Phase:04 TESTING

Project-02:AI-Enhanced Data Accuracy in CRM Systems Using AI Data Analysts

College Name:Dr.Ambedkar Institute Technology(Dr.AIT).

RAVI K R **CAN_36049468**

AI-Enhanced Data Accuracy in CRM Systems Using AI Data Analysts

CODE:

```
# Import libraries
import pandas as pd
import numpy as np
import re
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import DBSCAN
from sklearn.impute import KNNImputer

# 1. Load Dataset
df = pd.read_csv("Twitter Scraping Tweets Dataset.csv")

# Show basic info
print("Initial Dataset Shape:", df.shape)
print("Sample Data:\n", df.head())

# 2. Clean Phone Numbers
df['Phone'] = df['Phone'].astype(str).apply(lambda x: re.sub(r'\D', '', x))

# 3. Remove Rows with Missing or Invalid Emails
def is_valid_email(email):
    return bool(re.match(r"[^@]+@[^@]+\.[^@]+", str(email)))

df['Valid_Email'] = df['Email'].apply(is_valid_email)
df = df[df['Valid_Email'] == True]
df.drop(columns=['Valid_Email'], inplace=True)

# 4. Supervised Learning for Record Validity Prediction
# Example: Predict if record is valid based on name and country fields

# Add feature: name length
df['name_length'] = df['Name'].apply(lambda x: len(str(x)))

# Generate label for learning (1 for valid if purchase > 0, else 0)
df['is_valid'] = df['Purchase_Amount'].apply(lambda x: 1 if x > 0 else 0)

# Select features
features = ['name_length']
X = df[features]
y = df['is_valid']
```

```
# Split and train
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = LogisticRegression()
model.fit(X_train, y_train)
print("Supervised Model Accuracy:", model.score(X_test, y_test))

# 5. Duplicate Detection Using DBSCAN Clustering on Names
vectorizer = TfidfVectorizer(stop_words='english')
X_name = vectorizer.fit_transform(df['Name'].astype(str))

clustering_model = DBSCAN(eps=0.5, min_samples=2, metric='cosine')
clusters = clustering_model.fit_predict(X_name)
df['Cluster_Label'] = clusters

print("Number of Potential Duplicate Groups (Clusters):", len(set(clusters)) - (1 if -1 in
clusters else 0))

# 6. Missing Value Imputation (KNN Imputer)
imputer = KNNImputer(n_neighbors=3)
df[['Purchase_Amount']] = imputer.fit_transform(df[['Purchase_Amount']])

# 7. Output Cleaned Data
cleaned_df = df.drop_duplicates(subset=["Email", "Phone"])
print("Cleaned Dataset Shape:", cleaned_df.shape)

# Save final cleaned data
cleaned_df.to_csv("cleaned_crm_data.csv", index=False)
print("Cleaned data saved as 'cleaned_crm_data.csv'")
```

RESULT:

```
Initial Dataset Shape: (500, 10)
Sample Data:
    Unnamed: 0                                user_name  \
0           0                              Aravindh S
1           1                          Gbest Bulk SMS
2           2                      Kalyanashis Mahanty
3           3  Network Palava - Free, Cheap Data Daily.
4           4                                JayRoy IN

                  user_location  \
0              Tamil Nadu, India
1                  Abuja Nigeria
2  Barabhum ,  West Bengal, India
3                 Lagos, Nigeria
4                            NaN
```

```
                                  user_description  user_verified  \
0                       Nemophilist 🌎 90's KID 😉 😜          False
1  Providers of Bulk SMS | Data Bundle | Airtime ...          False
2  School teacher by profession and I am everythi...          False
3  Get Cheap, Free Browsing Solutions at https://...          False
4         Proud Indian \n\n(RT's r not endorsements)          False

              date                                               text  \
0  14-03-2022 12:38  And I am hearing new stories that within a sma...
1  14-03-2022 12:21  Do U have excess Airtime in ur line and will l...
2  14-03-2022 12:05  @airtelindia @Airtel_Presence bye bye for now....
3  14-03-2022 11:59  Working NapsternetV Configuration Files Downlo...
4  14-03-2022 11:55  @airtelnews @airtelindia @Airtel_Presence @air...

  hashtags              source    label
0      NaN   Twitter for iPhone        0
1      NaN  Twitter for Android  neutral
2      NaN  Twitter for Android        0
3  ['mtn']    Revive Social App  neutral
4      NaN      Twitter Web App        0
```

DOCUMENT OF ABOVE CODE:

CODE TESTING CRM -PHASE-04.ipynb