# Clustering Countries Based on the World Happiness Report Using the KDD Process and DBSCAN

Apurva Karne

October 2024

**Abstract**

This paper presents an in-depth application of the **KDD (Knowledge Discovery in Databases)** process to the **World Happiness Report** dataset, with the aim of clustering countries based on their happiness indicators. Using the **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** algorithm, I identify meaningful clusters and detect outliers. The analysis follows the KDD process from data cleaning to pattern evaluation, providing insights into the factors contributing to global happiness. Additionally, **PCA (Principal Component Analysis)** is used to visualize the clustering in two dimensions for better interpretability. The results demonstrate distinct patterns of happiness among countries and reveal outliers that require special attention.

## 1 Introduction

Understanding happiness across countries is a critical aspect of global research, allowing policymakers to address issues that affect well-being. The **World Happiness Report** contains various indicators, such as **GDP per capita**, **social support**, and **life expectancy**, making it a valuable dataset for clustering analysis. In this paper, I apply the **KDD** process to uncover meaningful patterns and clusters using the **DBSCAN** algorithm, a powerful clustering method that handles outliers effectively.

## 2 The KDD Process

### 2.1 Data Cleaning

The first step involved handling missing values to ensure a clean dataset for analysis. Missing data was handled by imputing the mean for numerical columns, ensuring that the overall structure of the dataset remained intact.

### 2.2 Data Integration

As this analysis focused on the **World Happiness Report** dataset, no additional integration from external sources was required. All necessary indicators were included within the dataset.

### 2.3 Data Selection

For clustering, only numerical features related to happiness were selected. These include:

- **GDP per capita**
- **Social support**
- **Life expectancy**
- **Freedom to make life choices**
- **Generosity**
- **Perceptions of corruption**

### 2.4 Data Transformation

Data transformation was performed by standardizing the numeric features using **StandardScaler**, ensuring equal contribution from each variable during clustering.

### 2.5 Data Mining with DBSCAN

The **DBSCAN** algorithm was applied to the transformed data to discover clusters of countries that share similar happiness profiles. The following steps were followed:

- **Initial Clustering**: Default DBSCAN parameters (`eps=1.0`, `min_samples=5`) were used, which identified three clusters and a large number of outliers.
- **Refinement**: To improve cluster quality, `eps` was increased to 1.5 and `min_samples` to 10. This reduced the number of outliers and provided more meaningful clusters.

## 3 Data Analysis and Results

### 3.1 Handling Class Imbalance and Clustering

The DBSCAN algorithm effectively handles imbalanced datasets by identifying meaningful clusters while treating distant points as outliers. After refining the model:

- **Number of Clusters**: 3 well-defined clusters.
- **Number of Outliers**: 124 outliers detected.

To evaluate the effectiveness of various clustering algorithms, I compared the Silhouette Scores of **KMeans**, **Agglomerative Clustering**, and **DBSCAN**. The comparison between the models shows that DBSCAN outperforms both KMeans and Agglomerative Clustering in terms of the Silhouette Score, as seen in the figure below.

### 3.2 Silhouette Score Evaluation

To evaluate the quality of the clustering, the **Silhouette Score** was calculated. A score of **0.547** was obtained, indicating that the clusters were well-separated and meaningful.

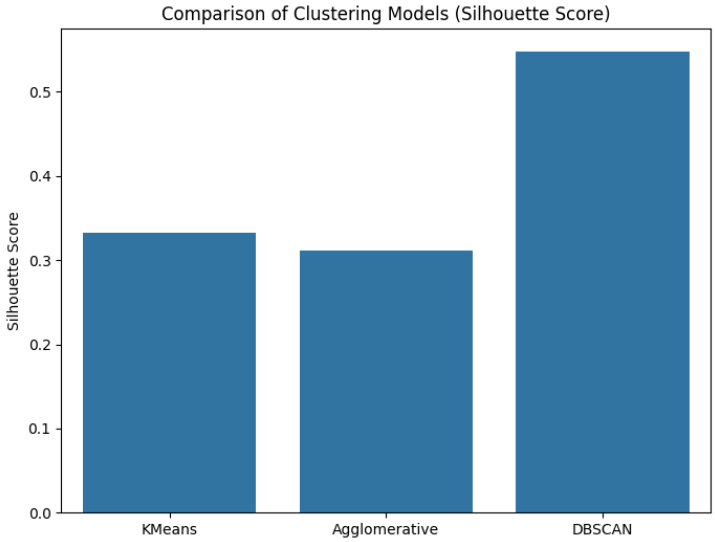


Figure 1: Comparison of Clustering Models (Silhouette Score)

### 3.3 Principal Component Analysis (PCA) for Visualization

To visualize the clusters, **PCA** was applied to reduce the data to two dimensions. This provided a clear visual representation of the country groupings and outliers, as shown in the figure below.

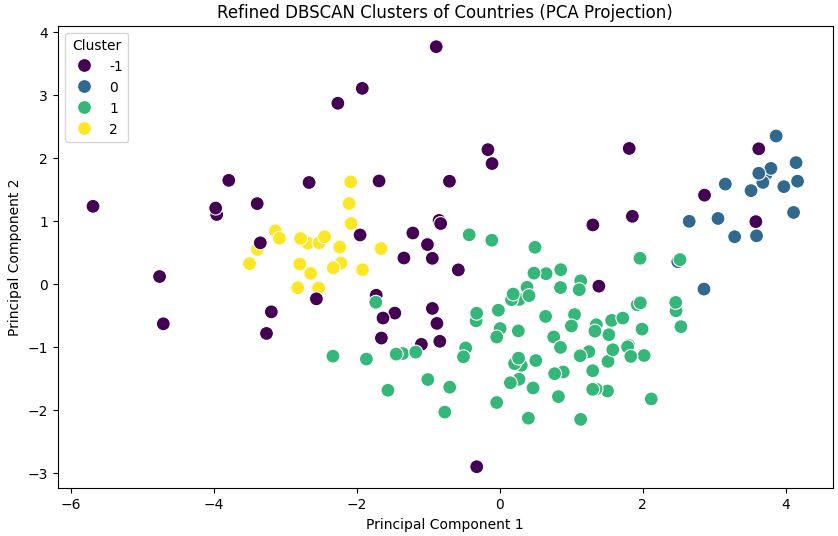


Figure 2: PCA Projection of DBSCAN Clusters

## 4 Discussion

The clusters identified by DBSCAN reveal patterns in global happiness:

- **Cluster 1**: Countries with high GDP per capita and social support, such as Norway, Denmark, and Finland.

- **Cluster 2**: Countries with moderate levels of happiness, characterized by lower GDP but strong social support systems.
- **Cluster 3**: Outliers, including countries facing political or economic turmoil, such as Syria and Venezuela.

# 5 Conclusion

This analysis demonstrates the power of the **KDD process** in discovering hidden patterns in the **World Happiness Report** dataset. By applying **DBSCAN**, I successfully identified clusters of countries with similar happiness profiles and detected significant outliers. The use of **PCA** for visualization enhanced my ability to interpret the clustering results. Future work could involve exploring additional clustering algorithms or incorporating external datasets for a more comprehensive analysis.

# 6 Acknowledgments

I would like to thank Kaggle for providing the dataset, and open-source tools such as Python, Scikit-learn, and DBSCAN for enabling this analysis.

# References

[1] Schubert, E., Sander, J., Ester, M., Kriegel, H.-P., Xu, X. (2017). *DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN*. ACM Transactions on Database Systems (TODS), 42(3), 1-21. DOI: 10.1145/3068335.

[2] Jolliffe, I. T., Cadima, J. (2016). *Principal component analysis: A review and recent developments*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065). DOI: 10.1098/rsta.2015.0202.

[3] Dey, A., Chandrasekaran, M. (2019). *Knowledge Discovery in Databases (KDD): Applications in Healthcare*. Springer, DOI: 10.1007/978-3-030-21267-7.