Employing PCA and K-Means Clustering for Insightful Data Exploration

The selected heart_disease dataset includes a multitude of variables including categorical, numeric and binary data types, each providing insight into various risk factors and associated indicators to cardiovascular health. In the pursuit of understanding the innate structures and potentially discovering any natural clustering in the data, an analysis method, principal component analysis (PCA) and k-Means clustering, was implemented.

Data Pre-processing

A thorough pre-processing phase was fundamental, given the heterogeneous nature of the variables in the dataset. Recognizing the diversity in the variables (numerical, categorical, and binary), appropriate pre-processing procedures were designed for each data type. Numerical variables were addressed with mean imputation for missing values and were scaled using Standard Scaler. Categorical variables were mode-imputed and one-hot encoded to transpose their nominal nature into a numerical format. Binary variables were similarly mode-imputed and subsequently scaled. These modified approaches ensured the retention of data integrity whilst mitigating the influence of any outlying variables during the modelling process.

Application of models

1. **Principal Component Analysis (PCA)**

   PCA was applied to navigate the high-dimensional space of the data by reducing it to a more manageable, still informative and lower-dimensional space. The number of components to retain is an important decision, carefully considered by observing the proportion of explained variance and ensuring a balance between data reduction and information retention.

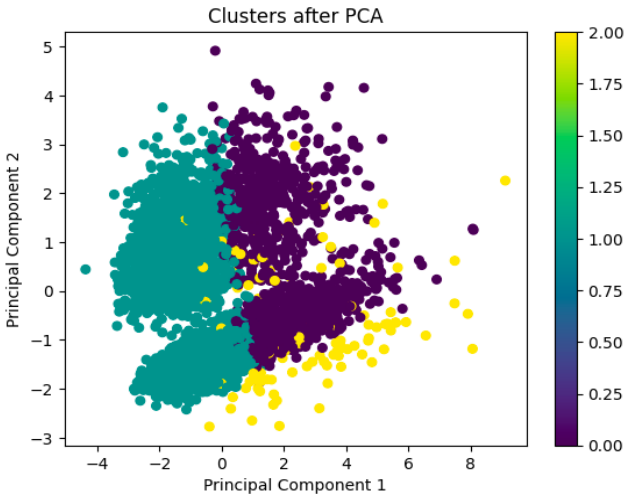
2. **K-Means Clustering**

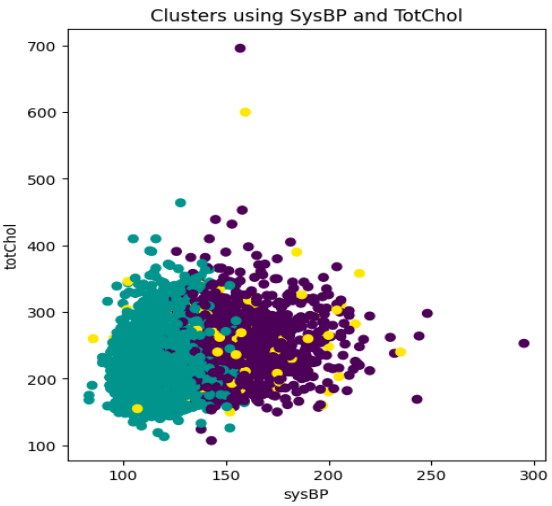
   k-Means clustering was engaged to distinguish any inherent groupings within the data. The critical choice of “k”, the number of clusters, was determined through the Elbow Method, providing a balance between minimizing variance within clusters without dramatically increasing model complexity.

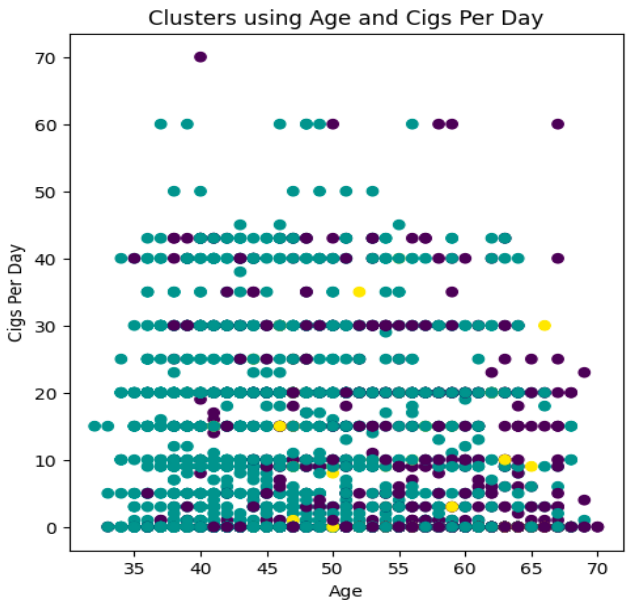
When performing PCA, the dimensionality was reduced significantly, collapsing the original feature space whilst retaining a significant portion of the original variance. Subsequent k-Means Clustering revealed distinct clusters within the data, potentially representing different risk profiles or disease-related demographic groups.

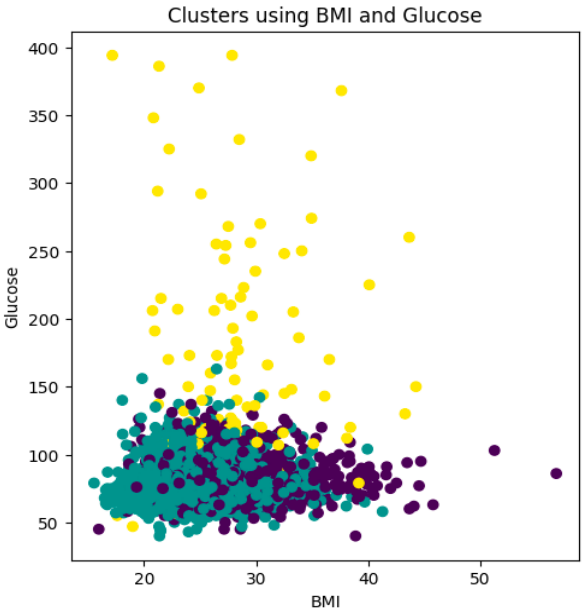
Visualizations

Scatterplots of the first two principal components, coloured according to cluster distributions, shows clear separation between clusters, although there are some overlapping points which suggest a reasonable degree of distinction between identified groups.









The combined application of PCA and k-Means Clustering has demonstrated a fascinating exploration of a dataset, allowing succinct visualization of its internal structure and facilitating the identification of potential clusters.