Medical Cost Personal Dataset

Dataset:

The “Medical Cost Personal Dataset” by “Packt Publishing” is implemented on the novel methodology for the paper[1]. The dataset contains tabular data with 7 features and over 1338 data points. The data is originally kept private and is referenced in the published book “Machine Learning with R” by Brett Lantz which provides an introduction to machine learning using R. The dataset contains the personal information of individuals seeking insurance along with the monthly charges involved. Further details of the features are given below.

Columns

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of the body, weights that are relatively high or low relative to height,
- objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

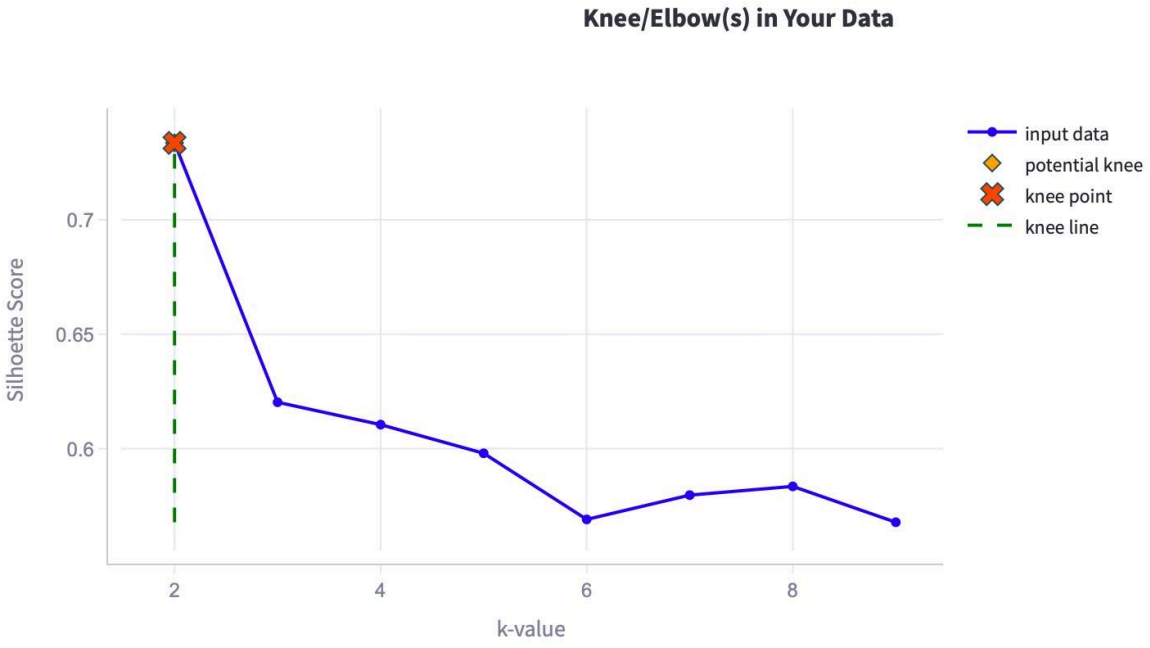
Pre-processing:

As the data was obtained from an open-source database, it was necessary to perform comprehensive data cleaning. Firstly, all the NaN values were eliminated and replaced with the median values of the respective columns. Additionally, certain features such as 'customer ID' were removed as they did not contribute towards customer segmentation. In order to facilitate the data processing by the model, the categorical data were transformed into numerical labels. Nevertheless, the dataset could be further modified based on the requirements of the client.

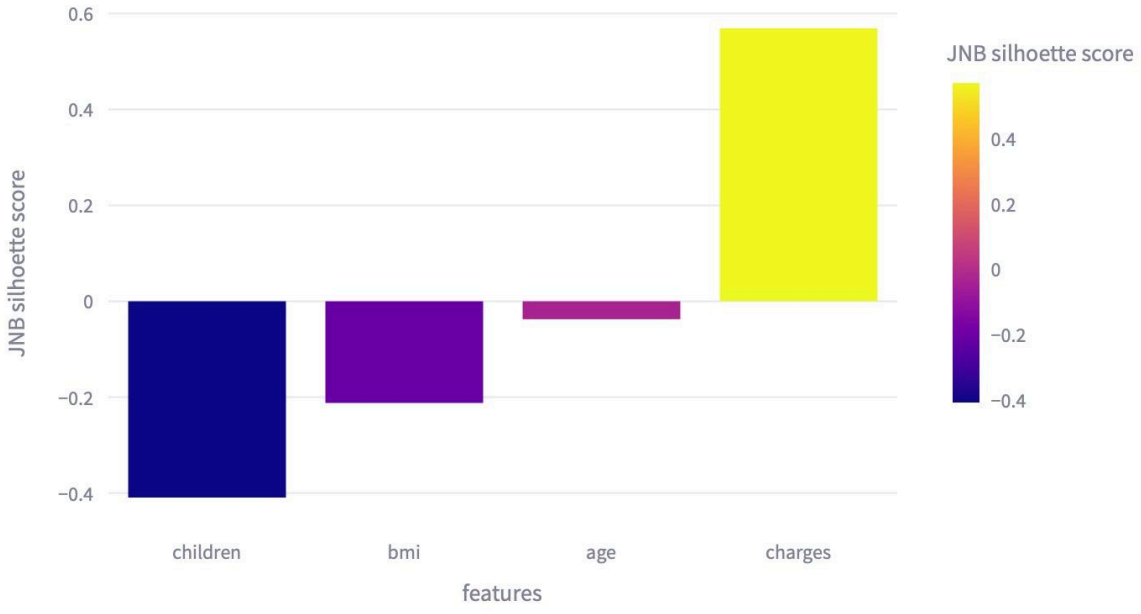
Methodology:

After initial processing, the attributes are categorised as numerical or categorical based on a user-defined threshold. Here, if the number of unique values is less than 5, the attribute is considered categorical. Out of the 7 features, 4 were classified as numerical, and the remaining were considered categorical. The numerical features were then clustered into two groups using JNB clustering and ordered based on their silhouette scores. The top 25% of these features are selected and used to create a feature matrix. In this case, the combination of "charges" and “age” is chosen for further analysis based on its silhouette score of 0.735. The entropy of all categorical features and the selected set was compared, and K-means clustering was used to generate all clusters of the selected numerical features.

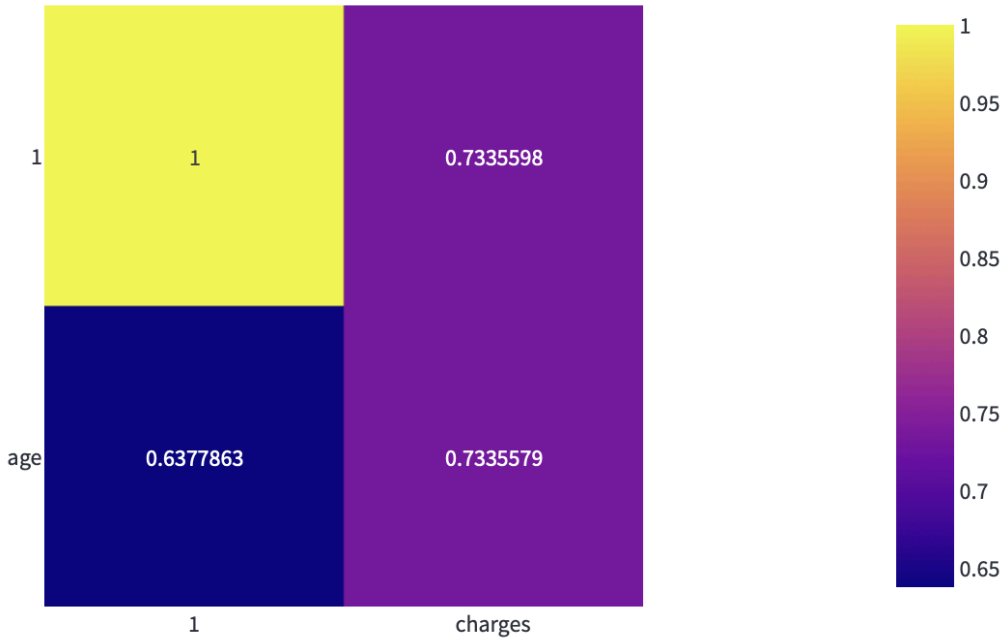
The joint entropy of these features was calculated against all categorical features, and the features were compared based on their entropy values. The figure showed that the maximum entropy value was 0.163 for “region”, and the minimum was 0.028 for "smoker". Depending on the user's requirements, one or more categorical features can be chosen.



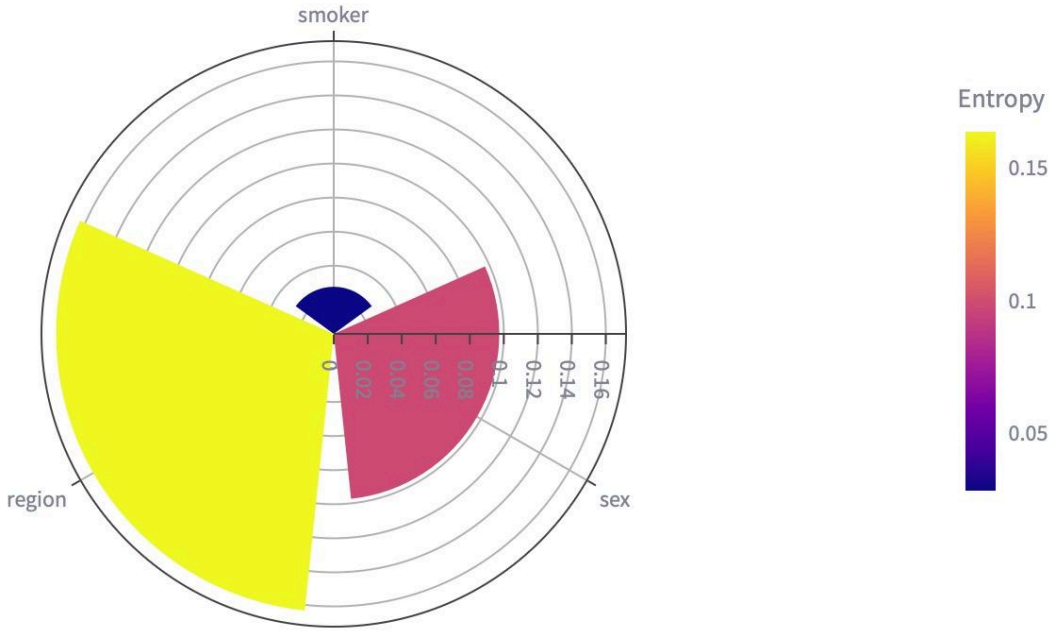
The Silhouette scores of JNB Clustering, based on the features [3]:



Feature matrix for case study 2 [4]



The Entropy of each categorical feature w.r.t. The selected numeric feature/s[5]



Reference Link

1. https://www.kaggle.com/datasets/mirichoi0218/insurance