Histograms of all numerical variables

Diagram

Description automatically generated with low confidence

Density plot of all numerical variables

Diagram, engineering drawing

Description automatically generated

Box plot

Diagram

Description automatically generated

Note for boxplots: Since there are many outliers, the data is very skewed, to reduce the skewness, we apply standard scalar to the data to standardize it.

Correlation matrix:

Chart, histogram

Description automatically generated

Life expectancy is positively correlated with income composition of resources (0.680) and schooling (0.747), and negatively correlated with adult mortality (-0.684) and thinness 5-9 years (-0.397).

- Adult mortality is negatively correlated with life expectancy (-0.684), Hepatitis B immunization coverage (-0.191), and income composition of resources (-0.475).

- Infant deaths are positively correlated with under-five deaths (0.999) and thinness 1-19 years (0.513), and negatively correlated with life expectancy (-0.234) and Hepatitis B immunization coverage (-0.409).

- Alcohol consumption per capita is positively correlated with life expectancy (0.268) and schooling (0.306), and negatively correlated with thinness 1-19 years (-0.326) and thinness 5-9 years (-0.291).

- Percentage expenditure on health is positively correlated with life expectancy (0.455), Hepatitis B immunization coverage (0.215), and GDP (0.822).

Life expectancy in categorical variables

Chart, box and whisker chart

Description automatically generated

Applied 4 models: Linear Regression, Kneighbors Regressor, Decision Tree, Random Forest

Using the dataset, different models were trained and evaluated to predict life expectancy. The performance of each model is summarized below:

- Linear Regression: R² score of 0.82, mean squared error of 17.34

- Decision Tree: R² score of 0.91, mean squared error of 9.71

- Random Forest: R² score of 0.95, mean squared error of 6.29

- KNN: R² score of 0.79, mean squared error of 20.21

The Random Forest model has the highest R² score (0.95) and the lowest mean squared error (6.29), making it the best-performing model for predicting life expectancy in this dataset.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Some Summary Statistics to add:

The average life expectancy across all countries is 71.19 years, with a standard deviation of 5.78 years.

- Adult mortality has an average of 133.64 deaths per 1,000 population and a standard deviation of 77.52 deaths.

- The average number of infant deaths is 63.46, with a standard deviation of 217.35 deaths.

- Alcohol consumption per capita has an average of 1.99 liters, with a standard deviation of 2.41 liters.

- The average percentage expenditure on health is 493.97, with a standard deviation of 1169.13.

- Hepatitis B immunization coverage has an average of 84.11%, with a standard deviation of 21.98%.

- The average number of measles cases is 4442.91, with a standard deviation of 13958.12 cases.

- The average Body Mass Index (BMI) is 19.43, with a standard deviation of 4.51.

- The average under-five deaths is 85.59, with a standard deviation of 21.22 deaths.

- Polio immunization coverage has an average of 86.19%, with a standard deviation of 19.46%.