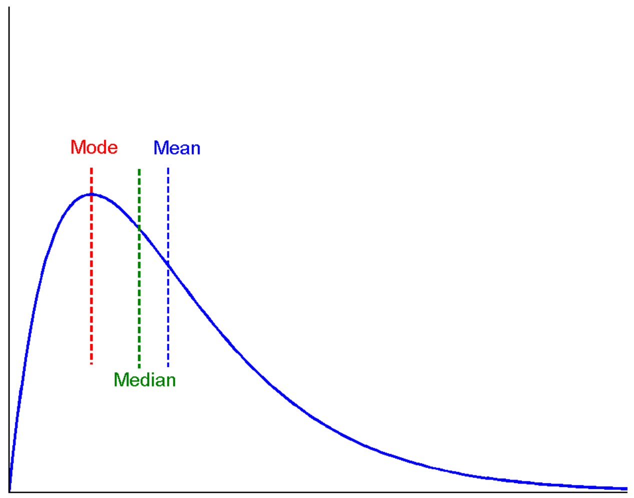
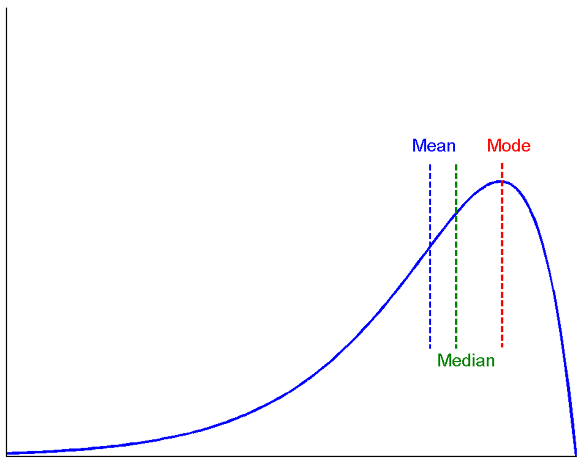
Ques 2. Positively skewed.

Skewness tells us to which side the mean leans in a probability distribution curve. In positively skewed distribution, the tail is the right side of the curve. Here, mean is the right-most of the three measures (mean, median and mode). Mode is the highest frequency point of the data. Median is in between mean and mode.



Ques 3. Negatively skewed:
In negatively skewed distribution, the tail is the left side of the curve. Here, mean is the left-most of the three measures (mean, median and mode). Mode is the highest frequency point of the data. Median is in between mean and mode.



Ques 5. ETL -
ETL stands for Extract, Transform , Load and it is the process involved in Data warehousing.

1) Extract - involves obtaining or extracting data from different sources. This part of the ETL process includes parsing of the data and checking if the data meets the expected pattern or structure for data storage purpose.
2) Transform - In this step, data is transformed into a particular format and data quality which can be easily processed and stored. It also involves cleaning up of data like removing duplicates, fixing different formats, handling null values, sorting and aggregating of the data.
3) Load - This step includes loading data into database or data warehouse for storage purposes. During this phase, various constraints like the referential integrity of the data, uniqueness of columns, etc are defined in the database schema and triggered while inserting the data into the database.

Ques 8. Chronological order:

step1. Prepare training data
step2. Set aside hold-out set
step3.teach classifier
step4.save model parameters
step5. verification with testing data
step6.input unlabeled new data into model
step7. output guesses on new data

Ques 10. Information Gain
Information gain is used to decide the splitting criteria. It helps in comparing how mixed or pure the cluster results are, before and after splitting.

Information gain is often used to decide which of the attributes are the most relevant, so they can be tested near the root of the decision tree i.e., it is used to build decision trees by choosing relevant attributes required to be investigated first during the decision making process.

Ques.11. Fixing gradient descent
RMSE is the difference between the values actually observed and the values predicted by the model. In the given example, we observe that RMSE has been continuously increasing after

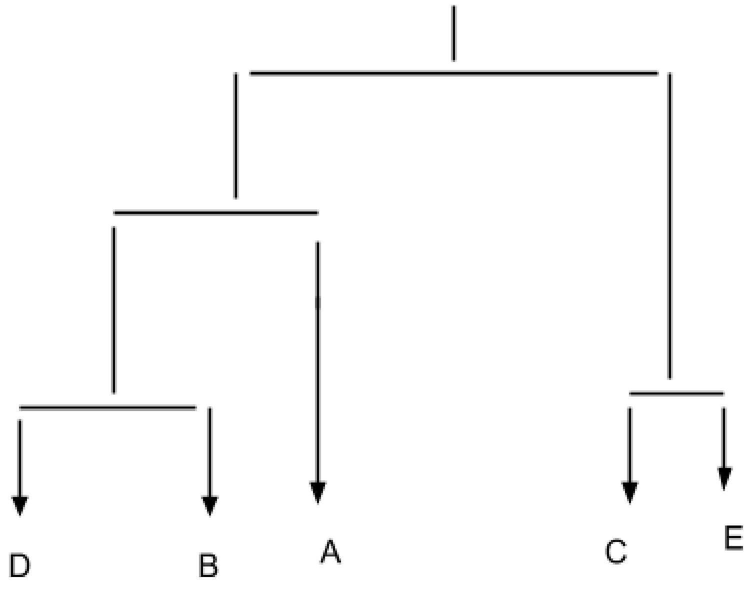
each iteration. To fix this, the **step-size parameter** has to be made small, so that more variables can be considered to train the model using linear regression.

Ques13. Square or Triangle?

1) k-1 : when k=1, we need to consider one nearest neighbour of the unknown circle. We see that triangle is the nearest to the circle. Therefore, ans = triangle

2) k=5: when k=5, we need to consider 5 nearest neighbours of the unknown circle and choose the symbol which is repeated the most. We see that the unknown circle has 3 squares and 2 triangles. Therefore, ans = square

Ques14. Dendrogram



Ques 16. Jaccard Index = (A intersection B)/ (A union B)

= 2 / (5+4-2) = 2/7 = 0.2857