Academic Year	Module	Assessment Number	Assessment Type
2024	5CS037 Concepts and Technologies of AI	Final	Report

Title: Report on Regression Task

Student Id : 2408624
Student Name : Prayas Shrestha
Section : L5CG15
Module Leader : Mr. Siman Giri
Tutor : Mr. Siman Giri

Abstract

Objective: This report aims to project national historical patterns of energy use based on demographic, economic, and energy production data.

Approach of Study: In the dataset, named'owid-energy-data.csv', energy consumption records have been observed worldwide, including conducting exploratory data analysis (EDA) and building linear regression and KNN regression models, followed by hyper-parameter optimization as well as feature selection.

Key findings: R-squared, MSE,was the metrics used to evaluate the varying performances of the models. The performance of the KNN model improved after feature selection and hyperparameter tuning when compared to the Linear regression model.

Conclusion: The regression model successfully predicts energy consumption and improves accuracy through feature selection.

1. Introduction

1.1 Problem Statement

The main aim of the study is predicting total energy consumption trends from a dataset that consists of values for energy use, economic indicators, and demographic variables.

1.2 Dataset

Dataset: `owid-energy-data.csv`

- Source: Our World in Data
- Information on energy production and consumption across countries.
- Includes attributes, such as GDP, population, renewable energy share, and total energy consumption.

1.3 Objective

The objective is to develop a predictive regression model for forecasting total energy consumption using relevant features.

2. Methodology

2.1 Data Preprocessing

- Deleted columns with >40% missing values.
- Imputed median for numerical columns to solve missing values.
- Duplicates were checked and removed.

2.2 Exploratory Data Analysis (EDA)

- Summary statistics computed.
- Visualizations include:
- Histograms depicting numerical distribution.

2.3 Model Development

-Two regression models were developed.

-Baseline Model: Linear Regression

-Complex Model: KNN Regression

2.4 Model Validation

The following metrics have been used:

R-squared (R^2): variance explained.

Mean Squared Error (MSE): measure of error magnitude.

3. Outcomes & Discoveries

3.1 Model Comparison

Model	R^2	MSE
Linear Regression	**0.53**	**3.407579546399219**
Knn regressor	**0.67**	**1.7242975657777304**

With lesser error, Decision Tree did better.

3.2 Hyper-parameter Tuning

Best parameters for Linear Regression: {'alpha': 10, 'solver': 'saga'}

Best cross-validated MSE for Linear Regression: 3.264146243647712

Best parameters for KNN Regressor: {'metric': 'manhattan', 'n_neighbors': 9, 'weights': 'uniform'}

Best cross-validated MSE for KNN Regressor: 1.6021608936318668

3.3 Feature Selection

Select_K

Best features selected: gdp, population, renewable_share, electricity_generation

3.4 Final Model

Best model: KNN Regression

Retrained using selected features & tuned hyperparameters.
R^2 went down from 0.67 → 0.64.

4. Discussion

4.1 Model Performance

KNN Regression performed better on predictive accuracy.
Linear Regression failed because the data had non-linear relationships.

4.2 Effect of Hyper-parameter Tuning and Feature Selection

Feature selection improved interpretability and therefore reduced overfitting.
Hyperparameter tuning further tuned KNN.

4.3 Challenges

Data had so many missing values that it needed to be very handled.
Feature selection improved the performance of the model.

4.4 Future Work

Experiment with ensemble models(Random Forest, Gradient Boosting).
Add more economic and environmental indicators.

5. Conclusion

KNN Regression outperformed Linear Regression in predicting variations in energy consumption.
The improvements in feature selection and tuning of hyperparameters led to improved performance of the model.