HyperionDev

Unsupervised Learning – K-Means Clustering

Model Answer Approach

Visit our website

Auto-graded task 1

This unsupervised learning approach begins by importing essential libraries, including `pandas`, `numpy`, and `matplotlib`, alongside clustering methods from `scikit-learn`. The dataset is loaded, and initial exploratory data analysis (EDA) is performed, which includes checking the dataset's structure, datatypes, missing values, and descriptive statistics. Non-numeric features like “country” are dropped to ensure that the dataset only contains features suitable for clustering.

A correlation heatmap and scatter plots are generated to explore relationships between features like child mortality and GDP per capita, offering insights into the data. A pair plot is also created to visually represent the distribution of features in relation to child mortality.

To prepare the data for clustering, the features are normalised using `MinMaxScaler`. K-means clustering is applied to the scaled dataset, and the elbow method and silhouette score are used to determine the optimal number of clusters, ensuring that the model is well-optimised for grouping similar data points.

The final K-means model is fitted with four clusters, and the silhouette coefficient is used to evaluate the performance of the model. The clusters are then visualised using scatter plots, comparing key features like child mortality and GDP per capita across different clusters. Finally, the predicted clusters are added to the original dataset for further analysis.

A possible pitfall in this approach could be failing to properly normalise the data before clustering, which could lead to biased clustering results due to differing feature scales. Additionally, choosing an inappropriate number of clusters might affect the accuracy of the model. It’s also important to ensure that the random state is consistently set for reproducible results.