Life Expectancy Prediction Using SVM Algorithm

Author: Adel.Ahmadi

1. Introduction

In this project we focus on classifying life expectancy into categories of Low, Medium, and High based on various socio-economic and healthcare indicators. Life expectancy is a crucial measure of public health and global development. Understanding its determinants is vital for policy-makers aiming to improve global health.

We use the dataset, sourced from the World Health Organization (WHO), containing cleaned and scaled data, anyone can access this dataset with this link:

- Country: Name of Country
- Year: Scaled Years, Original Years if 2000-2015
- Status: Scaled Status
- Life expectancy: Scaled Life Expectancy
- Adult Mortality: Scaled Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
- Infant deaths: Scaled Infant Deaths per 1000 infants
- Alcohol: Scaled Alcohol (per capita (15+) consumption in liters of pure alcohol)
- percentage expenditure: Scaled Percentage Expenditure (percent of GDP per capita spent on healthcare)
- Hepatitis B: Scaled Hepatitis B immunization coverage among 1-year-olds in percent
- Measles: Scaled Measles Cases (reported number of cases per 1000 people)
- BMI: Scaled Average BMI of the entire population
- under-five deaths: Scaled Number of under-five deaths per 1000 people
- Polio: Polio immunization coverage among 1-year-olds in percent
- Total expenditure: Scaled General government expenditure on health as a percentage of total government expenditure
- Diphtheria: Scaled Diphtheria immunization coverage among 1-year-olds in percent
- HIV/AIDS: Scaled Deaths per 1000 live births from HIV/AIDS
- GDP: Scaled GDP of each country in USD

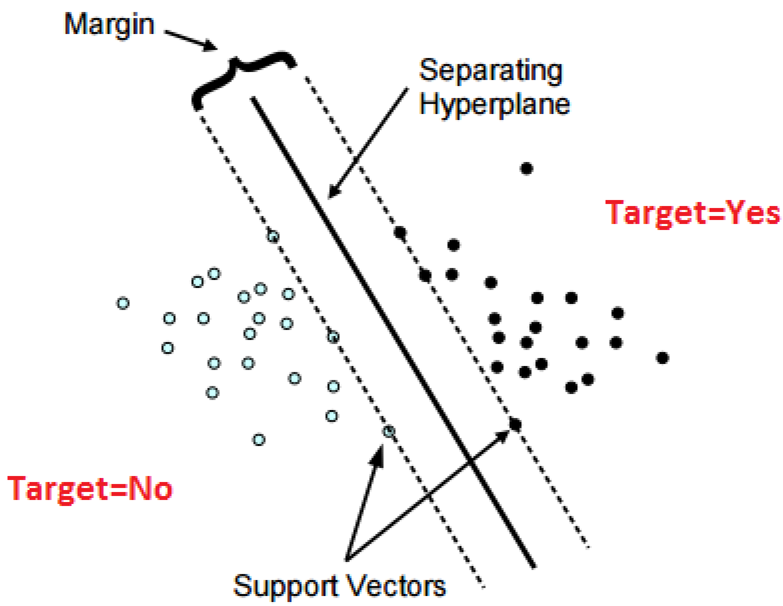
- Population: Scaled Total population of the country
- thinness 1-19 years: Scaled Prevalence of thinness among children and adolescents for Age 10 to 19 in percent
- thinness 5-9 years: Scaled Prevalence of thinness among children for Age 5 to 9 in percent
- Income composition of resources: Scaled Human Development Index (HDI) in terms of income composition of resources
- Schooling: Scaled number of years in school in years on average

Analyzing these features using machine learning (ML) techniques, like Support Vector Machines (SVM), allows for a data-driven understanding of life expectancy across countries.

The necessity of this analysis lies in its ability to provide insights into the socio-economic disparities that influence life expectancy. By identifying key factors through classification, governments and health organizations can better allocate resources and develop targeted strategies for improving public health.

2. SVM Algorithm Overview

Support Vector Machines (SVM) is a supervised machine-learning algorithm used for classification and regression. It works by finding the optimal hyperplane that separates data points of different classes in a high-dimensional space. The goal of SVM is to maximize the margin between data points of different classes.



Support Vector Machine

SVM Mathematics

In this section, we review some basic mathematics behind support vector machines (SVM).

- Hyperplane Equation

  $w.x + b = 0$

  w: $Weight\ vector$

  x: $Input\ feature\ vector$

  b: $Bias\ term$

- Hard Margin SVM Optimization (Used for Linearly separable data)

  $Minimize\ \frac{1}{2}||w||^2$

  $\frac{2}{||w||}$: $margin$

  $subject\ to$:

  $y_i(w.x_i + b) \geq 1, \forall i$

$y_i \in \{-1, 1\}:\ class\ labels$

- Soft Margin SVM Optimization (Used for Non-linearly separable data)

  $Minimize\ \frac{1}{2}||w||^2 + c\sum_{i=1}^{n} \xi_i$

  $subject\ to:$

  $y_i(w.x_i + b) \geq 1 - \xi_i, \forall i$

  $\xi_i \geq 0:\ Penalties\ for\ misclassification.$

  $C:\ Regularization\ parameters$

- Kernel Trick (Used for Non-linear data)

  $K(x_i, x_j) = \phi(x_i).\phi(x_j)$

  $\phi(x_i)\ is\ a\ mapping\ to\ higher - dimensional\ space$

- Decision Function

  $f(x) = sign(\sum_{i=1}^{n} a_i y_i K(x_i, x) + b$

  $Support\ vectors\ contribute\ to\ this\ sum\ (where\ a_i > 0)$

SVM Hyper-parameters

In this section, we review support vector machines (SVM) hyper-parameters that affect model accuracy etc …, later on next chapters, we try to use some algorithms to find optimal hyper-parameters and optimize our model.

- C

  Controls the trade-off between maximizing the margin and correctly classifying training examples.

- Kernel

  We can use different kernel functions (e.g., linear, polynomial, RBF).

- Gamma

  Controls how far the influence of a single data point reaches.

3. Code Implementation

The implementation involves few crucial steps:

1. **Data preprocessing:** Our dataset is cleaned, scaled and preprocessed as mentioned in Kaggle. We Just bin continuous target values for life expectancy into three categories (Low, Medium, High). We use One-hot encoding to handle categorical variables such as Country, etc.. .
2. **Train The SVM Model:** An SVM classifier is trained to predict the life expectancy categories using the Scikit learn package in python.
3. **Grid Search for Hyper-parameters:** A grid search using GridSearchCV package in python is employed to systematically evaluate different combinations of hyper-parameters (C, gamma, and kernel) to find the best-performing model.
4. **Model Evaluation:** After training, the model is evaluated using classification metrics such as precision, recall, and F1-score.

4. Hyper-parameter Tuning and Results

During hyperparameter tuning, we experimented with a range of values for C, gamma, and different kernels to find optimal hyper-parameters. The best model was obtained with a linear kernel, a C value of 0.1, and a gamma value of 1. This configuration provided the highest classification accuracy.

The overall accuracy was satisfactory (0.92), with a clear differentiation between the life expectancy categories. The results highlighted that socio-economic factors highly influence life expectancy value.

In conclusion, this project demonstrates the effectiveness of using SVM for classifying life expectancy based on socio-economic data.