Predictive Analysis of Water Use by Different Energy Sources: An Application of the KDD Process

Sri Vinay Appari

September 30, 2023

Contents

1 Abstract

This research paper focuses on the comprehensive analysis of the 'Predict Water Use' dataset from Kaggle. The objective is to predict the water consumption associated with different energy sources. Utilizing the Knowledge Discovery in Databases (KDD) methodology, we meticulously explore the dataset's attributes, preprocess the data, and apply an array of regression models. We conclude with key insights and actionable recommendations that have the potential to shape future research and strategies in the energy domain.

2 Introduction

With the world's increasing energy demand and the finite nature of freshwater resources, understanding the water footprint of different energy sources is paramount. This study aims to bridge the knowledge gap by providing a predictive model for water use based on various energy-related attributes. By employing the KDD methodology, a systematic and proven approach to data analysis, we aim to offer valuable insights and predictions that can aid policy-makers, researchers, and industry leaders.

3 Data Understanding

The dataset under investigation offers a rich set of attributes associated with different energy sources. Initial examination revealed the presence of both numerical and categorical attributes, each providing unique insights into the energy landscape.

3.1 Data Characteristics

The dataset is a culmination of research efforts and provides the following structure:

- **Numerical Features**: Comprising 22 attributes, these features offer quantitative insights into aspects such as greenhouse gas emissions, land use metrics, and various toxicity measures.
- **Categorical Features**: The 'Entity' column, representing different energy sources, offers qualitative information, aiding in categorizing the data.
- **Missing Data**: A challenge often faced in real-world datasets, certain columns present missing data. Notably, 'Agricultural land use', 'Urban land use', and 'Death rates' lacked data entries.

4 Data Pre-processing

Data pre-processing is the backbone of any analytical task. Ensuring the data's quality and integrity directly impacts the efficacy of predictive models.

4.1 Handling Missing Values

An imperative step in the pre-processing pipeline is addressing missing data. Our approach was twofold: entirely removing features with 100% missing values, ensuring no informational loss, and employing median imputation for the 'Total land use' attribute, preserving the central tendency of the feature.

4.2 Encoding and Scaling

Categorical data, while informative, needs translation into a numerical format for machine learning tasks. The 'Entity' column underwent label encoding, converting energy sources into a set of numerical labels. Furthermore, given the varied scales of different features, standard scaling was employed, ensuring all attributes have a consistent scale, thereby aiding in the convergence of algorithms and enhancing model interpretability.

5 Feature Selection

In the realm of data science, not all features are created equal. Our endeavor was to identify the most influential attributes, shaping the predictive landscape.

Using a Random Forest model, a tree-based ensemble method known for its robustness and ability to capture intricate patterns, we ascertained the importance scores of features. 'Uranium', 'Metal and mineral requirements', and 'Freshwater eutrophication' emerged as paramount attributes, playing a pivotal role in water use predictions.

6 Modeling and Evaluation

The crux of our research was to develop a predictive model that can accurately forecast water use based on various attributes. We embarked on this journey by first establishing a baseline using a simple linear regression model. This model, while rudimentary, provided a benchmark against which more sophisticated models were evaluated.

Ridge Regression, a linear regression variant that incorporates L2 regularization, and the Random Forest Regressor, a tree-based ensemble technique, were the chosen models. Performance evaluation revealed the efficacy of simpler models on the dataset, with the Ridge Regression closely mirroring the performance of the baseline.

7 Conclusion and Recommendations

The interplay between energy production and water use is intricate. Through our research, we have shed light on this relationship, providing a predictive model and actionable insights. For practitioners and policymakers, focusing on the influential features identified can yield better strategies for sustainable energy production. While our models offer robust predictions, as more data becomes available, there is potential for further refinement and accuracy enhancement.

8 References

1. Impacts of Energy Production Dataset, Kaggle