Unit-4

✱ 18/sept/2023 Data Analytics Techniques

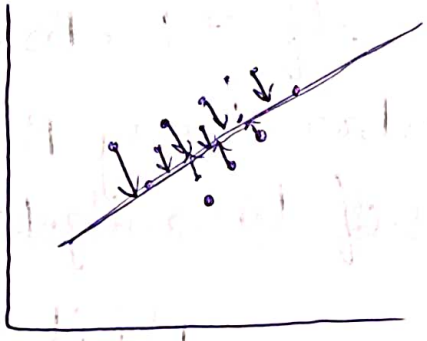
✱ Regression analysis :- is a statical method to model the relationship b/w a dependent and independent var with one or more independent var. (explain further ourself example etc)

① dependent var (target var) :-
② independent var (predictor) :-
③ Outliers :-
④ multi collinearity :-
⑤ under fitting.
⑥ Over fitting.

Types of Regression

① Linear Regression ✓ } syllabus
② logistic " ✓
③ Polynomial "
④ Support Vector " ——— ✓ } syllabus
⑤ Decision Tree "
⑥ Random Forest "
⑦ Ridge regression.
⑧ Lasso Regression.

⊛ Linear Regression:-



$$y = mx + C$$

$$y = a_2 x + a_1$$

where

$$a_2 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$a_1 = \frac{\sum y \sum x^2 - \sum x \sum xy}{n\sum x^2 - (\sum x)^2}$$

Here y = dependent var (target var).
x = independent var (predictor).
a, b are linear coefficient.
a_2 = slope
a_1 = intercept

18/sept/2023

⊛ Logistic Regression:-

$$\text{odds} = \frac{\hat{p}}{1-\hat{p}}$$

p is the proportion of customers in the population of the customers from which the sample was drawn who would respond yes to the questio

$n = 250$

$x = 210$

$$\hat{p} = \frac{210}{250} = 0.84$$

$$\text{odds (yes)} = \frac{\hat{p}}{1-\hat{p}} =$$

model for logistic Regression:- In simple linear Regression we modelled the μ_y of the response var y as a ~~non~~linear func^n of the explainatory func^n.

$$\boxed{\mu = \beta_0 + \beta_1 x}$$

when y is just 1 (success) or 0 (failure) the mean is the probability p of a success.

Logistic Regression models the mean p in terms of an explainatory var x.

We might try to relate p and x as in a simple linear regression

$$p = \beta_0 + \beta_1 x$$

Unfortunately this is not a good model whenever $\beta_1 \neq 0$ extreme values of x will give values of $\beta_0 + \beta_1 x$ that fall out side the range of possible values of p $(0 \leq p \leq 1)$.

The logistic regression solution to this difficulty is to transform the odds $\left(\frac{p}{1-p}\right)$ using the natural logarithmic, we use the term log odds or logit for this transformation.

We use y for response var so for women

$$y = \log(\text{odds})$$
$$y = \log(1.5694)$$
$$y = 0.4507$$

Count	User NO	Yes	Total
Women	209	328	537
	38.92%.	61.08%.	
men	298	234	532
	56.02%.	43.98%.	
Total	507	562	1069

$$\underset{(\text{women})}{\text{odds}} = \frac{p}{1-p} = \frac{.6108}{1-.6108} = 1.5694$$

$$\text{odds (men)} = \frac{.4398}{1-.4398} = 0.785$$

For men

$$y = \log(\text{odds})$$
$$y = \log(0.785)$$
$$y = -0.2427$$

For women

$$y = \log(\text{odds})$$
$$y = \log(1.5694)$$
$$y = 0.4507$$

In these expression for the log odds we used y as observed value of the response var. The log odds of using instagram. We are now ready to build the logistic regression model.

We model the log odds as a linear func^n of the explainatory var.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

For our insta example, there are $n = 1069$ young persons in the sample the explainator var is sex, which we have coded using an indicator var $x = 1$ for women
$x = 0$ for men

the response var y is also an indicator val thus each person either a insta user or not. Think of a process of selecting

a young person at random and recording y and x. The model says that the prob. P, that this person is an insta user can depend upon user sex (x=0, x=1).

There are two possible values for P_{women} and P_{men}

$$\log\left(\frac{P_{women}}{1 - P_{women}}\right)$$

the logistic reg. model specifies the relationship b/w P and x bcoz there are only two values for women

$$\log\left(\frac{P_{women}}{1 - P_{women}}\right) = \beta_0 + \beta_1(1) \quad \text{—— (1)}$$

and

$$\log\left(\frac{P_{men}}{1 - P_{men}}\right) = \beta_0 + 0 \quad \text{—— (2)}$$

(x=0)

$$\beta_0 = \log\left(\frac{-0.2419}{1 + 0.2419}\right)$$

$$\beta_0 = \log\left(\frac{-0.2419}{1.2419}\right)$$

$$\beta_0 = -0.2419$$

$$\beta_1 = 0.6926$$

18 | sept | 2023

$\log(\text{odds}) = -0.2419 + 0.6926x$

Ans.

The slope of in this logistic reg. model is the differen b/w log(odds) for men and log(odds) for women.

$$= \log\left(\frac{P_{women}}{1-P_{women}}\right) - \log\left(\frac{P_{men}}{1-P_{men}}\right)$$

$$= \log(1.5692) - \log(0.785)$$

Or

$$= \log\left(\frac{\text{odds}_{women}}{\text{odds}_{men}}\right) = \log\left(\frac{1.5694}{0.785}\right)$$

$$= \log(1.9992)$$

$$= 0.30085 \quad \text{Ans.}$$

~~Odds/women = 1.99 x odds/men~~

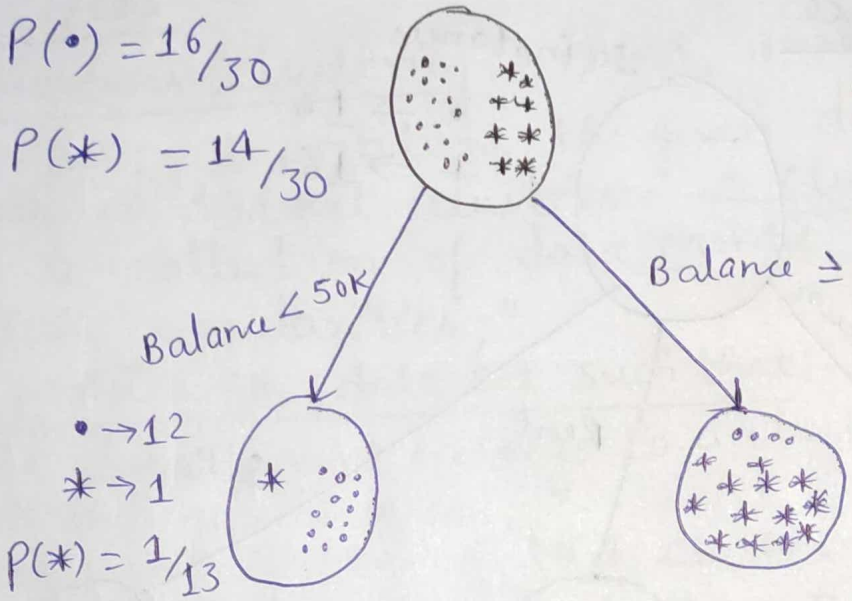
20|sept|2023

Decision Tree

27/Sept/2023

→ Entropy (impurity)

Q:- Consider an ex where we are predicting a decisiontree where a loan can be right off or not . — —



$$E = \sum_{i=1}^{K} (-f_i * \log f_i)$$

$$E(\text{parent}) = -\frac{16}{30} \times \log\frac{16}{30} - \frac{14}{30} \log\frac{14}{30}$$

$$= 0.99$$

$$E(\text{Balance} < 50) = -\frac{12}{13} \log_2\frac{12}{13} - \frac{1}{13} \log\frac{1}{13}$$

$$= 0.39$$

$$E(\text{Balance} \geq 50K) = -\frac{4}{17} \log\frac{4}{17} - \frac{13}{17} \log\frac{13}{17} = 0.79$$

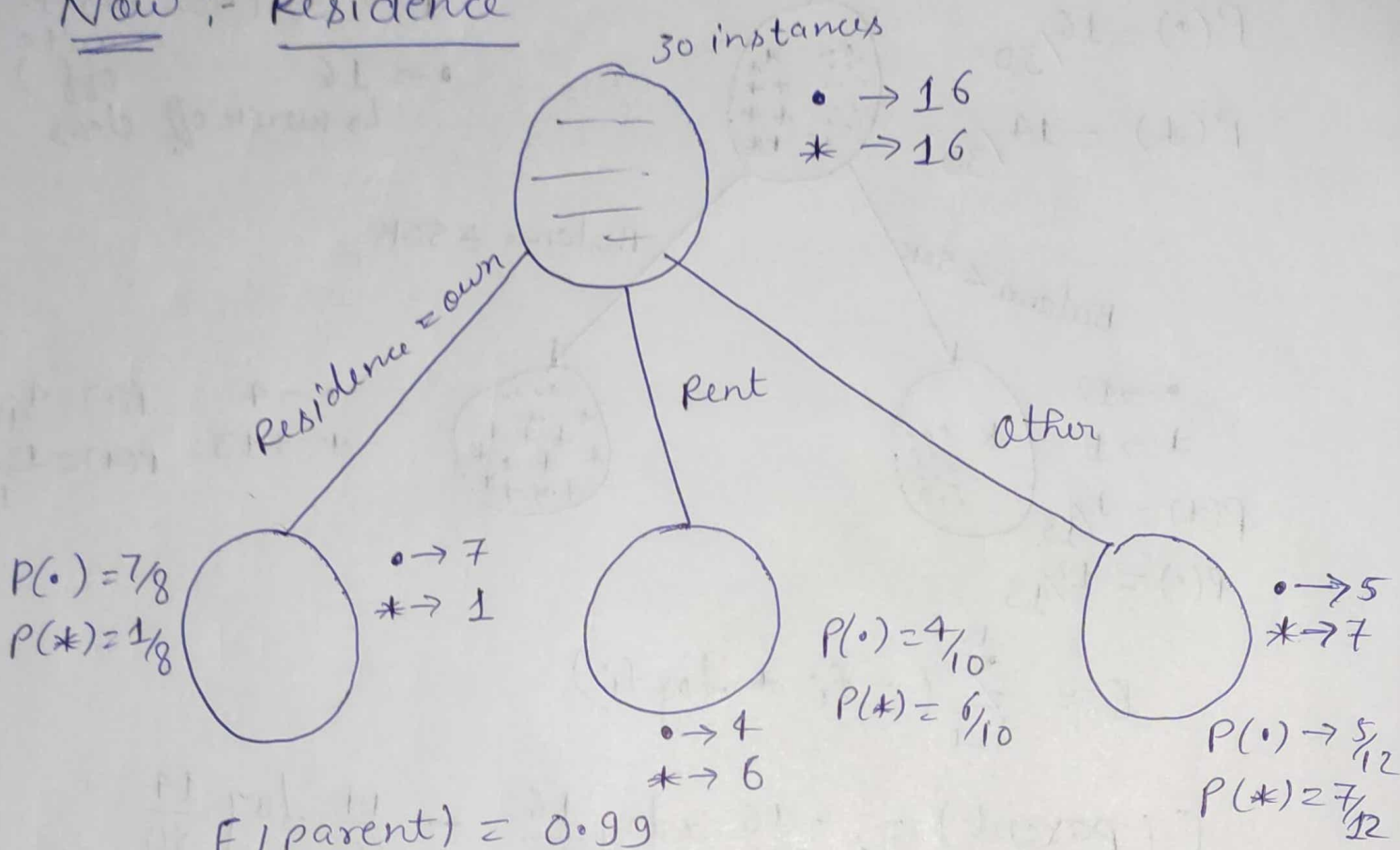
$$E(\text{Balance}) = \frac{13}{30} \times 0.39 + \frac{17}{30} \times 0.79$$

$$= 0.62$$

$$IG(\text{Parent}, \text{Balance}) = E(\text{Parent}) - E(\text{Balance})$$

$$= 0.99 - 0.62 = 0.37$$

Now :- Residence



$$E(\text{parent}) = 0.99$$

$$E(\text{own}) = -\frac{7}{8}\log\frac{7}{8} - \frac{1}{8}\log\frac{1}{8} = 0.543$$

$$E(\text{Rent}) = -\frac{4}{10}\log\frac{4}{10} - \frac{6}{10}\log\frac{6}{10} = 0.97$$

$$E(\text{other}) = -\frac{5}{12}\log\frac{5}{12} - \frac{7}{12}\log\frac{7}{12} =$$

$$= -0.416 \log(0.416) - 0.583 \log 0.583$$

$$= -0.416 \times (-1.26) + 0.583 \times 0.77$$

$$= 0.52416 + 0.44891 \Rightarrow 0.978$$

$= + 0.875 \times 0.77 + 0.125 \times 3$

$= 0.67375 + 0.375$

$= 1.04875$

IGr = 0.13
(Info Gain)

11/Oct/2023

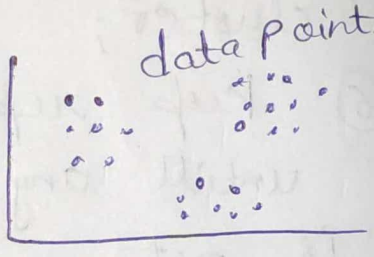
⊛ K-means clustering :- is an unsupervised clustering technique; It partitions the given dataset into K predefined distinct clusters. "cluster is defined as a collection of data points exibiting certain similarities;"

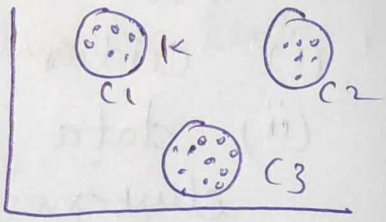
→ It partition the data set such that;-

→ Each data point belongs to a cluster with the nearest mean;

→ Data point belonging to 1 cluster have higher degree of similarity.

→ Data point belonging to different clusters have higher degree of dis-similarity;





After clustering.

⊛ K-means clustering Algo;-

① choose the no. of clusters;

② (i) Randomly select any K data points as cluster centers;

(ii) select cluster centers such that they are as far as possible from each other;

11/Oct/2023

③ (i) Calculate the distance b/w each data point and each cluster center, (ii) the distance may be cal. either by using given distance function or by using Euclidean distance formula;

④ Assign each data point to some ~~sub~~ cluster, a data point is assigned to that cluster whose center is nearest to that data point;

⑤ Recompute the center of newly formed clusters; the center of a cluster is computed by taking mean of all the data points contained in that cluster;

⑥ Keep repeating the procedure from step 3 to 5 untill any of the following stopping criteria is met

(i) center of newly formed clusters don't change

(ii) data points remain present in the present clusters.

(iii) max no. of iterations are reached;

* Advantages of K-means clustering :-

① It is relatively efficient with time complexity $O(nkt)$ n = no. of instances
k = no. of clusters
t = no. of iterations

② It often terminates at local optimum;

③ Techniques such as simulated annealing or genetic algo may be used to find the global optimum;

⊛ Disadvantages of K-means clustering -

① It requires to specify the no of cluster k in advance;
② It can't handle noisy data and outliers;
③ It is not suitable to identify clusters with non convex shapes;

Problem 1:- cluster the following 8 points with (x,y) representing points in 3 clusters; distance func^n b/w two point $a=(x_1,y_2)$, $b=(x_2,y_2)$

$$P(a,b) = |x_2-x_1| + |y_2-y_1|$$

Sol^n:-

A_1 (2,10)
A_2 (2,5)
A_3 (8,4)
A_4 (5,8)
A_5 (7,5)
A_6 = (6,4)
A_7 = (1,2)
A_8 = (4,9)

Use K-means algo to find the 3 cluster center after the 2 iteration:- initial cluster centre are A_1, A_4, A_7;

① Itr:- K = 3

Given Points	Distance from centre A_1 (2,10)	Distance from A_4 (5,8)	Distance from A_7 (1,2)	Point belongs
A_1 (2,10)	0	5	9	C_1
A_2 (2,5)	5	6	4	C_3
A_3 (8,4)	12	7	9	C_2
A_4 (5,8)	5	0	10	C_2
A_5 (7,5)	10	5	9	C_2
A_6 = (6,4)	10	5	7	C_2
A_7 (1,2)	9	10	0	C_3
A_8 (4,9)	3	2	10	C_2

After first iteration :-

new center

Cluster -1 (A_1) $\overset{\text{mean}}{=}$ (2, 10)

Cluster -2 (A_3, A_4, A_5, A_6, A_8) $\overset{\text{mean}}{=}$ (6, 6)

Cluster -3 (A_2, A_7) = mean (1.5, 3.5)

Iteration 2 :-

Points	center (2,10) point	C2 (6,6)	C3 (1.5, 3.5)	Point belongs
A_1 (2,10)	0	8	7	C1
A_2 (2,5)	5	5	2	C3
A_3 (8,4)	12	4	7	C2
A_4 (5,8)	5	3	8	C2
A_5 (7,5)	10	2	7	C2
A_6 (6,4)	10	2	5	C2
A_7 (1,2)	9	9	2	C3
A_8 (4,9)	3	5	8	C1

C_1 = A_1 (2,10), A_8 (4,9)

C_2 = (A_3, A_4, A_5, A_6)

C_3 = (A_7, A_2)

Ans.

calculate mean, get center from that

18 | Oct | 2023

Unit-2 :- Imp Question

Q1-① data visualization types, benefits, features of tableau ;

② V Lookup function explain with help of example.

③ What is a pivot table? How to create a pivot table in excel online ?

④ Write all steps to create a dashboard in excel.

⑤ Brief notes on Power BI - visualization ;

⊛ Unit-4 Principle component Analysis (PCA)

Dimensionality Reduction in Machine Learning

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observation or possibly correlated vars into a set of values of linearly uncorrelated vars called principal component ;

We will discuss PCA with the help of an example.

① Prob definition :- given data in table reduce the dim from 2 to 1 using the PCA algo ;

18/oct/2023

feature	Ex1	Ex2	~~Ex3~~	Ex3	Ex4	mean
X_1	4	8	13	13	7	8
X_2	11	4	~~8~~	5	14	8.5

Soln:- Step1 :- cal mean of the X_1 and X_2;

$$\bar{X}_1 = \frac{32}{4} = 8$$

$$\bar{X}_2 = 8.5$$

Step2 :- cal covariance table

$$\begin{bmatrix} \overbrace{cov(X_1, X_1)}^{var(X_1)} & cov(X_1, X_2) \\ cov(X_2, X_1) & \underbrace{cov(X_2, X_2)}_{Var(X_2)} \end{bmatrix}$$

$$cov(X_1, X_1) = VAR(X_1)$$

$$= \frac{1}{N-1} \sum_{K=1}^{N} (X_{1K} - \bar{X}_1)^2$$

$$= \frac{1}{4-1} \left[(4-8)^2 + (8-8)^2 + (13-8)^2 + (7-8)^2 \right]$$

$$= \frac{1}{3} [42]$$

$$cov(X_1, X_1) = 14$$

$$\text{cov}(x_1, x_2) = \frac{1}{N-1} \sum_{k=1}^{N} (x_{1k} - \bar{x}_1)(x_{2k} - \bar{x}_2)$$

$$= \frac{1}{4-1} \Big[(4-8) \times (11-8.5) + (8-8) \times (4-8.$$
$$+ (13-8) \times (5-8.5) + (7-8) \times (14-8.5$$
$$\Big]$$

$$= -11$$

$$\text{Cov}(X_2, X_1) = \text{Cov}(X_1, X_2) = -11$$

$$\text{var}(X_2) = \frac{1}{N-1} \sum_{k=1}^{N} (X_{2k} - \bar{X}_2)^2$$

$$= \frac{1}{3} \Big[(11-8.5)^2 + (4-8.5)^2 + (5-8.5)^2$$
$$+ (14-8.5)^2 \Big]$$

$$= \frac{1}{3} \Big[(2.5)^2 + (-4.5)^2 + (-3.5)^2$$
$$+ (5.5)^2 \Big]$$

$$= \frac{1}{3} \Big[6.25 + 20.25 + 12.25 + 30.25 \Big]$$

$$= \frac{69}{3} = 23 \text{ Ans}$$

Covariance matrix

$$S = \begin{vmatrix} \text{cov}(x_1, x_2) & \text{cov}(x_1, x_2) \\ \text{cov}(x_2, x_1) & \text{cov}(x_2, x_2) \end{vmatrix}$$

$$S = \begin{vmatrix} 14 & -11 \\ -11 & 23 \end{vmatrix}$$ Ans

Step-3 i- Eigen vectors of the covariance mat

S :-

$$|S - \lambda I| = 0$$

$$\begin{vmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{vmatrix} = 0$$

$$(14-\lambda)(23-\lambda) - 121 = 0$$

$$322 - 14\lambda - 23\lambda + \lambda^2 = 121$$

$$\lambda^2 - 37\lambda + 322 - 121 = 0$$

$$\lambda^2 - 37\lambda + 201 = 0$$

$$\boxed{\lambda = \lambda_1, \lambda_2 = 30.3849, 6.6151}$$

Step-4 Eigen Vector

$$U = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = (S - \lambda_1 I)\, U$$

$$= \begin{bmatrix} 14-\lambda_1 & -11 \\ -11 & 23-\lambda_1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} (14-\lambda_1)u_1 - 11u_2 \\ -11u_1 + (23-\lambda_1)u_2 \end{bmatrix}$$

$$(14-\lambda_1)u_1 - 11u_2 = 0$$

$$-11u_1 + (23-\lambda_1)u_2 = 0$$

$$\frac{u_1}{11} = \frac{u_2}{14-\lambda_1} = t \quad \text{(using delta, gradient descent Rule)}$$

$$u_1 = 11t$$
$$u_2 = (14-\lambda_1)t \quad \Big\} \; t=1$$

$$U_1 = \begin{bmatrix} 11 \\ 14-\lambda_1 \end{bmatrix} \quad \text{first Eigen vectors}$$

$$\|U_1\| = \sqrt{11^2 + (14-\lambda_1)^2}$$
$$= \sqrt{11^2 + (14-30.3849)^2}$$
$$\|U_1\| = 19.7348$$

a unit eigen vector corresponding to λ_1

$$e_1 = \begin{bmatrix} 11/\|U_1\| \\ (14-\lambda_1)/\|U_1\| \end{bmatrix}$$

$$= \begin{bmatrix} 11/19.7348 \\ (14-30.3849)/19.7348 \end{bmatrix}$$

$$= \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix}$$

A unit eigen vector corresponding to λ_2 can be calculated in similar way.

$$e_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix}$$

Step 5 i- Computation of first PCA

Let $\begin{bmatrix} x_{1K} \\ x_{2K} \end{bmatrix}$ be the k^{th} sample, the first PCA

$$e_1^T \begin{bmatrix} x_{1K} - \bar{x}_1 \\ x_{2K} - \bar{x}_2 \end{bmatrix} = \begin{bmatrix} 0.5574 & -0.8303 \end{bmatrix} \times \begin{bmatrix} x_{1K} - \bar{x}_1 \\ x_{2K} - \bar{x}_2 \end{bmatrix}$$

$$= 0.5574(x_{1K} - \bar{x}_1) - 0.8303(x_{1K} - \bar{x}_2)$$

We know $\begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} = \begin{bmatrix} 4 \\ 11 \end{bmatrix}$

$$= \begin{bmatrix} 0.5574 & -0.8303 \end{bmatrix} \begin{bmatrix} x_{11} - \bar{x}_1 \\ x_{21} - \bar{x}_2 \end{bmatrix}$$

$$= 0.5574(x_{11} - \bar{x}_1) - 0.8303(x_{21} - \bar{x}_2)$$

$$= 0.5574(4 - 8) - 0.8303(11 - 8.5)$$

$$= -4.30535$$

X_1	4	8	13	7
X_2	11	4	5	14
PCA	-4.3052	3.7361	5.0928	-5.1238

25/Oct/2023

⊛ AdaBoost :-

AdaBoost works by weightening the instances in the training data set based on the accuracy of prev classification. Boosting has now become a pofular strategy for dealing binary classification issues, these algos used predetermine power by..-

Boosting algo work on the idea of first building a model on the training dataset and building a second model to correct the faults in first model. This technique is repeated untill the mistake are reduced and dataset is accurately predicted.

Types of 3 types of boosting algo:-

① Adaboost
② Gradient descent
③ Xtreme gradient descent algo;

*Understanding the working of the Adaboosting Classifier algo:-

There are 7 steps

① The image (following table) represent the Adaboost algo example

Row No	Grender	Age	income	illness	sample wt
1.	Male	41	40,000	Yes	1/5
2.	Male	54	30,000	No	1/5
3.	Female	42	25000	No	1/5
4.	Female	40	60,000	Yes	1/5
5.	male	46	50,000	Yes	1/5

$w(x_i, y_i) = \frac{1}{N}$, $i = 1, 2, \ldots, N$

Since we have 5 points

Sample weight $= \frac{1}{5} = 0.20$

Step2 :- We will examine how will "gender" classifies the samples followed by how the var (age, income) categorize the samples. We will make a decision stump for each characteristics and compute each tree's gini index. Our first stump will be the tree with lowest gini index.

Let's suppose "gender" has the lowest gini index in our dataset this will be our first stump;

Step3 :-

$$\frac{1}{2} \log \frac{1 - \text{total Error}}{\text{total Error}}$$

using this approach (step 2) we will now determine the "Amount of say" or "importance" or "influence" For this classifier in categorizing the data points

$$\text{performance stump } (\alpha) = \frac{1}{2} \log \frac{1 - \text{total Error}}{\text{total Error}}$$

$$\alpha = \frac{1}{2} \log \frac{1 - 0.2}{0.2}$$

$$\alpha = \frac{1}{2} \log \left(\frac{0.8}{0.2} \right) = \frac{1}{2} \log_e (4)$$

$$\alpha = 0.69$$

α = zero suprusent flaw less;

$\alpha = 1$ bad

Step 4 :- The weights of incorrect forecast will be increased while the weights of successful predictions will be droped. When we create our next model after updating the weights we will assign greater weights to the points with higher weights;

After determining the classifier significance and total error we must update the weights using the following formula

$$\text{New sample weight} = \text{old weight} * e^{\pm\alpha}$$

when the sample is successfully identified the amount of, say, α will be negative

Table same

New Sample weight	
0.1004	α = -
0.1004	α = -
0.1004	α = -
0.3988	α = +
0.1004	α = -

We Know that the entire sum of sample weight must be equal to 1;

in our case sum = 0.8004;

New sample weight

0.1004 / 0.8004 = 0.1254

" = 0.1254

" = 0.1254

0.3988/0.8004 = 0.4982

0.1004/0.8004 = 0.1254

+1

Step 5:- We must now create a ~~fake~~ fresh dataset to see whether or not the mistake have decreased. To do this we will delete the 'sample weights' and "New sample weights" and split our datapoints into buckets based on the new sample weights;

Buckets	New sample weight
0 – 0.1254	
0.1254 – 0.2508	deleted
0.2508 – 0.3762	
0.3762 – 0.8744	
0.8744 – 0.9998	

Step 6:- Row No. Gender Income Illness

We are nearly there, the method now choses the random values ranging from 0-1 bcoz in property categorized records have greater sample weight the likelyhood of picking them is relatedly high. Assume the five random nos chosen by our algo are (0.38, 0.26, 0.98, 0.40, 0.55)

We will match random no.s with bucket and create our new database which is written below:

Row NO	Gender	age income	illness
1	Female	40	Yes
2	male	54	NO
3	Female	42	NO
4	Female	40	Yes
5	Female	40	Yes

This is our new data set and we can see that the datapoints that was incorrectly categorized has been picked 3 times it has a greater wt;

Step 7:- This is now our new dataset and we must repeat all of the preceding stages i.e given each data point on equal equal. Determine the stump that best classifies the new group of samples by giving their gini index and ~~the~~ picking the one with lowest gini index. To update the prior sample weights,