**SVR**

*1. Introduction*

This documentation outlines the process of analyzing a dataset on life expectancy and implementing machine learning models to predict life expectancy based on various factors. The dataset contains information on life expectancy and several socio-economic and health-related factors from different countries.

*2. Dataset Overview*

The dataset is loaded from a CSV file named 'Life Expectancy Data.csv'. It includes the following columns:

* *Year: the year that the data was observed*
* *Life expectancy: the number of years that an average person can be expected to live*
* *Adult Mortality: this is the probability that a person aged 15 years old will die before they reach the age of 60 (in a population per 1000 people)*
* *Infant deaths: the number of infant deaths (in a population per 1000 people)*
* *Alcohol: pure alcohol consumption per capita of persons aged over 15 in Litres*
* *percentage expenditure: the percentage of a country's GDP per capita expended on healthcare*
* *Hepatitis B: the percentage of Hepatitis B immunization among children aged 1 year old*
* *Measles: the number of Measles cases reported (in a population per 1000 people)*
* *BMI: the average body mass index value of the country's population*
* *under-five deaths: the number of deaths (in a population per 1000 people) of children under 5 years of age*
* *Polio: the percentage of Polio immunization among children aged 1 year old*
* *Total expenditure: the percentage of total government expenditure allocated to general healthcare expenditure.*
* *Diphtheria: the percentage of Diphtheria, Tetanus, Toxoid, Prtussis (DTP3) immunization among children aged 1 year old*
* *HIV/AIDS: the number of HIV/AIDS deaths (per 1000 live births) of children under 4 years of age.*
* *GDP: the Gross Domestic Product value per capita calculated in USD*
* *Population: the country's population*
* *thinness 1-19 years: the percentage thinness of children aged between 1 and 19 years*
* *thinness 5-9 years: the percentage thinness of children aged between 5 and 9 years*
* *Income composition of resources: the Human Development Index value (0-1) in reference to a country's configuration of resources*
* *Schooling: the mean number of years of schooling*

***3. Data Preprocessing***

- Handling Missing Values: Missing values are filled using the mean of each column.

- Feature Engineering: New features are created, such as 'Alcohol\_affect\_Mortality', 'Vaccine\_Coverage\_Score', and 'Child Mortality'.

- One-Hot Encoding: The 'Country' column is one-hot encoded to convert categorical data into numerical form.

- Standardization: Features are standardized to ensure that each feature contributes equally to the analysis.

*4. Exploratory Data Analysis (EDA)*

- Correlation Matrix: A heatmap is generated to visualize the correlation between different features. The correlation matrix helps identify relationships between variables.

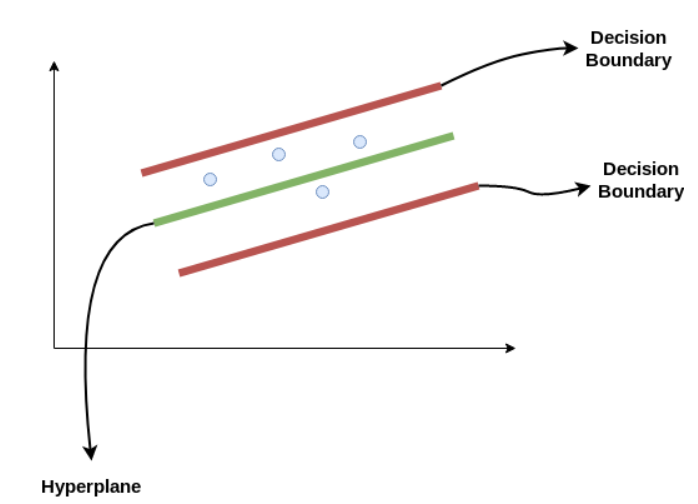
*5. Model Implementation: Support Vector Regression (SVR)*

*A diagram of a diagram of a space

Description automatically generatedA diagram of a graph

Description automatically generated*

*-Support Vector Regression (SVR) is a type of* Support Vector Machine (SVM) algorithm used for regression tasks. It is particularly effective for datasets with complex relationships between features and target variables. SVR works by finding a hyperplane in a high-dimensional space that best fits the training data, while also maximizing the margin, i.e., the distance between the hyperplane and the nearest data points (support vectors).



*5.1Hyperparameters Tuning:*

GridSearchCV is employed to optimize the SVR model by searching through a specified parameter grid and selecting the combination of hyperparameters that yields the best performance. The following hyperparameters are tuned:

* C: Penalty parameter of the error term. It controls the trade-off between maximizing the margin and minimizing the error. Higher values of C result in a smaller margin but fewer margin violations.
* Gamma: Kernel coefficient for 'rbf' and 'poly' kernels. It defines the influence of each training example. Higher values of gamma lead to a more complex decision boundary.
* Kernel: Specifies the kernel type to be used in the algorithm. The 'linear' kernel computes the dot product of the feature vectors, while the 'rbf' kernel uses a Gaussian radial basis function.

*5.3 Model Evaluation:*

The SVR model is evaluated using the following metrics:

* Mean Squared Error (MSE): It measures the average squared difference between the actual and predicted values. Lower MSE indicates better model performance.
* Mean Absolute Error (MAE): It calculates the average absolute difference between the actual and predicted values. MAE is less sensitive to outliers compared to MSE.
* R-squared (R2): It quantifies the proportion of the variance in the target variable that is predictable from the independent variables. R2 ranges from 0 to 1, where higher values indicate better model fit.
* Cross-Validation R-squared Scores: Cross-validation is performed to assess the model's generalization performance by splitting the dataset into multiple subsets. The average R-squared score across different folds is calculated.

A group of mathematical equations

Description automatically generated

*6. Results and Visualization*

* After hyperparameter tuning, the SVR model's performance is compared to the baseline SVR model without hyperparameter optimization. The optimized SVR model demonstrates improved accuracy metrics, including MSE, MAE, and R2 scores. The results highlight the effectiveness of hyperparameter tuning in enhancing the SVR model's predictive capabilities.
* Overall, SVR proves to be a powerful algorithm for regression tasks, particularly when dealing with non-linear relationships and high-dimensional data. Its ability to capture complex patterns in the data makes it a valuable tool for predicting life expectancy based on various socio-economic and health-related factors.
* - Visualization: Scatter plots are generated to visualize actual vs. predicted life expectancy and distribution of residuals for each model.