Data Analysis is an aspect of data science & data analytics. That is all about analysing data for different kinds of purposes. The data analyses process involves inspecting, cleaning, transforming & modelling data to draw useful insights from it.

Types of data analysis:-

(i) Descriptive analysis: The goal of it is to describe or summarize a set of data. It is the very first analyses performed in a data analysis process. It generates simple summaries of data sample & measurements.
It involves some common descriptive statistics like measurement of central tendency, variability, frequency & position.

(ii) Diagnostive Analysis: It seeks why did this happen. By taking a more depth look at data to uncover hidden patterns.
It typically comes from descriptive analysis, taking initial findings & investigating why certain pattern in data happen. It may involve analysing other related data sources, including past data, to reveal more insights into current data set.

* Central limit Theorem: It states that the distribution of data sample means approx. a normal distribution as the sample size gets larger, regardless of the population distribution.

It is ideal for exploring patterns in data to explain ~~exploring~~ anamolies.

(iii) EDA : It involves exploring data & finding relationship b/w variables that were previously unknown. It is useful for discovering new connections within the data & forming various hypothesis. It drives design, planning data collection.

(iv) Inferential Data Analysis : It involves using a small data sample to infer information about a large population of data. The estimated data is a sepresentation of data population & gives a measure of uncertainity to your estimation. The accuracy of inference depends heavily on sampling scheme. If the sample is not the representation of the population, the generalization will not be accurate. This is known as central limit.

(v) Predictive Analysis :-
It involves using the historical or current data to find patterns & make predictions about the future.

(vi) Casual Analysis: ~~It's a~~ It searches for the cause & effect of relationship b/w variable & it focused on finding the reason of a correlation.

(vii) Mechanistic Analysis: It is used to understand the exact changes in the variable that lead to other changes in other variables.

(viii) Prescriptive Analysis: It compile insights from other previous data & determine actions that can be taken to prepare for predictive trends / patterns.

Ques. You are building a ML model to classify whether an email is spam or not.
~~You~~ After training your model you test it on the testing model dataset & confusion matrix given below is generated

	Actual	
Pred.	90	30
	20	160

TP	FP
FN	TN

① Calculate the accuracy of the model. What does the above matrix tell you about the model general performance.

② Calculate the precession, recall & f1 score for the model w.r.to identifying the spam email (use spam class as the +ve class).

Explain the significance of these metrices in the context of spam classification.

3. Compute the Mathew correlation coefficient for the model.
What does this parameter indicate about the quality of the classification?
How this is diff from other parameter like accuracy?

4. Calculate the G-means for the model.
What does this parameter tell you about the balance b/w sensitivity & specificity?

5. Based on the calculated parameters, access the strengths & weaknesses of the model.

6. If you have to improve the model which parameter would you focus & why.

① Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$

$= \frac{90 + 160}{90 + 160 + 30 + 20} = \frac{250}{300} = \frac{5}{6} = 83.33\%$

② Precision = $\frac{TP}{TP + FP} = \frac{90}{90 + 30} = \frac{90}{120} = \frac{3}{4} = 75\%$

Recall = $\frac{TP}{TP + FN} = \frac{90}{90 + 20} = \frac{90}{110} = \frac{9}{11} = 81.81\%$

F1 score = $\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot (0.75)(0.81)}{0.75 + 0.81} = 0.77\%$

③ MCC = $\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

$= \frac{(90 \times 160)(30 \times 20)}{\sqrt{(90 + 30)(90 + 20)(160 + 30)(160 + 20)}}$

④ ~~G-Means~~ = Sensitivity = $\frac{TP}{TP + FN}$

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