Leave-One-Out Cross-Validation):-

→ It is an extreme case of K-fold cross-validation, where
K → no. of data points in the dataset
→ For a dataset with n samples, LOOCV:
- Uses n-1 samples for training
- Uses the 1 remaining sample for testing.

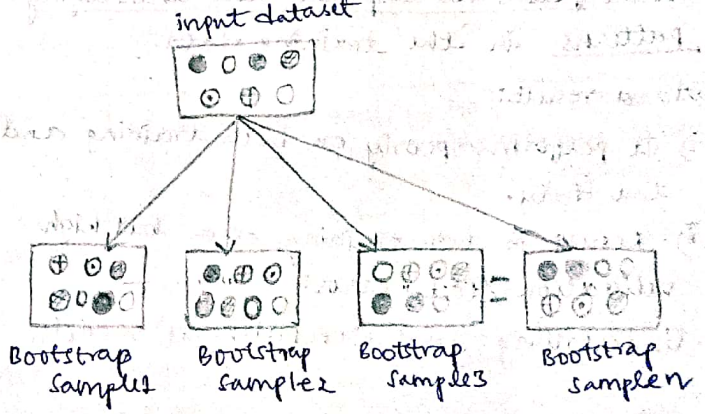
→ This process is repeated n times, each time leaving out a different sample as the test data.

LOOCV: Leave one out cross validation.



Bootstrap sampling:-

→ It uses technique of simple random sampling with replacement (SRSWR).
→ A bootstrap sample is a random sample taken from a dataset with replacement, is usually the same size as the original dataset.
→ Bootstrap sampling involves creating multiple new datasets (called bootstrap samples) by randomly selecting data points from the original dataset with replacement. This means a data point can appear multiple times in a sample, or not at all.
→ This technique is particularly useful in case of input data sets of small size, i.e having very less number of data instances.



Overfitting and Underfitting:-

In machine learning (ML), overfitting and underfitting are two common issues that affect the performance of models.

1. Overfitting:-

Def:- The model learns the training data too well, including its noise and outliers.

As a result:-

i) It performs very well on training data.

ii) Poor performance on unseen/test data.

2. Underfitting:-

Def:- The model is too simple to capture the underlying structure of the data.

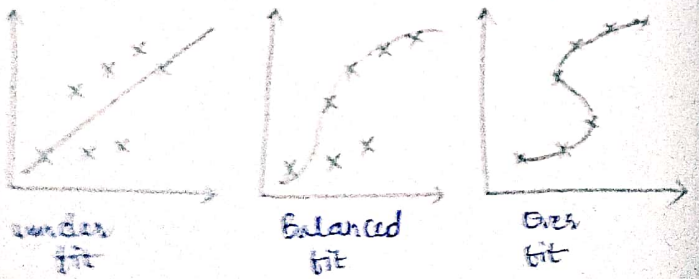
Under fitting occurs when a machine learning model fails to capture the underlying patterns in the training data.

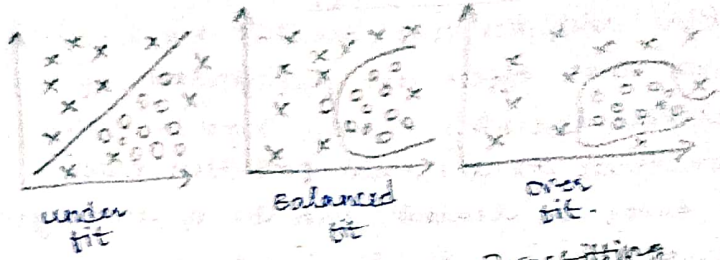
As a result:-

i) It performs poorly on both training and test data.

ii) Results in low training error but high validation/test error.

Underfitting and Overfitting of models.





Key differences between Overfitting and Underfitting.

Aspect	Overfitting	Underfitting
Definition	Models learns the training data too well, including noise.	model fails to learn the underlying pattern in data.
Model complexity	Too complex	Too simple.
Training error	very low	high
Test error	High (poor generalization)	High.
cause	Too many parameters, not enough data, overtraining.	Too few parameters, insufficient training, oversimplification.
Performance	Excellent on training data but poor on unseen data.	Poor on both training and test data.
Symptoms	High variance	High bias.
Solution	Simplify the model, use regularization get more data.	Increase model complexity, improve features, train longer.
Bias-variance	low bias, high variance.	High bias, low variance.

Bias-variance trade off:-

→ Bias and variance are two key sources of error that effect the performance of machine learning models. Together, they contribute to the total prediction error.

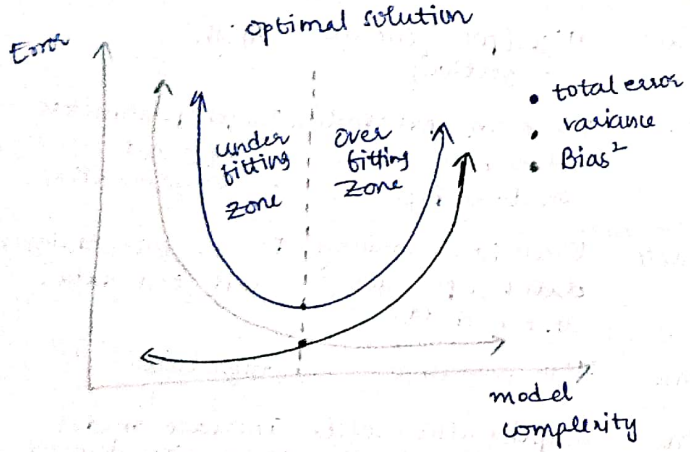
→ error in learning can be of two types.

1. errors due to 'bias' and
2. error due to 'variance'.

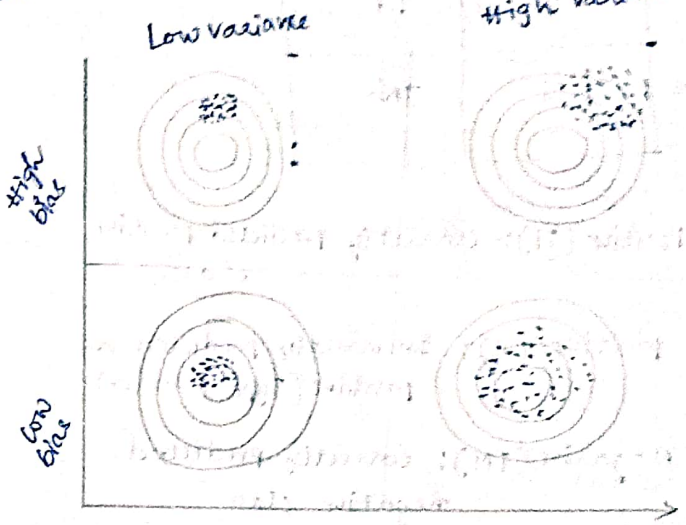
→ Errors due to 'bias'; it is due to underfitting of the model.

→ Errors due to variance: occurs when a model is too sensitive to the training data- it learns the noise or random fluctuations in the training set rather than the actual pattern.

→ Total Error = Bias^2 + variance + Irreducible error.



As can be observed in figure the best solution is to have a model with low bias as well as low variance.



Bias-variance trade-off.

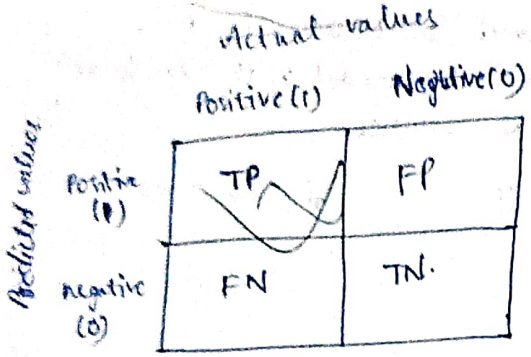
Evaluating performance of a model:-

→ confusion Matrix

confusion matrix is a performance measurement for the machine learning classification problems where the output can be two or more classes. It is a table with combinations of predicted and actual values.

→ A matrix containing correct and incorrect predictions, in the form of TPs, FPs, FNs and TNs is known as confusion matrix..

→ confusion matrix:- It is extremely useful for measuring the Recall, Precision, Accuracy and AUCROC curves.



- True Positive (TP):- correctly predicted positive class.
- False positive(FP): Incorrectly predicted as positive (Type I error)
- True Negative (TN): correctly predicted negative class.
- False Negative (FN): Incorrectly predicted as negative (Type II error).

Accuracy :-

model accuracy is given by total number of correct classifications (either as the class of interest, i.e True positive or as not the class of interest, i.e True negative) divided by total number of classifications done.

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

Recall(Sensitivity or True Positive Rate):-

→ Sensitivity of a model measures the proportion of TP examples or positive cases which were correctly classified.

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

Precision:-

→ Precision; gives the proportion of positive predictions which are truly positive.

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

F-measure:-

→ F-measure is another measure of model performance which combines the precision and recall. It takes the harmonic mean of precision and recall as calculated as

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

Specificity :-

→ also known as True Negative Rate (TNR).
→ It measures the proportion of negative examples which have been correctly classified.

$$\text{Specificity} = \frac{TN}{TN+FP}$$

Example:-

→ the confusion matrix of the win/loss prediction of cricket match problem to be as below:-

	Actual win	Actual loss
Predicted win	85	4
Predicted loss	2	9

→ Total count of TPs = 85

" " FPs = 4

" " FNs = 2

" " TNs = 9

model

$$\text{Accuracy} = \frac{85+9}{85+4+2+9} = 0.94.$$

$$\text{Recall} = \frac{85}{85+4} = 0.977\%$$

$$\text{Precision} = \frac{85}{85+4} = 0.955\%$$

$$\text{F-measure} = \frac{2 \times 0.95 \times 0.97}{0.95+0.97} = 96.6\%$$

$$\text{Specificity} = \frac{9}{9+4} = 0.69$$

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

In content of the above confusion matrix.

$$\text{Error rate} = \frac{4+2}{85+4+2+9} = \frac{6}{100} = 6\%$$

$$= 1 - \text{Model accuracy}$$

$$= 1 - 0.94$$

$$= 0.06.$$

→ In certain learning problems it is crucial to have extremely low number of FN cases.

→ example, if a tumor is malignant but wrongly classified as benign by the classifier, then the repercussion of such misclassification is fatal.

→ It does not matter if higher number of tumours which are benign are wrongly classified as malignant.

→ In these problems there are some measures of model performance which are more important than accuracy.

→ Two such critical measurements are Sensitivity and Specificity of the model.

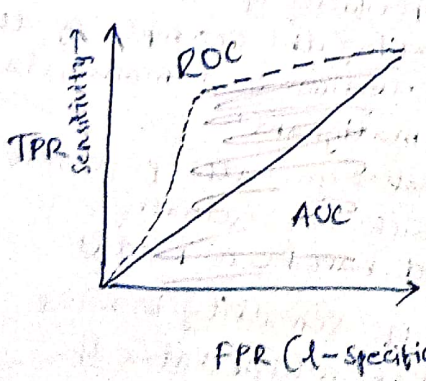
→ So, again taking the example of the malignancy prediction of tumours, class of intrest is 'malignant'.

→ Sensitivity measure gives the proportion of tumours which are actually malignant and have been predicted

→ A high value of sensitivity is more desirable than a high value of accuracy.

Receiver Operating characteristic (ROC) curves:-

→ ROC - Receiver operating characteristics,
→ It helps in visualizing the performance of a classification model.
→ In the ROC curve, the FP rate is plotted (in horizontal axis) against true positive rate (in the vertical axis) at different classification thresholds.
→ AUC - ROC curve is a performance measurement metrics for the classification problems at various threshold settings.
→ ROC is a probability curve, and
→ AUC represents the degree or measure of seperability.
→ It tells how much model is capable of distinguishing between classes, higher the AUC, better the model is at Predicting.



0.5 - 0.6 → Almost no predictive ability
0.6 → 0.7 → weak predictive ability
0.7 - 0.8 → Fair predictive ability
0.8 - 0.9 → good predictive ability
0.9 - 1.0 → excellent predictive ability

Supervised learning - regression:-

→ To asses the performance of a regression model, several metrics are used to quantify the difference between predicted and actual values.
Common metrics used for regression:-
- Mean squared Error (MSE),
- Root mean squared Error (RMSE),
- Mean Absolute Error (MAE), and
- R-squared (R^2)

Unsupervised learning - clustering:-
Performance metrics used for clustering:
- Silhouette Score:-

→ Measures how well each data point fits within its assigned cluster, considering both cohesion (similarity within the cluster) and separation (dissimilarity to other clusters)
→ **elbow method:**
→ Performance metrics used for Association analysis:-
- Support,
- confidence and
- lift

What is Feature?

→ Attribute of a data set that is used in a machine learning process.

→ certain machine learning practitioners consider only those attributes which are meaningful to a machine learning problem as features...

What is feature Engineering?

An Important pre-processing step for machine learning having two major elements –

→ feature transformation.

→ feature subset selection (or simply feature selection).

Feature transformation :- transforms into a new set of features.

Two variants of feature transformation –

→ feature construction.

→ feature extraction.

Feature construction :-

Discovers missing information about relationship between features and augments the feature space by creating additional features.

→ Say, there are 'n' feature or dimensions in a dataset.

→ After feature construction, 'm' more features get added.

→ So at the end the data set will become 'n + m' dimensional.

apartment length	apartment breadth	apartment price
80	59	23,60,000
54	45	12,15,000

⇒

apartment length	apartment breadth	apartment area	apartment price
80	59	4720	23,60,000
54	45	2,430	12,15,000.

Feature construction:-

⇒ Encoding categorical (ordinal) variables.

→ required in order to transform ordinal variable to a numeric variable.

→ eg; say in a data set there are three variables- science marks, maths marks and grade as shown.

→ A feature 'num-grade' can be created by mapping a numeric value against each ordinal value of the original feature 'grade'. numeric values 1, 2, 3 and 4 are mapped against ordinal grade values A, B, C and D.

marks-science	marks-maths	grade
78	75	B
56	62	C
87	90	A
91	95	A
45	42	D
62	57	B

(a)

"	"	num-grade
78	75	2
56	62	3
87	90	1
91	95	1
45	42	4
62	57	2

(b)

⇒ Transforming numeric (continuous) features to categorical:-

→ Required in order to transform numeric variable to a categorical variable.

→ Needed when we want to treat a regression problem as a classification problem.

→ For example, in content of [illegible] price prediction problem, the original data set has a numerical feature apartment-price, which needs to be predicted. However, it can be transformed to a categorical variables price grade.

apartment area	apartment price.
4,720	23,60,000
2,430	12,15,000
4,368	21,84,000
3,969	19,84,500
6,142	30,71,000
7,912	39,56,000

(a)

apartment area	price-grade
4,720	2
2,430	1
4,368	2
3,969	1
6,142	3
7,912	3

(c)

apartment area	price-grade.
4,720	medium
2,430	low
4,368	medium
3,969	low
6,142	high
7,912	high

(b)

⇒ Text specific feature construction:-

→ Text data is inherently unstructured in nature.

→ All machine learning models need numerical data as input. So the text data in the data sets need to be transformed into numerical features. Done following a process known as vectorization.

→ Vectorization consolidates word occurrences in all documents belonging to the corpus in the form of bag-of-words:

→ There are three major steps that are followed:

- Tokenize
- Count
- Normalize.

→ Generates a matrix, known as document-term matrix (also known as term-document matrix).