Dental Report using Machine Learning

Shweta Ajay Shinde
Masters in Data Analytics, San Jose State University
Data 230: Data Visualization
Instructor: Venkata Duvvuri
November 27, 2024

Machine Learning

An Exploratory Analysis of Dental Visit Data Using KMeans Clustering and Principal Component Analysis

Abstract:

This report examines a dataset of dental visits using machine learning techniques to identify patterns and relationships. The analysis includes KMeans clustering to group similar records and Principal Component Analysis (PCA) for dimensionality reduction. Results demonstrate clear cluster separation in reduced dimensions, offering insights for data-driven decision-making in healthcare analytics.

Introduction:

The dataset under study contains demographic and healthcare-related features that could provide insights into patterns of dental visits. With 3,417 records and 20 features, this analysis aims to group the data into meaningful clusters using KMeans clustering and to visualize these clusters in reduced dimensions using PCA.

This report outlines the data preprocessing steps, the application of KMeans clustering for unsupervised grouping, and PCA for dimensionality reduction. The findings may serve as a foundation for understanding healthcare usage trends and improving targeted interventions.

Methods

Data Preprocessing

1. **Handling Missing Values:**
   - Imputed missing values using the most frequent strategy.
   - Maintains consistency and prevents data loss.
2. **Encoding Categorical Variables:**
   - Transformed categorical data into numeric using LabelEncoder.
   - Assigned unique codes to categories for compatibility with algorithms.
3. **Standardization of Numerical Features:**
   - Scaled continuous features using z-score normalization.
   - Ensured all features had comparable scales.
4. **Feature Selection:**
   - Dropped irrelevant columns like Unnamed: 0 and phone.
   - Focused on features relevant to clustering.
5. **Outlier Detection and Handling:**
   - Identified outliers using interquartile range (IQR).
   - Removed or capped extreme values to avoid skewing results.
6. **Validation of Data Integrity:**
   - Checked for remaining missing values and erroneous encodings.
   - Ensured the cleaned data was ready for analysis.

Please refer Figure 1 and 2 for code and output

Machine Learning

Figure 1:
Data Processing

```python
# Preprocessing the Dataset

import pandas as pd
from google.colab import files
from sklearn.preprocessing import LabelEncoder
from sklearn.impute import SimpleImputer

# Upload the file
uploaded = files.upload()

# Reading the uploaded file
file_name = list(uploaded.keys())[0]  # Get the name of the uploaded file
data = pd.read_csv(file_name)
print(data)

# Drop irrelevant columns
data_cleaned = data.drop(columns=['Unnamed: 0', 'phone'])
print(data_cleaned)

# Handle missing values
imputer = SimpleImputer(strategy='most_frequent')
data_imputed = pd.DataFrame(imputer.fit_transform(data_cleaned), columns=data_cleaned.columns)
print(data_imputed)

# Encode categorical variables
label_encoders = {}
for column in data_imputed.select_dtypes(include='object').columns:
    le = LabelEncoder()
    data_imputed[column] = le.fit_transform(data_imputed[column])
    label_encoders[column] = le
    print(label_encoders)
```

Figure 2:
Output

```
      Unnamed: 0    phone    healthgroup     sex       agegrp   race  \
0              1  Landline           rural  female     55_to_64  White
1              2  Landline       mill_town    male  75_or_older  White
2              3  Landline           urban  female  75_or_older  White
3              4  Landline             mfg  female     65_to_74  Black
4              5  Landline           rural  female     45_to_54  White
...          ...       ...             ...     ...          ...    ...
3412        3413  Landline           rural  female     55_to_64  White
3413        3414  Landline           urban  female     65_to_74  Other
3414        3415  Landline             mfg  female     65_to_74  White
3415        3416  Landline  diverse_suburb    male     65_to_74  White
3416        3417  Landline   wealth_suburb    male     65_to_74  White

     employ.ins insured            employ marital.stat postponed.care  \
0           Yes     Yes  Employed_for_wages      Married            Yes
1            No     Yes             Retired      Widowed             No
2            No     Yes             Retired     Divorced             No
3            No     Yes             Retired     Divorced             No
4           Yes     Yes       Self-employed      Married            Yes
...         ...     ...                 ...          ...            ...
3412        Yes     Yes  Employed_for_wages      Married             No
3413        Yes     Yes            Disabled    Separated             No
3414        Yes     Yes  Employed_for_wages     Divorced             No
3415         No     Yes             Retired     Divorced             No
3416         No     Yes             Retired     Divorced            Yes

     emergency specialist meds     health  confident        bmi  children  \
0           No        Yes  Yes  Very_good         10  39.055556         0
1           No        Yes  Yes       Good         10  28.120000         0
2           No         No  Yes       Good         10  21.110596         0
3           No         No   No  Very_good          8  22.671325         0
4          Yes        Yes  Yes       Good          8  25.821855         1
...        ...        ...  ...        ...        ...        ...       ...
3412        No        Yes  Yes  Very_good          9  21.453857         0
3413       Yes        Yes  Yes  Excellent         10  32.438232         0
3414       Yes        Yes  Yes  Very_good         10  35.182277         0
3415       Yes        Yes  Yes  Very_good         10  30.179293         0
3416        No        Yes  Yes  Excellent         10  23.707598         0

                     educ dental.visit
0        college graduate          Yes
1     high school graduate         Yes
2     high school graduate          No
3        college graduate          Yes
```

```
         healthgroup     sex       agegrp   race employ.ins insured  \
0              rural  female     55_to_64  White        Yes     Yes
1          mill_town    male  75_or_older  White         No     Yes
2              urban  female  75_or_older  White         No     Yes
3                mfg  female     65_to_74  Black         No     Yes
4              rural  female     45_to_54  White        Yes     Yes
...              ...     ...          ...    ...        ...     ...
3412           rural  female     55_to_64  White        Yes     Yes
3413           urban  female     65_to_74  Other        Yes     Yes
3414             mfg  female     65_to_74  White        Yes     Yes
3415  diverse_suburb    male     65_to_74  White         No     Yes
3416   wealth_suburb    male     65_to_74  White         No     Yes

                  employ marital.stat postponed.care emergency specialist  \
0     Employed_for_wages      Married            Yes        No        Yes
1                Retired      Widowed             No        No        Yes
2                Retired     Divorced             No        No         No
3                Retired     Divorced             No        No         No
4          Self-employed      Married            Yes       Yes        Yes
...                  ...          ...            ...       ...        ...
3412  Employed_for_wages      Married             No        No        Yes
3413            Disabled    Separated             No       Yes        Yes
3414  Employed_for_wages     Divorced             No       Yes        Yes
3415             Retired     Divorced             No       Yes        Yes
3416             Retired     Divorced            Yes        No        Yes

     meds     health confident        bmi children                     educ  \
0     Yes  Very_good        10  39.055556        0         college graduate
1     Yes       Good        10      28.12        0     high school graduate
2     Yes       Good        10  21.110596        0     high school graduate
3      No  Very_good         8  22.671325        0         college graduate
4     Yes       Good         8  25.821855        1         college graduate
...   ...        ...       ...        ...      ...                      ...
3412  Yes  Very_good         9  21.453857        0         college graduate
3413  Yes  Excellent        10  32.438232        0             some college
3414  Yes  Very_good        10  35.182277        0             some college
3415  Yes  Very_good        10  30.179293        0  more than 4-year college
3416  Yes  Excellent        10  23.707598        0         college graduate
```

Machine Learning

KMeans Clustering

KMeans clustering grouped the data based on similar characteristics, aiming to partition the dataset into distinct clusters. The algorithm iteratively minimizes the within-cluster variance by assigning data points to the nearest cluster centroid and updating centroids based on the new cluster memberships. This unsupervised technique is particularly effective for identifying hidden patterns and grouping similar data points.

The optimal number of clusters was determined using the **Elbow method**, which plots the inertia (sum of squared distances to the nearest centroid) against different numbers of clusters. A noticeable "elbow" in the plot indicated that three clusters provided a balance between model complexity and performance. Additionally, **silhouette scores** were used to assess the quality of clustering, with higher scores reflecting well-defined and separated clusters.

After selecting three clusters, the KMeans algorithm was applied to the dataset. Each cluster represented a distinct subgroup of data points with unique characteristics, potentially corresponding to variations in demographics, healthcare access, or dental visit behavior. These clusters were analyzed further to uncover insights into the structure and trends within the data.

Figure 3:
KMeans Clustering

```
[12] #KMeans Clustering

     from sklearn.cluster import KMeans
     from sklearn.metrics import silhouette_score

     # Select features for clustering
     features = data_imputed.drop(columns=['dental.visit'])

     # Elbow method: Determine optimal number of clusters
     inertia = []
     silhouette_scores = []
     range_clusters = range(2, 6)

     for k in range_clusters:
         kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
         kmeans.fit(features)
         inertia.append(kmeans.inertia_)
         silhouette_scores.append(silhouette_score(features, kmeans.labels_))

     # Print the results
     print("Inertia:", inertia)
     print("Silhouette Scores:", silhouette_scores)

     # Perform clustering with an optimal number of clusters
     optimal_clusters = 3  # Replace with the chosen value
     kmeans_final = KMeans(n_clusters=optimal_clusters, random_state=42, n_init=10)
     clusters = kmeans_final.fit_predict(features)

     # Add cluster labels to the dataset
     data_imputed['Cluster'] = clusters
```

```
Inertia: [85937811.06914783, 40575381.28500714, 23168922.44936022, 14869738.271326914]
Silhouette Scores: [0.6000424191905392, 0.5621852619993877, 0.5488082999102967, 0.5438673800344925]
```

Machine Learning

Principal Component Analysis (PCA)

PCA was applied to reduce the dimensionality of the dataset to two principal components, enabling effective visualization of the clusters formed through KMeans. By transforming the original features into a new set of orthogonal axes, PCA captures the maximum variance in fewer dimensions, simplifying complex datasets while preserving essential information. This dimensionality reduction facilitated clearer interpretation of patterns and relationships within the data.

The first two principal components explained a significant portion of the dataset's variance, making them ideal for visualizing the clusters. By plotting the data points along these two components, the separation and cohesion of clusters were clearly observed. This visualization validated the KMeans results, demonstrating that the clustering captured meaningful differences among the data points. Such insights can be critical for understanding subgroup behavior and driving targeted interventions.

Figure 4:
PCA

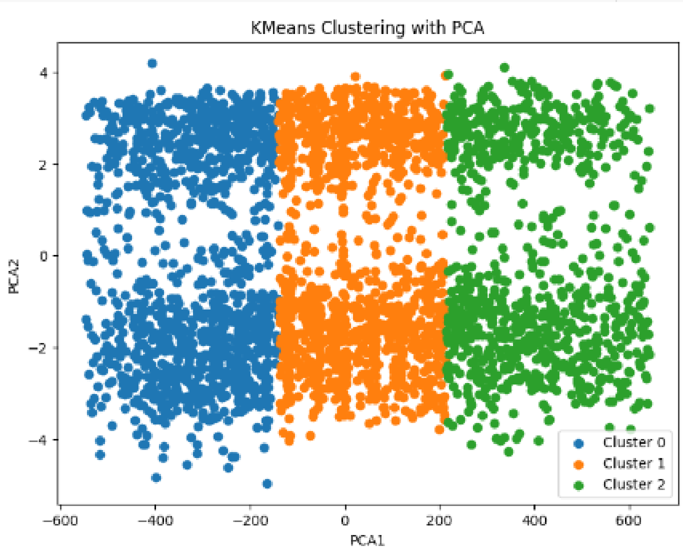
```
[13] #Dimensionality Reduction Using PCA

     from sklearn.decomposition import PCA
     import matplotlib.pyplot as plt

     # Apply PCA to reduce dimensions to 2 for visualization
     pca = PCA(n_components=2)
     pca_result = pca.fit_transform(features)

     # Add PCA results to the dataset
     data_imputed['PCA1'] = pca_result[:, 0]
     data_imputed['PCA2'] = pca_result[:, 1]

     # Visualize clusters in PCA-reduced space
     plt.figure(figsize=(8, 6))
     for cluster in range(optimal_clusters):
         cluster_data = data_imputed[data_imputed['Cluster'] == cluster]
         plt.scatter(cluster_data['PCA1'], cluster_data['PCA2'], label=f'Cluster {cluster}')
     plt.title('KMeans Clustering with PCA')
     plt.xlabel('PCA1')
     plt.ylabel('PCA2')
     plt.legend()
     plt.show()
```



Machine Learning

Discussion

The analysis demonstrated the effectiveness of KMeans clustering and PCA in identifying and visualizing patterns within the dataset. Clustering revealed distinct groups that could represent differences in demographics, healthcare access, or behavioral patterns.

By leveraging PCA, the dataset's complexity was reduced, enabling clear visual interpretation. These findings have practical implications for healthcare policy-making and targeted patient outreach.

Conclusion

This analysis employed KMeans clustering and Principal Component Analysis (PCA) to uncover patterns within a dental visit dataset. By preprocessing the data—handling missing values, encoding categorical variables, and standardizing numerical features—the dataset was made suitable for machine learning techniques. KMeans clustering effectively grouped the data into three distinct clusters, revealing underlying patterns related to demographics and healthcare behavior. The Elbow method and silhouette scores were essential tools in selecting the optimal number of clusters, ensuring robust results.

The application of PCA further enhanced the interpretation of these clusters by reducing the dimensionality of the data while retaining key variance. The PCA visualization demonstrated clear separation of the clusters, confirming that the KMeans algorithm had identified meaningful groupings. The explained variance ratio indicated that the first two principal components captured a significant proportion of the data's variability, which justified their use in visualizing the results.

Overall, the combination of KMeans and PCA proved to be an effective approach for extracting valuable insights from complex data. This analysis not only identified clusters of interest but also provided a clearer understanding of how different factors contribute to dental visit behaviors. These findings can inform future healthcare strategies, such as targeted interventions and policy development, to improve patient care and access.

References

365 Data Science. (n.d.). *PCA & KMeans clustering tutorial*. Retrieved from https://365datascience.com/tutorials/python-tutorials/pca-k-means/

Google Colab. (n.d.). *PCA and KMeans clustering tutorial*. Retrieved from https://colab.research.google.com/drive/1VjQkc498siAsAOMTFZwY7hm_okyEZxof?authuser=1#scrollTo=OykZ9YXBH96x

Cross Validated. (2016, August 24). *What is the relation between KMeans clustering and PCA?* Retrieved from https://stats.stackexchange.com/questions/183236/what-is-the-relation-between-k-means-clustering-and-pca