Data Mining

Suicide Prediction System

MILESTONE 2: Data Preparation/Collection & Cleaning

By Group7

Atharv Kadam、Om Kokate、Yuni Wu

Professor

Alfonso G. Bastias, Ph.D.

Abstract

The **Suicide Prediction System** project aims to analyze environmental and socioeconomic factors that influence suicide rates across different countries. By examining variables such as GDP, weather conditions, happiness indices, and temperature, the project seeks to identify patterns and trends that may indicate higher suicide risks. The project employs data science techniques, including data collection, cleaning, and visualization, to provide insights that can aid mental health professionals and organizations in developing proactive measures to prevent suicides. This report summarizes the data collection and preparation methods, data cleaning processes, and the visualizations created to illustrate our findings.

1. Data Collection

Data for this project was collected from multiple reputable sources to ensure reliability and comprehensiveness. The following datasets were utilized:

- **GDP**: Economic data obtained from the World Bank API.
- **Weather**: Climate data sourced from OpenWeatherMap API, providing temperature and weather conditions.
- **Happiness Index**: Global happiness scores acquired from the World Happiness Report.
- **Suicide Rates**: Historical suicide rates data retrieved from the WHO database.
- **Population**: Population statistics for each country obtained from the United Nations database.

The data was collected using web scraping and APIs to ensure up-to-date and accurate information. Each dataset was saved in CSV format for further processing.

2. Data Preparation

The collected datasets were merged and pre-processed to create a comprehensive dataset for analysis. The preparation steps included:

- **Merging Datasets**: Using pandas in Python, datasets were combined based on a common identifier (e.g., country name).
- **Data Structuring**: Ensuring consistent column naming and data types across all merged datasets.

3. Data Cleaning

Data cleaning was crucial to maintain the integrity of the analysis. The following steps were taken:

- **Handling Missing Values**:
    - Missing values were addressed by either imputation (filling in missing data with the mean or median) or removal of rows where critical data was absent.
- **Removing Outliers**:
    - Outliers were identified using Z-scores and were removed or adjusted based on their influence on the analysis.
- **Ensuring Consistency**:
    - Data formats (e.g., date formats, numerical precision) were standardized to ensure consistency across the dataset.
- **Data Quality Checks**:
    - Basic statistics (mean, median, mode) were calculated to verify data distribution and identify any anomalies.

4. Data Visualizations

Data visualizations were created using libraries such as Matplotlib and Seaborn to illustrate relationships and trends within the data. The following types of visualizations were included:

1. **Bar Charts**: To show the suicide rate across the countries.
2. **Scatter Plots**: To illustrate correlations between population and suicide rates.
3. **Box Plots**: To visualize the distribution of suicide rates and identify outliers.
4. **Heatmaps**: To display the correlation between these factors.

Insights from Visualizations

Each visualization provided insights into the trends and factors influencing suicide rates. For instance, the scatter plot indicated a negative correlation between happiness index scores and suicide rates, suggesting that higher happiness scores are associated with lower suicide rates.