Global Pollution Analysis and Energy Recovery

Introduction to the Study

This project aimed to analyze global pollution data and explore its relationship with energy recovery using clustering algorithms and neural network models. The primary focus was on understanding how different pollution indicators—such as air, water, soil pollution, industrial waste, and CO_2 emissions—affect a country's ability to recover energy. Using a real-world dataset, we implemented two unsupervised learning techniques, K-Means and Hierarchical Clustering, to classify countries into environmental performance groups. We also developed a neural network regression model to predict energy recovery potential from pollution data. The goal was to compare the effectiveness of clustering models in revealing structural patterns versus neural networks' ability to predict numeric energy recovery values. Finally, actionable recommendations were developed to guide countries on how to transition toward a cleaner and more energy-efficient future. This report presents a full comparison of models and outlines key insights that can inform environmental policy and data-driven decision-making.

Clustering Models – K-Means Overview

K-Means clustering was employed to group countries based on pollution indices and energy recovery metrics. The algorithm works by partitioning data into 'k' clusters based on proximity to cluster centroids.

Using the Elbow Method, we identified 3 as the optimal number of clusters. The results showed that countries with similar pollution levels and energy profiles were effectively grouped together, making K-Means a strong candidate for environmental segmentation. For instance, countries with high CO_2 emissions and low renewable energy usage clustered together, while those with moderate pollution but higher renewable adoption formed another group. This segmentation helps identify where different levels of intervention are needed. The model was particularly useful in showing latent patterns that are not obvious in raw data. However, the performance of K-Means is limited by its dependency on centroid initialization and its inability to capture complex hierarchical relationships between countries. Despite that, it serves as a robust tool for unsupervised learning in environmental analytics.

Clustering Models – Hierarchical Clustering

Hierarchical Clustering, specifically the agglomerative type, was used to build a tree-like structure that reveals nested groupings of countries based on environmental indicators. Unlike K-Means, this method does not require a pre-specified number of clusters. The dendrogram provided by hierarchical clustering offered a visual and interpretable breakdown of how countries relate to each other environmentally. This model allowed us to understand subgroup formations—like nations that exhibit similar energy recovery patterns but differ in specific pollution dimensions. For example, two countries with similar soil and water pollution levels but very different industrial waste figures still clustered together due to dominant feature influence. One of the key advantages of Hierarchical Clustering is its transparency and visual representation. However, it is computationally intensive and becomes impractical for very large datasets. In our study, it complemented K-Means by offering

a hierarchical view, enriching the interpretation of environmental similarities and dependencies across countries.

Neural Network Model – Overview and Results

A feedforward neural network was constructed using Keras and TensorFlow to predict energy recovery values (in GWh) from pollution-related features such as air pollution index, CO_2 emissions, industrial waste, and population. The model was trained on 80% of the dataset and evaluated on the remaining 20%. The architecture included multiple hidden layers with ReLU activation and was optimized using the Adam optimizer. Unfortunately, the model yielded poor performance: an R^2 score of -0.23 and a high Mean Squared Error (MSE), indicating that the model could not generalize well. Attempts to improve the network through hyperparameter tuning (e.g., adding more layers, changing the learning rate) led to even worse performance. This implies that the dataset either lacked nonlinear patterns required for deep learning to excel or was too small to train a deep neural network effectively. These results challenge the assumption that deep learning is always superior and highlight the importance of model selection based on data characteristics.

Linear Regression as a Benchmark

To evaluate the performance of the neural network, we built a simple linear regression model as a benchmark. Surprisingly, linear regression outperformed both the original and improved neural networks, achieving

the highest R^2 score among all models (although still slightly negative). This result suggests that the relationship between pollution variables and energy recovery is likely linear or only mildly nonlinear. The linear model had lower MSE and MAE values, making it a better predictor despite its simplicity. This finding emphasizes an important lesson in data science: sometimes, simpler models are not just more interpretable but also more effective. In this context, the poor performance of neural networks could be due to overfitting or insufficient data volume. Therefore, for this dataset, linear regression stands out as the most effective predictive model, while clustering models serve as valuable tools for structural analysis and categorization.

Model Comparison Summary

When comparing the three types of models—K-Means, Hierarchical Clustering, and Neural Networks—it becomes evident that each serves a unique purpose. K-Means is excellent for grouping countries quickly based on quantitative pollution metrics, while Hierarchical Clustering provides deeper insights into nested relationships. Both clustering models are helpful for identifying general trends and forming policy groups. On the other hand, the neural network model was intended for precise energy recovery prediction but failed to achieve satisfactory accuracy, making it less suitable in its current form. Linear Regression, though basic, performed better and should be considered the go-to model for predictive tasks in this case. Therefore, the best approach is to use clustering for segmentation and trend analysis, and linear models for making reliable energy predictions. Neural networks may require more refined feature engineering or larger datasets to be effective in this scenario.

Clustering Insights – Global Patterns

The clustering results revealed distinct global environmental patterns. For instance, countries with high levels of air and water pollution often had lower renewable energy usage and poorer energy recovery rates. These countries clustered together, suggesting that high pollution and energy inefficiency are mutually reinforcing issues. Conversely, countries with moderate pollution but strong renewable adoption showed higher energy recovery. Clustering also highlighted geographic trends: several countries in similar climate zones or industrial stages tended to cluster together, regardless of political or economic factors. This shows the power of data-driven insights in challenging or reinforcing existing geographic assumptions. Such findings can help international organizations prioritize which regions to support for environmental reforms, and which countries could serve as models or case studies for others in the same cluster.

Actionable Insight – Pollution Control

Based on clustering, it is evident that countries with extreme pollution levels need urgent policy intervention. Governments should focus on reducing industrial waste, improving air quality monitoring systems, and enhancing public awareness about pollution. Countries in high-pollution clusters can adopt best practices from countries in cleaner clusters, such as better waste management protocols, green taxation, and strict environmental regulation. Data also revealed that water pollution often co-occurs with poor energy recovery, suggesting a deeper environmental interconnection. This means environmental policies should be holistic—addressing multiple forms of pollution at once, not in silos. Clustering also allows for benchmarking: countries can compare themselves with others in the same cluster and track their progress over time.

Actionable Insight – Energy Recovery

The insights from prediction models show that energy recovery is closely tied to pollution control and renewable energy adoption. Countries can improve their recovery rates by investing in technologies like waste-to-energy conversion, carbon capture, and bioenergy systems. Neural network results, while not precise, suggested that features like CO_2 emissions and population had predictive influence, indicating these should be prioritized in environmental strategies. Countries with low energy recovery despite high pollution may lack the infrastructure to capitalize on energy recovery opportunities. Hence, policymakers should allocate funds to develop such technologies, especially in countries with high waste and emission levels. Furthermore, cross-border collaborations on renewable technology can accelerate the energy transition and enable global sustainability.

Cross-Model Strategy

An integrated strategy should combine the strengths of all models. Clustering helps in understanding country types, while linear regression helps in forecasting energy recovery potential. Together, they enable both macro and micro-level policy planning. For example, policymakers can use clustering to group countries into intervention zones (e.g., urgent reform, moderate reform), and then use regression models within each zone to predict the impact of potential changes. This layered approach ensures that strategies are both scalable and specific. Instead of relying on a one-size-fits-all model, this method tailors actions to real-world environmental and economic contexts.

Limitations of the Study

Despite meaningful insights, the study has limitations. The dataset contains only 200 samples, which restricts the complexity of the machine learning models. The neural network's underperformance is likely due to this small sample size. Additionally, the dataset lacks geographic identifiers beyond country names, which could provide richer context for environmental patterns. The absence of certain variables—like policy strength, climate type, or energy pricing—also limits the model’s explanatory power. Furthermore, clustering does not imply causation, so further research would be required to establish concrete cause-effect relationships between pollution and energy recovery. Still, this study offers a strong foundation for data-driven environmental analysis.

Conclusion and Recommendations

This project presents a comprehensive approach to analyzing global pollution and energy recovery trends using both clustering and predictive models. The results show that clustering is effective for identifying macro patterns, while linear regression currently outperforms more complex models for prediction. Based on the findings, we recommend: (1) investing in renewable energy to boost recovery rates, (2) adopting integrated pollution control strategies, and (3) using clustering to benchmark countries and track progress. This hybrid method can be extended with more features, larger datasets, and further model tuning to become a practical decision-support system for governments, NGOs, and environmental think tanks. Future work should also explore time-series modeling and deep learning with richer datasets.