SABANCI UNIVERSITY DSA 210 TERM PROJECT PHASE 3

ZİŞAN YEŞİL 32587

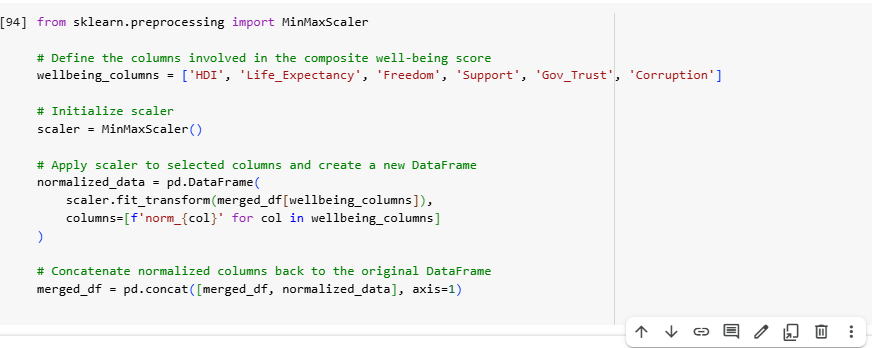
The Data Behind Women Leaders and National Well-Being

In this phase, I applied supervised machine learning models to predict the composite well-being score I constructed earlier. My goal was not only to evaluate how accurately well-being can be predicted but also to understand the relative influence of women's political empowerment and participation compared to traditional structural indicators like GDP and education.

1. **Feature Engineering and Composite Score Creation**

To measure national well-being in a multidimensional way, I created a composite score using six normalized indicators: HDI, Life Expectancy, Freedom, Social Support, Government Trust, and Corruption. I inverted the normalized corruption values so that higher values consistently reflected better outcomes. I then averaged these six indicators to create a single **Wellbeing Score** for each country-year entry. This approach was inspired by the methodology proposed in *The Composite Global Well-Being Index (CGWBI)* by Battista and Almond (2016).

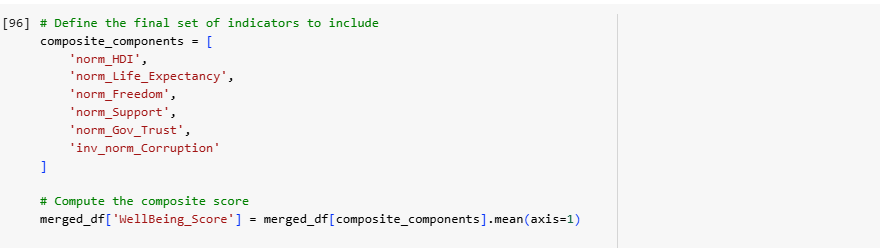
These indicators were chosen to reflect a holistic view of human well-being, going beyond economic output. All variables were normalized using Min-Max scaling, and the Corruption variable was inverted so that higher scores consistently represent better outcomes across all features. This composite score serves as the target for the machine learning models in the following analysis.



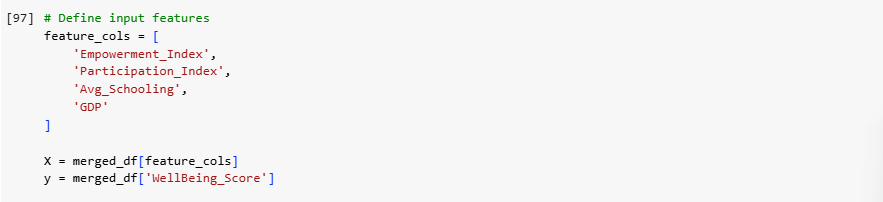
The normalized corruption score was inverted (1 - norm\_Corruption) so that higher values indicate lower corruption. This keeps the direction consistent with other features in the composite well-being score, where higher values always represent better conditions.



The final composite score was calculated by taking the average of six normalized indicators: HDI, Life Expectancy, Freedom, Support, Government Trust, and the inverted Corruption score. This score represents an overall measure of national well-being, where higher values indicate better outcomes.



The model uses four input features to predict the composite well-being score: Empowerment\_Index and Participation\_Index (representing women’s political leadership), along with Avg\_Schooling and GDP (as structural development controls). These variables form the input matrix X, while the target variable y is the previously constructed WellBeing\_Score.

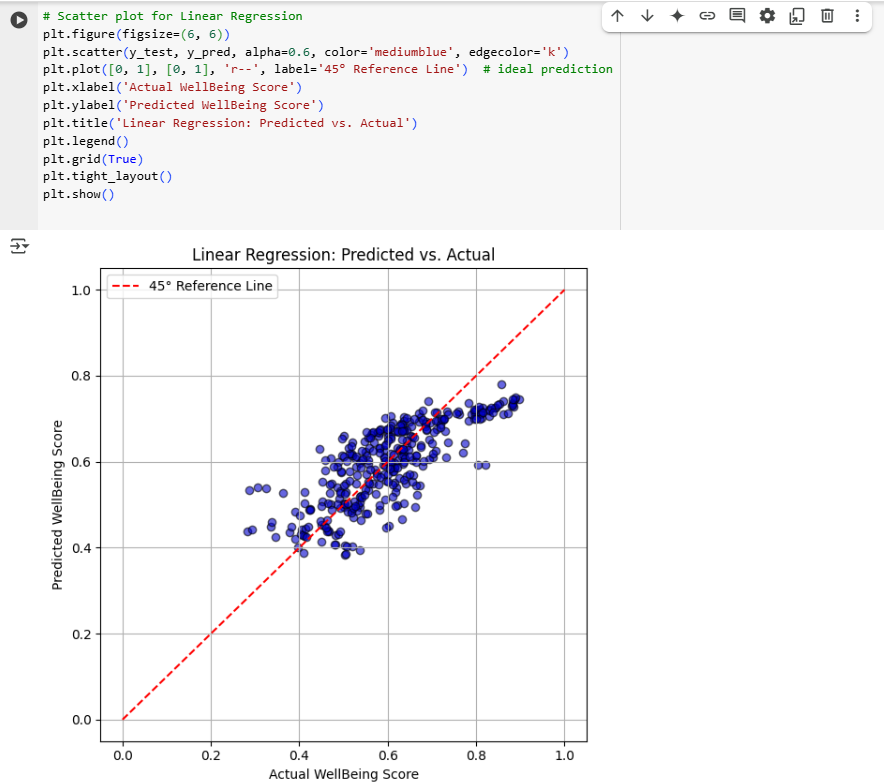


1. **Linear Regression**



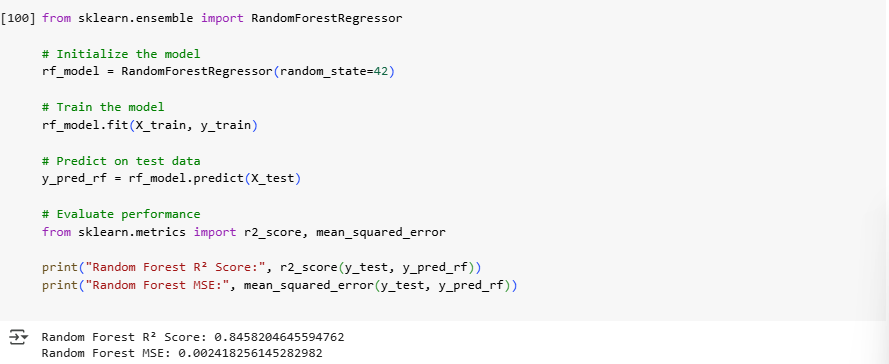
A multiple linear regression model was trained to predict the composite well-being score using four features: Empowerment\_Index, Participation\_Index, Avg\_Schooling, and GDP. The dataset was split into training and test sets (80/20), and the model’s performance was evaluated on the test set using the R² score and Mean Squared Error (MSE).

The resulting R² score of approximately 0.59 which indicates that the model explains about 59% of the variance in national well-being. This suggests that women's leadership, when combined with structural factors like education and economic performance, has a meaningful association with overall well-being.



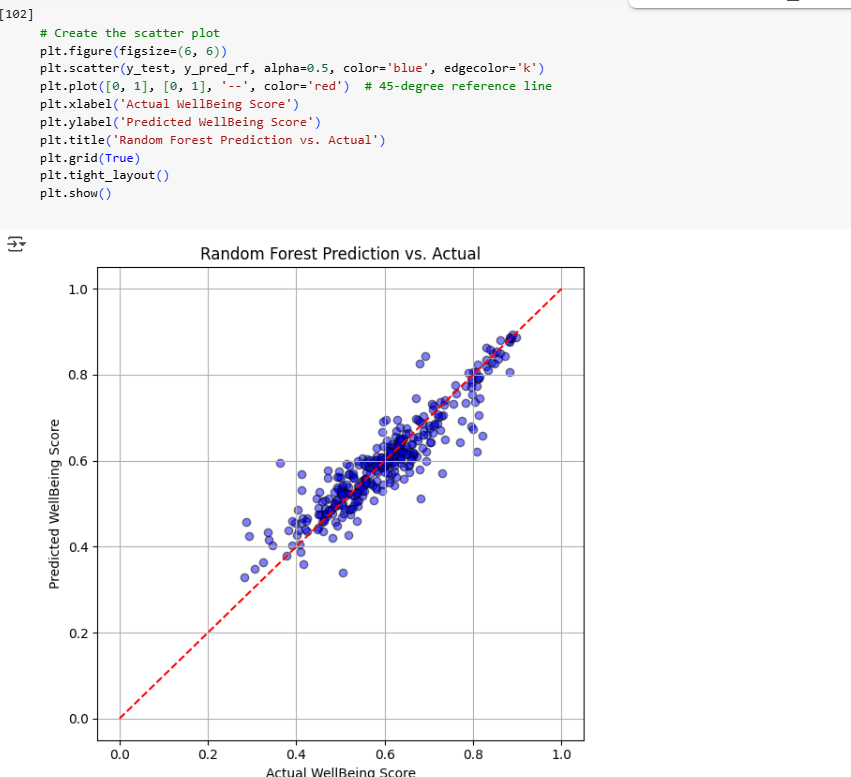
This scatter plot shows how well the Linear Regression model predicts the Composite Well-Being Score. Points closer to the red 45° line represent better predictions. While there is a clear positive trend, the spread indicates the model has limited accuracy and may miss some complexity in the data.

1. **Random Forest**



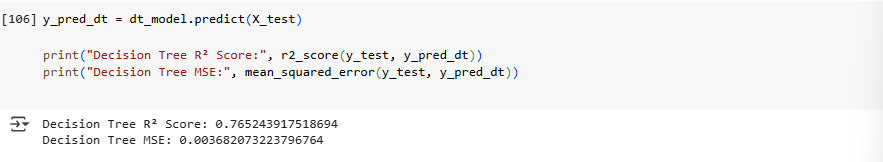
To capture potential nonlinear relationships between features and well-being, a Random Forest Regressor was trained using the same four predictors: Empowerment\_Index, Participation\_Index, Avg\_Schooling, and GDP. Random Forests, being ensemble models based on decision trees, are well-suited for modeling complex interactions without requiring explicit feature transformations.

The model achieved a high R² score of approximately 0.85, significantly outperforming the linear regression model. This indicates that 85% of the variance in the well-being score can be explained by the selected features, suggesting a strong predictive relationship, especially when allowing for nonlinear effects.

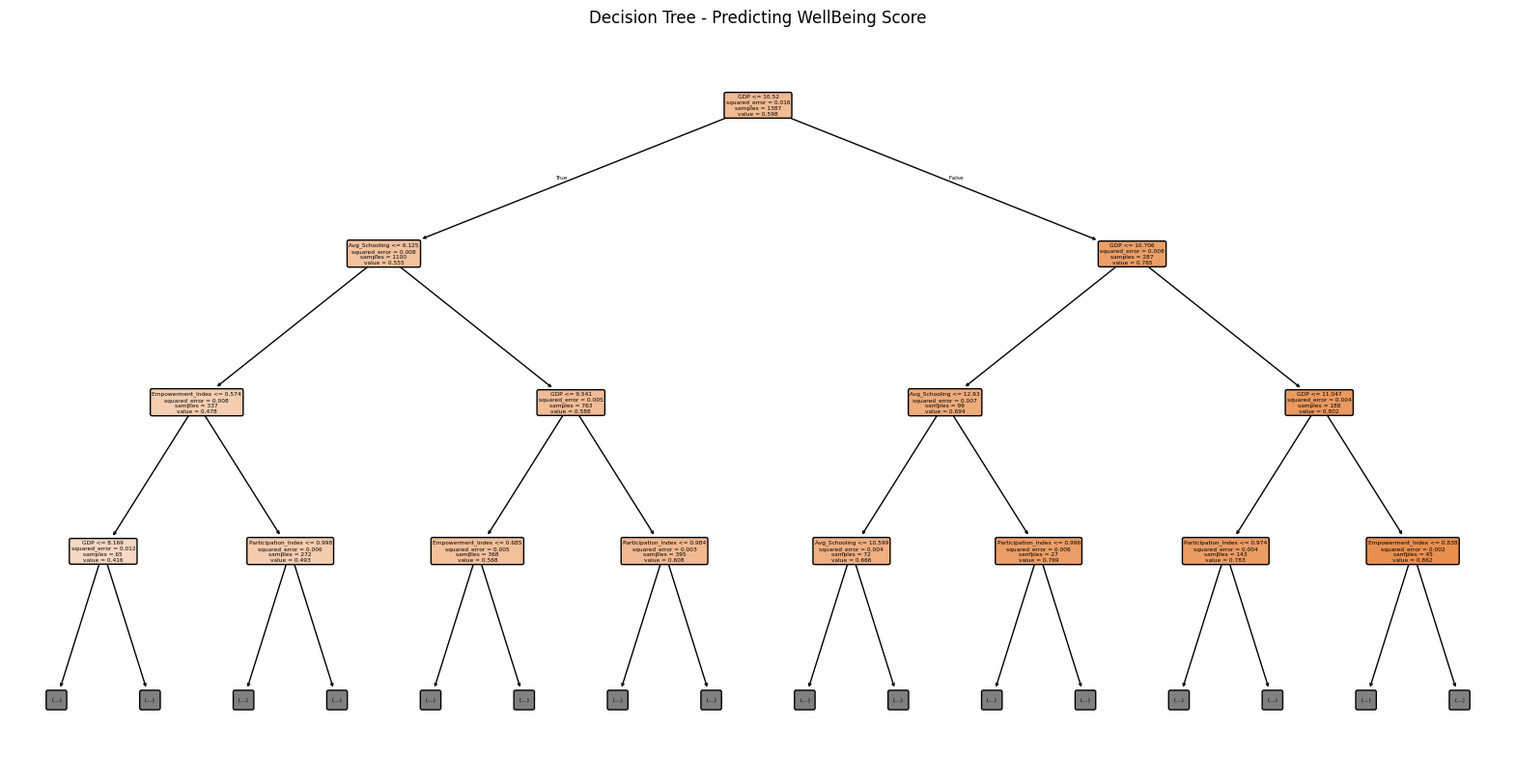


This scatter plot compares the Random Forest model’s predicted well-being scores to the actual values from the test set. The red dashed line represents the ideal 45-degree line where predicted values would perfectly match actual scores. Most data points cluster closely around this line, indicating a strong predictive performance. This visual reinforces the model’s high R² score and supports the conclusion that women’s leadership, combined with structural indicators, can meaningfully predict national well-being.

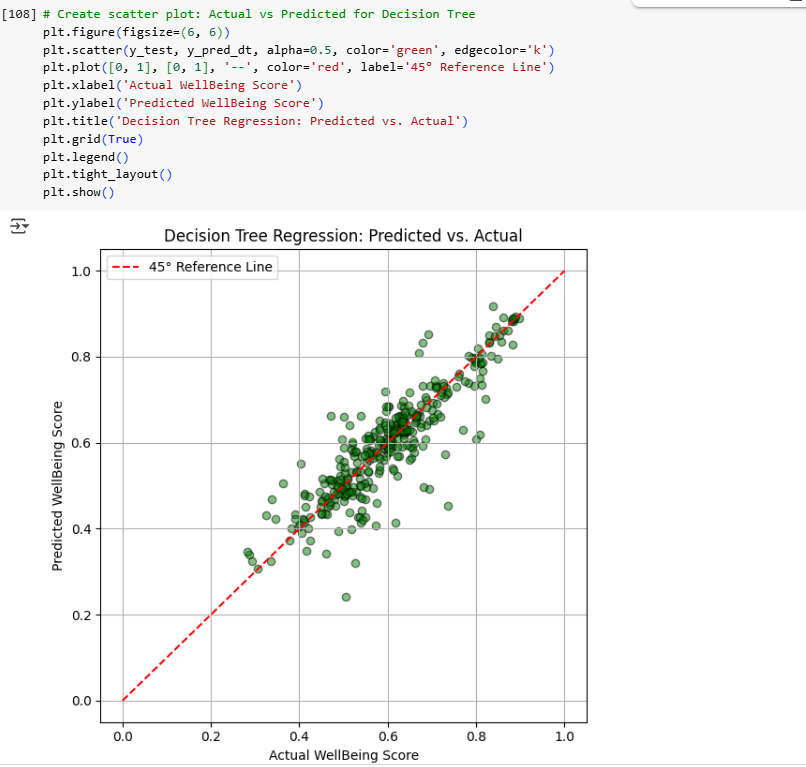
1. **Decision Tree**



This code evaluates the performance of the Decision Tree Regressor by predicting the well-being scores for the test set and calculating the R² Score and Mean Squared Error (MSE). An R² Score of approximately 0.77 indicates that the model explains a substantial portion of the variance in well-being scores, while the low MSE confirms good prediction accuracy.



This visualization displays the structure of the trained Decision Tree Regressor used to predict the composite Well-Being Score. Each node represents a decision based on a feature (e.g., GDP or Avg\_Schooling), and branches split the data accordingly. The tree helps interpret how different variables and thresholds contribute to predicting well-being, with the depth limited to 3 for clarity.

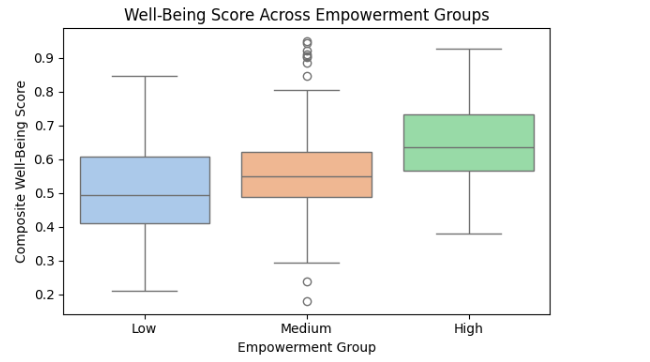


This scatter plot shows how well the Decision Tree Regressor predicts the composite Well-Being Score. Each point compares actual vs. predicted values, and the red dashed line represents perfect predictions. Most points are close to the line, indicating decent performance, but with more variance than the Random Forest model.

1. **Comparing Well-Being Across Levels of Women’s Empowerment**

To examine how well-being or freedom varies with different levels of women’s empowerment, the continuous Empowerment\_Index was categorized into three ordinal groups: Low (≤ 0.6), Medium (0.6–0.8), and High (> 0.8). This transformation makes it easier to compare well-being outcomes across clearly defined empowerment levels.

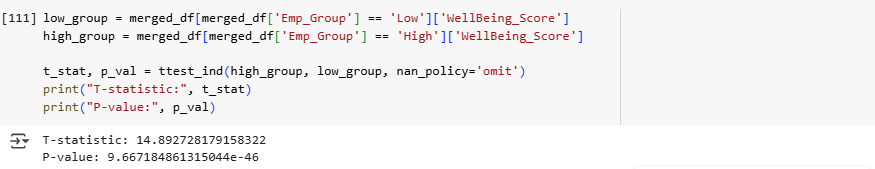




This boxplot visualizes how the Composite Well-Being Score varies across Low, Medium, and High empowerment groups.

* As empowerment increases, both the median and distribution of well-being rise.
* The boxplot supports the statistical findings: higher empowerment aligns with greater national well-being.

The visual distribution mirrors the t-test results: higher empowerment groups enjoy visibly better well-being.



This performs an independent samples t-test to determine whether there's a statistically significant difference in Well-Being Score between the countries with Low and High levels of women's empowerment.

The T-statistic of ~14.89 indicates a large difference between means.

The p-value (< 0.0001) is far below 0.05, confirming the difference is statistically significant.

So, we can say that countries with higher women's empowerment scores tend to have significantly higher well-being scores.

**Conclusion**

Through this machine learning analysis, I confirmed that women’s political empowerment and participation are important contributors to national well-being. Across all models, including Linear Regression, Random Forest, and Decision Tree, leadership indicators consistently played a meaningful role in predicting composite well-being.

The statistical tests and group comparisons further supported this finding, showing significantly higher well-being and freedom scores in countries with higher levels of women’s political empowerment. These results reinforce the central idea of my project: **political empowerment of women is not only a measure of gender equality but a key factor in building happier, healthier societies**.