
Assignment on Practical Tutorial

course: Data Mining (Statistical Data Analysis)
VIII

code: Stat Lab - 412

Name: Sourav Basak

Roll: 132

Date: 3.12.2025

Q1) Data given,

Year	month	Day	TEM	MXT	HUM	SLP	MNT
1980	1	1	12.5	26.0	12.2	1013.5	69
1980	1	2	14.4	21.2	9.5	1014.0	9.5
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
2022	3	31	24.6	34.6	25.6	77	100.6

SSH	CLD
7.9	0
0.8	6
:	:
7.1	2

For missing value treatment:

`data.isnull().sum()`

Output:

Year	0	WIS	0
Month	0	RIN	2
Day	0	SSH	0
TEM	1	CLD	0
MXT	55		
MNT	71		
HUM	0		
SLP	2		

From result, we can see there is 1, 55, 71, 2, 2 missing value present in columns called MXT, MNT, SLP, RIN respectively.

For missing value imputation we used `mean` & `median`.

Code:

```
import numpy as np
x = np.mean(data["TEM"])
y = np.mean(data["MXT"])
z = np.mean(data["MNT"])
s = np.mean(data["SLP"])
q = np.mean(data["RIN"])
data["TEM"].fillna(x, inplace = TRUE)
data["MXT"].fillna(y, inplace = TRUE)
data["MNT"].fillna(z, inplace = TRUE)
data["SLP"].fillna(s, inplace = TRUE)
data["RIN"].fillna(q, inplace = TRUE)
data.isnull().sum
```

Output:

		MXT	0	RIN	0
Year	0	MNT	0	SSH	0
Month	0	HUM	0	CLD	0
Day	0	SLP	0		
TEM	0	WIS	0		

For exploratory analysis:

```
import seaborn as sns
```

```
import matplotlib.pyplot as plt
```

```
sns.lineplot(x=data["Year"], y=data["TEM"])
```

```
plt.show()
```

```
sns.lineplot(x=data["Year"], y=data["SLP"])
```

```
plt.show()
```

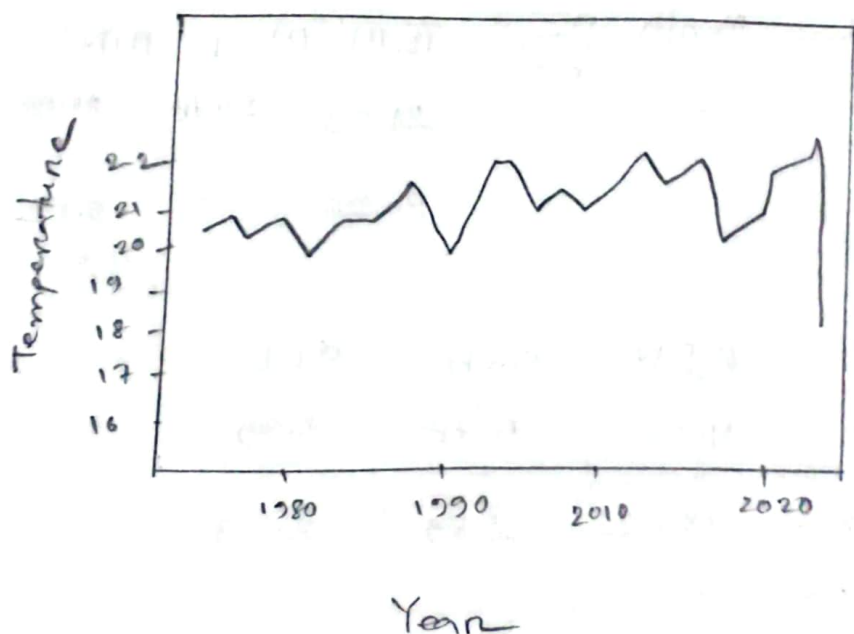


Fig: Year vs. Temperature

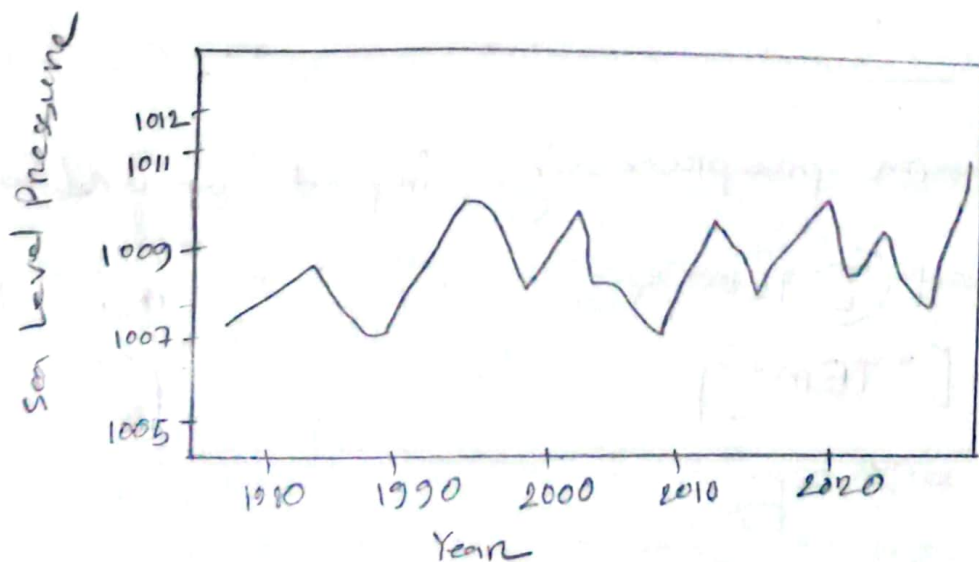


Fig: Year vs. Sea Level Pressure

As from both figure we can see Temperature upward trend but SLP remain more to stationary.

For summary statistics:

`data.describe()`

Output:

	Year	month	Day	TEM	MXT	MNT	HUM	SLP
mean	-	-	-	21.03	31.78	20.99	78.71	1007
std	-	-	-	3.96	3.9	5.82	8.10	6.3

	WIS	RIN	SSH	CLD
mean	5.47	4.59	5.78	3.99
std	3.52	13.52	2.89	2.53

For scaling:

from sklearn.preprocessing import StandardScaler

`X = data.drop(["TEM", "Year", "Month", "Day"], axis=1)`

`Y = data["TEM"]`

`X_scaled = X.copy`

for col in X.columns:

$$X_scaled[col] = 0.1 + 0.8 * ((X[col] - X[col].min()) / ((X[col].max() - X[col].min())))$$

output:

X_scaled

MX T	MNT	HUM	SLP	WIS	RIN	SSH	CLD
0.43	0.32	0.49	0.85	0.166	0.10	0.63	0.1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.667	0.71	0.60	0.81	0.266	0.1	0.58	0.3

For fitting neural network:

```
import pandas as pd
from sklearn.neural_network import MLPRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error

model = MLPRegressor(hidden_layer_sizes = (64, 64),
                      activation = "relu", solver = "adam",
                      learning_rate = "adaptive", max_iter = 2000,
                      random_state = 132132)
```

```
X_train, X_test, Y_train, Y_test = train_test_split(
    X_scaled, Y, test_size = 0.2,
    random_state = 210446)
```

```
model.fit(X_train, Y_train)
```

```
Y_pred = model.predict(X_test)
```

```
Y_train_pred = model.predict(X_train)
```

```
Y_test_pred = model.predict(X_test)
```

```
R2_train = r2_score(Y_train, Y_train_pred)
```

```
R2_train = mean_squared_error(Y_train, Y_train_pred)
```

```
R2_test = r2_score(Y_test, Y_test_pred)
```

```
mse_test = mean_squared_error(Y_test, Y_test_pred)
```

```
print("R2_train")
```

```
print("R2_test")
```

Output:

Training performance

$$R^2 = 0.9748$$

Testing Performance

$$R^2 = 0.97$$

Interpretation: Test Training acc = 0.97 means

97% predicted true positive as it is true.

For Scatterplot of maximum rainfall :

```
X = data.drop(["RIN", "Year", "Month", "Day"])
```

```
Y = data["RIN"]
```

```
X_Scale = X_scaled.apply(X_scaled)
```

```
Model = MLP.fit(X_train, Y_train)
```

```
Y_train_pred = Model.predict(X_train)
```

```
Y_test_pred = Model.predict(X_test)
```

```
def scatter_plot(actual, predicted, title):
```

```
    plt.figure(figsize=(6,6))
```

```
    plt.scatter(actual, predicted, alpha=0.6, edgecolor="k",
```

```
    plt.show())
```

Figure:

Actual vs. Predicted (Training data)

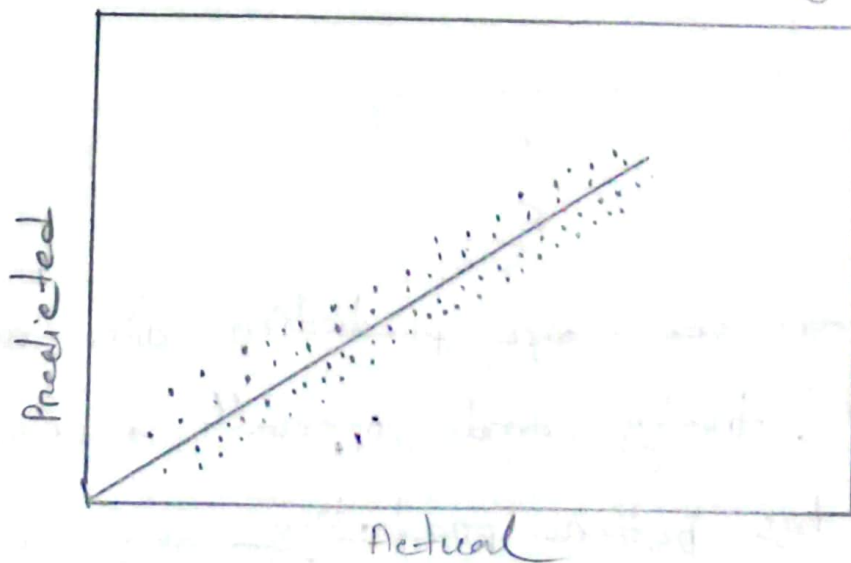


Fig: 1

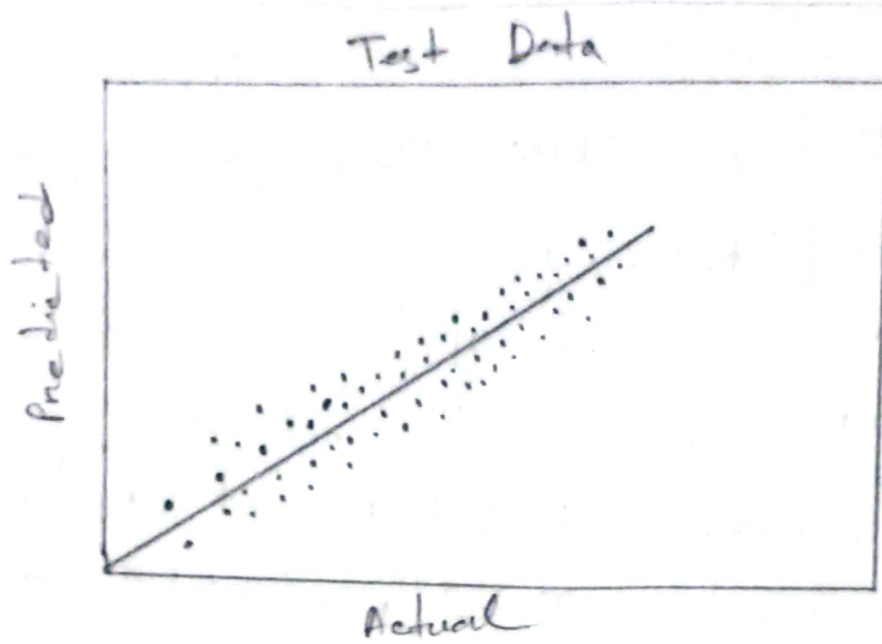


Fig: 2

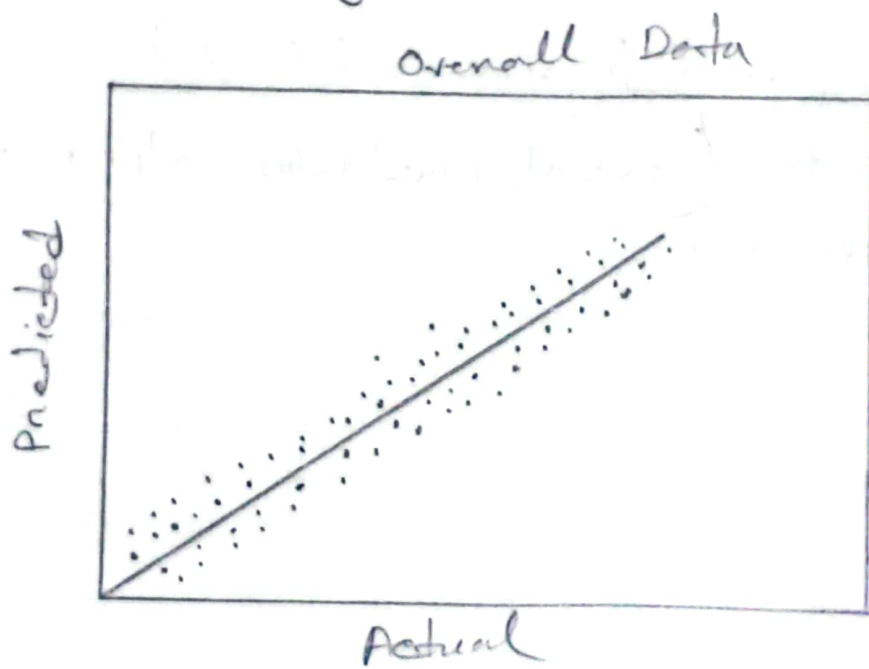


Fig: 3

As we can see our prediction fits actual data in both test, training and overall cases, so, Neural network has better predicting and pattern recognition ability for this dataset.