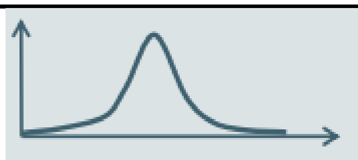
FORMULA SHEET OF STATISTICS

S.N.	STATISTIC	FORMULA	STATEMENTS
1.	**Mean**		The mean is the most widely spread measure of central tendency. It is the simple average of the dataset.
(a)	Population Mean (μ)	$\frac{\sum_{i=1}^{n} x_i}{N}$ **OR** $\frac{x_1 + x_2 + x_3 + \ldots \ldots + x_{N-1} + x_n}{N}$ (Where, N=size of population)	
(b)	Sample Mean ($\bar{x}$)	$\frac{\sum_{i=1}^{n} x_i}{n}$ (Where, n=size of sample)	
2.	**Median (M)**	if n is odd, $\left(\frac{n+1}{2}\right)^{th}$ if n is even, $\frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$	The median is the midpoint of the ordered dataset.
3.	**Mode**		The mode is the value that **occurs most often**. A dataset can have 0 modes, 1 mode or multiple modes.
4.	**Skewness**	$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sqrt[3]{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}}$	Skewness indicates whether the data is concentrated on one side or not. If the data is Right concentrated = Positive skewness Left concentrated = Negative skewness No concentrated = Zero skewness
(a)	Positive skewness	Mean > Median & Mode	
(b)	Negative skewness	Mode > Median & Mean	
(c)	Zero skewness	Mean = Median = Mode	
5..	**Variance**		Variance measures the dispersion of a set of data points around their mean.
(a)	Population Variance (σ^2)	$\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$ (Where, μ=mean of population)	
(b)	Sample Variance (s^2)	$\frac{\sum_{i=1}^{n}(x_i - \bar{x}))^2}{n-1}$ (Where, $\bar{x}$ =mean of sample)	

FORMULA SHEET OF STATISTICS

6.	**Standard Deviation**		Standard deviation is the most common measure of variability for a single Dataset.
(a)	Population Standard Deviation (σ)	$\sqrt{\sigma^2}$	
(b)	Sample Standard Deviation (s)	$\sqrt{s^2}$	
7.	**Coefficient of Variance (Cv)**	$\frac{\textbf{Standard Deviation}}{\textbf{Mean}}$	Cv is the most common measure of variability for two or more dataset.
8.	**Covariance**		The two variables are correlated and the main statistic to measure this correlation is called **covariance**. Covariance can take on values from -∞ to +∞.
(a)	Sample Formula	$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})*(y_i - \bar{y})}{n-1}$	
(b)	Population Formula	$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)*(y_i - \mu_y)}{N}$	
9.	**Correlation coefficient**	$\frac{Cov(x,y)}{Stdev(x) * Stdev(y)}$	Correlation adjusts covariance, so that the relationship between the two variables becomes easy and intuitive to interpret. Correlation coefficient is between -1 ≤ Correlation coefficient ≤ 1 Correlation of 1 (perfect positive corelation) Correlation of 0 (variables are independent) Correlation of -1 (perfect negative corelation)

FORMULA SHEET OF STATISTICS

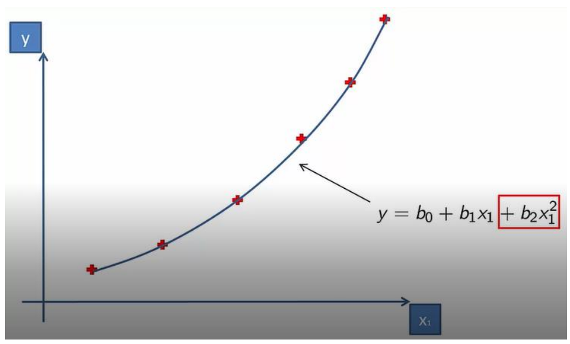
10.	**Probability distribution**		Distribution is a function that shows the possible values for a variable and how often they occur.
(a)	Normal distribution	$N \sim (\mu\,, \sigma^2)$	
(b)	Standard normal distribution	$N \sim (0\,, 1\,)$	The Standard Normal distribution is a particular case of the Normal distribution. It has a mean of 0 and a standard deviation of 1.
(c)	Z-score	$\dfrac{x-\mu}{\sigma}$	Variable which converts normal distribution variable to standard normal distribution is called z-score.
11.	**Confidence intervals**		A confidence interval is an interval within which we are confident (with a certain percentage of confidence) the population parameter will fall. General formula of Confidence Interval: **C.I.=** **[** $\bar{x} \pm ME$ **]** Where ME is the margin of error.
(a)	Confidence Level	**(1-α)** $0 \leq \alpha \leq 1$	
(b)	Margin of Error (ME)	$reliability\ factor * \dfrac{\sigma}{\sqrt{n}}$ **or** $reliability\ factor * \dfrac{s}{\sqrt{n}}$	
(c)	Standard Error	$\dfrac{\sigma}{\sqrt{n}}$ **or** $\dfrac{S}{\sqrt{n}}$	
(d)	Population variance known (Z)	$\bar{x} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$	
(e)	Population variance unknown (T)	$\bar{x} \pm t_{n-1,\alpha/2}\dfrac{s}{\sqrt{n}}$	
(f)	For two population means with dependent	$\bar{d} \pm t_{n-1,\alpha/2}\dfrac{s_d}{\sqrt{n}}$	
(g)	Student's T distribution	$t_{n-1,\alpha}=\dfrac{\bar{x}-\mu}{s/\sqrt{n}}$	

FORMULA SHEET OF STATISTICS

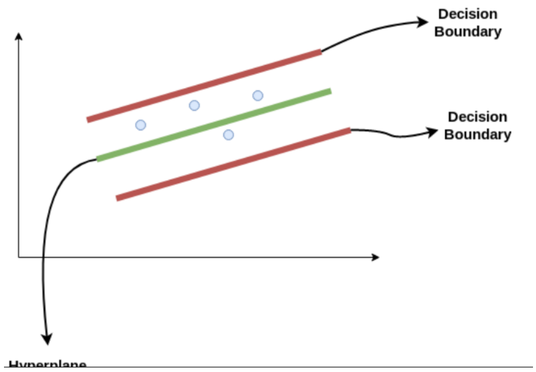
12.	**Hypothesis Testing**		A hypothesis is an idea that can be tested NULL HYPOTHESIS → (H_o) ALTERNATIVE HYPOTHESIS → $(H_1\ OR\ H_A)$
(a)	TYPE I ERROR	REJECTING TRUE NULL HYPOTHESIS OR FALSE POSITIVE	
(b)	TYPE II ERROR	ACCEPTING A FALSE NULL HYPOTHESIS OR FALSE NEGETIVE	
(c)	P value	P-value is the smallest level of significance at which we can still reject the null hypothesis, given the observed sample statistic.	
13.	**Advance Statistics**		(SST = SSR + SSE)  **<u>OLS</u> → Ordinary least squares** OLS stands for min SSE Or we can say that lower error. $\min \sum_{i=1}^{n} e_i^2$
(a)	Linear Regression	$\hat{\boldsymbol{y}} = \boldsymbol{b_0} + \boldsymbol{b_1 x_1}$ Where, $\hat{y} = \ Predicted\ /\ inferred\ value$ $b_0 = \ Intercept$	
(b)	Sum of squares total	$\sum_{i=1}^{n} (y_i - \bar{y})^2$	
(c)	Sum of squares regression	$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	
(d)	Sum of squares error	$\sum_{i=1}^{n} e_i^2$	
(e)	R-Square	$\frac{\boldsymbol{SSR}}{\boldsymbol{SST}}$	
(f)	Multivariate linear regression	$\hat{\boldsymbol{y}} = \boldsymbol{b_0} + \boldsymbol{b_1 x_1} + \boldsymbol{b_2 x_2} + \cdots + \boldsymbol{b_k x_k}$	when we have multiple variables then we use multivariate analysis.
(g)	Adjusted R-Square	$\overline{\boldsymbol{R^2}} < \boldsymbol{R^2}$	The adjusted R-squared penalizes excessive use of variables. adjusted R square always use in multilinear regression.

FORMULA SHEET OF STATISTICS

14.	**Principle Component Analysis (PCA)**		The objective of PCA is to reduce the dimensionality of the data while retaining as much of the variation in the original dataset as possible.
(a)	Covariance of X&Y	$cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$	
(b)	Identical Matrix	**C- λI=0**	
15.	**Slope (m)**	$m = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$	Y = mx+c (equation of regression line) Where m=slope The slope indicates the steepness of a line.
16.	**CHI-SQUARE TEST (x^2)**	$x^2 = \frac{\sum (O_i - E_i)^2}{E_i}$	The Chi-Square test is a statistical procedure for determining the difference between observed and expected data. This test can also be used to determine whether it correlates to the categorical variables in our data.
17.	**Gradient Descent**		Gradient descent is an algorithm that finds the best fit line for given training data set.
(a)	Mean squared error/ Cost function/ OLS	$\frac{1}{n}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$	The Mean Squared Error measures how close a regression line is to a set of data points.
18.	**Polynomial Regression**	$y = b_0 + b_1x_1 + b_2x_1^2 + \ldots + b_nx_1^n$	Polynomial regression is a type of regression analysis in which the relationship between the independent variable (x) and the dependent variable (y) is modelled as an nth degree polynomial.



FORMULA SHEET OF STATISTICS

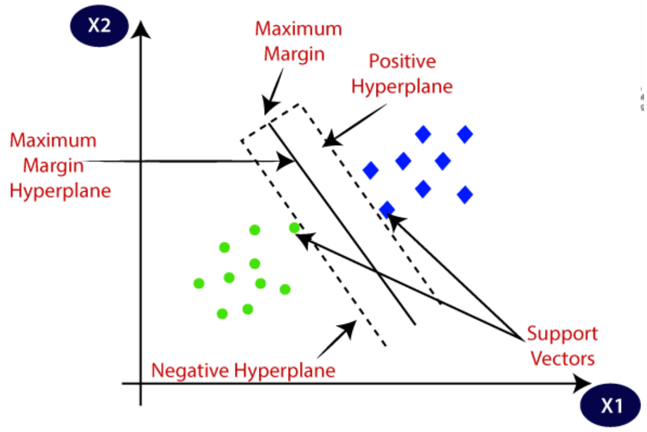
19.	**Bias & Variance**		Bias refers to the Training Data Variance refers to the Testing data. Low bias, High variance -- overfitting problem High bias, Low variance -- underfitting problem Low bias, Low variance = Good Model
(a)	**Total error**	Bias^2+Variance+Irreducible error	
20.	**Simple Vector Regression (SVR)**		 Main aim in SVR is to decide a decision boundary at 'a' distance from the original hyperplane such that data points closest to the hyperplane or the support vectors are within that boundary line.
(a)	Hyperplane equation	**Y=wx+b**	
(b)	Decision boundary	**wx+b=a (for positoive side)** **wx+b=-a (for negative side)**	
(c)	Hyperplane satisfy SVR	**-a < Y- wx+b < +a**	
21.	**Regularization**		Regularization is a technique used in machine learning to prevent overfitting and improve the generalization ability of a model.
(a)	Lasso Regression (L1 regularization)	**Loss** + $\alpha \lVert w \rVert$ $\lVert w \rVert = w_1+w_2+w_3+...+w_n$	Scaled down to 0 (due to we are eliminate the feature) this also called as FEATURE ELIMINATION TECHNIQEU
(b)	Ridge Regression (L2 regularization)	**Loss** + $\alpha \lVert w \rVert^2$ $\lVert w \rVert^2 = w_1^2+w_2^2+w_3^2+...+w_n^2$	Scaled down high coefficient to low coefficient but it not 0 (scale down never happened to 0).
(c)	Elastic net Regression	**Loss** + $\alpha_1 \lVert w \rVert + \alpha_2 \lVert w \rVert^2$	It uses the penalties from both the lasso and ridge techniques to regularize regression models

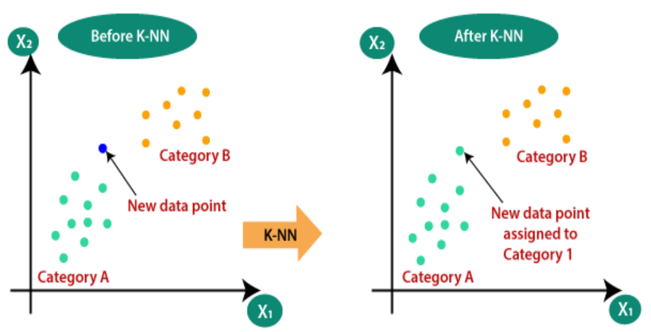
FORMULA SHEET OF STATISTICS

22.	**Feature Scaling**		Feature scaling is a data preprocessing technique that is used to bring different features of a dataset onto a similar scale.
(a)	Normalization	$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$	Normalization typically refers to scaling a set of values to a range between 0 and 1. This is useful when the range of the data varies widely, as it puts all the values on a similar scale.
(b)	Standardization	$$X' = \frac{X - \mu}{\sigma}$$	Standardization involves transforming a set of values to have a mean of 0 and a standard deviation of 1. This is useful when the distribution of the data is normal or approximately normal.
23.	**Confusion Metrix**		It is a matrix that compares the predicted labels of a model with the actual labels in the test dataset, and provides a summary of the model's performance
(a)	Model Accuracy	$\frac{TP + TN}{TOTAL}$	
(b)	Error Rate	**1 - Accuracy** **Or** $\frac{FP + FN}{TOTAL}$	
(c)	Recall	$\frac{TP}{Actual\,Yes}$ **Or** $\frac{TP}{TP + FN}$	
(d)	Precision	$\frac{TP}{Predicted\,Yes}$	
(e)	F1 score	$\frac{2.(Precission * Recall)}{((Precission + Recall)}$	
24.	**Logistic Regression**	$$Max \sum_{i=1}^{n} y_i w_i^T x_i$$	It is a statistical method for analysing a dataset in which there are one or more independent variable that determine an outcome. Logistic regression also called as logit or maxent

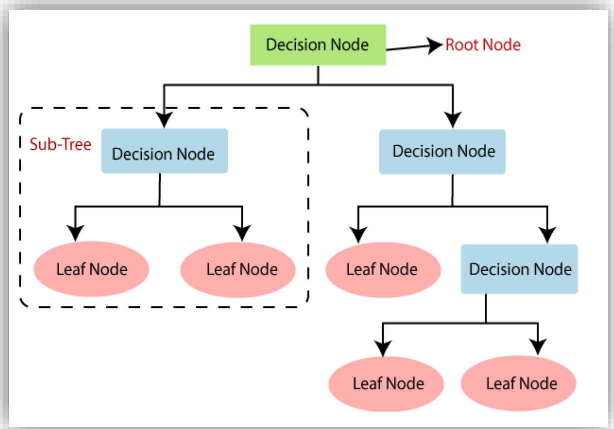
		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

FORMULA SHEET OF STATISTICS

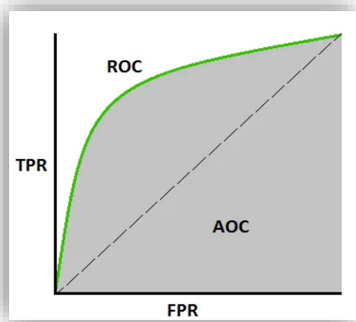
<table>
<tr><td>25.</td><td>Sigmoid Function</td><td>$$P = \frac{1}{1+e^{-y}}$$ $$ln\left(\frac{p}{1-p}\right) = b_0 + b_1 \cdot x$$</td><td>When the data has outlier then logistic regression algorithm would misclassify, that's why we use probability concept using the help probability function which is called (SIGMOID FUNCTION).</td></tr>
<tr><td>26.</td><td colspan="2">Support Vector Machines (SVM)</td><td rowspan="3">It is a supervised machine learning problem where we try to find a hyperplane that best separates the two classes.


Two type of support vector machine ->
1-Linear SVM => When the data is perfectly linearly separable only then we can use Linear SVM.
2-Non-Linear SVM=> if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data</td></tr>
<tr><td>(a)</td><td>For Linear SVM</td><td>$$\max(w^*, b^*)\ \frac{2}{\|w\|}$$</td></tr>
<tr><td>(b)</td><td>For Non-Linear SVM</td><td>$$\min\left(w^*, b^*\right)\ \frac{\|w\|}{2} + c\sum_{i=1}^{n}\zeta_i$$</td></tr>
<tr><td>27.</td><td colspan="2">K-Nearest Neighbor (KNN)</td><td rowspan="3">K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories

</td></tr>
<tr><td>(a)</td><td>Euclidean distance</td><td>$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$</td></tr>
<tr><td>(b)</td><td>Manhatten distance</td><td>distance = absolute sum ||xi-yi||</td></tr>
</table>

FORMULA SHEET OF STATISTICS

28.	**Naïve Bayes**		It is a supervised machine learning algorithm, which is used for classification tasks, like text classification.
(a)	BERNOULLI NAVE BAYE'S VARIANCE	$P(x = x) = p^x \cdot (1-p)^{1-x}$	
(b)	GAUSSION NAVE BAYE'S VARIANCE	$f_{(x)} = \frac{1}{\sqrt{2\Pi}\sigma} e \frac{-1}{2\sigma^2}(x-\mu)^2$ $-\infty < x < \infty$	
(c)	MULTINOMIAL NAVE BAYE'S VARIANCE	$P(x_1 = x_i \dots x_k)$ $\frac{n!}{x_1! \dots x_k} ! P^{1^{x_1 \dots}} P_k^{x_k}$	
(d)	Conditional Probability	P(A|B) = $\frac{P(A \cap B)}{P(B)}$	Conditional probability is a measure of the probability of an event occurring, given that another event has already occurred.
(e)	Baye's Theorem	P(A|B) = $\frac{P(B\|A) * P(A)}{P(B)}$ P(A|B)=Posterior probability P(B|A) = Likelihood P(A) =Prior Probability P(B) = Marginal likelihood	Bayes' Theorem states that the conditional probability of an event, based on the occurrence of another event, is equal to the likelihood of the second event given the first event multiplied by the probability of the first event.
29.	**Decision Tree**		It is a tree-structured classifier where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. 
(a)	Information Gain	$-\frac{P}{P+N}\log_2\left(\frac{P}{P+N}\right) - \frac{N}{P+N}\log_2\left(\frac{N}{P+N}\right)$	
(b)	Entropy E(A)	$\sum_{i=1}^{v} \frac{P_i + N_i}{p + N}(P_i N_i)$	
(c)	Gain	I.G. - E(A)	

FORMULA SHEET OF STATISTICS

(d)	GINI Index	$I_G = 1 - \sum_{j=1}^{c} p_j^2$	
(e)	Entropy	$I_H = -\sum_{j=1}^{c} p_j log_2(p_j)$	
30.	**Ensemble Technique**		**Ensemble methods** is a machine learning technique that combines several base models in order to produce one optimal predictive model. Ensemble Method is split into three types - 1. Bagging --> Random Forest 2. Boosting --> (a) ADABOOST (b) GRADIENT BOSSTING (c) XGBOOST 3. Voting (stacking)
(A)	**ADABOOST**		
(a)	Performance of Stump	$\frac{1}{2}\log_e\left(\frac{1 - Total\ Error}{Total\ Error}\right)$	
(b)	New sample weight	$old\ weight\ *\ e^{\pm Amount\ of\ say\ (\alpha)}$	
(B)	**Gradient Boosting**		
(a)	F(x)	$h_0(x) + l_1 h_1(x) + l_2 h_2(x) + \cdots + l_n h_n(x)$ **Where,** l = learning rate	
31.	**AUC & ROC Curve**		It will visualization for confusion matrix (classification) AUC & ROC is graph to measure the accuracy of the dataset 
(a)	True positive rate (TPR) Or Sensitivity	$\frac{TP}{TP+FN}$	
(b)	False positive rate (FPR) Or Specificity	$\frac{FP}{FP+TN}$	

FORMULA SHEET OF STATISTICS

32.	**Clustering**		Clustering is an unsupervised machine learning method of identifying and grouping similar data points in larger datasets without concern for the specific outcome. Clustering (sometimes called cluster analysis) is usually used to classify data into structures that are more easily understood and manipulated.
(a)	Euclidean Distance	$\sqrt{(X_0 - X_c)^2 + (Y_0 - Y_c)^2}$	Euclidean distance is used in many machine learning algorithms as a default distance metric to measure the similarity between two recorded observations. It works on the principle of the Pythagoras theorem and signifies the shortest distance between two points.
(b)	Within-Cluster Sum of Square (WCSS)	$\sum_{p_i=1 \in C_j}^{p_m} distance(C_j, p_i)^2$	WCSS is defined as the sum of the squared distance between each member of the cluster and its centroid.