Academic Year	Module	Assignment Number	Assignment Type
2	5CS037/HJ1: Concepts and Technologies of AI	Final Project	Individual Coursework

Regression Analysis Report

Student Id	: 2408008
Student Name	: Nayanika Dubey
Section	: L5CG14
Module Leader	: Mr. Siman Giri
Tutor	: Ms. Sunita Parajuli
Submitted on	: 11-02-2025 (February 11, 2025)

Table of Contents

Abstract

The goal of this research project was to use a dataset related to global energy statistics to create a regression model that would estimate the proportion of renewable energy in total final energy consumption. Starting from scratch, the study created a linear regression model and compared its performance to pre-made models from the sklearn package, such as support vector regression (SVR), ridge regression, and linear regression. These early models provided a starting point for improvement. The main goal was to optimize the best model, in this case the random forest regression model, which required feature selection and hyperparameter adjustment. To find the ideal values for parameters like the number of trees in the forest, the maximum tree depth, and the minimum samples needed for node splitting, the hyperparameter optimization procedure used a grid search technique. The most important predictors, which included gdp_per_capita, primary_energy_consumption_per_capita_(kwh/person), low-carbon_electricity_(%_electricity), and electricity_from_renewables_(twh), were chosen through feature selection using the SelectFromModel method. On the test set, the final model, which was constructed with the chosen features and optimized hyperparameters, produced an R-squared (R^2) score of 0.504, a Mean Absolute Error (MAE) of 14.22, and a Root Mean Squared Error (RMSE) of 21.67. These findings demonstrated how well the optimized random forest model performed in comparison to the initial models, highlighting the advantages of feature selection and hyperparameter adjustment for improving prediction accuracy.

Introduction

Automated analytical model building for data analysis, focusing on identifying patterns from data, learning from it, and making decisions with minimal human intervention is Machine Learning.

In general, machine learning problems fall under one of the two:

- Supervised Learning
- Unsupervised Learning

Supervised Learning trains a computer algorithm based on input data that has been labeled for a particular task/ output. Model training continues until it can detect an underlying pattern or connection between the provided input data and the output labels, in a manner that allows it to provide accurate labelling results for unseen data. There are several ways to approach supervised learning, however in this section of the final assessment we will focus on Regression.

Regression refers to the mathematical approach of finding relationships between variables. In contrast to classification models, regression models provide numerical values, as well as continuous values for both independent and dependent variables.

Problem Context

The Problem Statement

The project's focus is to implement linear regression with Python and Scikit-Learn. The dataset "Global Data on Sustainable Energy Consumption" is used for this project, to build various models that can predict the values accurately.

Dataset Overview and Connection of Dataset to SDG goal:

"Global Data on Sustainable Energy Consumption" dataset holds over 3000 rows of unique data related to energy consumption relevant under various features, and was obtained from Kaggle.

This dataset directly correlates to Affordable and Clean Energy, Sustainable Development Goal 7.

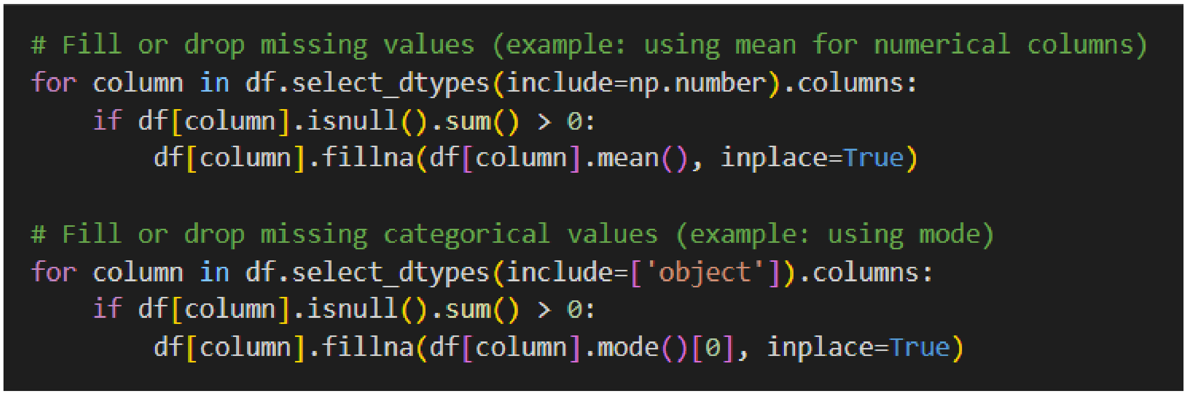
Various indicators within the dataset directly reflect the progress towards universal energy availability, which when analysed can help us assess the progress towards Goal 7 and the areas for improvement. For example, the dataset measures the measure of renewable energy sources in energy consumption, which helps us reflect on clean and sustainable energy usage.

Analytical strategy

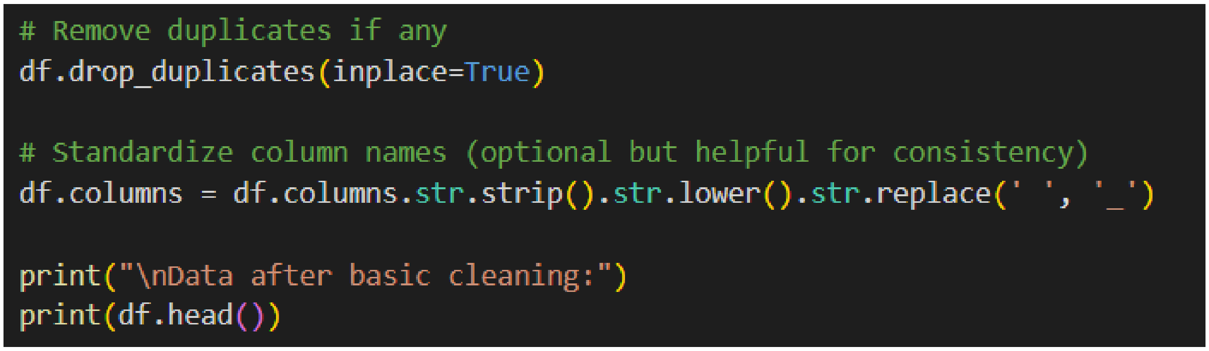
The primary analytical strategy is to build, train, and evaluate regression models to predict the 'renewable_energy_share_in_the_total_final_energy_consumption_(%)' using a variety of explanatory variables. This process involves data preprocessing, model building and implementation (fitting), model evaluation, hyperparameter tuning, feature selection, and final model building using the best features and optimized hyperparameters.

Data Preprocessing

The dataset, post loading, is inspected using various functions like df.head(), df.describe() and df.info(). The column names are standardized to lowercase, underscore separated values for better flow in implementation. After achieving a basic understanding of the dataset, null values are handled by replacing them with mean for numerical values and mode for categorical ones. Feature data is normalized by subtracting the mean and dividing by the standard deviation.



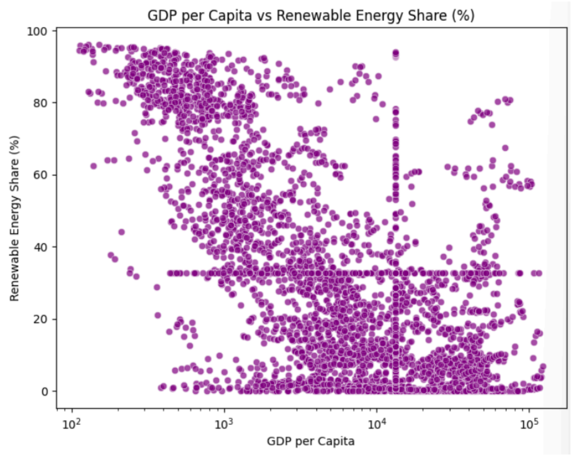
```python
# Fill or drop missing values (example: using mean for numerical columns)
for column in df.select_dtypes(include=np.number).columns:
    if df[column].isnull().sum() > 0:
        df[column].fillna(df[column].mean(), inplace=True)

# Fill or drop missing categorical values (example: using mode)
for column in df.select_dtypes(include=['object']).columns:
    if df[column].isnull().sum() > 0:
        df[column].fillna(df[column].mode()[0], inplace=True)
```

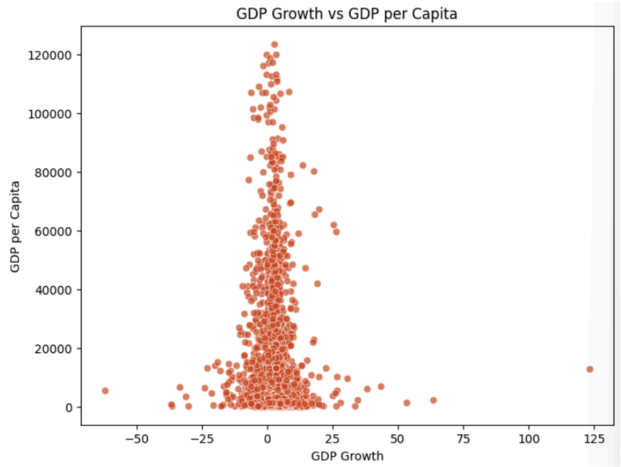


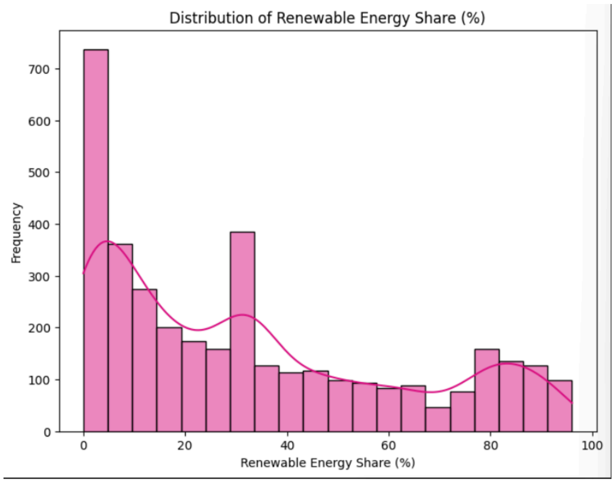
```python
# Remove duplicates if any
df.drop_duplicates(inplace=True)

# Standardize column names (optional but helpful for consistency)
df.columns = df.columns.str.strip().str.lower().str.replace(' ', '_')

print("\nData after basic cleaning:")
print(df.head())
```

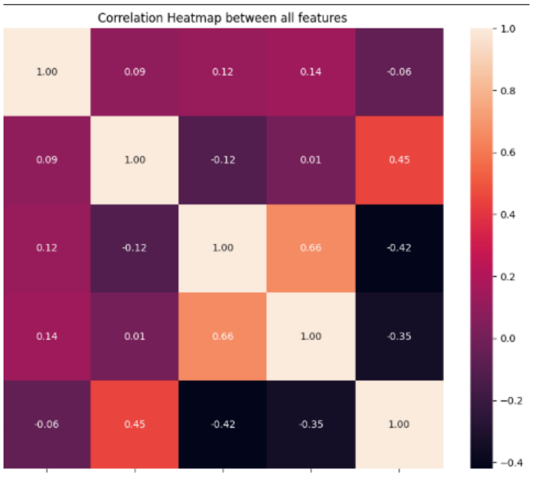
Visualization

Scatter plots, histograms, and heatmaps are used to visualize data and explore relationships between variables. Scatter plots are used to show the relationship between GDP per capita and renewable energy share, Low-carbon electricity and renewable energy share, Primary energy consumption and renewable energy share, GDP growth and GDP per capita, Electricity from fossil fuels and renewable energy share, and Electricity from nuclear and renewable energy share. Similarly, the distribution of the dependent variable “renewable energy share in the total final energy consumption” is visualized using a histogram. The correlation heatmap is used to visually see the correlation between several key features (Labels are included in the .ipynb file).









1.00	0.09	0.12	0.14	-0.06
0.09	1.00	-0.12	0.01	0.45
0.12	-0.12	1.00	0.66	-0.42
0.14	0.01	0.66	1.00	-0.35
-0.06	0.45	-0.42	-0.35	1.00

Model Development

A total of 5 regression models are build in this project, namely a linear regression model from scratch, a pre-build linear regression model to compare with the performance of model from scratch, a Ridge Regression Model, a Support Vector Regression Model, and Random Forest Regression Model.

Model from scratch:

A linear regression model, as practised in our tutorials and workshops, was built from scratch. Mean Squared Error was defined to compute the cost function. Gradient Descent Algorithm was used to minimize cost function and estimate model parameters. The model was trained, and then tested using evaluation metrics like MSE, RMSE, and R-squared.

Pre-Built models:

Linear Regression, Ridge Regression, Support Vector Regression (SVR) and Random Forest Regression were built and fitted using sklearn. These models were used to do detailed comparative analyses.

The goal of the model-building process was to evaluate and compare how different regression models performed on the given dataset.

Evaluation Metric and Results

The models were evaluated based on their MSE, RMSE, and R-squared errors.

Mean Squared Error:

It refers to the average of the difference between the actual value and predicted value.

Root Mean Squared Error

It gives the root of the MSE, bringing it to the same metric unit as the target variables.

R-squared

It is the ratio of sum of squared error to total sum of squared deviation from the mean subtracted from 1. Negative R-squared shows a poorly fit model.

Model Metrics:

Below are the evaluation metrics of the different models:

	MSE	**RMSE**	**R-squared**
Linear Regression (Scratch)	595.7954611634183	24.408921753396204	0.3709284054184322
Linear Regression	689.2191656938644	26.252983938856634	0.18526827779661592
Ridge Regression	660.0179577511349	25.690814657210364	0.2196986653619244
Support Vector Regression	407.105817820794	20.176863428709478	0.5187568184668296
Random Forest Regression	35.743494201702674	5.978586304612711	0.957747317489057

Based on the above observations, we can draw the conclusion that Random Forest Regression did the best job at the Regression task when compared to the other models, as it has a R-squared value closer to 1 standing at 0.95.

Model Enhancement

Hyperparameter Tuning

Hyperparameters refer to the parameter used during machine learning that is chosen before the training process starts and isn’t learned by the model, but affects its learning process and performance.

RandomizedSearchCV was used to tune the 'alpha' hyperparameter for Ridge Regression, and the 'C', 'epsilon', and 'kernel' hyperparameters for SVR, while GridSearchCV was used to tune the 'max_depth', 'min_samples_split', 'n_estimators', and 'min_samples_leaf' hyperparameters for Random Forest Regression.

Feature Selection

The features considered for selection across all the models were: 'gdp_per_capita', 'primary_energy_consumption_per_capita_(kwh/person)', 'electricity_from_renewables_(twh)', and 'low_carbon_electricity_(%_electricity)'.

Recursive Feature Elimination (RFE) was used to select the top 3 features for Ridge Regression and SVR, and SelectFromModel was used to pick features based on importance scores for Random Forest Regression.

Final Model Building

The final model was built with the initial performance comparison, hyperparameter tuning and feature selection in mind. A Random Forest Model was built with the following hyperparameters:

- **n_estimators:** 50 was chosen as the number of trees in the forest.
- **max_depth:** 15 was the best value for maximum depth of the trees.
- **min_samples_split:** 3 was the minimum number of samples required to split an internal node.
- **min_samples_leaf:** 2 was the minimum number of samples required to be at a leaf node.

The features selected for the model were:

- **'low-carbon_electricity_(%_electricity)**
- **electricity_from_renewables_(twh)**

The updated model returned the evaluation metrics as follows:

- Mean Absolute Error (MAE): 14.22315794306771

- Root Mean Squared Error (RMSE): 21.669506695345472

- R-squared (R^2) Value: 0.5042063793885416

While the R-squared value reduces significantly compared to the initial model, it doesn't necessarily reflect that the model is bad. A R-Squared value closer to 1 reflects that the model captures most of the patterns leading to better prediction. However, a value so close to 1 could also reflect that the model is memorizing noise or that irrelevant features are included. Hence with hyperparameter tuning and feature selection we are reducing the chances of these issues.

Conclusive remarks

The regression task successfully developed a predictive model for renewable energy share using a random forest regressor. Through steps like hyperparameter tuning as well as feature selection, the final model achieved a reasonable level of accuracy, demonstrating that these methods can improve the performance of regression models. Based on the features provided, the final random forest model is a viable method for estimating the share of renewable energy and represents a considerable improvement over the other models in this research. The R-squared number shows that the model does not fully account for some of the variance in the dependent variable, therefore there is still room for improvement.