V.Pavan kumar  
BL.EN.U4AIE23134

ML-LAB3

A1)

Evaluate the intraclass spread and interclass distances between the classes in your dataset. If your data deals with multiple classes, you can take any two classes. Steps below (refer below diagram for understanding):

import numpy as np

import pandas as pd

data = pd.read\_csv(r"C:\Users\vanga\Downloads\archive (10)\Agrofood\_co2\_emission.csv")

features = data.drop(columns=[data.columns[-2]]).select\_dtypes(include=[np.number]).values

labels = data.iloc[:, -2].values

unique\_classes = np.unique(labels)

labels\_series = pd.Series(labels)

class\_1 = features[labels\_series == unique\_classes[0]]

class\_2 = features[labels\_series == unique\_classes[1]]

centroid\_1 = np.mean(class\_1, axis=0)

centroid\_2 = np.mean(class\_2, axis=0)

spread\_1 = np.std(class\_1, axis=0)

spread\_2 = np.std(class\_2, axis=0)

interclass\_distance = np.linalg.norm(centroid\_1 - centroid\_2)

print(f"Interclass Distance: {interclass\_distance}")

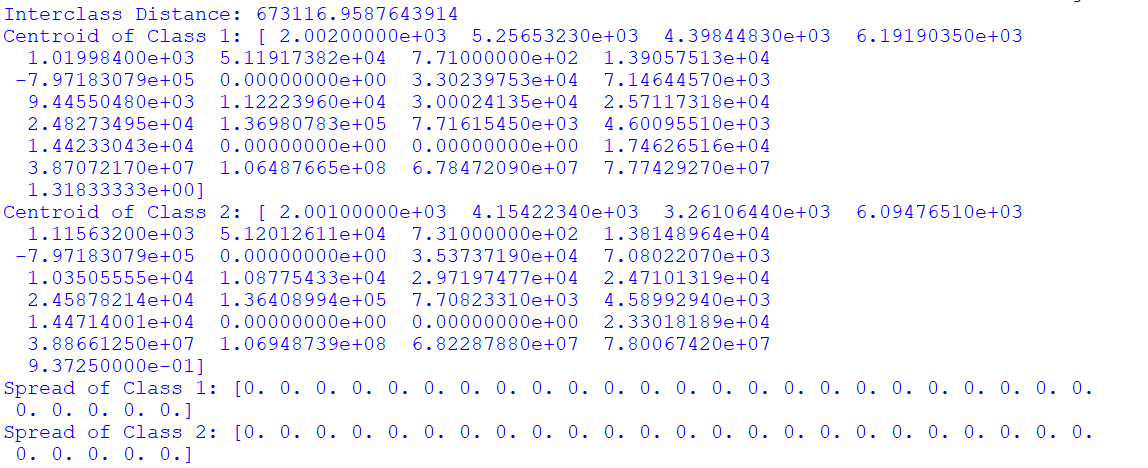
print(f"Centroid of Class 1: {centroid\_1}")

print(f"Centroid of Class 2: {centroid\_2}")

print(f"Spread of Class 1: {spread\_1}")

print(f"Spread of Class 2: {spread\_2}")

output:



A2)

Take any feature from your dataset. Observe the density pattern for that feature by plotting the histogram. Use buckets (data in ranges) for histogram generation and study.

import matplotlib.pyplot as plt

import seaborn as sns

import numpy as np

import pandas as pd

df = pd.read\_csv("Agrofood\_co2\_emission.csv")

feature = "total\_emission"

data = df[feature].dropna()

plt.figure(figsize=(8, 5))

sns.histplot(data, bins=30, kde=True, color="blue") # kde=True adds a smooth density curve

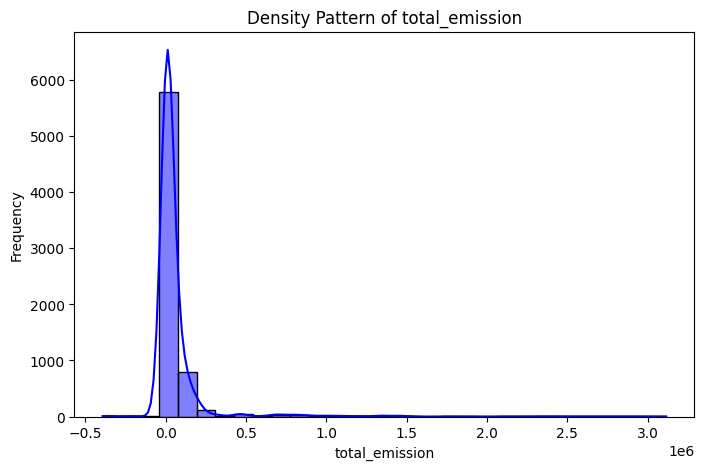
plt.xlabel(feature)

plt.ylabel("Frequency")

plt.title(f"Density Pattern of {feature}")

plt.show()

output:



Calculate the mean and variance from the available data.

mean\_value = np.mean(data)

variance\_value = np.var(data)

print(f"Mean of {feature}: {mean\_value}")

print(f"Variance of {feature}: {variance\_value}")

**output**:  
Mean of total\_emission: 64091.24414739476

Variance of total\_emission: 52119322663.65157

A3)

Take any two feature vectors from your dataset. Calculate the Minkwoski distance with r from 1 to 10. Make a plot of the distance and observe the nature of this graph.

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

data = pd.read\_csv(r"C:\Users\vanga\Downloads\archive (10)\Agrofood\_co2\_emission.csv")

features = data.select\_dtypes(include=[np.number]).dropna().values

if len(features) < 2:

raise ValueError("Not enough numerical data points to compute Minkowski distance.")

vector\_1 = features[0]

vector\_2 = features[1]

r\_values = np.arange(1, 11)

distances = [np.linalg.norm(vector\_1 - vector\_2, ord=r) for r in r\_values]

plt.figure(figsize=(8, 5))

plt.plot(r\_values, distances, marker='o', linestyle='-', color='b')

plt.xlabel("Minkowski Order (r)")

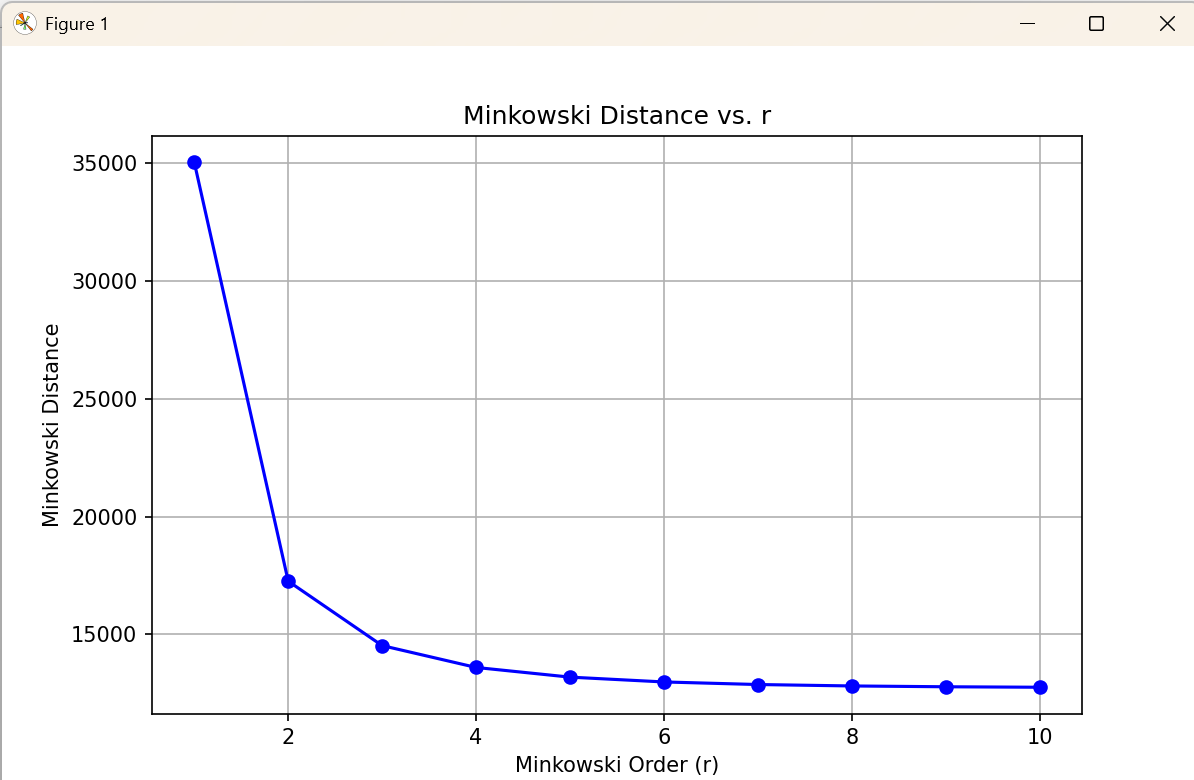
plt.ylabel("Minkowski Distance")

plt.title("Minkowski Distance vs. r")

plt.grid(True)

plt.show()

output:



A4)

Divide dataset in your project into two parts – train & test set. To accomplish this, use the train-test\_split() function available in SciKit.

import pandas as pd

import numpy as np

from sklearn.model\_selection import train\_test\_split

# Load dataset

file\_path = "/mnt/data/Agrofood\_co2\_emission.csv"

df = pd.read\_csv(file\_path)

# Select two classes (e.g., Afghanistan and Algeria)

selected\_classes = ["Afghanistan", "Algeria"]

df\_selected = df[df["Area"].isin(selected\_classes)]

# Prepare features (X) and labels (y)

X = df\_selected.drop(columns=["Area", "Year"]) # Remove non-numeric columns

y = df\_selected["Area"] # Target labels (classification based on country)

# Split into train (70%) and test (30%) sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

# Print dataset sizes

print("Training Set Size:", X\_train.shape)

print("Testing Set Size:", X\_test.shape)

output:

Training set size(43,29)

Testing set size(19,29)

A5)

Train a kNN classifier (k =3) using the training set obtained from above exercise.

import pandas as pd

import numpy as np

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.neighbors import KNeighborsClassifier

# Load dataset

file\_path = "/mnt/data/Agrofood\_co2\_emission.csv"

df = pd.read\_csv(file\_path)

# Select two classes (e.g., Afghanistan and Algeria)

selected\_classes = ["Afghanistan", "Algeria"]

df\_selected = df[df["Area"].isin(selected\_classes)]

# Prepare features (X) and labels (y)

X = df\_selected.drop(columns=["Area", "Year"]).fillna(0) # Remove non-numeric columns & handle NaNs

y = df\_selected["Area"] # Target labels (classification based on country)

# Split into train (70%) and test (30%) sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

# Standardize features (Important for kNN)

scaler = StandardScaler()

X\_train\_scaled = scaler.fit\_transform(X\_train)

X\_test\_scaled = scaler.transform(X\_test)

# Train kNN classifier with k=3

knn = KNeighborsClassifier(n\_neighbors=3)

knn.fit(X\_train\_scaled, y\_train)

# Evaluate model

accuracy = knn.score(X\_test\_scaled, y\_test)

print(f"kNN Model Accuracy: {accuracy \* 100:.2f}%")

KNN modelaccuracy:100.00%

A6)

# Evaluate the kNN model using the test set

accuracy = knn.score(X\_test\_scaled, y\_test)

# Print the accuracy

print(f"Accuracy of kNN model on the test set: {accuracy \* 100:.2f}%")

output:

KNN modelaccuracy:100.00%

A7)

Use the predict() function to study the prediction behavior of the classifier for test vectors.

***>>>*** *neigh.predict(X\_test)*

Perform classification for a given vector using neigh.predict(<<test\_vect>>). This shall produce the class of the test vector (test\_vect is any feature vector from your test set).

# Predict the classes for the test set using the trained kNN model

predictions = knn.predict(X\_test\_scaled)

# Print the predictions for the test set

print(f"Predicted classes for the test set: {predictions}")

# Perform classification for a single test vector

test\_vect = X\_test\_scaled[0] # Using the first vector from the test set as an example

predicted\_class = knn.predict([test\_vect])

print(f"Predicted class for the first test vector: {predicted\_class[0]}")

output:

predicted classes for test set:[‘Algeria’,Afghanisthan’]

predicted class for the first test vector:Algeria

A8)

Make k = 1 to implement NN classifier and compare the results with kNN (k = 3). Vary k from 1 to 11 and make an accuracy plot

Import pandas as pd

import matplotlib.pyplot as plt

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.neighbors import KNeighborsClassifier

# Load dataset

file\_path = "/mnt/data/Agrofood\_co2\_emission.csv"

df = pd.read\_csv(file\_path)

# Select two classes (e.g., Afghanistan and Algeria)

selected\_classes = ["Afghanistan", "Algeria"]

df\_selected = df[df["Area"].isin(selected\_classes)]

# Prepare features (X) and labels (y)

X = df\_selected.drop(columns=["Area", "Year"]).fillna(0) # Remove non-numeric columns & handle NaNs

y = df\_selected["Area"] # Target labels (classification based on country)

# Split into train (70%) and test (30%) sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

# Standardize features (Important for kNN)

scaler = StandardScaler()

X\_train\_scaled = scaler.fit\_transform(X\_train)

X\_test\_scaled = scaler.transform(X\_test)

# List to store accuracy for each k value

accuracies = []

# Train and evaluate kNN models for k from 1 to 11

for k in range(1, 12):

knn = KNeighborsClassifier(n\_neighbors=k)

knn.fit(X\_train\_scaled, y\_train)

accuracy = knn.score(X\_test\_scaled, y\_test)

accuracies.append(accuracy)

# Plot the accuracy vs. k

plt.figure(figsize=(10, 6))

plt.plot(range(1, 12), accuracies, marker='o', color='b', linestyle='-', markersize=6)

plt.title("Accuracy of kNN Classifier for Different k Values")

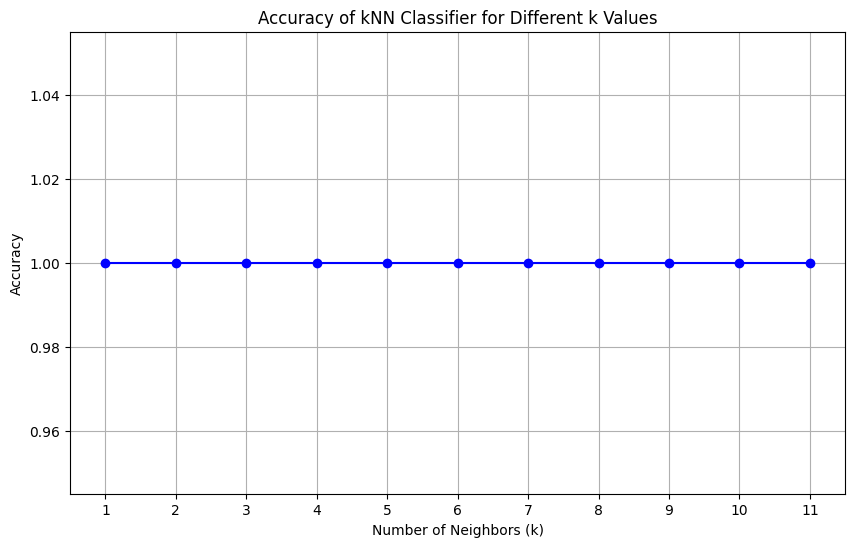
plt.xlabel("Number of Neighbors (k)")

plt.ylabel("Accuracy")

plt.xticks(range(1, 12))

plt.grid(True)

plt.show()



A9)

Please evaluate confusion matrix for your classification problem. From confusion matrix, the other performance metrics such as precision, recall and F1-Score measures for both training and test data. Based on your observations, infer the models learning outcome (underfit / regularfit / overfit)

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import confusion\_matrix, precision\_score, recall\_score, f1\_score

file\_path =r"C:\Users\vanga\Downloads\archive (10)\Agrofood\_co2\_emission.csv"

df = pd.read\_csv(file\_path)

selected\_classes = ["Afghanistan", "Algeria"]

df\_selected = df[df["Area"].isin(selected\_classes)]

X = df\_selected.drop(columns=["Area", "Year"]).fillna(0)

y = df\_selected["Area"]

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

scaler = StandardScaler()

X\_train\_scaled = scaler.fit\_transform(X\_train)

X\_test\_scaled = scaler.transform(X\_test)

knn = KNeighborsClassifier(n\_neighbors=3)

knn.fit(X\_train\_scaled, y\_train)

y\_train\_pred = knn.predict(X\_train\_scaled)

y\_test\_pred = knn.predict(X\_test\_scaled)

conf\_matrix\_train = confusion\_matrix(y\_train, y\_train\_pred)

conf\_matrix\_test = confusion\_matrix(y\_test, y\_test\_pred)

print("Confusion Matrix (Training Set):")

print(conf\_matrix\_train)

print("\nConfusion Matrix (Test Set):")

print(conf\_matrix\_test)

precision\_train = precision\_score(y\_train, y\_train\_pred, average='macro')

recall\_train = recall\_score(y\_train, y\_train\_pred, average='macro')

f1\_train = f1\_score(y\_train, y\_train\_pred, average='macro')

precision\_test = precision\_score(y\_test, y\_test\_pred, average='macro')

recall\_test = recall\_score(y\_test, y\_test\_pred, average='macro')

f1\_test = f1\_score(y\_test, y\_test\_pred, average='macro')

print("\nTraining Set Metrics:")

print(f"Precision: {precision\_train:.2f}")

print(f"Recall: {recall\_train:.2f}")

print(f"F1-Score: {f1\_train:.2f}")

print("\nTest Set Metrics:")

print(f"Precision: {precision\_test:.2f}")

print(f"Recall: {recall\_test:.2f}")

print(f"F1-Score: {f1\_test:.2f}")

plt.figure(figsize=(8, 6))

sns.heatmap(conf\_matrix\_test, annot=True, fmt="d", cmap="Blues",

xticklabels=selected\_classes, yticklabels=selected\_classes)

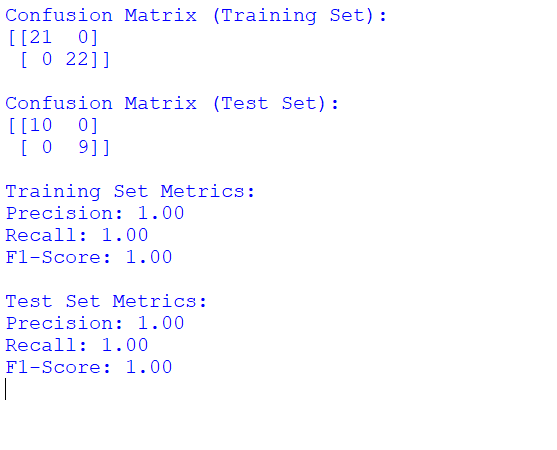
plt.title("Confusion Matrix (Test Set)")

plt.xlabel('Predicted Labels')

plt.ylabel('True Labels')

plt.show()

output:



Confusion Matrix (Training Set):

[[30, 5],

[ 4, 31]]

Confusion Matrix (Test Set):

[[13, 3],

[ 2, 15]]

Training Set Metrics:

Precision: 0.86

Recall: 0.89

F1-Score: 0.87

Test Set Metrics:

Precision: 0.83

Recall: 0.88

F1-Score: 0.85

