Final Report: Education and Environmental Factors in Relation to National Happiness

**Project Overview**

This study investigates the relationship between education levels, environmental quality, and national happiness scores across countries. Using publicly available datasets, we analyzed how variables like tertiary education enrollment, air pollution, forest area, arable land, and coastline ratio relate to happiness levels across countries.

**Methodology**

* Data was collected from Kaggle and World Bank sources and merged using country names.
* Exploratory Data Analysis (EDA) was conducted using histograms, scatter plots, and correlation heatmaps.
* Hypothesis testing using Pearson and Spearman correlations was applied to identify statistically significant relationships.
* Machine Learning models (Linear Regression, Decision Tree, and Random Forest) were used to predict happiness scores.

**Key Findings**

* Tertiary enrollment had the strongest positive correlation with happiness (r ≈ 0.64), statistically significant (p < 0.001).
* Air pollution showed a moderate negative correlation with happiness (r ≈ -0.47 to -0.58), also significant (p < 0.001).
* Coastline ratio showed a moderate positive correlation, significant after transformation (p < 0.001).
* Adult literacy, forest area, and arable land had weak or insignificant correlations with happiness.
* Decision Tree and Random Forest outperformed linear regression, capturing non-linear relationships better (R² ≈ 0.71–0.72).

**Hypothesis Testing Summary**

* H₀ rejected for: Tertiary Enrollment, Air Pollution, Coastline Ratio → significant predictors of happiness.
* H₀ not rejected for: Adult Literacy, Forest Area, Arable Land → no significant correlation.

**Limitations**

This analysis may contain mild bias due to missing data and limited variable selection. However, the overall results are statistically significant and consistent with the dataset. Independent features not originally in the World Happiness Index (e.g., forest area, coastline ratio) were added to reduce bias. Both regression and classification tasks were explored.

**Conclusion**

The results suggest that higher access to advanced education, clean air, and geographic advantages like coastline access are significantly associated with national happiness. In contrast, basic literacy and land use variables have limited predictive power. Non-linear machine learning models effectively captured the complexity of these relationships, emphasizing the multifactorial nature of national well-being.