Data Analysis and Visualization Final Report Predictive Medical Analysis

Bilal Ahmad
Umair Imran
Najam ul Islam
Shahzad Waris

December 4, 2024

Abstract

This report details the methodology and results of a predictive analysis project aimed at forecasting public health indicators using cluster-based time series modeling. By clustering countries based on key indicators and training models on these clusters, the project demonstrates an efficient approach to data-driven forecasting. Models such as ARIMA, SARIMA, LSTM, and Prophet were employed, with interactive visualizations created to support insights. Future enhancements and suggestions for real-time data integration are also provided.

1 Introduction

Time series forecasting is critical in public health to predict future trends and plan interventions. This project utilizes data scraped from the WHO website, spanning 2000 to 2024, and employs cluster-based modeling to group countries by similar characteristics. After clustering, time series models are trained to predict key indicators, including population and disease prevalence.

2 Data Gathering

The dataset for this project was collected by scraping the World Health Organization (WHO) website using Python and Selenium. The data spans from 2000 to 2024, covering key public health indicators such as population, health expenditure, and disease prevalence.

2.1 Challenges Faced

- **Dynamic Content**: Many pages used JavaScript to load data dynamically, requiring careful handling of page rendering.

- **Rate Limiting**: Frequent requests were blocked, necessitating the use of delays and retries to avoid IP bans.
- **Data Cleaning**: Extracted data often contained inconsistencies, such as missing values, duplicate entries, or mismatched units.
- **Data Normalization**: Population values lacked consistent units (e.g., million or billion), requiring manual adjustments based on context and external references.

3 Data Preprocessing

3.1 Dataset Description

The dataset includes the following columns: Name, health expenditure, WHO region, World Bank income level, population growth rate, year, population, life expectancy, health life expectancy, Number of new HIV infections and Suicide deaths. Each country is represented across 24 years.

3.2 Preprocessing Steps

- Handled missing values using statistical and advanced imputation techniques.
- Encoded categorical columns like `name` using binary encoding, as these columns are nominal.
- Encoded ordinal columns such as `WHO region` and `world bank income level` using label encoding.
- Standardized numerical columns for clustering and time series modeling.
- Imputed missing data using MICE (Multiple Imputation by Chained Equations) and Random Forest models for specific columns based on their nature and dependencies.

3.2.1 MICE (Multiple Imputation by Chained Equations)

- **Health Expenditure:** Imputed using MICE to capture interdependencies with features like population size and regional characteristics. This method effectively estimates continuous data with complex relationships.
- **Population:** Missing population data was iteratively imputed using MICE, leveraging correlated features such as health expenditure and life expectancy to ensure robust estimations.
- **Number of New HIV Infections:** Used MICE to impute this column by considering patterns associated with health expenditure and population demographics, ensuring consistency in data relationships.
- **Suicide Deaths:** Imputed using MICE, as this approach utilizes correlated indicators like population size and regional health metrics, providing reliable approximations for sparse data.

3.2.2 Random Forest for Imputation

- **Population Growth Rate (%):** Imputed using a Random Forest model trained on features like population size, life expectancy, and other health indicators. Random Forest excels at capturing non-linear relationships and interactions among variables.
- **Life Expectancy:** Missing values in the life expectancy column were imputed using a Random Forest model. This method effectively captures the complex, non-linear relationships between features such as health expenditure, income level, and lifestyle indicators (e.g., tobacco use).

4 Methodology

4.1 Clustering

To group countries with similar trends, clustering was applied to various indicators such as population, tobacco use, and alcohol consumption. The following models were evaluated:

- **KMeans**: A simple and scalable clustering algorithm suitable for structured datasets.
- **DBSCAN**: A density-based clustering method effective in identifying outliers and handling non-linear cluster shapes.
- **Autoencoder-based Clustering**: A deep learning approach for capturing high-dimensional and non-linear patterns in data.

Cluster-based modeling enhances efficiency and interpretability by grouping countries with similar trends, which simplifies downstream analysis and modeling.

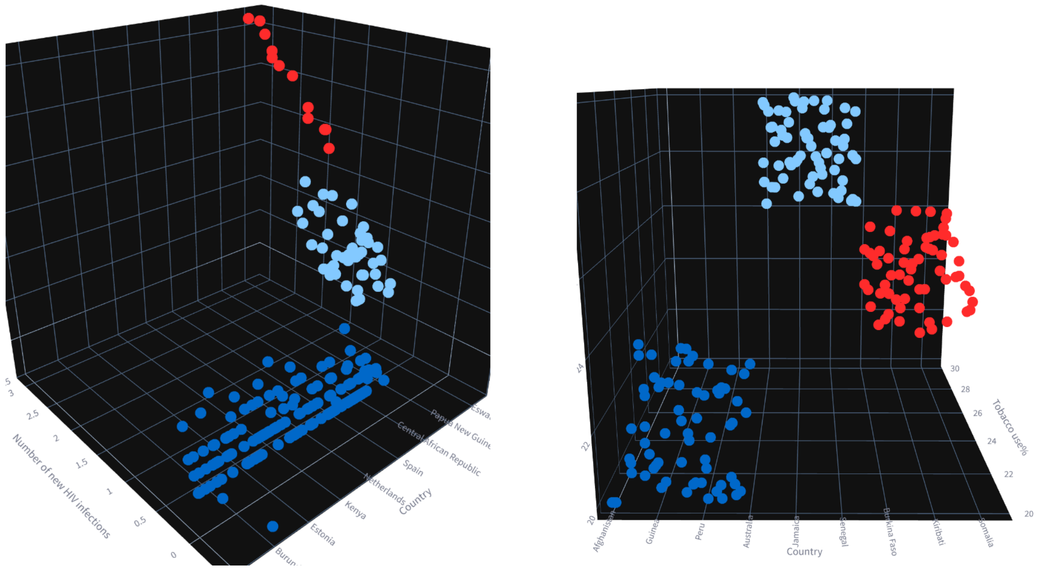
4.1.1 Best Models for Each Indicator

For each indicator, the model achieving the highest silhouette score was selected as the best. The results are as follows:

- **Number of New HIV Infections:** Autoencoders with a silhouette score of 0.8572.
- **Tobacco Use (%):** DBSCAN with a silhouette score of 0.837.
- **Alcohol Consumption:** DBSCAN with a silhouette score of 0.7947.
- **Population:** Autoencoders with a silhouette score of 0.7818.
- **Prevalence of Hypertension (%):** Autoencoders with a silhouette score of 0.8084.

4.1.2 Clustering Visualizations

The visualizations below show the clusters assigned for each indicator. Each plot highlights how countries were grouped based on their similarities, with outliers identified where applicable. These visualizations assist in understanding the inherent patterns and variations among countries.



(a) Cluster Based on HIV

(b) Cluster based on Tobacoo

Figure 1: Clustered Assigned by models

4.2 Time Series Modeling

To forecast indicators, various time series models were applied. These models are suited to different types of patterns and complexities in the data:

- **ARIMA:** Best for univariate and linear trends, capturing simple relationships.
- **SARIMA:** Effective in handling seasonality in time series data.
- **LSTM:** A neural network-based model, suitable for capturing long-term and complex dependencies.
- **Prophet:** A flexible model, well-suited for data with missing values and abrupt trend changes.

For example, the LSTM model was used to predict the population of Afghanistan country until 2050, showing how long-term trends are captured effectively. The forecasted population is displayed in the line graph below, illustrating the model's accuracy and trend prediction.

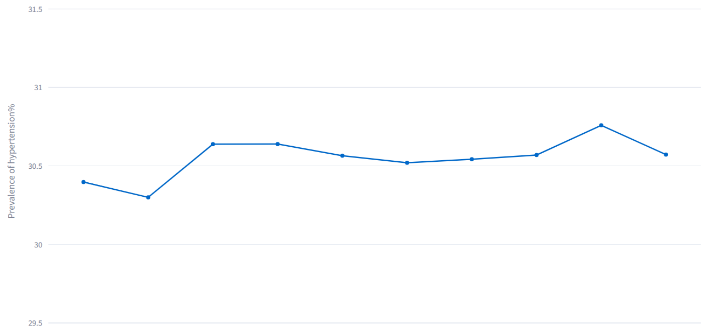


Figure 2: LSTM prediction of the Hypertension over time in Afghanistan

5 Results and Evaluation

5.1 Evaluation Metrics

- Clustering Metrics: Silhouette score. It evaluates both the compactness and separation of clusters, making it ideal for health data where well-defined clusters help identify patterns such as disease prevalence or healthcare resource distribution. The score ranges from -1 to 1, with higher values indicating better clustering quality.

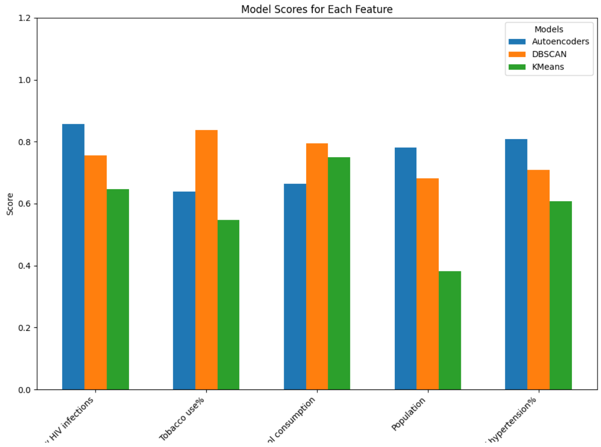


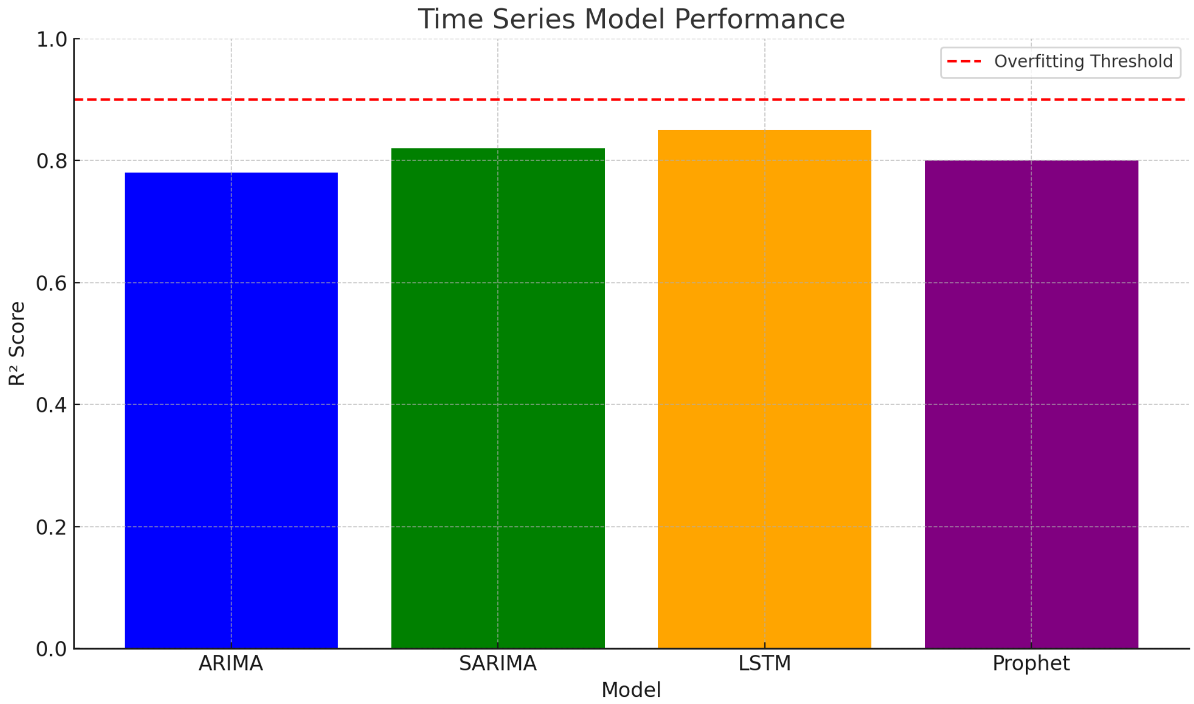
Figure 3: Comparison of different clustering models based on Silhouetee score

- Time Series Metrics: R^2 score, RMSE, and MAE.

Below are the evaluation metrics for each model:

- **ARIMA:** R^2 score = 0.78
- **SARIMA:** R^2 score = 0.82
- **LSTM:** R^2 score = 0.85
- **Prophet:** R^2 score = 0.80

The performance of each time series model is visualized below, showing their comparative effectiveness in forecasting indicators.



Time Series Model Performance

6 Future Scope

- Integrating real-time data feeds for updated predictions.
- Exploring advanced neural networks like Transformer-based models.
- Incorporating external factors like economic indicators or news sentiment.

7 References

- World Health Organization Website
- Python Libraries: `pandas`, `scikit-learn`, `tensorflow`, `statsmodels`.