# Lab-1

Date :
Page No.: 1

Write python code, consider filename as "housing.csv"

i) To load .csv file into the data frame

```
import pandas as pd
df = pd.read_csv("housing.csv")
```

ii) To display information of all columns

```
print(df)
```

iii) To display statistical information of all numerical

```
print(df.describe())
```

iv) To display the count of unique labels for "Ocean Proximity" column

```
print(df["Ocean Proximity"].value_counts())
```

v) To display which attributes in dataset have missing values count greater than zero

```
missing_values = df.isnull().sum()
missing_columns = missing_values[missing_values>0]
print(missing_columns)
```

Date :
Page No.:

Fa both datasets Pabetes and adult income

1) which column in dataset had missing values How did you handle them.

Dropped the missing values with the help of dropna since the number of rows with missing values were less comparitive

2) what were the categorical column and how did you encode it?

diabetes.csv : ['Gender', 'CLASS']
adult.csv : ['workclass', 'education', 'marital-status', 'occupation', 'relationship', 'race', 'gender', 'native-country', 'income']

and we used LabelEncoder() to encode it

3) what is difference Min max scaling and standardization? when we are over the other

Both Min-Max scaling and standardization are feature scaling techniques

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad X' = \frac{X - \mu}{\sigma}$$

we use min max when we do not have significant outliers
and we use standardization when we have outliers

PS
4/3/25

# Lab-2

Date :
Page No.: 3

Demostrate steps to build ml model that predicts median house price

1. perform describe and info
describe the dataset

2. Plot histogram of each feature
median_income is right skewed
median_house_value is right skewed

3. Write difference between testset random and stratified.
we take random record from the df
in stratified we target a particular feature

4. what does housing prices wrt to location show
closer to the coast has more prices
inland areas have lower prices
this is due high population

5. Find most correlated columns
Ocean proximity is correlated to numerical values using .get_ columns then found correlation

6. list features that could be combined to improve correlation
rooms per household, bedrooms_per_rooms and population_per household.

# Lab - 3

Date :
Page No. 4

① 

| $x_i$ | $y_i$ |
|---|---|
| 1 | 2 |
| 2 | 4 |
| 3 | 5 |
| 4 | 9 |

$$\text{slope} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$\bar{y} = 5$

$\bar{x} = 2.5$

$$\text{slope} = \frac{4.5 + 0.5 + 0 + 6}{2.25 + 0.25 + 0.25 + 2.25} = \frac{11}{5} = 2.2$$

$b = \bar{y} - m\bar{x}$

$= 5 - (2.2 \times 2.5)$

$= 5 - 5.5 = -0.5$

$y = 2.2x - 0.5$

②

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\beta = ((X^T X)^{-1} X^T) y$$

$$= \begin{bmatrix} 1.5 & -0.5 \\ -0.5 & 0.2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0.5 & 0 & -0.5 \\ -0.3 & -0.1 & 0.1 & 0.3 \end{bmatrix}$$

$$= \begin{bmatrix} -0.5 \\ 2.2 \end{bmatrix}$$

$y = -0.5 + 2.2x$

Date : 17/3
Page No.:

1 considering canada_per_capita_income.csv salary.csv did you perform data preprocessing

Yes missing values were calculated by using simple imputer. and label encoder

2 did you visulize regression line for the csv file

Yes, the regression line was plotted. The plot shows a strong linear relationship between year and per capita income meaning that as the year increases.

3 The predicted salary is printed in the script depends on the trained model's coefficient.

4 Did you encode categorical variables for 1000_companies.csv

Ye

Yes, it was done. Column was encoding using Label Encoder()

# Lab-4,5

Date: 24/3/25
Page No.: 6

1. Binary classification with logistic regressions.

a) $P(y=1|x) = \frac{1}{1+e^{-(a_0+a_1x)}}$

b) $P(y=1|x=7) = \frac{1}{1+e^{(-0.5+0.8(7))}}$

$= \frac{1}{1+e^{-0.6}}$

$= \frac{1}{1+0.5488} = 0.645$

c) predicted class : It is pass since it is above 0.5

2. applying softmax

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

$P_1 = \frac{e^2}{11.107} = 0.665$

$P_2 = \frac{e^1}{11.107} = 0.245$

$P_3 = \frac{e^0}{11.107} = 0.090$

Date : 24/3/25
Page No. :

1 For iris..csv data set.

The accuracy was 98% indicating it classifies efficiently

The confusion matrix shows no miscalculation

2. for petrol-consumption.csv dataset

regression tree predicts based on features

It predicts for continious values

MAE = 90.60
MSE = 16851.60
RMSE = 129.81

→ NO = 4 time
yes = 1 time.

(Ex) $H(S) = -\left(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) = 0.72$

$H(Hot) = -\left(\frac{3}{4}(-0.42) + \frac{1}{4}\times(-200)\right)$

$H(Cool) = 0$ = 0.81

it branches at as.

Date: 7/4/25
Page No.: 8

# Lab-6

| Person | Age | Salary | Target | Distance | |
|---|---|---|---|---|---|
| A | 18 | 50 | N | 52.79 | |
| B | 23 | 55 | N | 46.86 | |
| C | 24 | 70 | N | 31.95 | (2) |
| D | 41 | 60 | Y | 40.45 | (3) |
| E | 43 | 70 | Y | 31.04 | (1) |
| F | 38 | 40 | Y | 60.08 | |

[Y, N, Y] so final prediction is Y

(1) For Iris dataset : How to choose the value in KNN

We try different values of K and then based on the accuracy and the value of K which has the best accuracy is chosen. this helps avoiding overfitting and underfitting. best K is the one which has accuracy rate is highest and accuracy error rate is lowest

(2) For Diabetes dataset : what is the purpose of feature scaling & How to perform?

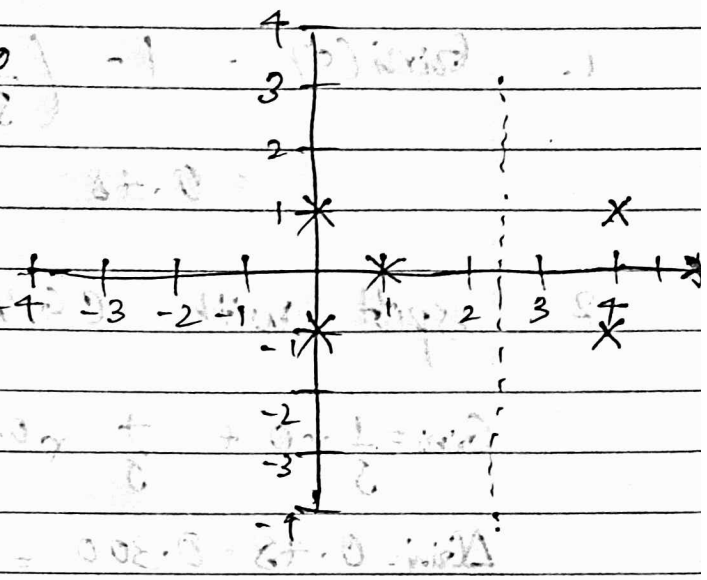
In the diabetes dataset, the feature are measured in different units and scales. Some feature have values in hundend other decimals. If we do not scale the largest data entity will dominate. Hence scaling helps all features contribute equally It is done by standardization and normalization

Date : 21/4/20
Page No.: 9

# Lab - 7

positive class lies to the right $x = 3$

negative class point are left of $x = 2$




hyperplane : 2.5

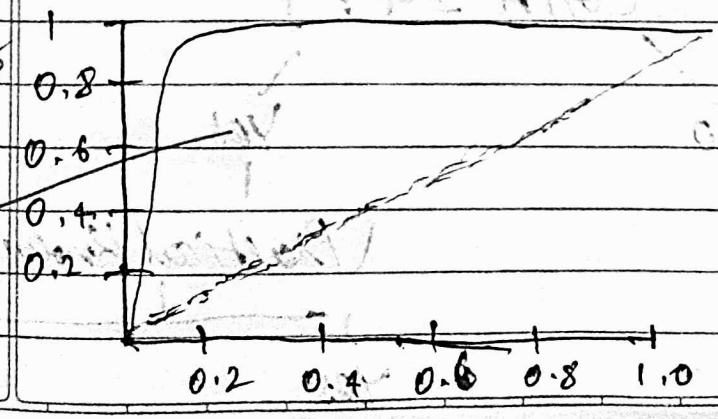
1. For Iris dataset

Linear Kernel
Accuracy : 1.0
confusion matrix
[[10 0 0
0 9 0
0 0 11]]

RBF
Accuracy : 1.0
Confusion Matrix :
[[10 0 0
0 9 0
0 0 11]]

2. For letter dataset

accuracy : 0.95025
AUC : 1.00



Date : 6/5/25
Page No. : 10

# Lab - 8

1. $Gini(S) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 1 - 0.36 - 0.16$

$= 0.48$

2. split with CGPA

Y –
N –

$Gini = \frac{1}{5} \times 0 + \frac{4}{5} \times 0.375 = 0.300$

$1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2$

$\Delta Gini = 0.48 - 0.300 = 0.18$

3. growing CGPA ≥ 9 node based on interaction

→ $Gini = \frac{2}{4} \cdot 0 + \frac{2}{4} 05 = 0.25$

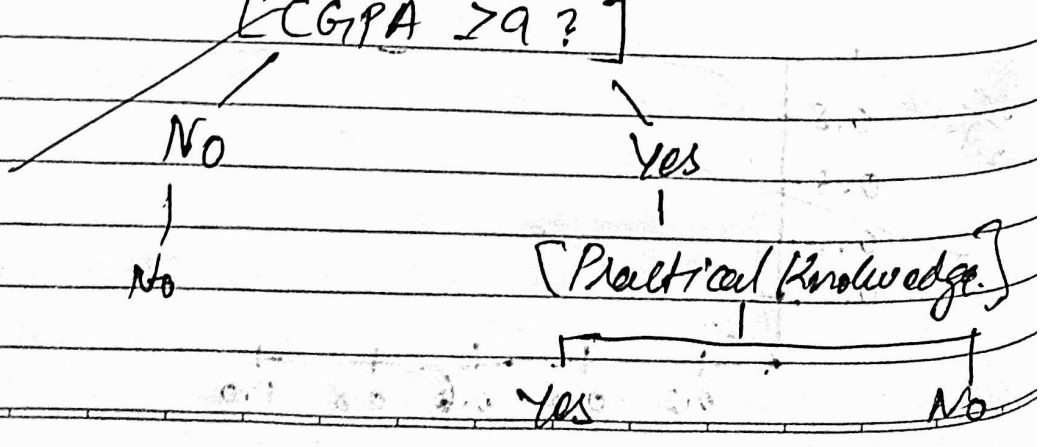
$\Delta Gini = 0.375 - 0.25 = 0.125$

→ based on practical knowledge

$Gini = \frac{3}{4} \cdot 0 + \frac{1}{4} \cdot 0 = 0$

$\Delta Gini = 0.375 - 0 = 0.375$ (best)

final cast



1. Gini = $1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2$ = $1 - 0.64 - 0.04$

= 0.32

2. Gini = $\frac{2}{5} \times 0 + \frac{3}{5} \times 0.444 = 0.2664$

$\Delta$Gini = 0.32 - 0.2664 = 0.0536

3 growing the "no" branch with other features

→ CGPA :

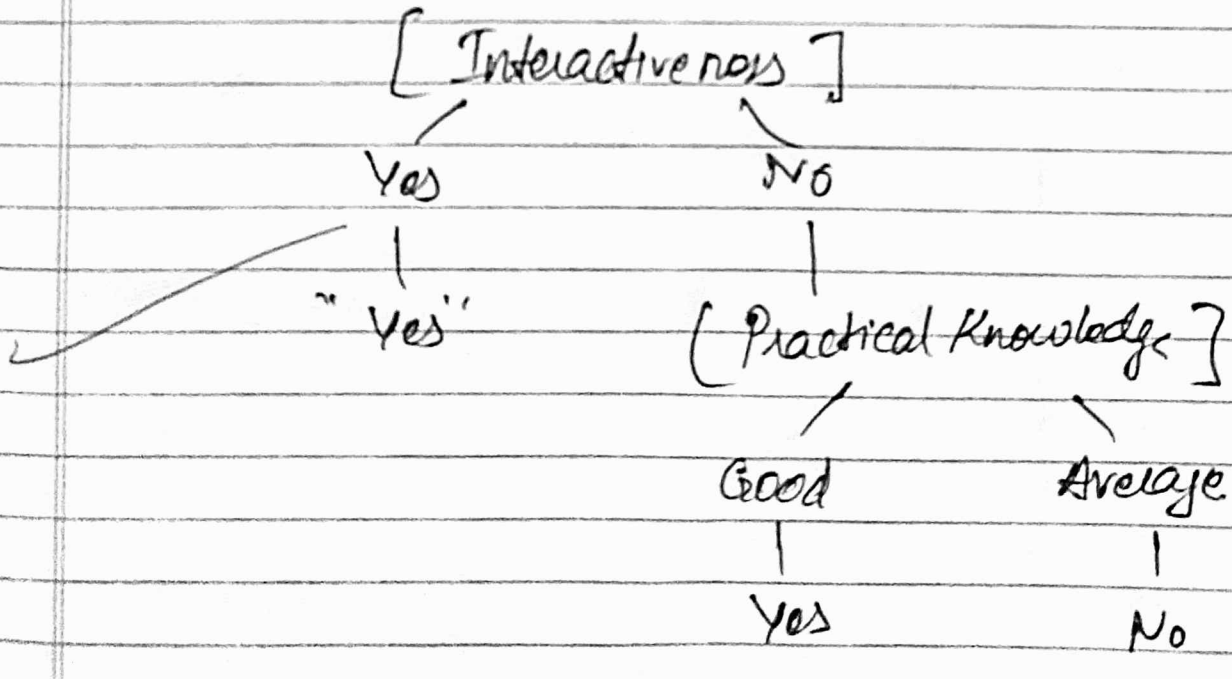
Weighted Gini : $\frac{1}{3} \times 0 + \frac{2}{3} \times 0.50$ . 0333

$\Delta$Gini = 0.111

→ Practical Knowledge :

weighted gini : 0

$\Delta$Gini : 0.444 (best)



| RF | n_estimators | mean_accuracy |
|---|---|---|
| 1 | 10 | 0.9667 |
| 2 | 50 | 0.9667 |
| 3 | 100 | 0.9667 |
| 4 | 150 | 0.9667 |
| 5 | 200 | 0.9667 |

The best can be chosen with n_estimators

Date: 5/5/25
Page No.: 13

# Lab - 9

1. →

| CGPA | Actual Job | |
|---|---|---|
| ≥9 | +1 | 1/6 |
| <9 | +1 | 1/6 |
| >9 | −1 | 1/6 |
| <9 | −1 | 1/6 |
| ≥9 | +1 | 1/6 |
| ≥9 | +1 | 1/6 |

$$\varepsilon_{CGPA} = 2 \times \frac{1}{6} = 0.333$$

$$\alpha_{CGPA} = 0.347 = \frac{1}{2} \ln \left[\frac{1-0.333}{0.333}\right]$$

$$w_{CGPA} = \frac{1}{6} \times 4 \times e^{-0.347} + \frac{1}{6} \times 2 \times e^{0.347}$$

$$= 0.9428$$

$$wt_{+ve} = \frac{1/6 \times e^{-0.347}}{0.9428} \qquad wt_{-ve} = \frac{1/6 \times e^{+0.347}}{0.9428}$$

$$= 0.1249 \qquad = 0.2501$$

→

| Interactiveness | Actual offa | |
|---|---|---|
| Yes | Yes | 0.1249 |
| No | Yes | 0.2501 |
| No | No | 0.2501 |
| No | No | 0.1249 |
| Yes | Yes | 0.1249 |
| Yes | Yes | 0.1249 |

$\varepsilon_{\text{Interactiveness}} = 0.2501$

$\lambda = \frac{1}{2} \ln\left[\frac{1-0.2501}{0.2501}\right] = 0.5490$

$Z_{\text{Interactiveness}} = 0.1249 * 4 \times e^{-0.549} + 0.2501 \times 1 \times e^{-0.549} + 0.2501 \times 1 \times e^{0.549}$

$\approx 0.866$

$wt_{+ve} = \frac{0.1249 * e^{-0.549}}{0.866} = 0.0832$

$wt_{+ve} = \frac{0.2501 \times e^{-0.549}}{0.866} = 0.1664$

$wt_{-ve} = \frac{0.2501 \times e^{0.549}}{0.866} = 0.5001$

→ No instance is misclassified by Platical Knowledge.

→

| Com Skill | Actual Jobopar | weight |
|---|---|---|
| Good | Yes | 0.0832 |
| moderate | Yes | 0.5001 |
| moderate | No | 0.1662 |
| Good | No | 0.0832 |
| Moderate | Yes | 0.0832 |
| Moderate | Yes | 0.0832 |

$\varepsilon_{cs} = 0.7497$

$$\alpha = \frac{1}{2} \ln\left[\frac{1 - 0.7497}{0.7497}\right]$$

$= -0.5485$

$$Z_{cs} = 0.0832 \times 1 \times e^{0.7485} + 0.1667 \times 1 \times e^{0.5485} + 0.5001 \times 1 \times e^{-0.5485} + 0.0832 \times 3 \times e^{-0.5485}$$

$= 0.866$

| CGPA | Interaction | Communication | Avg | Pred |
|---|---|---|---|---|
| Yes | Yes | Yes | 0.3485 | Y |
| No | No | No | 0 | N |
| Yes | No | No | 0.317 | Y |
| No | No | Yes | -0.5485 | N |
| Yes | Yes | No | 0.896 | Y |
| Yes | Yes | No | 0.896 | Y |

Ans 5/6

2.

| Adaboost | n_estimators | mean_accuracy |
|---|---|---|
| 1 | 10 | 0.820377 |
| 2 | 50 | 0.830023 |
| 3 | 100 | 0.832050 |
| 4 | 150 | 0.832296 |
| 5 | 200 | 0.832624 |

Best

The best accuracy is from 200 estimators

Date: 12/5/25
Page No.: 16

# Lab-10

| Record | c-1 dist | c2 dist | assignment |
|---|---|---|---|
| 1,1 | 0 | 7.21 | C1 |
| 1.5, 2 | 1.12 | 6.12 | C1 |
| 3, 4 | 3.61 | 3.61 | C1 |
| 5, 7 | 7.21 | 0 | C2 |
| 3.5, 5 | 4.12 | 2.5 | C2 |
| 4.5, 5 | 5.31 | 2.06 | C2 |
| 3.5, 4.5 | 4.30 | 2.92 | C2 |

$C_1 = [1.83, 2.33]$
$C_2 = [7.12, 5.37]$

| Record | C1 | C2 | assignment |
|---|---|---|---|
| 1,1 | 1.57 | 5.37 | C1 |
| 1.5, 2 | 0.47 | 4.27 | C1 |
| 3, 4 | 2.04 | 1.77 | C2 |
| 5, 7 | 5.64 | 1.85 | C2 |
| 3.5, 5 | 3.15 | 0.72 | C2 |
| 4.5, 5 | 3.78 | 0.53 | C2 |
| 3.5, 4.4 | 2.74 | 1.07 | C2 |

C1 = [1.25, 1.5]
C2 = [3.9, 5.1]

The elbow plot shows a sharp elbow at K=3, indicating that three clusters is the optimal choice for the petal length of iris dataset

Date: 12/5/25
Page No.: 17

# Lab-11

| | | | | | |
|---|---|---|---|---|---|
| $x_1$ | 4 | 8 | 13 | 7 | $\bar{x}_1 = 8$ |
| $x_2$ | 11 | 4 | 5 | 14 | $\bar{x}_2 = 8.5$ |

$$B = \begin{bmatrix} -4 & 0 & 5 & -1 \\ 2.5 & -4.5 & -3.5 & 5.5 \end{bmatrix}$$

$$S = \frac{1}{N-1}[B B^T]$$

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

$$\lambda^2 - \text{Trace}(S)\lambda + \det(S) = 0$$
$$\lambda^2 - 37\lambda + 201 = 0$$

$$\lambda = 30.3849,\ 6.6151$$

$$\begin{bmatrix} -16.3849 & -11 \\ -11 & -7.3849 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} -11/-16.3849 \\ 1 \end{bmatrix}$$

$$\hat{u}^T = \hat{u}^T = \begin{bmatrix} -0.5574 & 0.8303 \end{bmatrix}$$

$$\begin{bmatrix} -0.5574 & 0.8303 \end{bmatrix} \begin{bmatrix} -4 & 0 & 5 & -1 \\ 2.5 & -4.5 & -3.5 & 5.5 \end{bmatrix}$$

$$= \begin{bmatrix} 9.305 & -3.736 & -5.693 & +5.124 \end{bmatrix}$$

Date :
Page No.:

for heart.csv dataset

Logistic regression: 0.9014
SVM: 0.8526
Random forest = 0.8525

After PCA

Logistic regression : 0.8689
SVM: 0.8689
Random forest : 0.8852

This shows that with a slight tradeoff with accuracy and lesser computation the model yields similar result.