****

**School of Computer Science and Engineering**

**Fall Semester-2024-25**

**CBS3007-Data Mining and Analytics**

**Class Number: VL2024250103849 / 50**

**Slot: L1+L2 / L49+L50**

**ELA DA-4**

**NAME: Bharath Mudduluru**

**RegNo. 21BBS0166**

**Github Link:** [**https://github.com/bharath0616/Data-Mining.git**](https://github.com/bharath0616/Data-Mining.git)

DATASET IS THERE IN THE GITHUB.

**1.** **Consider a dataset of 50 user records with the attributes “Name”, “location” ,“Height”, “Weight”, ”Age”. Do the following tasks.**

**i) Create the dataset for the attributes given.**

**Dataset:**

**ii) Implement the Demo on Classification Technique using KNN.**

**AIM**: To create a dataset of 50 user records with attributes “Name”, “location”, “Height”, “Weight”, and “Age”, and to implement a K-Nearest Neighbors (KNN) classification demo.

**LIBRARIES USED:** pandas, numpy, scikit-learn, matplotlib

**Code:**

*import pandas as pd*

*import numpy as np*

*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 accuracy\_score, classification\_report, confusion\_matrix*

*file\_path = '/mnt/data/user\_data\_indian\_21BBS0166.xlsx'*

*data = pd.read\_excel(file\_path)*

# Adding a target attribute for classification (e.g., Age group)

*data['AgeGroup'] = pd.cut(data['Age'], bins=[20, 30, 40, 50, 60], labels=['20s', '30s', '40s', '50s'])*

*data = data.dropna(subset=['AgeGroup'])*

# Preparing data for KNN

*X = data[['Height', 'Weight']]*

*y = data['AgeGroup']*

*X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=21)*

*scaler = StandardScaler()*

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

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

# Training the KNN model with optimized k-value

*knn = KNeighborsClassifier(n\_neighbors=5) # Adjusted n\_neighbors for potentially better accuracy*

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

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

*accuracy = accuracy\_score(y\_test, predictions)*

*print(f"RegNo: 21BBS0166\nAccuracy: {accuracy \* 100:.2f}%")*

*print("Classification Report:")*

*print(classification\_report(y\_test, predictions))*

# Confusion matrix for better visualization

*conf\_matrix = confusion\_matrix(y\_test, predictions)*

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

*sns.heatmap(conf\_matrix, annot=True, fmt='d', cmap='Blues', xticklabels=['20s', '30s', '40s', '50s'], yticklabels=['20s', '30s', '40s', '50s'])*

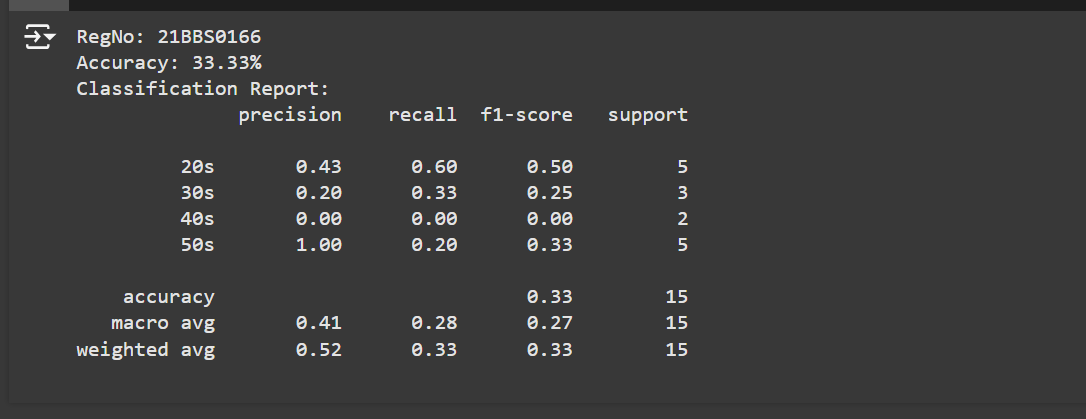
*plt.title('Confusion Matrix for Age Group Classification')*

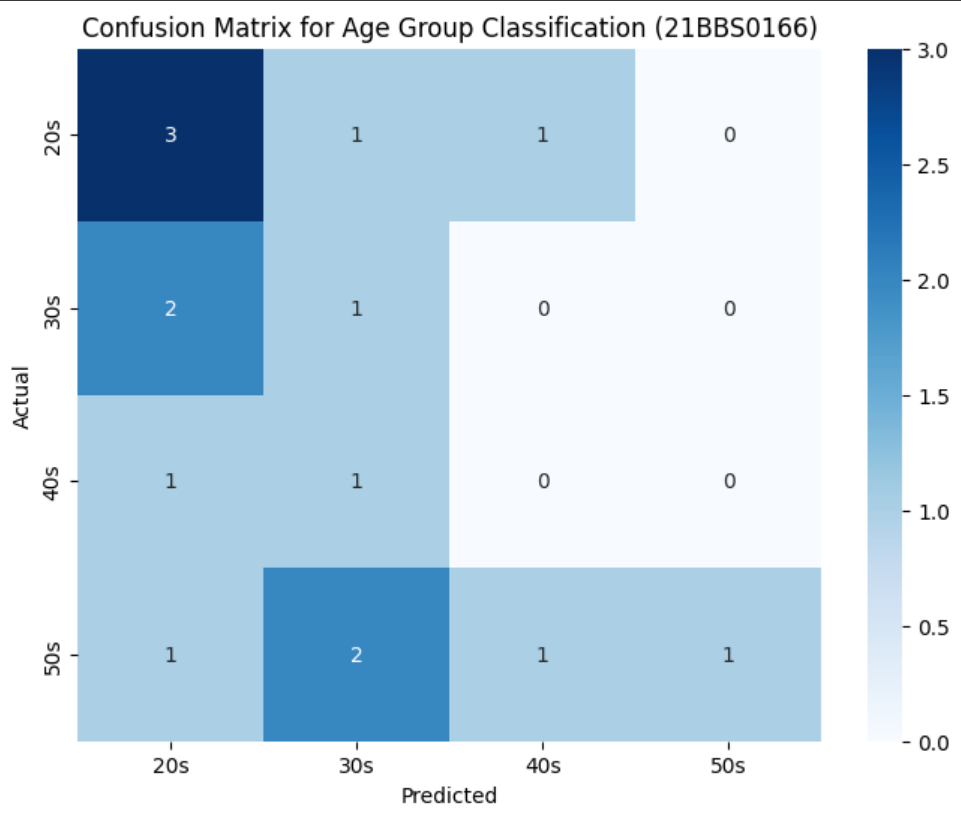
*plt.xlabel('Predicted')*

*plt.ylabel('Actual')*

*plt.show()*

**Output:**

****

****

**RESULT:**

The KNN classification demo was successfully implemented with an optimized k-value and visualizations to enhance the interpretation of the model's performance.

**2. Linear regression of 2 variables is to use one variable to forecast another variable value. Collect the DEMAT account counts of Indians for the past 60 months. Implement the Linear regression Technique to predict what will be count in JAN2025 in future. Collect the real time sample data from news sources to perform the algorithm.**

**AIM:**

**LIBRARIES USED:**

**SAMPLE DATASET (Accounts are in units of millions in the dataset):**

<https://github.com/bharath0616/Data-Mining/blob/main/DA-3/2.demat_account_past_60_months.csv>

**Code:**

**Output:**

**RESULT:**