**Report on the GDP Data Model**

**Introduction**

This comprehensive report presents a detailed analysis of GDP (Gross Domestic Product) data from various countries around the world. The primary objective of this analysis is to construct predictive models for GDP per capita and gain insights into the factors that significantly influence a country's economic performance. The analysis encompasses multiple stages, including data preprocessing, missing value handling, feature selection, and the development of predictive models using three distinct algorithms: Linear Regression, Artificial Neural Networks (ANN), and XGBoost.

**Data Understanding**

The dataset employed for this analysis is stored in a CSV file named "gdpWorld.csv." It contains an array of information on various features that might be correlated with a country's GDP per capita. The key columns in the dataset include:

Population: The total population of the country.

Area (sq. mi.): The total land area of the country in square miles.

Pop. Density (per sq. mi.): Population density per square mile.

Coastline (coast/area ratio): The ratio of coastline length to land area.

Net migration: Net migration rate (immigration rate minus emigration rate).

Infant mortality (per 1000 births): The number of infant deaths per 1000 births.

GDP ($ per capita): The GDP per capita of the country (target variable).

Literacy (%): The literacy rate in the country.

Phones (per 1000): The number of phone lines per 1000 people.

Arable (%): The percentage of land suitable for agriculture.

Crops (%): The percentage of land used for crop cultivation.

Other (%): The percentage of land not suitable for agriculture.

Climate: A categorical variable representing the climate type.

Birthrate: The number of births per 1000 people.

Deathrate: The number of deaths per 1000 people.

Agriculture: The percentage of the economy based on agriculture.

Industry: The percentage of the economy based on industry.

Service: The percentage of the economy based on services.

**Data Preprocessing**

Dataset Dimensions:

Number of Rows: 227

Number of Columns: 18

Handling Missing Values:

Missing values in the dataset were systematically addressed by imputing them with the median value of their respective columns. This process ensured that the data was complete and ready for subsequent analysis and modeling.

**Feature Selection:**

Correlation Analysis

A correlation matrix was generated to unravel the relationships between the various features and the target variable, GDP per capita. Features displaying a correlation greater than 0.25 were judiciously selected for further investigation. This subset of features was chosen based on its potential influence on a country's GDP.

**8 Features were selected for building the model**

**Models Used:**

The core of this analysis centers around the development of three predictive models, each employing distinct algorithms:

**1. Linear Regression Model**

Linear Regression is a fundamental and widely used statistical method for modeling the relationship between a dependent variable (in this case, GDP per capita) and one or more independent variables (features). It assumes a linear relationship between the variables and aims to find the best-fitting linear equation that describes this relationship.

Model Description:

Linear Regression uses a linear equation of the form:

Y=a+bX

Where:

Y is the dependent variable (GDP per capita).

X is one of the selected independent features.

a is the intercept.

b is the coefficient that represents the relationship between $X$ and $Y$.

Performance Metrics:

Mean Squared Error (MSE): Measures the average squared difference between predicted and actual values. Lower MSE indicates better model fit.

Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual values. It is more robust to outliers.

R-squared (R2): Represents the proportion of variance in the dependent variable explained by the independent variables. R2 values range from 0 to 1, with higher values indicating a better fit.

**2. Artificial Neural Networks (ANN)**

Artificial Neural Networks (ANN) are a class of machine learning models inspired by the human brain's neural structure. ANNs consist of interconnected nodes or neurons organized in layers, and they are capable of learning complex patterns and relationships in data.

Model Description:

ANN comprises multiple layers, including input, hidden, and output layers.

Each neuron in a layer is connected to neurons in the adjacent layers through weighted connections.

Neurons in hidden layers apply activation functions to their inputs, allowing them to model complex, nonlinear relationships.

During training, ANNs adjust the weights of connections to minimize the prediction error.

Performance Metrics:

Mean Squared Error (MSE): Measures the average squared difference between predicted and actual values.

Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual values.

**3. XGBoost Model**

XGBoost (Extreme Gradient Boosting) is an advanced gradient boosting algorithm designed for both regression and classification tasks. It has gained popularity for its exceptional predictive performance and robustness.

Model Description:

XGBoost builds an ensemble of decision trees, combining their predictions to make a final prediction.

It uses a boosting technique, where each tree corrects the errors of the previous one.

XGBoost employs various hyperparameters to control tree depth, learning rate, and the number of trees in the ensemble.

Performance Metrics:

Mean Squared Error (MSE): Measures the average squared difference between predicted and actual values.

Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual values.

R-squared (R2): Represents the proportion of variance in the dependent variable explained by the independent variables.

**Modelling**

The core of this analysis centers around the development of three predictive models, each employing distinct algorithms:

**1. Linear Regression**

A Linear Regression model was meticulously constructed to predict GDP per capita using the selected features. The model's performance was assessed using key metrics:

Mean Squared Error (MSE): 19360890.548763532

Mean Absolute Error (MAE): 3286.9206387112918

R-squared (R2): 3286.9206387112918

**2. Artificial Neural Networks (ANN)**

An Artificial Neural Network (ANN) architecture with three hidden layers was thoughtfully engineered to predict GDP per capita. The model was diligently compiled, utilizing mean absolute error as the loss function and the Adam optimizer. It underwent training for 50 epochs. The evaluation metrics encompassed:

Mean Squared Error (MSE): 11320880.0

Mean Absolute Error (MAE): 2395.72998046875

**3. XGBoost**

An XGBoost model, fortified with hyperparameter tuning, was meticulously crafted to predict GDP per capita. The GridSearchCV technique was employed to optimize hyperparameters. The model's performance was evaluated using the following metrics:

Mean Squared Error (MSE): 6620582.688281867

Mean Absolute Error (MAE): 1867.823260763417

R-squared (R2): 0.9203831288162106

**Conclusion**

In summary, this in-depth analysis revealed that the XGBoost model outperformed the other two models (Linear Regression and ANN) in predicting GDP per capita. The XGBoost model demonstrated the lowest Mean Squared Error (MSE) and Mean Absolute Error (MAE), suggesting its suitability for modeling the complex relationship between the selected features and a country's economic performance.