Name – Gaurang Sonkavde

Class – M.Tech in AI and ML

Subject – Unsupervised Learning

PRN Number – 212900028

Assignment 1

Assignment 1

Q1) **Explain the Importance of Machine Learning Algorithms. In detail explain the importance of Weka Tool. Any of the ML algorithm implementation with the Tool. List Pros and Cons of the tool. (2+2+3+1=8)**

Ans -

Importance of Machine Learning Algorithm

- Machine Learning is basically the same thing as it's name suggests , what happens in Machine learning is the Algorithms which we apply will provide ability to learn on its own based on the experience it goes through. Also it can make predictions based on patterns.
- One of the most important thing in Machine learning is Dataset's ,Machine Learning is not a new term infact this term was introduced was back in 1952 by Arther Samuel , But it didn't gained that much popularity one of the reason being Cost of Computing and later the Data which used to be Generated at that time was not that much humongous as it is now.
- Also Machine learning forms to be most important aspect of Artificial Intelligence (AI) , Over here we don't need to do explicit Programing it learns from experience and improve on their own by learning through the Dataset which is given to it , The main goal is to make Machines more human like.
- Machine Learning is subset of Artificial Intelligence and Deep Learning is Subset of Machine Learning.
- Machine Learning is Classified into three Types Based on it's applications, The three type of Machine Learning Algorithms are :

1) **Supervised Learning**
2) **Unsupervised Learning**
3) **Reinforcement Learning**

1) Supervised Learning

Supervised Learning has Presence of Data and Associated Labels for the Data included

Supervised Learning is us the User or Programer training the machine with data and telling us what the data is given or in other words we can say we tell ML Algorithm what it has seen.

The foundation of Supervised Learning is **$y = f(x)$** where y = output variable and x = input variable.

Here there is a mapping done at the output as a function of Input Variable

The Supervised Learning can be Classified in :

1) Regression and 2) Classification

Regression – Prediction of Future Value from Past Data.

Eg : Cost of House , Estimating Weather Conditions , Stock Market , etc

Classification – Categorization of Items using Data

Eg : Classification of Animals (Cats , Dogs)

2) Unsupervised Learning

Goal of the Algorithm is to find a structure preset in the Data.

Nothing apart from Input Variable are Present we need to do assumptions in Unsupervised Learning.

No labels only Random Images, Data and predict what it has given.

We have to guess the output with help of input given in this Algorithm.

Unsupervised Learning can be Classified as:

1) Clustering and 2) Association

Clustering : Grouping on input variable's with simillar characteristics.

Eg : Color of Car , Model of Car , etc

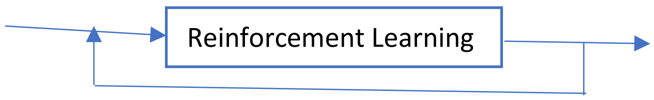
Association : Mapping Association based on Data .

Eg : E-Commerce Website (if bought a book Algorithm will recommend to buy a pen.

Video Streaming Service (If watched a Thriller Suspense Movie then Algorithm will recommend another Thriller Suspense Movie)

3) Reiforcement Learning

Reinforcement Learning is a learning which works on Feedback.



Trained based on the feedback from learning. ,Algorithm works towards the Rewards.

Machine Learning Algorithm is forced to look towards rewards and would chase rewards if it does all it is supposed to do.

Example – Dog will give handshake to get reward if not then it will miss reward.

Weka Tool

- WEKA elobrates to Waikato Enviorment for Knowledge Analysis
- It is a machine learning software which was developed at the University of Waikato, New Zealand.
- The program is written in Java.
- It has a gathered type of Collection of visualization tools and also algorithms which will be used for data analysis and predictive modeling along with Graphical User Interface(GUI).

Importance of Weka Tool

- Weka supports various Data mining tasks such as data pre-processing, clustering, classification, regressing, visualization and feature selection.
- The Data which we get at start generaly has a huge amount of noise . we need to consider noise which can affect output, for example, null values, Duplicate values, empty space, irrelevant fields like that. All this things can lead to missleading output so we should clear that noise data.
- **Preprocessor** will be used to clean the noisy data.
- If data is noisy we can’t do further steps for analysis , So first we find what are the problems occurring in our data, ABT (Analyze Base Table) in our data we can view from the excel file, and use some excel filter menus to find details of the data,
- Data preprocessing has four stages:
  1) Data Cleaning 2) Data Integration 3) Data Reduction,
4) Data Transformation.

Importance in Application Part

It provides great support to the overall process of data mining such as:
Access to databases, exploration and selection of data along with data processing:

• **Preprocess :** It has functionality aimed at importation, transformation (application of filters) and data extraction.

• **Visualize :** functionality aimed at the visualization of data using graphic techniques. 2. Predictive and descriptive modelling. Compiles a wide range of data mining procedures for the knowledge models:

• **Classify (classification and regression):** predictive modelling (supervised learning).

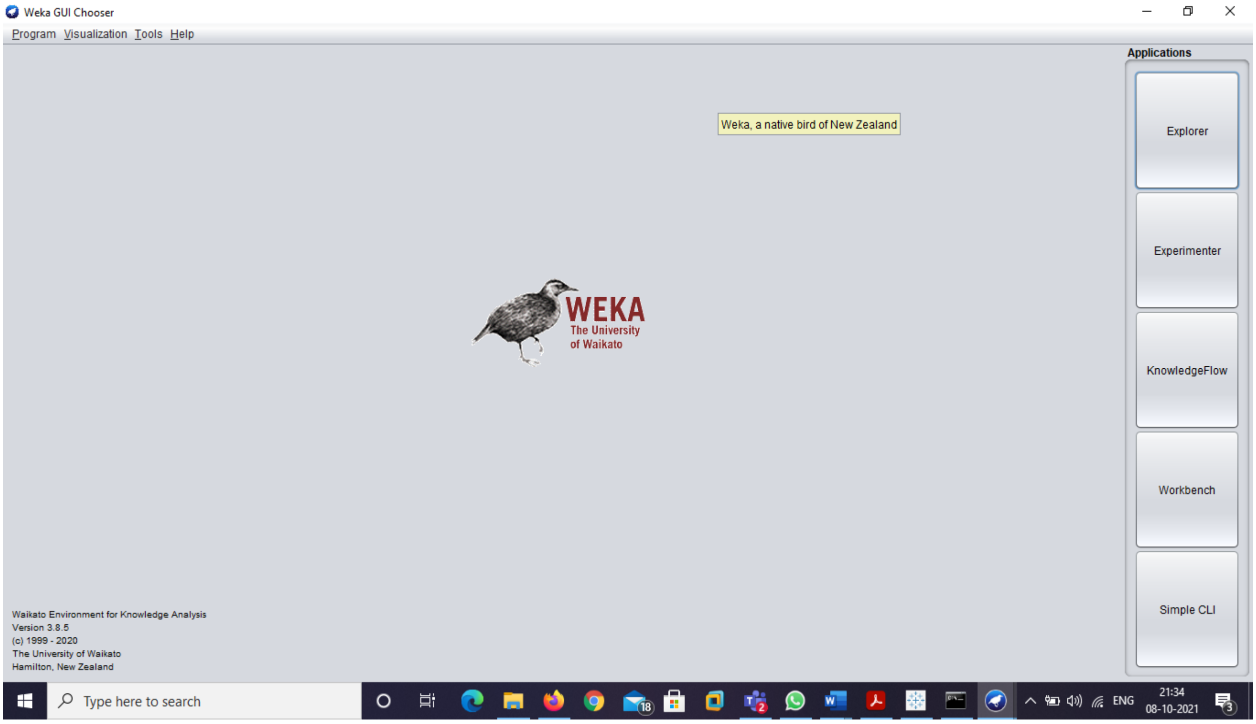
• **Cluster (grouping) and Associate (association rules**): descriptive modelling (unsupervised learning).

• **Select attributes :** selection of predictive attributes.

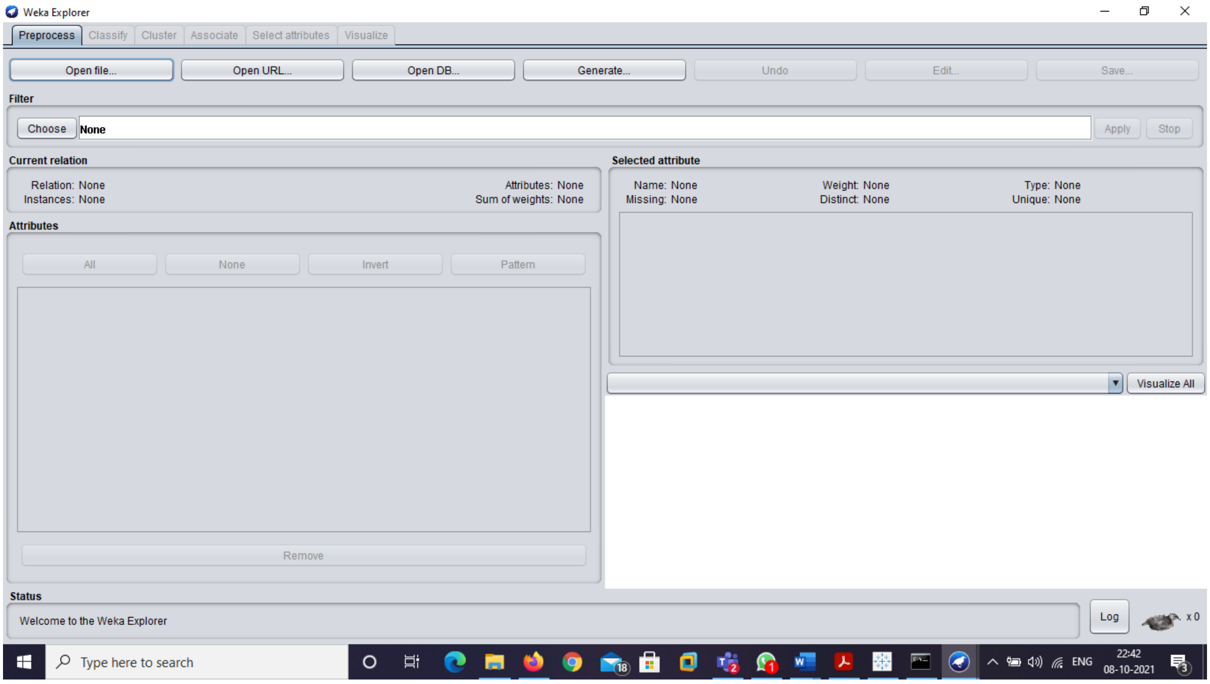
Lastly, it is worth highlighting the possibility of system extensibility : it allows the user to modify Weka by integrating new functionality developed in Java code, using its structure and object oriented functional design.

Implementation of Weka Tool with ML Algorithm (Linear Regression)

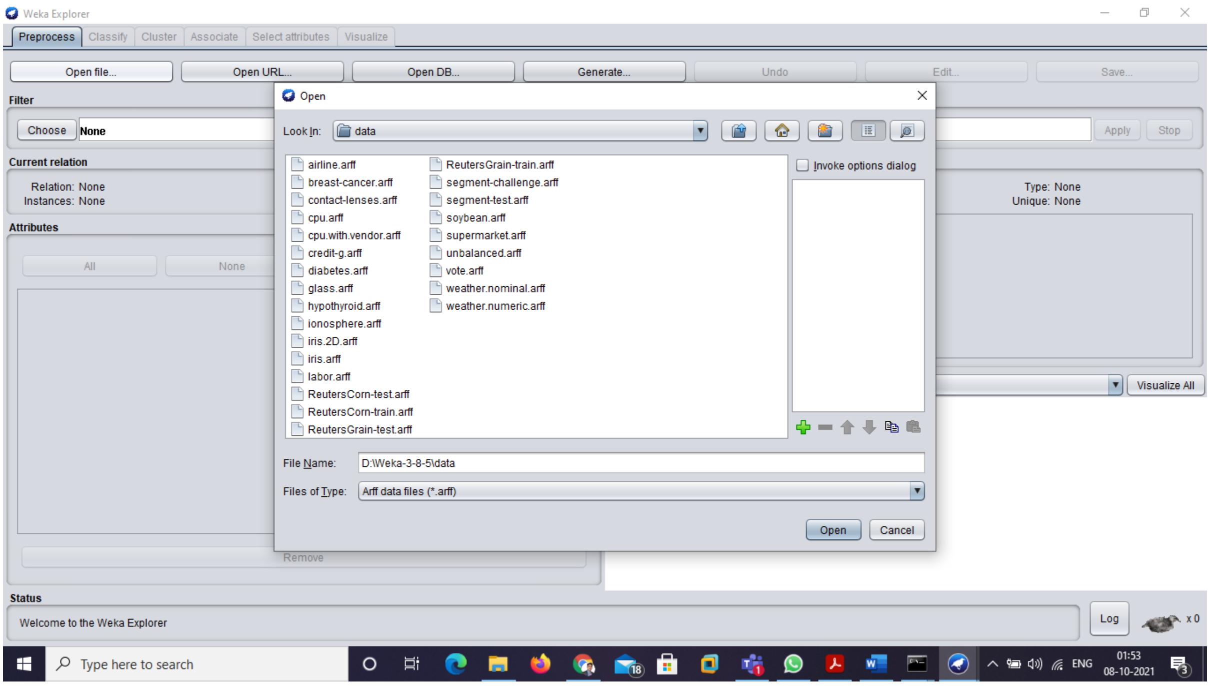
Step 1 -Open Weka Tool and choose Explorer.



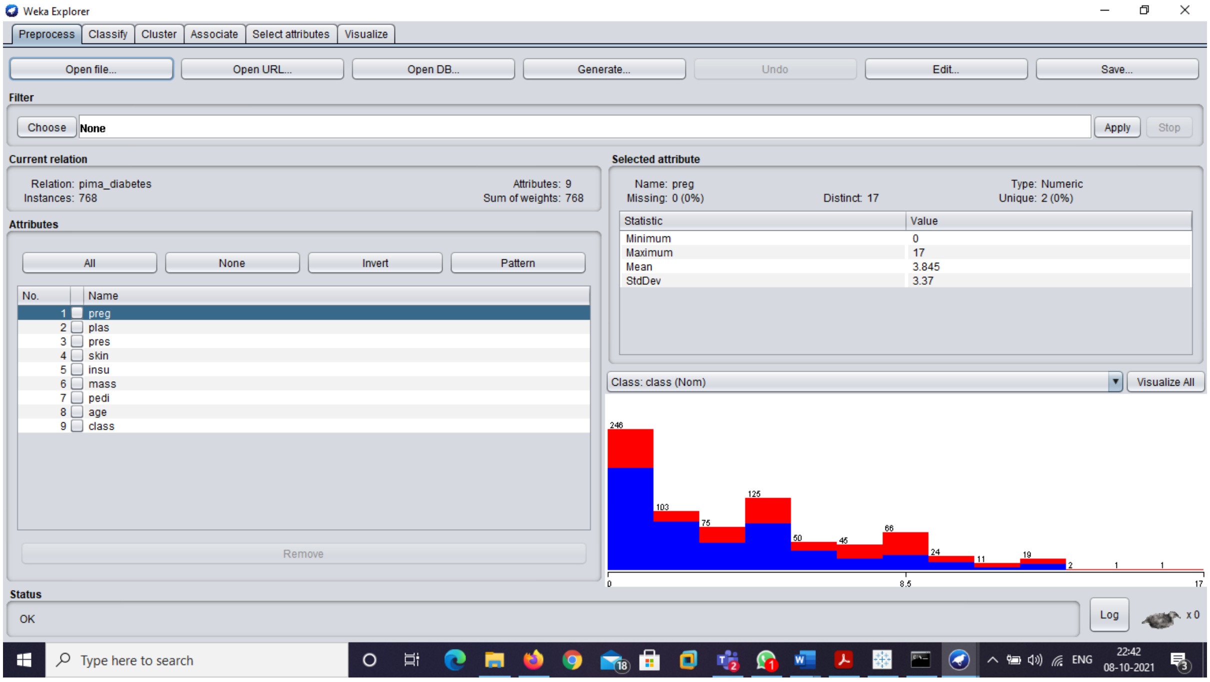
Step 2 : Click on Open File in Preprocess



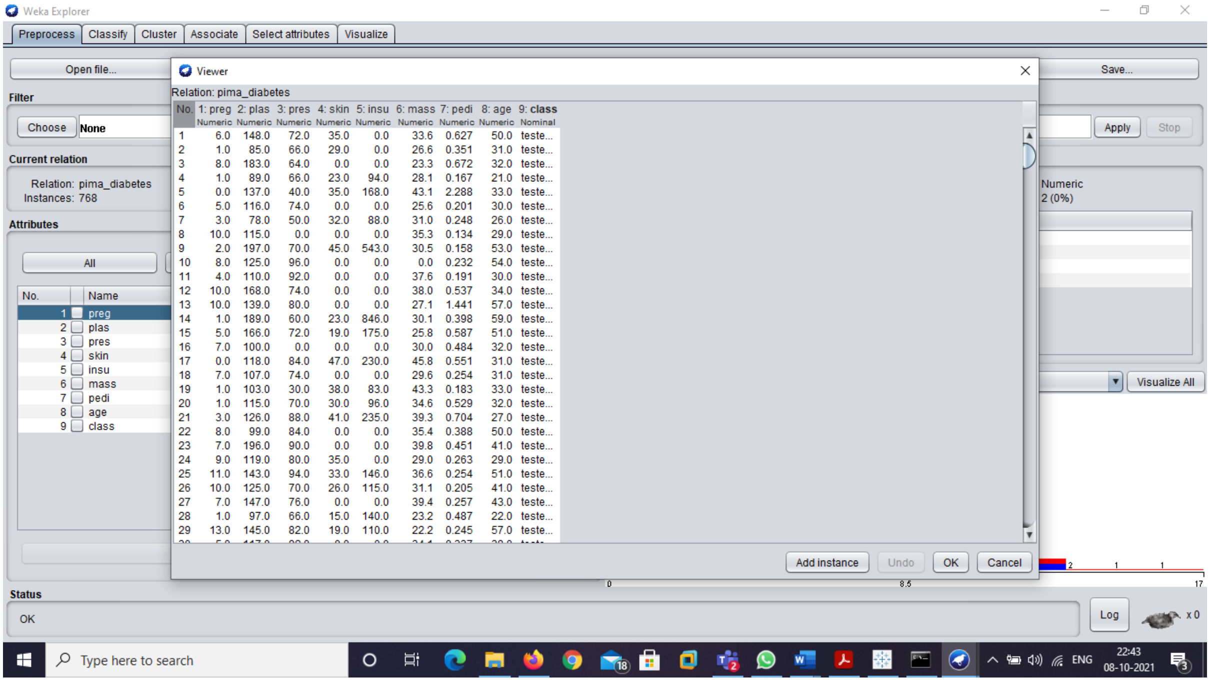
Step 3 : Choose Dataset (Here we have choose Diabetes Dataset)



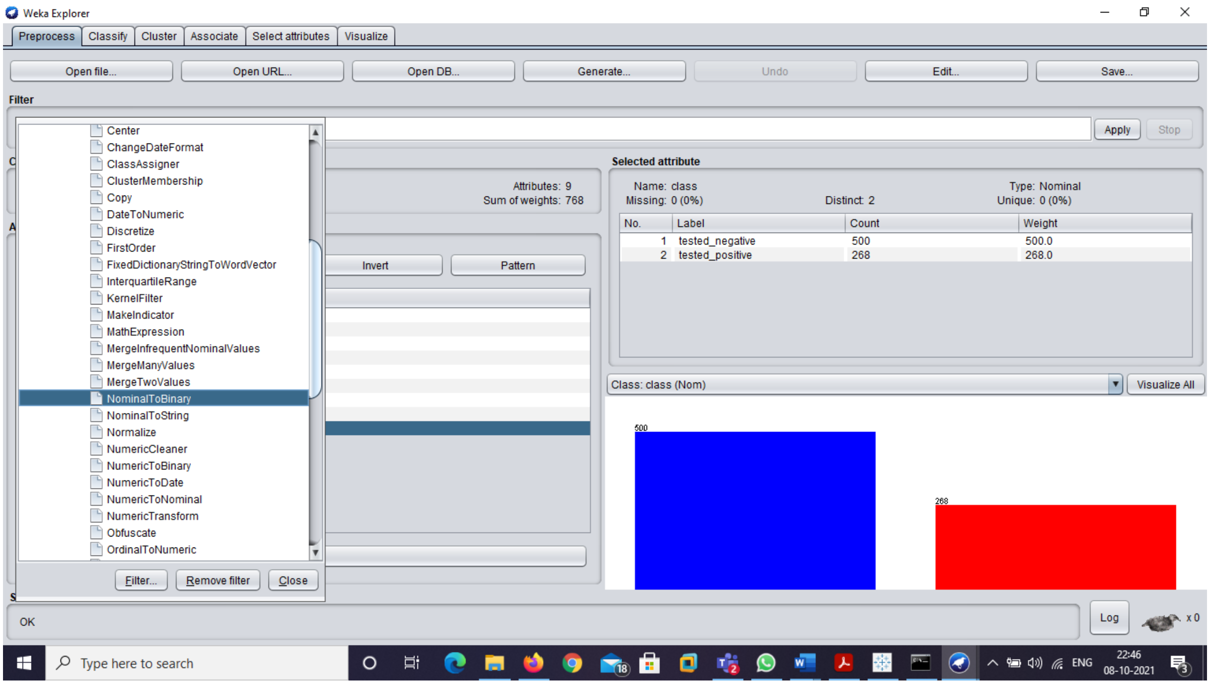
Step 4 : We have loaded Dataset but we cant use Linear Regression because it cant work on Nominal Class variables)



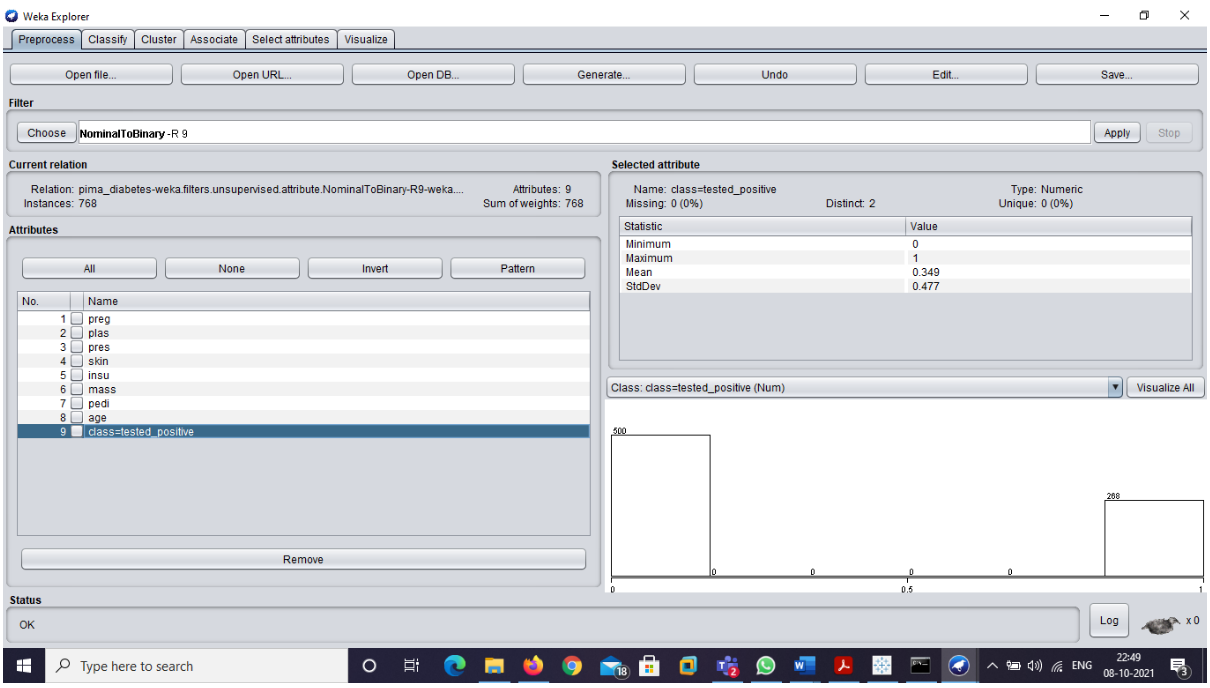
Here is Diabetes Dataset



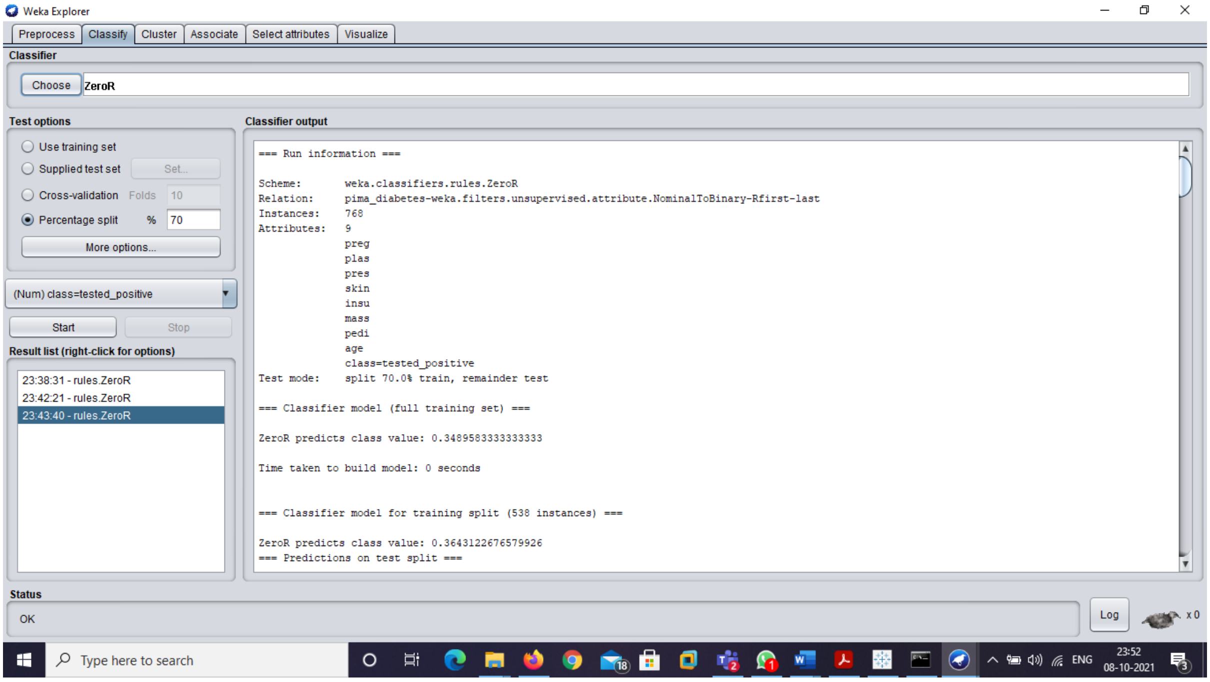
We cannot work because Class is nominal variable hence we select an Filter "Nominal to Binary to convert and after that use for Linear Regression

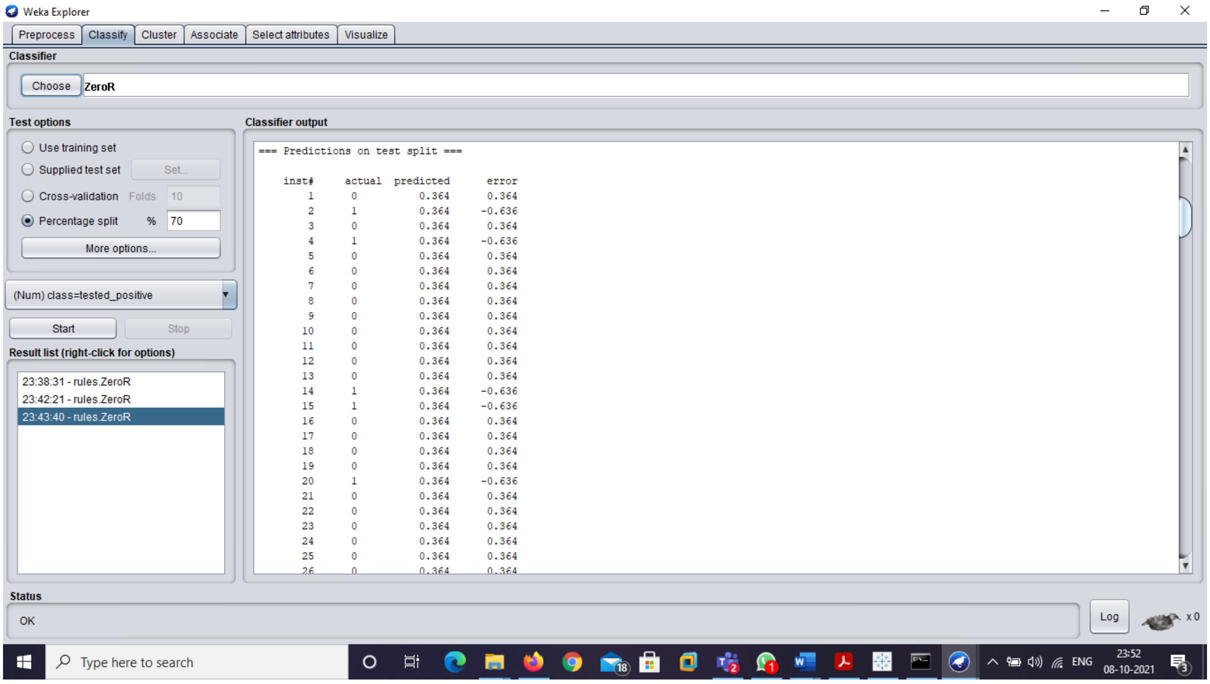


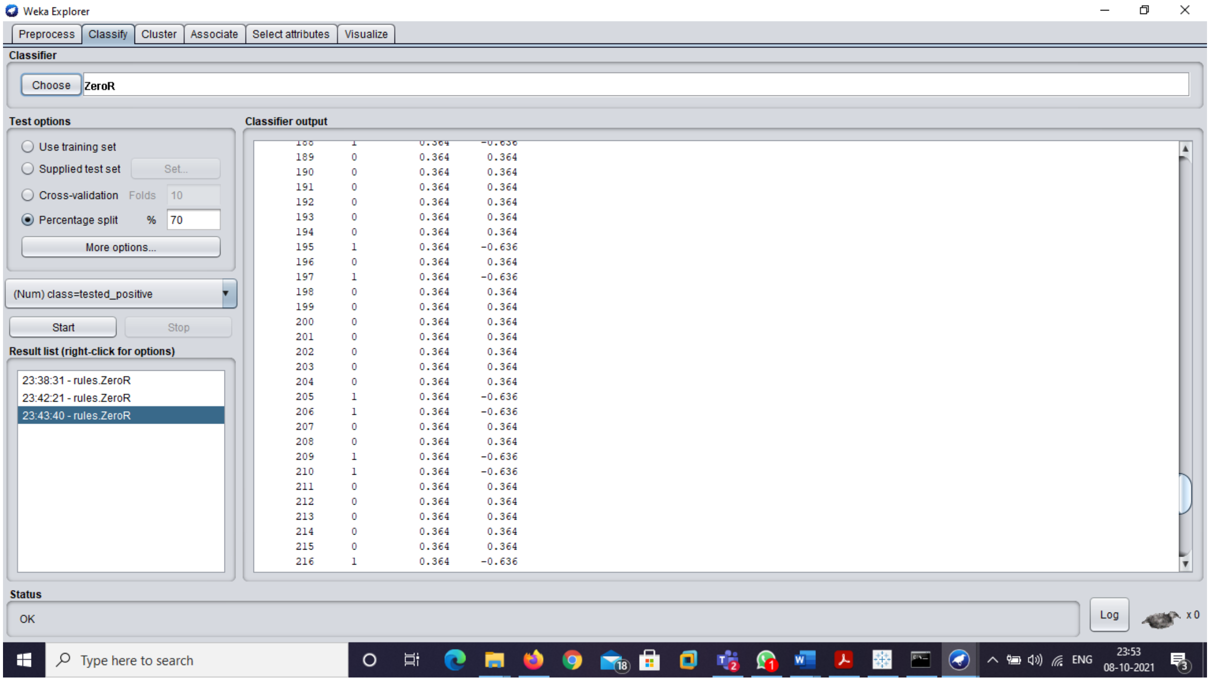
We apply Filter and here we can see that it has been converted and hence can be used for Linear Regression

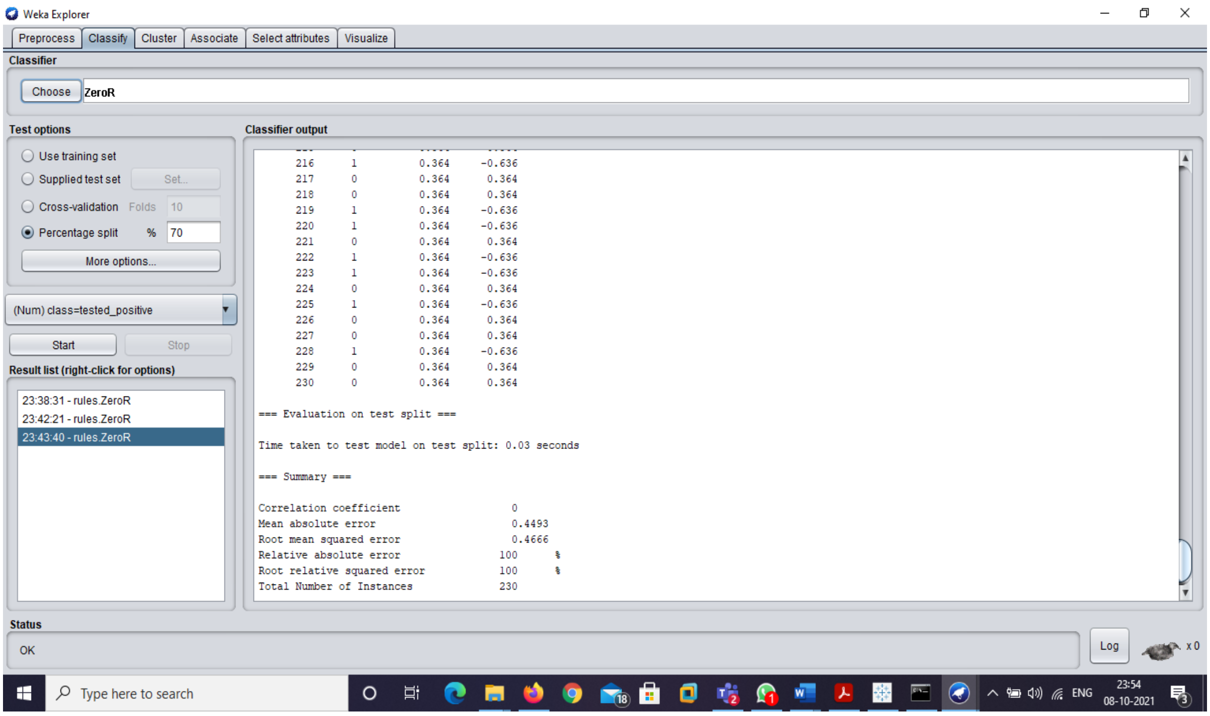


Output of Linear Regression in Classifier Window









Linear Regression

Linear Regression is a method to predict dependent variable (Y) based on value of Independent variable X

It can be used in cases where we want to predict continuous quatity

$$y = mx + c$$

Where , y = Dependant Variable x = independat variable m = slope c = intercept

Pro's and Con's of Weka Tool

Pro's :

1) Most important Advantage of Weka is it is most easy to use Data Mining Tool and can implement Machine Learning Algorithms easily.

2) It is easy for Weka Tool for Analyze Cluster and Classify Data.

3) Weka is a Tool that has whole range of data preparation, feature selection and data mining algorithms are integrated.

Con's :

1) Weka does not implement the newest Techniques

2) The documentation of Free Weka is quite limited as there is limited knowledge gain.

3) The free GUI does not implement all the possible options.

1) **List the importance of Dimensity Reduction Technique's . Apply PCA (Principal Component Analysis) on the following Dataset State the Limitations of PCA.**

Ans :

Dimensionality Reduction is the process of reducing number of variables or features in Dataset.

Features is an input to the Model , basically Feature is a Coloumn of the data which has information in it which will be used to build models

Here we will be doing Features Reduction which will decrease the numbers of Dimensions based on Accuracy of Model not getting much hampered.

This is done for reducing time to train the model ,while learning we use small amount of Dataset so it doesn't take much time to Build a model

But in Real World the Datasets which we will get will be huge in size and will take more time to train the model. This Problwm while building Models is called "Curse of Dimensionality".

So to reduce the Time for Model Building we reduce some Dimensions or Features which does not make a huge impact on Accuracy of the Algorithm

For eg : 100 Dimensions → 94% Accuracy

50 Dimensions → 90 % Accuracy

Dimensions are reduced and so is Time and this only Drops Accuracy by 4% but saves huge amount of Time so it is Feasable to reduce Dimensions.

Also Dimensity Reduction will remove multi-collinearity which will be useful for ML Model.

There are 2 Components o Dimensity Reduction

Feature Selection : We will find Subset of Original Set of Variables of Features which will help to model the problem.

This is done using three ways which are Filter , Wrapper , Embedded

Feature Extraction : This will reduce the Dimensions in the Data.

There are various methods of Dimensity Reduction such as:

1) Principal Component Analysis
2) Linear Discriminant Analysis
3) Generalized Discriminant Analysis

1) **Principal component Analysis (PCA)**

   Principal Component Analysis was termed as and developed by Karl Pearson.
   Suppose we want to reduce a dimension on Feature, then by observing we can reduce Dimension by seing that there is less loss and we don't get low variance.

Limitations of PCA

1) Here we readuce some Feature or Dimensions so that causes some amount of loss in Data.
2) PCA may find Linear Corelations which is not always good for our Model.
3) When a High variance dataset is given in structured Format, sometimes a Structured Format is found in Low Variance Also.
4) We must do Data Standardization before PCA.

The PCA Sum on dataset is given on Next Page :

Mean of X = 3

Mean of Y = 5

Q2] (cont) Apply PCA (Principal Component Analysis) on following dataset

X	Y
1	4
2	3
3	4
4	6
5	8

STEP 1 - Find out mean of X & Y

Mean = Sum of all / Total no

$$\text{Mean of } X = \frac{1+2+3+4+5}{5} = \frac{15}{5}$$

$$\boxed{\bar{X} = 3}$$

$$\text{Mean of } Y = \frac{4+3+4+6+8}{5} = \frac{25}{5}$$

$$\boxed{\bar{Y} = 5}$$

STEP 2 - Find Covariance Matrix

$$C = \begin{bmatrix} Cov(x,x) & Cov(x,y) \\ Cov(y,x) & Cov(y,y) \end{bmatrix}$$

$$\boxed{Cov(x,x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(x_i - \bar{x})}{n-1}}$$

$$\boxed{Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}}$$

$$\boxed{Cov(y,y) = \frac{(y_i - \bar{y})(y_i - \bar{y})}{n-1}}$$

$$\boxed{Cov(y,x) = \sum_{i=1}^{n} \frac{(y_i - \bar{y})(x_i - \bar{x})}{n-1}}$$

Cov(x,x)

$(x_i - \bar{x})(x_i - \bar{x}) = (x_i - \bar{x})^2$

x	y	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	4	1-3 = (-2)	4
2	3	2-3 = (-1)	1
3	4	3-3 = 0	0
4	6	4-3 = 1	1
5	8	5-3 = 2	4
$\bar{x} = 3$	$\bar{y} = 5$	$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

$Cov(x,x) = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$

$$Cov(x,x) = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$$

$$= \frac{4+1+0+1+4}{5-1} = \frac{10}{4}$$

$$\boxed{Cov(x,x) = 2.5}$$

Cov(x,y)

x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x - \bar{x})(y_i - \bar{y})$
1	4	1-3 = (-2)	4-5 = (-1)	(-2)×(-1) = 2
2	3	2-3 = (-1)	3-5 = (-2)	(-1)×(-2) = 2
3	4	3-3 = 0	4-5 = (-1)	0 × (-1) = 0
4	6	4-3 = 1	6-5 = 1	1 × 1 = 1
5	8	5-3 = 2	8-5 = 3	2 × 3 = 6
$\bar{x} = 3$	$\bar{y} = 5$			$\Sigma(x - \bar{x})(y_i - \bar{y})$

$$Cov(x,y) \; \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$= \frac{2+2+0+1+6}{5-1} = \frac{11}{4}$$

$$\boxed{Cov(x,y) = 2.75}$$

Cov(x,x) = 2.5 **Cov(x,y) = 2.75**

$$Cov(y,x) = \frac{\sum_{i=1}^{n} (y_i-\bar{y})(x_i-\bar{x})}{n-1}$$

Covariance of (x,y) & $Cov(y,x)$ will be same

$$\therefore \boxed{Cov(y,x) = 2.75}$$

$$Cov(y,y) = \frac{\sum(y_i-\bar{y})(y_i-\bar{y})}{n-1} = \frac{\sum(y_i-\bar{y})^2}{n-1}$$

x	y	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$
1	4	4-5 = (-1)	1
2	3	3-5 = (-2)	4
3	4	4-5 = (-1)	1
4	6	6-5 = 1	1
5	8	8-5 = 3	9
$\bar{x} = 3$	$\bar{y} = 5$		$\sum(y_i - \bar{y})^2$

$$Cov(y,y) = \frac{\sum(y_i-\bar{y})^2}{n-1}$$

$$= \frac{(1+4+1+1+9)^2}{5-1} = \frac{16}{4}$$

$$\boxed{Cov(y,y) = 4}$$

$$C = \begin{bmatrix} Cov(x,x) & Cov(x,y) \\ Cov(y,x) & Cov(y,y) \end{bmatrix}$$

$$\boxed{C = \begin{bmatrix} 2.5 & 2.75 \\ 2.75 & 4 \end{bmatrix}}$$

Cov(y,x) = 2.75 Cov(y,y) = 4

C = [2.5 2.75]

[2.75 4]

STEP 3 : Calculate Eigen Value 2

$$\boxed{C - \lambda I = 0}$$

where,
C = Covariance Matrix
I = Identity Matrix
λ = Eigen Value

$$\begin{bmatrix} 2.5 & 2.75 \\ 2.75 & 4 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0$$

$$\begin{bmatrix} 2.5 & 2.75 \\ 2.75 & 4 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

$$\begin{bmatrix} 2.5 - \lambda & 2.75 \\ 2.75 & 4 - \lambda \end{bmatrix} = 0$$

Make Solving Eqn

$$(2.5 - \lambda)(4 - \lambda) - (2.75)(2.75) = 0$$

$$10 - 2.5\lambda - 4\lambda + \lambda^2 - 7.5625 = 0$$

$$\cancel{\lambda^2 - 6.5\lambda + 2.43}$$

$$\boxed{\lambda^2 - 6.5\lambda + 2.4375 = 0}$$ # Eqn

$a = 1$, $b = -6.5$, $c = 2.4375$

Solve Eqn

$$\lambda_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$ # for positive

$$= \frac{+6.5 + \sqrt{(-6.5)^2 - 4(1)(2.4375)}}{2(1)}$$

$$= \frac{6.5 + \sqrt{42.25 - 9.75}}{2(1)}$$

$$= \frac{6.5 + \sqrt{32.5}}{2}$$

$$= \frac{6.5 + 5.700}{2}$$

$$\boxed{\lambda_1 = 6.1}$$

Eigen Value(**Lambda 1) = 6.1**

$$\lambda_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

$$= \frac{6.5 - \sqrt{(6.5)^2 - 4(1)(2.4375)}}{2(1)}$$

$$= \frac{6.5 - \sqrt{42.25 - 9.75}}{2}$$

$$= \frac{6.5 - \sqrt{32.5}}{2}$$

$$= \frac{6.5 - 5.700}{2}$$

$$\boxed{\lambda_2 = 0.4}$$

Step 4 : Calculate Eigen Vector

$$CV = \lambda V$$

where,

CV = Covariance Matrix

λ = Eigen Value

V = Eigen Vector (To calculate) $V = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$

$$\cancel{CV} = \begin{bmatrix} 2.5 & 2.75 \\ 2.75 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \cancel{0.399}\ 0.4 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

$$2.5\, x_1 + 2.75\, x_1 = \cancel{0.399}\ 0.4\, x_1 \quad \text{—— (1)}$$

$$2.75\, x_1 + 4\, y_1 = 0.4\, y_1 \quad \text{—— (2)}$$

Eigen Value (Lambda 2) = 0.4

for eq^n ① 12

$2.5\,x_1 + 2.75\,y_1 = 0.4\,x_1$

$2.5\,x_1 - 0.4\,x_1 = -2.75\,y_1$

$y_1 = 1$

$2.1\,x_1 = -2.75(1)$

$x_1 = \dfrac{-2.75}{2.1}$

$\boxed{x_1 = -1.31}$ $(-1.309 \approx -1.31)$

Eigen Vector for λ_2 i.e 0.4 is $\begin{bmatrix} -1.31 \\ 1 \end{bmatrix}$

Eigen Vector for $\lambda = 6.1$,

$$\begin{bmatrix} 2.5 & 2.75 \\ 2.79 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 6.100 \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$$

$2.5\,x_2 + 2.75\,y_2 = 6.100 \begin{Bmatrix} x_1 \\ y_1 \end{Bmatrix}$ — ①

$2.75\,x_2 + 4\,y_2 = 6.1\,y_2$ — ②

for eq^n ①

$2.5\,x_2 + 2.5\,y_2 = 6.1\,x_2$

put $y_2 = 1$

$2.5\,x_2 + 2.5(1) = 6.1\,x_2$

Eigen vector (Lambda1) = 0.4 = [-1.31 , 4]

$2.75 - 6.1x = -2.5x$

$2.75 = 3.6x$

$x = \frac{2.75}{3.6} = 0.76$

Eigen Vector for λ_2 ie 6.1 is $\begin{bmatrix} 0.76 \\ 1 \end{bmatrix}$

Eigen Vector of λ_1 & λ_2 are

$\cancel{\begin{bmatrix} -1.31 \\ 1 \end{bmatrix}}$ & $\begin{bmatrix} 0.76 \\ 1 \end{bmatrix}$, $\begin{bmatrix} -1.31 \\ 1 \end{bmatrix}$

STEP 5 : New Eigen Vector

$\cancel{\lambda x} = \begin{bmatrix} 0.76 \\ 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$

$= \sqrt{(0.76)^2 + (1)^2}$

$= \sqrt{1.5776} = 1.2560$

New Eigen Vector

$x = \frac{0.76}{1.2560} = 0.6050$

$y = \frac{1}{1.2560} = 0.7961$

$\boxed{\lambda_1 = \begin{bmatrix} 0.6050 \\ 0.7961 \end{bmatrix}}$

Eigen Vector = lambda 1 = 6.1 = [0.76 ,1]

New Eigen Vector for Lambda 1 = [0.6050 , 0,7961]

For λ_2

$$\begin{bmatrix} -1.31 \\ 1 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

$$\sqrt{(-1.31)^2 + (1)^2} = \sqrt{2.7161} = 1.6481$$

New Eigen Vector

$$\frac{-1.31}{1.6481} = -0.7948$$

$$\frac{1}{1.6481} = 0.6067$$

$$\boxed{\lambda_2 = \begin{bmatrix} -0.7948 \\ 0.6067 \end{bmatrix}}$$

New Eigen Vector's are

$$\lambda_1 = \begin{bmatrix} 0.6050 \\ 0.7961 \end{bmatrix}, \quad \lambda_2 = \begin{bmatrix} -0.7948 \\ 0.6067 \end{bmatrix}$$

New Eigen vector for Lambda 2 = [-0.7948 , 0.6067]

New Eigen Vectors are

Lambda 1 = [0.6050] Lambda 2 = [-0.7948]

[0.7961] [0.6067]

****