Machine Learning Assignment

By:-

Joel John Joseph

2247116

4 MCA-A

Perform the following

1) Explain the dataset identified

Dataset Description

The chosen dataset has 25 columns where the last column is the target column to determine if a person suffer from chronic kidney disease or not.

Chronic_Kidney_Disease Dataset

Number of instances - 400

Number of attributes - 25

Dataset characteristics - Multivariate

Missing value exist.

Metadata [full form of columns available] + (data type)

bp - blood pressure - Numerical
sg - specific gravity - Numerical
al - albumin - Nominal
su - sugar - Nominal
rbc - red blood cells - Nominal
pc - pus cell - Nominal
pcc - pus cell clumps - Nominal
bc - bacteria - Nominal
bgr - blood glucose random - Numerical
bu - blood urea - Numerical
sc - serum creatinine - Numerical
sod - sodium - Numerical
pot - potassium - Numerical
hemo - hemoglobin - Numerical
pcv - packed cell volume - Numerical
wc - white blood cell count - Numerical
rc - red blood cell count - Numerical

htn - hypertension - Nominal
dm - diabetes mellitus - Nominal
cad - coronary artery disease - Nominal
appet - appetite - Nominal
pe - pedal edema - Nominal
ane - anemia, class - classification, age
(Nominal) (Nominal) (Numerical)

Attributes to be choosen to apply distance measures

bu - blood urea - Numerical
sc - serum creatinine - Numerical
age - Numerical
bp - blood pressure -

2) Apply the following distance measures for the dataset you have chosen

a) Euclidean Distance
b) Manhattan Distance
c) Minkowski Distance

Data Matrix

id	age	bp	bu	sc
0	48	80	36	1·2
1	7	50	18	0·8
2	62	80	53	1·8
3	48	70	56	3·8
4	51	80	26	1·4

Euclidean Distance Measure

$$d(p,q) = d(q,p) = \left[\sum_{i=1}^{n} (p_i - q_i)^2 \right]^{1/2}$$

Dissimilarity / Distance Matrix

	0	1	2	3	4
0	0				
1	53.89	0			
2	22.03	71.77	0		
3	22.51	59.45	17.58	0	
4	10.15	53.85	29.15	21.86	0

$$d(0,1) = d(1,0) = \sqrt{(7-48)^2 + (50-80)^2 + (18-36)^2 + (0.8-1.2)^2}$$

$$= 53.89$$

$$d(0,2) = d(2,0) = \sqrt{(62-48)^2 + (80-80)^2 + (53-36)^2 + (1.8-1.2)^2}$$

$$= 22.03$$

$$d(1,2) = d(2,1) = \sqrt{(62-7)^2 + (80-50)^2 + (53-18)^2 + (1.8-0.8)^2}$$

$$= 71.77$$

$$d(0,3) = d(3,0) = \sqrt{(48-48)^2 + (70-80)^2 + (56-36)^2 + (3.8-1.2)^2} =$$

$$= 22.51$$

$$d(1,3) = d(3,1) = \sqrt{(48-7)^2 + (70-50)^2 + (56-18)^2 + (3.8-0.8)^2}$$

$$= 59.45$$

$$d(3,2) = d(2,3) = \sqrt{(48-62)^2 + (70-80)^2 + (56-53)^2 + (3.8-1.8)^2}$$

$$= 17.58$$

$$d(4,0) = d(0,4) = \sqrt{(51-48)^2 + (80-80)^2 + (26-36)^2 + (1.4-1.2)^2}$$

$$= 10.15$$

$$d(4,1) = d(1,4) = \sqrt{(51-7)^2 + (80-50)^2 + (26-18)^2 + (1.4-0.8)^2}$$

$$= 53.85$$

d(4,2) d(2,4) $= \sqrt{(51-62)^2 + (80-80)^2 + (26-53)^2 + (1.4-1.8)^2}$

$= 29.15$

d(4,3) d(3,4) $= \sqrt{(51-48)^2 + (80-70)^2 + (26-56)^2 + (1.4-3.8)^2}$

$= 31.86$

Manhattan Distance Measure

$d(q,p) \; d(p,q) = \sum_{i=1}^{n} |p_i - q_i|$ [Sum of absolute difference]

	0	1	2	3	4
0	0				
1	89.4	0			
2	31.6	121	0		
3	32.6	102	29	0	
4	13.2	80.6	38.4	45.4	0

d(1,0) = d(0,1)

$= |7-48| + |50-80| + |18-36| + |0.8-1.2| = 89.4$

d(2,0) d(0,2) =

$= |62-48| + |80-80| + |53-36| + |1.8-1.2| = 31.6$

d(2,1) = d(1,2) =

$|62-7| + |80-50| + |53-18| + |1.8-0.8| = 121$

d(3,0) = d(0,3) =

$|48-48| + |70-80| + |56-36| + |3.8-1.2| = 32.6$

d(3,1) = d(1,3) =

$|48-7| + |70-50| + |56-18| + |3.8-0.8| = 102$

$d(3,2) = d(2,3) = |62-48| + |70-80| + |56-53| + |3.8-1.8|$

$= 29$

$d(0,4) = d(4,0) = |51-48| + |80-80| + |26-36| + |4.4-1.2|$

$= 13.2$

$d(1,4) = d(4,1) = |51-7| + |80-50| + |26-18| + |1.4-0.8|$

$= 82.6$

$d(2,4) = d(4,2) = |51-62| + |80-80| + |26-53| + |1.4-1.8|$

$= 38.4$

$d(3,4) = d(4,3) = |51-48| + |80-70| + |26-56| + |3.8-1.4|$

$= 45.4$

MINKOWISKI DISTANCE MEASURE

$$d(x,4) = d(4,x) = \left[\sum_{i=1}^{n} |x_i - y_i|^p \right]^{1/p}$$

if $p=2 \Rightarrow$ Eudidean distance Measure

if $p=1 \Rightarrow$ Manhattan distance Measure

$p \Rightarrow$ Order of the norm.

let $p=3$

$d(0,1) = d(1,0) = \{(48-7)^3 + (50-80)^3 + (36-18)^3 + (1.2-0.8)^3\}^{1/3}$

$= 46.69$

$d(2,0) = d(0,2) = \{(62-48)^3 + (80-80)^3 + (53-36)^3 + (1.8-1.2)^3\}^{1/3}$

$= 19.71$

$d(2,1) = d(1,2) = \{(62-7)^3 + (80-50)^3 + (53-18)^3 + (1.8-0.8)^3\}^{1/3}$

$= 61.82$

$d(3,0) = d(0,3) = \{(48-48)^3 + (70-80)^3 + (56-36)^3 + (3.8-1.2)^3\}^{1/3}$

$= 20.81$

$d(3,1) = d(1,3) = \{(48-7)^3 + (70-50)^3 + (56-18)^3 + (3.8-0.8)^3\}^{1/3}$

$= 50.89$

$d(3,2) = d(2,3) = \{(48-62)^3 + (70-80)^3 + (56-53)^3 + (3.8-1.8)^3\}^{1/3}$

$= 15.58$

$d(4,0) = d(0,4) = \{(51-48)^3 + (80-80)^3 + (36-36)^3 + (1.4-1.2)^3\}^{1/3}$

$= 10.09$

$d(4,1) = d(1,4) = \{(51-7)^3 + (80-70)^3 + (36-18)^3 + (1.4-0.8)^3\}^{1/3}$

$= 44.26$

$d(4,2) = d(2,4) = \{(51-62)^3 + (80-80)^3 + (36-53)^3 + (1.8-1.4)^3\}^{1/3}$

$= 27.59$

$d(4,3) = d(3,4) = \{(51-48)^3 + (80-70)^3 + (26-56)^3 + (1.4-3.8)^3\}^{1/3}$

$= 30.38$

	0	1	2	3	4
0	0				
1	46.69	0			
2	19.71	61.82	0		
3	20.81	50.89	15.58	0	
4	10.09	44.26	27.59	30.38	0

3) Use charts to visualize the distance Matrix

	0	1	2	3	4
0	0	53.9	22.0	22.5	10.1
1	53.9	0	76.8	59.5	53.9
2	22.0	71.8	0	17.6	29.2
3	22.5	59.5	17.6	0	31.9
4	10.1	53.9	29.2	31.9	0

Distance Matrix Using Euclidean Distance Measures.

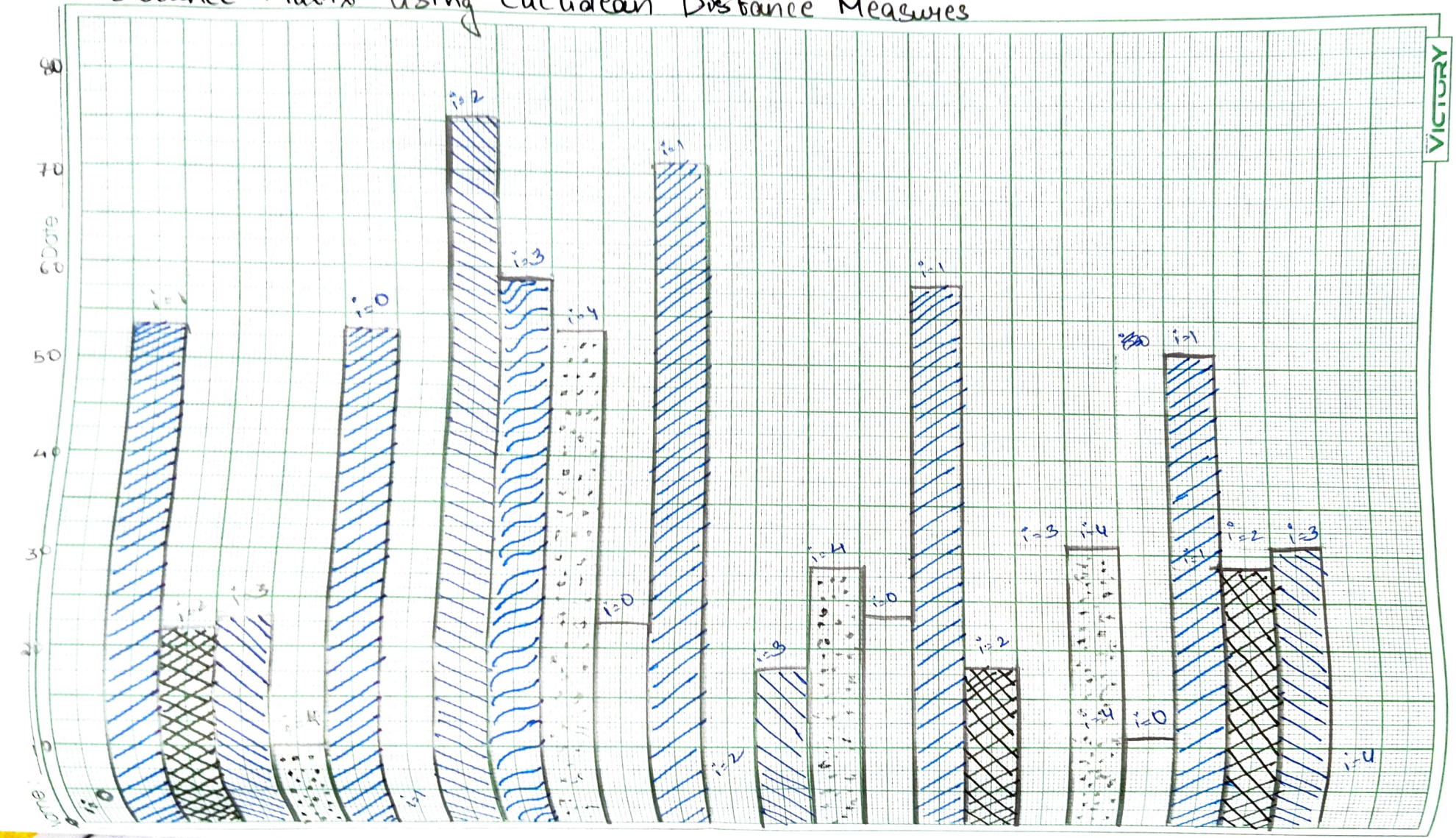
	0	1	2	3	4
0	0	89.4	31.6	32.6	13.2
1	89.4	0	121	102	82.6
2	31.6	121	0	29	38.4
3	32.6	102	29	0	45.4
4	13.2	82.6	38.4	45.4	0

Distance Matrix using Manhattan distance Measures

	0	1	2	3	4
0	0	46.7	19.7	20.8	10.1
1	46.7	0	61.8	50.9	44.3
2	19.7	61.8	0	15.6	27.6
3	20.8	50.9	15.6	0	30.4
4	10.1	44.3	27.6	30.4	0

Distance Matrix using Minkowski Distance Measures

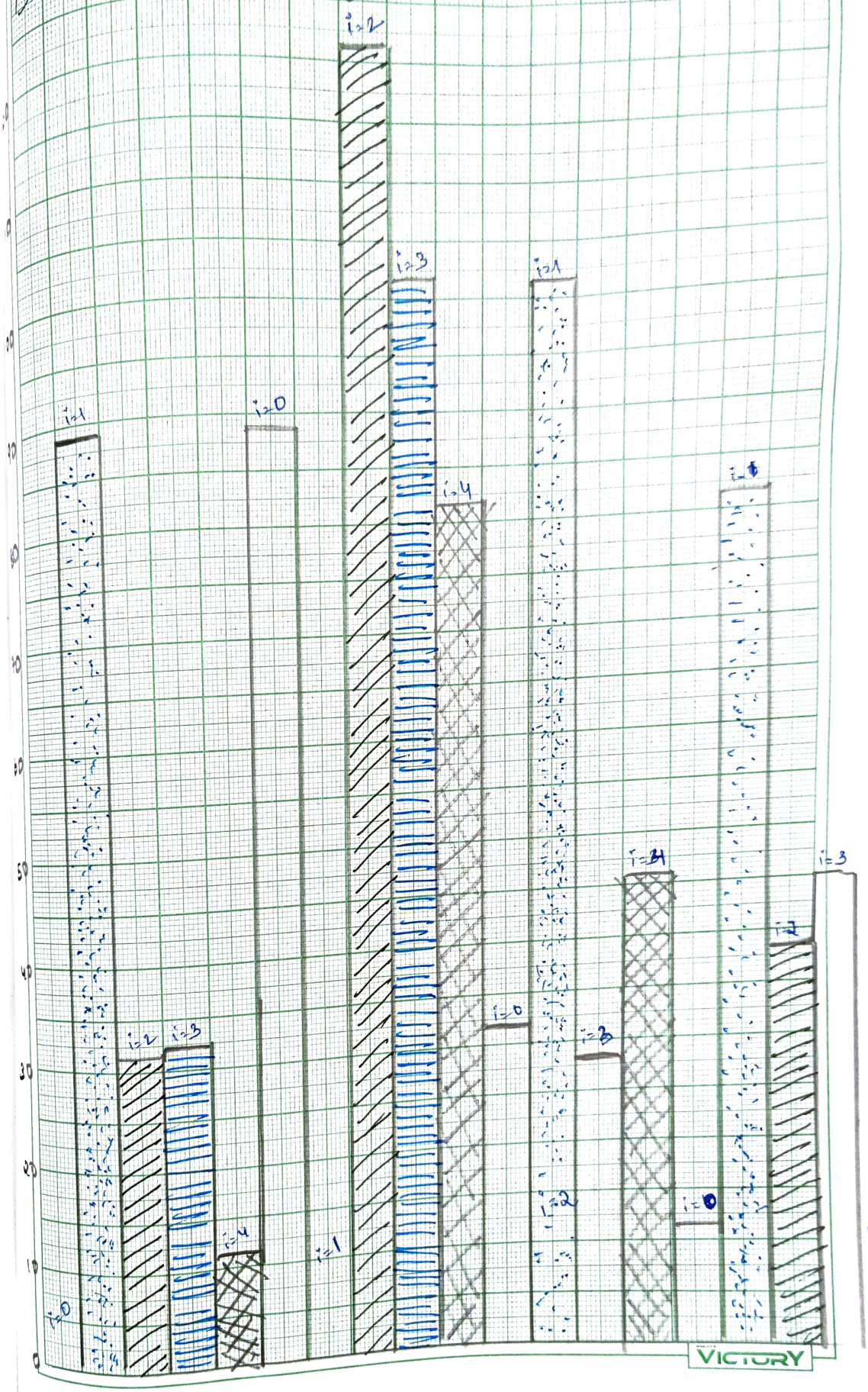
Distance Matrix Using Euclidean Distance Measures



Name :

Date

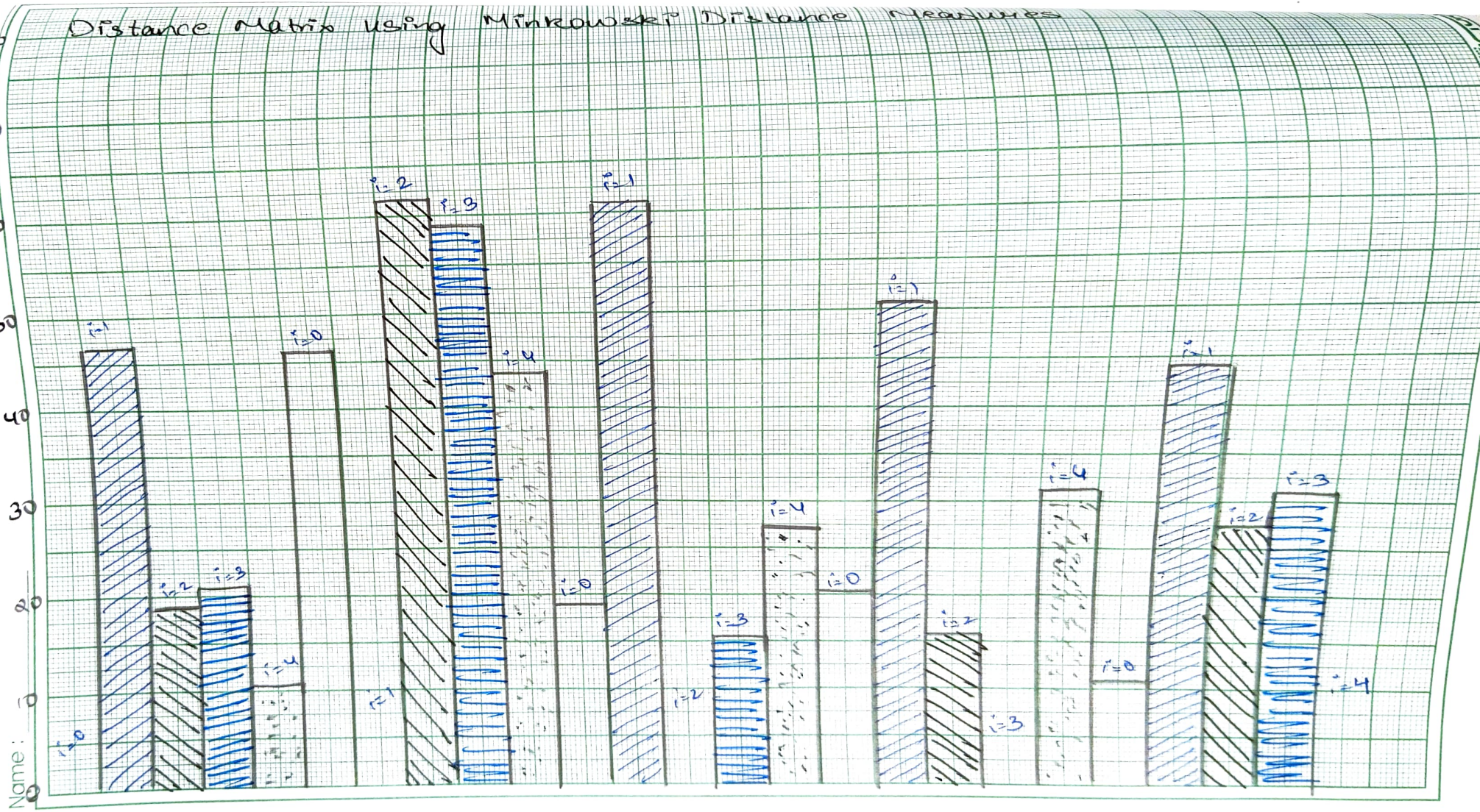
Distance Matrix Using Manhattan Distance Measures



Distance Matrix using Minkowski Distance Measures

Name :

Date



Regression Algorithm for your dataset.

predict the classification (CKD or not) using regression we would
e logistic regression for the dataset given

steps to follow the process

import the necessary libraries

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report
```

step2:-

load the dataset

```
kidney_data = pd.read_csv("kidney_disease.csv")
```

step3:-

Prepare the data

Seperate the features (attributes) and the target variable from the dataset:

```
X = kidney_data.drop('classification', axis=1)
Y = kidney_data['classification']
```

Remove the null values and clean the dataset as per requirement.

Step4:-

Split the data into training and testing set

Split the dataset into training and testing set to evaluate the performance of the logistic regression model.

```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2,
random_state=42)
```

Step 5:-

Create & train the logistic regression model

Create an instance of the logistic Regression model & fit it to the training set data.

```
logreg_model = logisticRegression()
logreg_model.fit(x_train, y_train)
```

Step6:-

Make predictions

Use the trained model to make predictions on the testing set.

```
y_pred = logreg_model.predict(x_test)
```

Step 7:- Evaluate the model

Access the performance of the logistic regression model using evaluation metric such as accuracy score & report.

```
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)
print("Accuracy:", accuracy)
print("Classification Report:\n", report)
```

Step 8:- Interpret the results

Predicted labels:- [1, 1, 0, 0, 0, 0, 1, 0 .... 0,]

1. The predicted labels represent the binary classification of whether a person has Chronic Kidney Disease (CKD) or not.
2. Label of 1 indicates that the model predicts the person has CKD, while a label of 0 suggest that the model predicts the person doesnot have CKD.