21st Century Democracy Metrics

Team: Number 4

Name: Daniel Felberg, Ei Tanaka, Ishani Makwana, Tharaka Maddineni

Columbian College of Arts and Sciences, The George Washington University

DATS 6103: Introduction to Data Mining

Instructor: Ning Rui

April 30, 2023

1. Introduction

In 2021, two Yale researchers, Yusuke Narita and Ayumi Sudo, sought to understand the relationship between democracy and various socioeconomic factors throughout the globe in the 21st century. Their study found that democracies, contrary to popular belief, seem to negatively impact economic growth, due to less investment, less trade, and slower productivity growth, overall. Furthermore, when looking at the impact of Covid-19 throughout 2020, their analysis also determined that there was a negative correlation between democracy and Covid-19 deaths. This meant that more democratic countries had, in general, more deaths as a result of the pandemic. To determine whether these findings could potentially be replicated, our group decided to also use the *Varieties of Democracy* dataset, the scoring system used to compare how democratic (or undemocratic) a given country was. In doing so, our goal was to investigate a variety of background factors, and what impact, if any, they might have on democracy.

1.1 Background Research

Having established a generalized research topic, we researched what the existing literature on the subject focuses on. Not surprisingly, the debate of what factors may potentially contribute towards democratization has gone on since the mid-20th century, and remains largely unanswered to this day. That being said, some research has been conducted to support a few basic theories. Starting with Lipset (1959), among others, who were proponents of *Modernization Theory*, they suggested that wealthier and more developed countries were better able to make use of their resources, and as a result, they would develop more resilient democratic systems. This theory, however, eventually lost traction, as it was seen as too generalized, and not really applicable in several countries (Tipps, 1973).

Another way of examining the effect that economic development has on democracy was performed by Barceló and Rosas (2021). Rather than directly examining more commonly observed economic factors, such as Lipset had, they utilized historical evidence of agricultural development as a

measure of urbanization, and consequently, development, also known as the *potato productivity shock* (Nunn, Qian, 2011). This measurement is particularly interesting, as it brings together different aspects that may be useful to ascertain factors that may contribute towards democratization, and "gauges how agricultural productivity-induced changes in urbanization affect a country's democratization" (Barceló, Rosas, 2021). Their study concludes that there is a "causal effect of economic development on democratization that is more credible than previously identified".

Moving on to other factors that potentially influence democracy, there is also research to suggest that democracy can contribute towards improved health systems, and consequently, the general health of their population (Franco, Alvarez-Dardet, Ruiz, 2004). Another variable that has been examined is the impact that education levels have on the overall democracy of a country. In their study, which also used the Freedom House index to measure countries' democracies, Alemán and Kim (2015) find that a positive correlation exists between the levels of education and democratization, with a particularly strong effect taking place in poorer countries.

1.2. Data sets

The research utilizes two datasets: Country-Year: V-Dem Full+Others by Varieties of Democracy and World Development Indicators by World Bank. The first dataset provides detailed information on countries' democratic institutions, such as civil rights, electoral processes, and freedom of the press. It also includes data on gender equality, corruption, and socioeconomic indicators like income inequality and poverty. In addition, the V-Dem dataset is based on expert assessments, making it an ideal source for cross-national comparisons of political systems. The second dataset is the World Development Indicators dataset by the World Bank, which includes information on economic and social indicators for over 200 countries, including poverty, education, health, and economic growth. This dataset is collected from various sources, including national statistical agencies and international organizations. It provides a comprehensive view of the development status of countries worldwide, allowing for a comparative

analysis of development trends. Combining these two datasets, the research explores the relationship between democratic institutions and socioeconomic background factors.

1.3. Research Questions / SMART Questions

This research aims to investigate the causes of democratization with a focus on three SMART questions, all based on previous research summarized above investigating various factors that may contribute to the democratization of countries. The first question is whether education levels impact a country's democratic institutions. It is hypothesized that education can improve the quality of citizens' political knowledge and critical thinking skills, leading to greater civic engagement and democratic participation. The second question explores whether citizens' economic welfare impacts a country's democratic institutions. It is hypothesized that a higher standard of living, including better access to basic needs such as food, housing, and healthcare, can lead to more significant support for democratic institutions. Finally, the third question investigates whether healthcare impacts a country's democratic institutions. It is hypothesized that access to healthcare can improve citizens' overall health and well-being, leading to greater engagement and support for democratic institutions. By examining these questions, this research aims to contribute to a better understanding of the factors that contribute to democratization.

2. EDA

The initial step of our study was to perform exploratory data analysis (EDA) to prepare and clean the data. The data set was analyzed to identify any missing values or outliers, which were then addressed through imputation or removal. Descriptive statistics, such as mean, median, mode, and standard deviation, were calculated for each variable to understand the distribution and central tendency of the data. We also used data visualization techniques such as histograms, box plots, scatterplots, etc., to identify any patterns or relationships between variables. Correlation analysis was also conducted to identify the strength and direction of the relationships between variables. This phase was crucial in

identifying any potential data quality issues and understanding the distribution and relationships of the variables, which provided a foundation for the subsequent modeling phase.

2.1. Data Preparation and Cleaning

In this study, we combined two datasets, Varieties of Democracy (V-Dem) and World Development Indicators (WDI) by the World Bank. Firstly, we imported the V-Dem dataset using the URL of the zip file and the pandas read_csv function. We created a list of variables of interest, and then created a subset of the dataset containing only the rows from 2000 onwards. We also calculated a new 'democracy_index' column by taking the mean of the five democracy variables. Next, we imported the WDI dataset from an online source and reformatted it by creating a dictionary for country_id and removing some unnecessary columns. We then merged the two datasets using country_name and year columns. After identifying the countries with less than 22 years of data, we filled the missing values by the median of the same country name or political region. Finally, the cleaned and prepared dataset was saved in CSV format for further analysis.

2.2. Data Structure & Data Types

The following table shows the dataset after cleaning for this project, which has 3740 entries and 30 columns. The variables of interest include the country name, country id, year, democracy index, and various socio-economic indicators. The data types for these variables range from object to int64 and float64. The non-null counts for all the variables are 3740, indicating that there are no missing values in the dataset. This dataset provides a comprehensive overview of the socio-economic indicators and their relationship with democracy index across countries from the year 2000 to 2021. The following variables are of notable importance to our analysis:

- “democracy_index”: Combined average index of the five different indices of democracy used by the V-Dem Project to compare democracy across different countries. The five indices are as follows:

    - *Electoral Democracy*: measures free and fair elections
    - *Liberal Democracy*: measures individual and minority rights
    - *Participatory Democracy*: measures active citizen participation
    - *Deliberative Democracy*: measures respectful and reasonable dialogue
    - *Egalitarian Democracy*: measures equal treatment
- "AccessToCleanCooking": % of population with access to clean fuels and
- "AdolescentFertility": births per 1,000 women (ages 15-19)
- "AgriForestFishValueAdded": % of GDP
- "FertilityRate": total births per woman
- "GNIPerCapita": Gross National Income per capita, Atlas method (current US$)
- "LifeExpectancy": at birth, total years
- "MobileSubscriptions": per 100 people
- "Under5Mortality": per 1,000 live births
- "PopulationGrowth": annual %
- "PrimarySchoolEnrollment": % gross

Table: Data Structure & Data Types

Range Index: 3740 entries Data Columns: total 30 columns		
Variable	Non-Null Count	Data Type
country_name	3740 non-null	object
country_id	3740 non-null	int
year	3740 non-null	int
democracy_index	3740 non-null	float
v2x_polyarchy	3740 non-null	float
v2x_libdem	3740 non-null	float
v2x_partipdem	3740 non-null	float
v2x_delibdem	3740 non-null	float
v2x_egaldem	3740 non-null	float
e_regionpol_6C	3740 non-null	int
AccessToCleanCooking	3740 non-null	float
AdolescentFertility	3740 non-null	float
AgriForestFishValueAdded	3740 non-null	float
CO2Emissions	3740 non-null	float
ExportsOfGoodsAndServices	3740 non-null	float
FertilityRate	3740 non-null	float
ForeignDirectInvestment	3740 non-null	float
GDP	3740 non-null	float
GDPGrowth	3740 non-null	float
GNIPerCapita	3740 non-null	float
MeaslesImmunization	3740 non-null	float
ImportsOfGoodsAndServices	3740 non-null	float
LifeExpectancy	3740 non-null	float
MobileSubscriptions	3740 non-null	float
Under5Mortality	3740 non-null	float
NetMigration	3740 non-null	float
PopulationGrowth	3740 non-null	float
HIVPrevalence	3740 non-null	float
PrimarySchoolEnrollment	3740 non-null	float

2.3. Descriptive Statistics

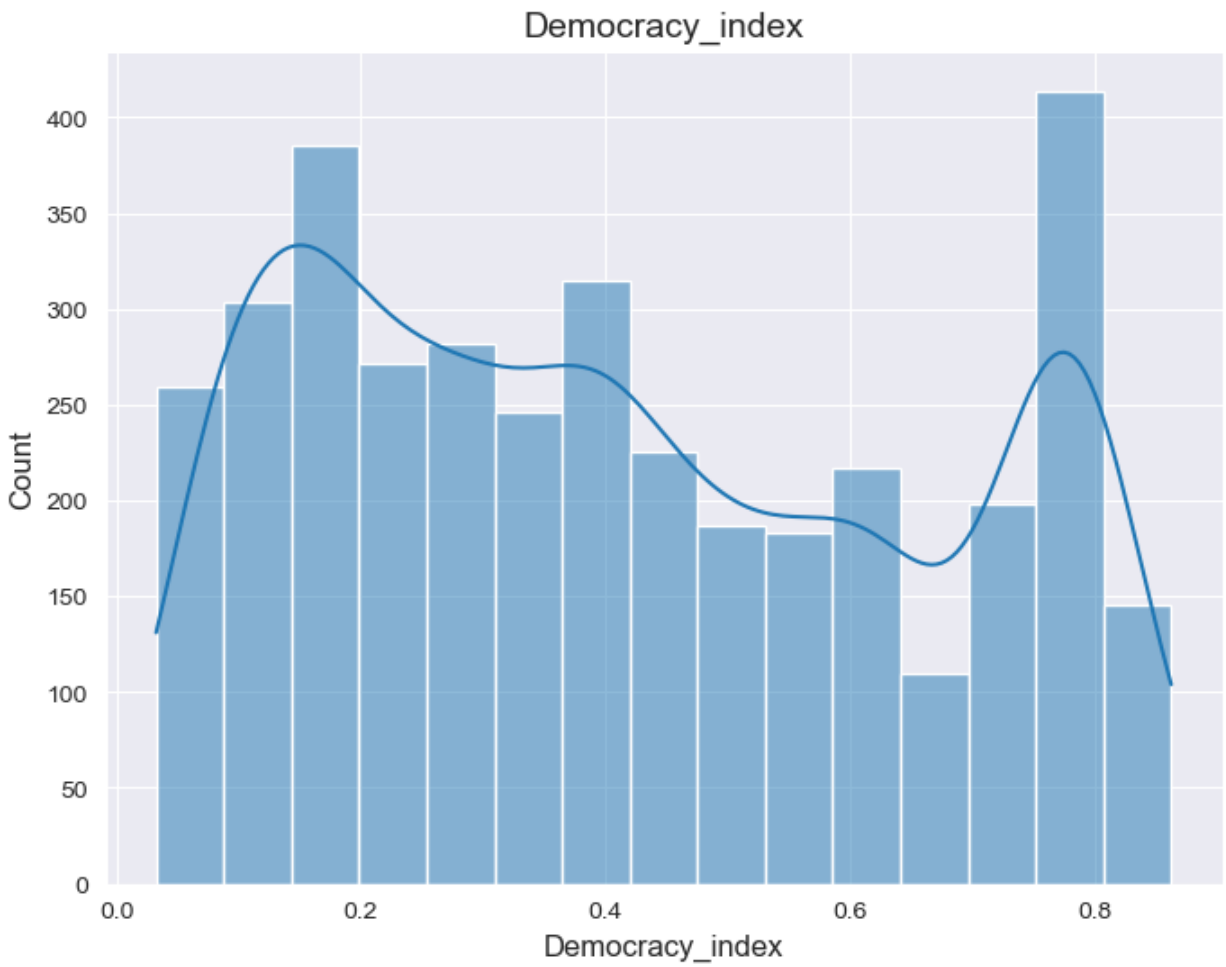
To have a clear understanding of the dataset and its comparison with other variables, We are going to take the mean, min, max of our variable of interest. We can see the mean of the democracy index is 0.42(Liberia) , min is 0.03(North Korea) and max is 0.86(Denmark). Other variables are summarized by the following table.

Table: Descriptive Statistics

	variable	mean	std	min	25%	50%	75%	max
outcomes	democracy_index	0.42	0.24	0.03 (North Korea)	0.2	0.39	0.63	0.86 (Denmark)
	v2x_polyarchy	0.52	0.26	0.01	0.29	0.53	0.77	0.92
	v2x_libdem	0.41	0.27	0.01	0.15	0.38	0.65	0.9
	v2x_partipdem	0.34	0.2	0.01	0.15	0.32	0.5	0.81
	v2x_delibdem	0.41	0.26	0.01	0.2	0.38	0.64	0.89
	v2x_egaldem	0.4	0.24	0.03	0.19	0.33	0.6	0.89
predictors	e_regionpol_6C	3.54	39.55	1	2	4	5	6
	AccessToCleanCooking	62.29	39.55	0.1	20.04	82.2	100	100
	AdolescentFertility	55.83	46.1	0.88	15.98	43.24	84.18	205.38
	AgriForestFishValueAdded	12.12	11.8	0.03	2.86	8.1	18.83	79.04
	CO2Emissions	4.4	5.58	0.02	0.57	2.51	6.26	47.65
	ExportsOfGoodsAndServices	41.41	30.42	0.46	23.08	33.94	51.38	228.99
	FertilityRate	2.96	1.57	0.84	1.71	2.43	4.05	7.73
	ForeignDirectInvestment	9671979975	38229432844	-3.30338E+11	95275914.68	762050000	4017746137	7.33827E+11
	GDP	3.78201E+11	1.53137E+12	75951210.69	8582484213	32253990479	1.94207E+11	2.33151E+13
	GDPGrowth	3.7	5.35	50.34	1.75	3.84	6.01	86.83
	GNIPerCapita	11474.49	17071.97	110	1180	3915	12995	104370
	MeaslesImmunization	85.38	15.2	16	80	92	96	99
	ImportsOfGoodsAndServices	46.16	27.53	0.3	28.75	39.25	56.39	221.01
	LifeExpectancy	69.74	9.13	41.96	63.31	71.76	76.87	85.39
	MobileSubscriptions	75.91	50.01	0	30.73	80.14	115.69	319.43
	Under5Mortality	39.58	41.85	2	8.1	22.2	60.12	228.5
	NetMigration	1078.31	192164.28	-2290411	-25378.75	-2234	15672	1479676
	PopulationGrowth	1.45	1.57	-6.85	0.47	1.33	2.37	19.36
	HIVPrevalence	1.54	4.03	0.1	0.1	0.2	1	29.8
	PrimarySchoolEnrollment	102.22	15.31	9.06	98.4	102.49	108.33	156.45

2.4. Distribution of the target variable

The distribution of the democracy index appears to be binomial, with one peak around 0.2 and another peak around 0.8. This indicates that there may be two distinct groups within the data, with some countries exhibiting higher levels of democracy than others. Given the non-linear relationship suggested by the bimodal distribution, a linear regression model may not be appropriate for modeling the relationship between the predictor variables and the democracy index. Instead, a tree-based model, such as a regression tree, random forest, or gradient boosting, may be more appropriate for capturing the non-linear patterns in the data. These models are capable of handling non-linear relationships and interactions between predictor variables, and can capture complex patterns in the data that may be missed by a linear model.

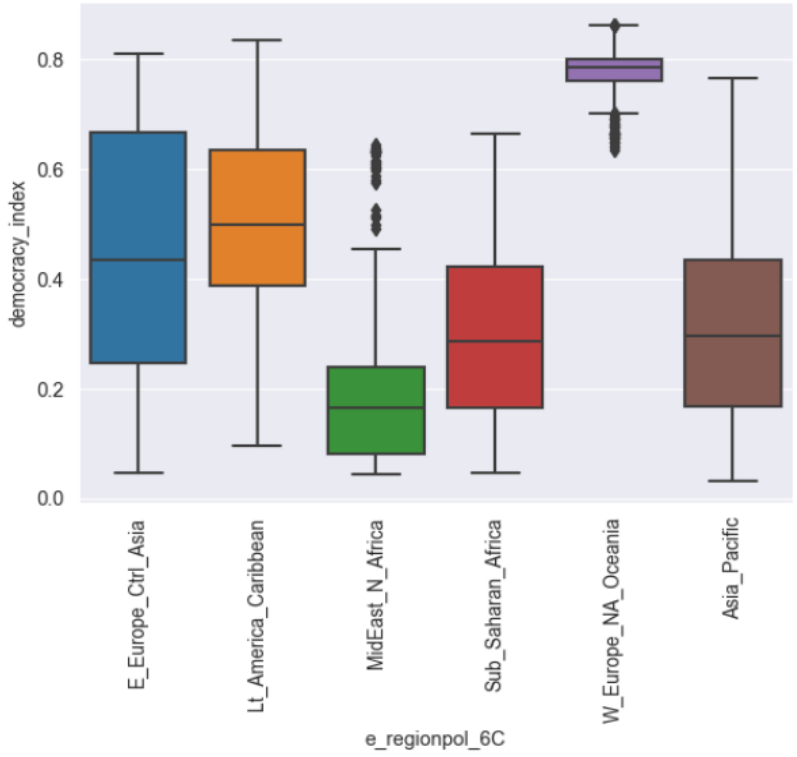


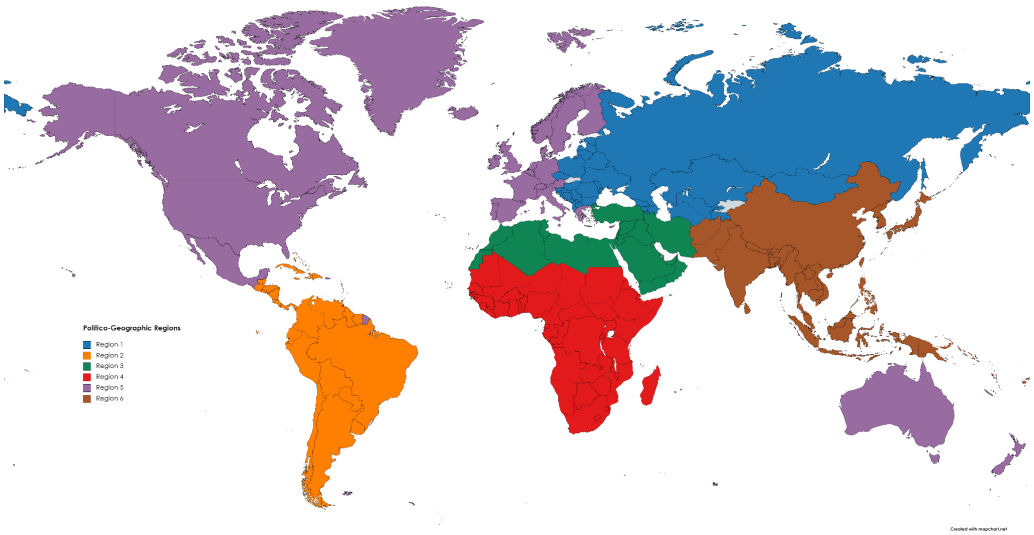
2.5. Visualization & Findings

2.5.1. Box plot Analysis

The box plot below shows the 20 year average graphical representation of the distribution of the democracy index across the different geo-political regions. It shows the median, interquartile range, and any potential outliers for each region (see the map below for reference). We can see that regional influence is a major factor in determining the level of democracy. For example, the median democracy index value for West Europe, North America, and Oceania is higher compared to other regions. Additionally, there are some potential outliers for some regions, indicating that there may be some countries with significantly higher or lower democracy index values compared to other countries in the

same region. For instance, those in the Middle East and North Africa may have lower scores. This observation is particularly significant as there are different factors to each region that affect the level of democracy. Understanding these factors is crucial in promoting and sustaining democracy. This information can be used to gain insights about the data, identify potential issues, or guide further analysis.

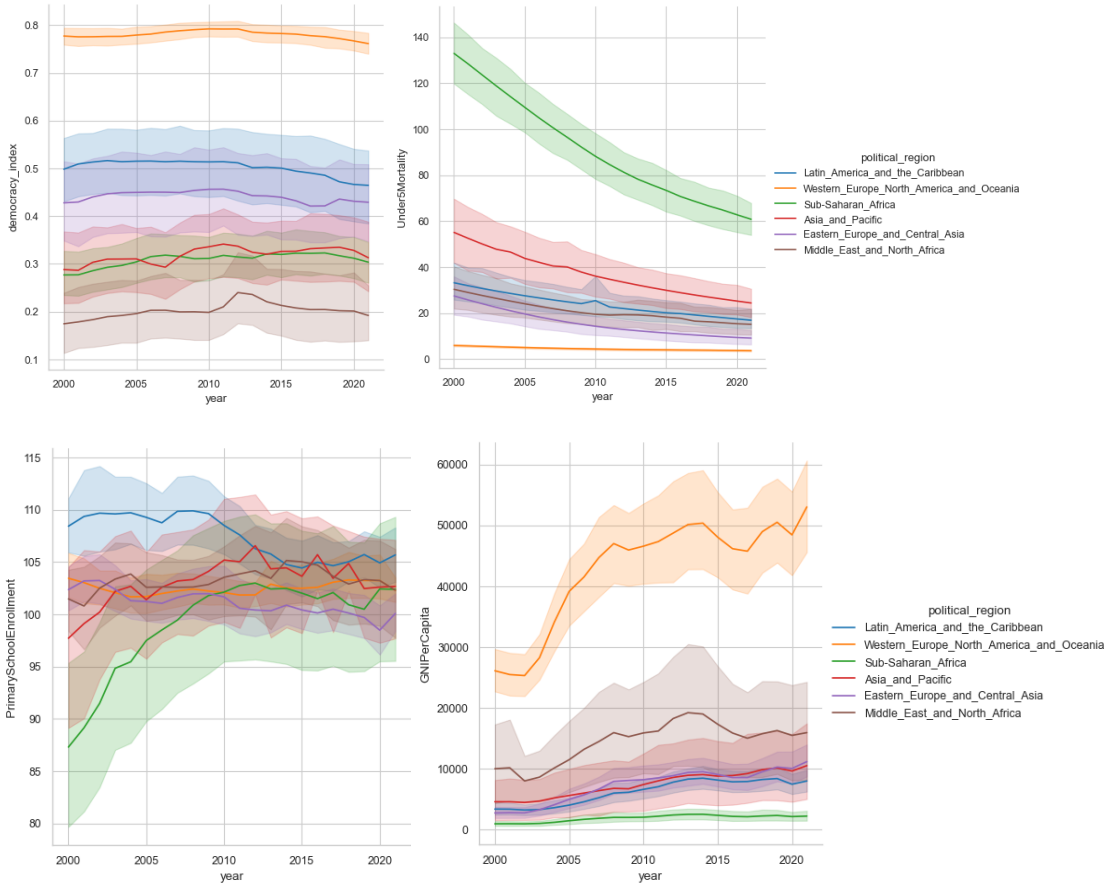




By looking at the world map, We can see that all the regions are indicated using different colors on the map, showing the aggregated version of the dataset. For example, the median democracy index value for West Europe, North America, and Oceania is higher compared to other regions.

2.5.2. Time Series Analysis

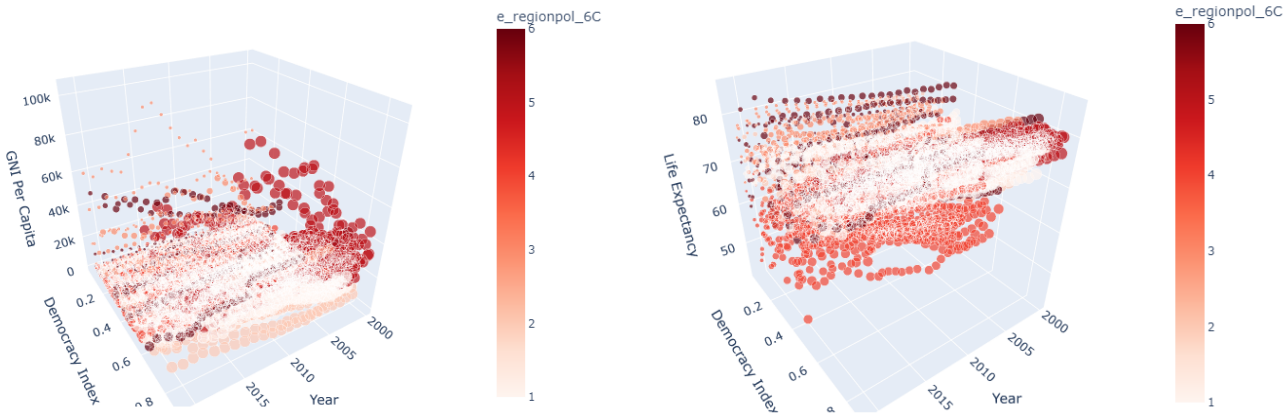
The boxplot above helps us obtain an overall picture of the democracy index by geo-political regions, but simultaneously, it also greatly simplifies the information in the data. To understand the impact that different factors have on democracy, it is necessary to measure the trend in each region throughout the observed period, as a time series analysis.



The line graphs above give us a better idea of how the democracy index (upper-left) as well as a few other relevant variables, which are child mortality, primary school enrollment, and GNI per capita (upper right, lower left, and lower right, respectively), have changed overtime. Most notably, we see that despite there

being a significant difference in the democracy index between geo-political regions, there does not seem to be a significant change in those values over time. At the same time, we see much more drastic changes occurring in all other factors listed above, meaning that perhaps it may be more difficult to ascertain a specific contributing factor towards the process of democratization. With that then, we decided to use more advanced plots that allow us to visualize changes in more than one variable over time.

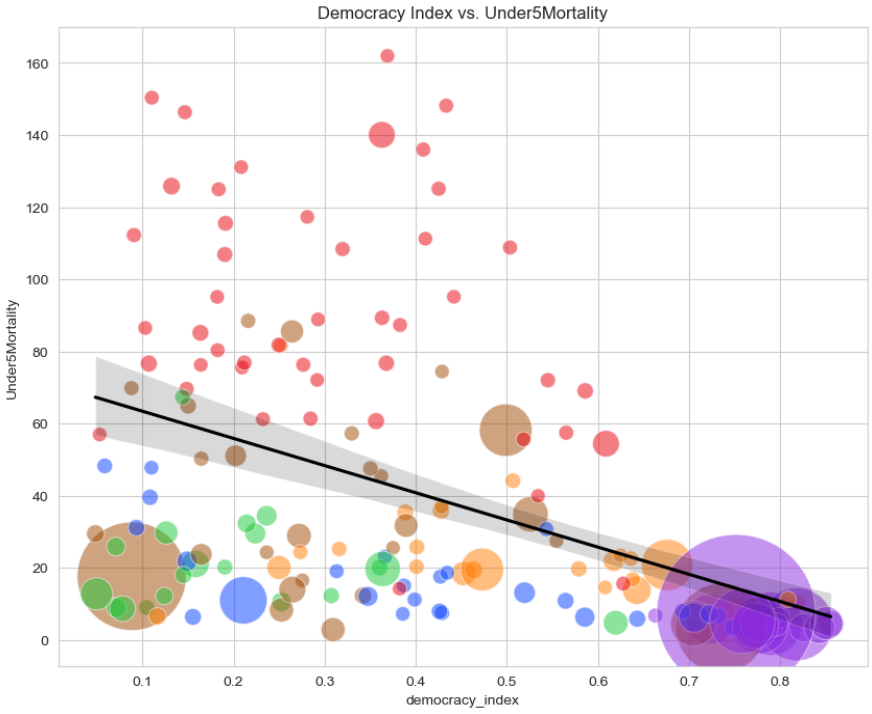
2.5.3. 3-D Scatter plot Analysis

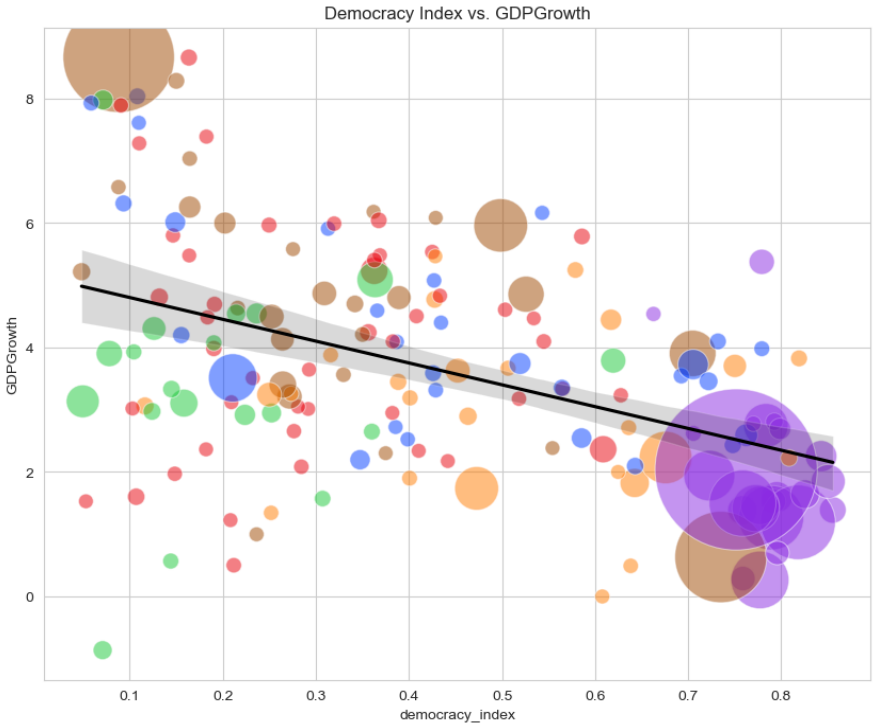


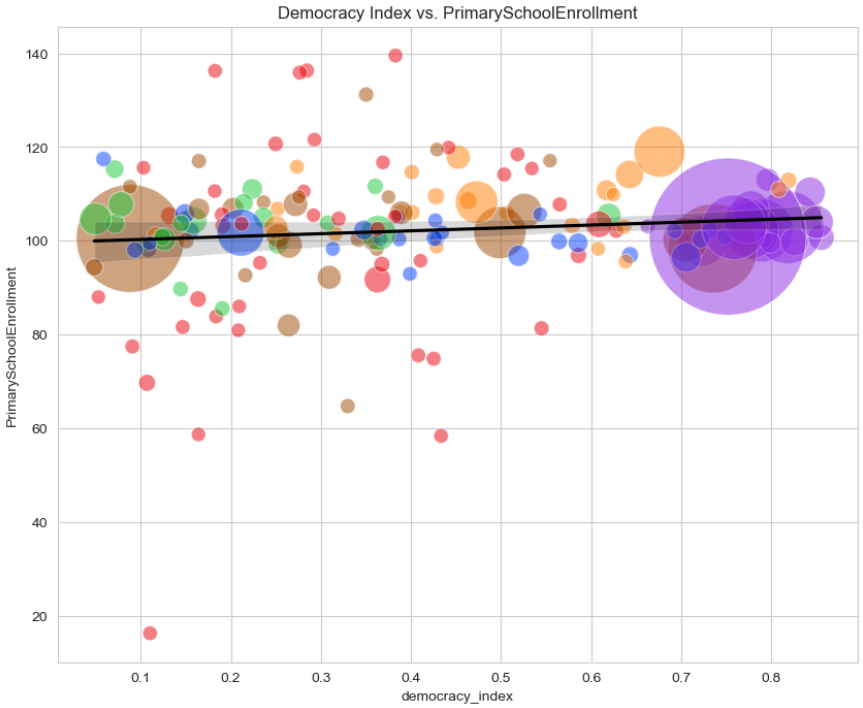
The two scatter plots above allow us to visualize the potential impact of GNI per capita (left), as well as life expectancy (right) on the democracy index of each country, over the course of 20 years. This type of visualization is helpful, as it allows us to obtain a more complete picture of the data, that includes each country, for each year, while also accounting for each region that they belong to. Granted, the screenshots above do not do these plots justice, which is why it helps to see them in the python code, where you are also able to manipulate the graphs, and visualize them from other angles. Regardless, we can observe the previous trends that we saw present in the line graphs, but now we also see that there are quite a few outliers for each region, for both the democracy index variable, as well as the GNI per capita and life expectancy variables, respectively.

2.5.4. Bubble Scatter Plot

Bubble plots were created to investigate the relationship between democracy_index and various background factors related to healthcare, economics, and education. The size of each bubble in the plots represents the GDP of each country. In the first plot, the y-axis represents the mortality rate for the under-five age group, and it shows a negative correlation between the democracy index and the mortality rate. This suggests that countries with higher democracy indices tend to have lower mortality rates for children under five. The second plot represents the GDP growth rate on the y-axis and reveals an exciting finding: countries with higher democracy indices have lower GDP growth rates in the 21st century. This suggests that economic factors might not be the primary drivers of democracy, and other factors, such as social or political factors, may also play a role. Finally, the last plot displays the elementary school participation rate on the y-axis and shows no strong correlation between the democracy index and primary school enrollment rate. This implies that education-related indicators may correlate less with the democracy index than health care or economic factors.



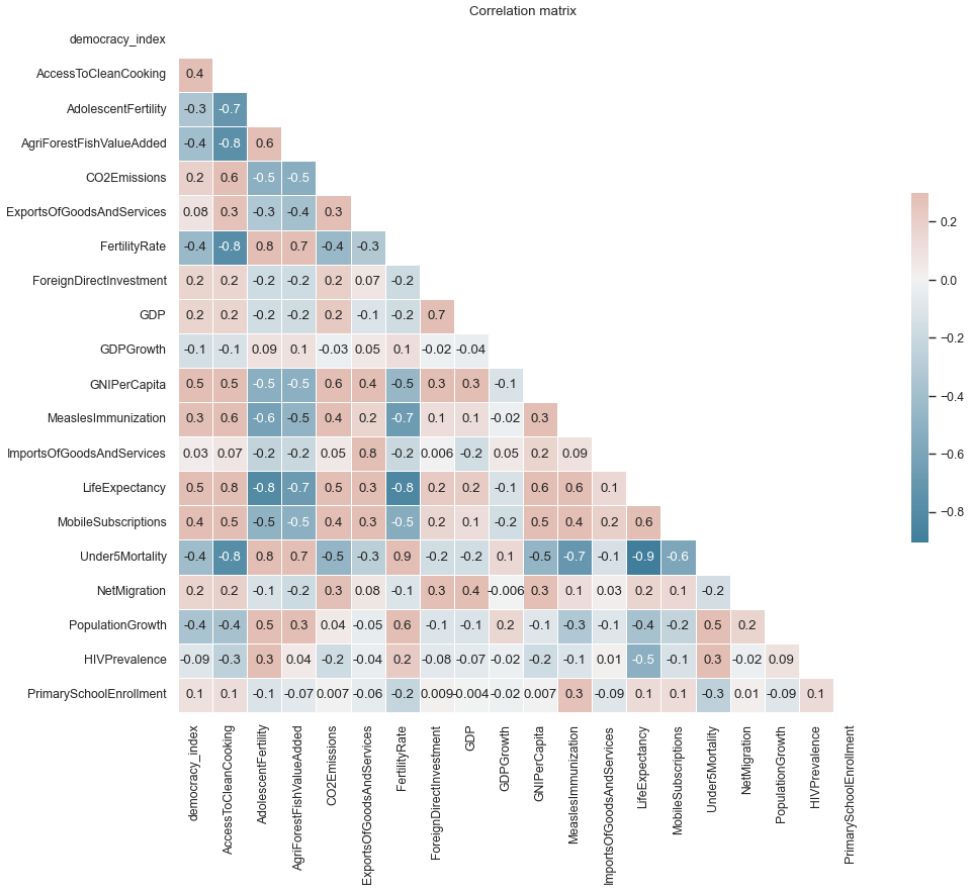

Democracy Index vs. GDPGrowth
GDPGrowth
democracy_index
8
6
4
2
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


Democracy Index vs. PrimarySchoolEnrollment
PrimarySchoolEnrollment
democracy_index
140
120
100
80
60
40
20
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

2.6. Correlation matrix:

The correlation matrix shows the correlation coefficients between all pairs of numerical variables in the dataset. The heatmap helps visualize the strength and direction of the linear relationship between pairs of variables. The diagonal line of the matrix represents the correlation of each variable with itself, which is always 1. The colors in the heatmap represent the correlation coefficient values, with more excellent colors indicating negative correlations and warmer colors indicating positive correlations. The annotation inside each square indicates the correlation coefficient value. High positive or negative correlation coefficients (close to 1 or -1) indicate a strong linear relationship between the variables, while values close to 0 indicate little to no linear relationship.

The correlation matrix can help identify potential multicollinearity issues between variables, which can affect regression analysis results. It can also help identify variables strongly correlated with the target variable, which can be helpful for feature selection.



	VIF	Features
0	4.451257	AccessToCleanCooking
1	7.948441	AdolescentFertility
2	4.056220	AgriForestFishValueAdded
3	2.462441	GNIPerCapita
4	4.727483	MobileSubscriptions
5	8.137766	Under5Mortality
6	2.476033	PopulationGrowth

The Result of VIF TEST

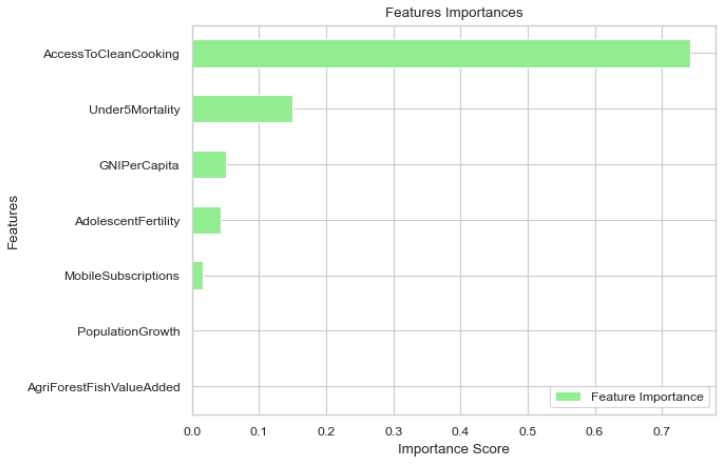
2.7. Feature Selection

We select a range of features related to healthcare, economic, and demographic factors to understand their impact on democratic institutions based on Pearson's correlation test. For healthcare factors, we include three features: access to clean fuels and technologies for cooking, adolescent fertility rate, and under-5 mortality rate. Access to clean fuels and technologies for cooking was chosen because it reflects the quality of living conditions and access to essential resources, which can impact the health and well-being of individuals. Adolescent fertility rate and under-5 mortality rate were chosen as indicators of reproductive health and child survival, which can have significant implications for a population's overall health and well-being. For economic factors, we select three features: GNI per capita, mobile cellular subscriptions per 100 people, and agriculture, forestry, and fishing value added as a percentage of GDP. GNI per capita reflects the overall economic welfare of citizens, while mobile cellular subscriptions can indicate access to technology and communication. Agriculture, forestry, and fishing value added as a percentage of GDP was chosen to measure economic diversity and development. Finally, we included one demographic feature: population growth rate, which can affect resource management and political stability. Using these features, we built three tree-based models: regression tree, random forest, and gradient boosting, to analyze the relationship between democratic institutions and the selected factors.

3. Model Building

Model building is a crucial step in understanding the contributing factors of democracy. In this study, we employ three tree-based models, namely regression tree, random forest, and gradient boosting, to investigate the relationship between democracy and its contributing factors (selected through the feature selection process). The regression tree is used to model the relationship between democracy index and the selected features, and to identify the most important factors that contribute to democracy. The random forest and gradient boosting models are used to further evaluate the significance of the selected features, and to predict the democracy index. The models will be evaluated based on their predictive power, feature importance, and interpretability. The results of this study will provide valuable insights into the factors that contribute to democracy and will help policymakers to make informed decisions about promoting democracy around the world.

3.1. **Regression Tree Model**



Regression Tree Matrics:

The result of the regression tree model suggests that the model has moderate predictive power, as it explains 49% of the variance in the target variable. The MSE value is relatively low, suggesting that the average squared difference between the predicted and actual values is low. The MAE value is also

relatively low, indicating that the average absolute difference between the predicted and actual values is low. The MSLE value suggests that the model's error is distributed logarithmically. The MedAE value indicates that the median absolute error is relatively low, suggesting that the model is relatively consistent in its predictions.

```
Decision/Regression Tree model evaluation metrics:
R^2: 0.49187094996036995 (49.2%)
MSE: 0.030391060560576433 (3.0%)
MAE: 0.1383268761814072 (13.8%)
MSLE: 0.016214299428986292 (1.6%)
MedAE: 0.12025500898472587 (12.0%)
```

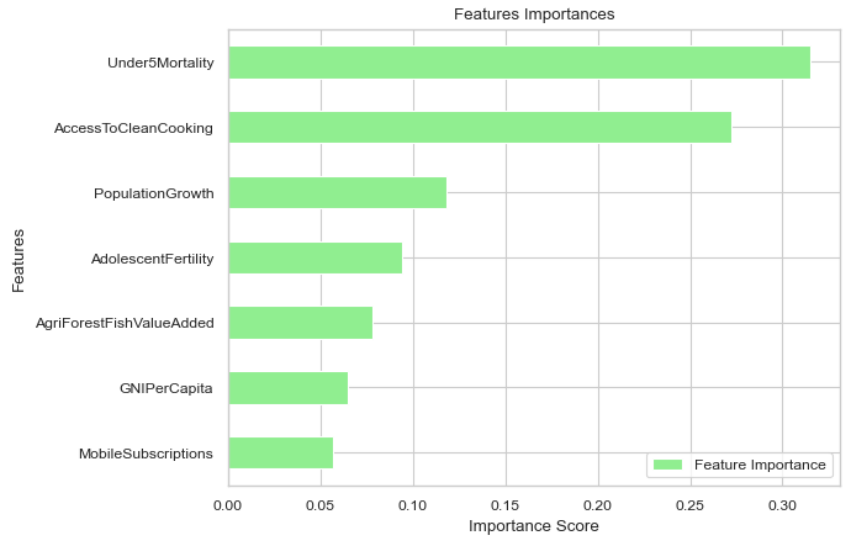
5 Fold cross-validation metrics for regression tree model:

```
5-fold X-Validation for regression model:
Cross-validation R^2 scores:
0.3744, 0.4732, 0.495, 0.5571, 0.4695,
Mean R^2: 0.4738
```

Finally, we performed 5-fold cross-validation on the decision tree model and calculated the R-squared values and mean R-squared values. The cross-validation R-squared values suggest that the model has moderate predictive power, although it is not as good as the original model.

3.2. Random Forest

The feature importance graph shows the importance score of each feature in the random forest model. The higher the importance score of a feature, the more impact it has on the model's output. In the graph, the features are sorted in descending order based on their importance score, with the most important feature at the top. By analyzing the graph, you can identify the features that have the highest impact on the model's predictions. These are the features that contribute the most to the overall performance of the model. You can also identify the features that have the least impact on the model's predictions and consider removing them from the dataset to simplify the model. It is important to note that feature importance is not the only factor to consider when selecting features for a model. Domain knowledge and correlation analysis should also be taken into consideration.



Random forest model evaluation metrics:

```
Random forest model evaluation metrics:
R^2: 0.9304659388892831 (93.0%)
MSE: 0.00415881332128877 (0.4%)
MAE: 0.045269606951871655 (4.5%)
MSLE: 0.002317157192869873 (0.2%)
MedAE: 0.030396000000000062 (3.0%)
```

The R^2 value of 0.9304 indicates that the model can explain 93.4% of the variance in the target variable, which means it is a good fit. The MSE value of 0.0041 shows the minimal difference between the predicted and actual values, indicating that the model's predictions are accurate. The MAE value of 0.0452 means that the average difference between the predicted and actual values is also tiny. The MSLE value of 0.0023 indicates that the model makes accurate predictions across the entire range of target values. Finally, the MedAE value of 0.0303 means that half of the absolute errors are smaller than this value, indicating that the model is reliable in predicting the target variable.

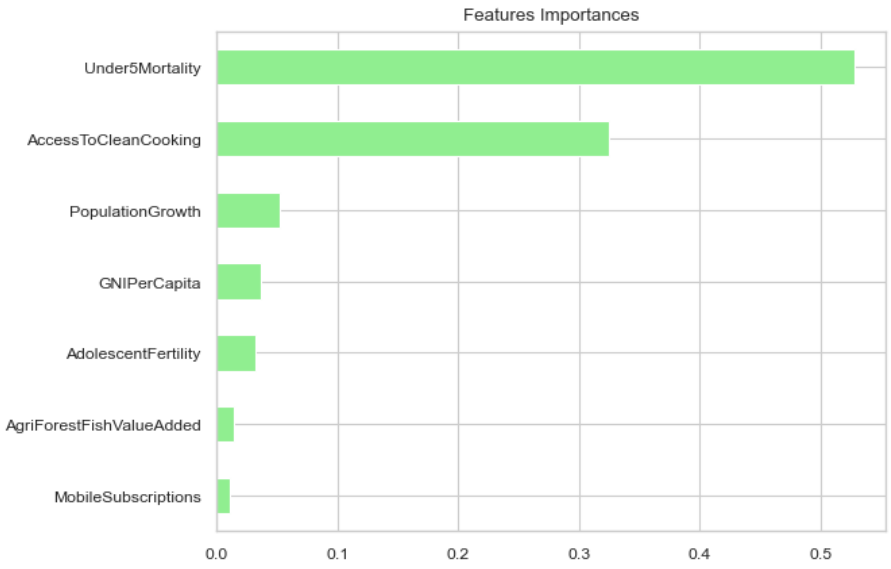
5 Fold cross-validation metrics for random forest model:

```
5-fold X-Validation for random forest model:
Cross-validation R^2 scores:
0.892, 0.8775, 0.8732, 0.908, 0.8936,
Mean R^2: 0.8889
```

Cross-validation is a technique used to evaluate the performance of a model on different subsets of data to check if the model is consistent across different samples. The R-squared value measures how well the model fits the data, with higher values indicating a better fit. In this case, the cross-validation results show that the model's performance is consistent across different subsets of data, with R-squared values ranging from 0.8732 to 0.908. However, the mean R-squared value across all folds is lower than the original R-squared value of the model, indicating that the model may have to overfit the data. Overfitting occurs when a model becomes too complex and fits the training data too closely, resulting in lower performance on new data. To address this issue, the model's hyperparameters may need to be adjusted, or alternative machine-learning techniques may need to be explored to improve the model's performance on new, unseen data.

3.2.1. Refined Random Forest Model

Additionally, we attempted to improve the model's performance by hyperparameter tuning with GridSearchCV() from sklearn.model_selection. 'n_estimators' hyperparameters: [25, 50, 75], 'min_samples_leaf': [0.1, 0.2, 0.3], 'max_depth': [2, 4, 6, None] and cross-validation (cv = 5) included. GridSearchCV() is then used to fit the new model to the training set of data, and evaluation metrics are then applied. However, after fitting this improved model to the train data, it turned out to be a middle-performing model with a significantly lower R-squared value than the original RF model. By analyzing the graph, we can identify the features with the highest impact on the model's predictions. These are the features that contribute the most to the overall performance of the model.



Random forest model with Hypertuning metrics:

```
Refined RF model evaluation metrics:
R^2: 0.5119488214983685 (51.2%)
MSE: 0.029190208513658045 (2.9%)
MAE: 0.143655877565372 (14.4%)
MSLE: 0.015541008096290302 (1.6%)
MedAE: 0.11881484368836123 (11.9%)
```

The metrics used to evaluate an enhanced random forest model developed with GridSearchCV. A more excellent value denotes a better fit, and the R2 value measures how well the model can account for the variation in the target variable. Lower numbers denote more remarkable performance. The MSE and MAE values measure the average squared and absolute discrepancies between the predicted and actual values. A lower value indicates that the model makes more accurate predictions across the range of target values. The MSLE value measures the model's error distribution on a logarithmic scale. A lower value indicates more incredible performance. The MedAE value is the median of the absolute errors between the anticipated and actual values. The refined random forest model has lower R^2, higher MSE, higher MAE, and higher MedAE values than the previous model, which means that the model's ability to explain the variation in the target variable has decreased, and the average squared and absolute differences

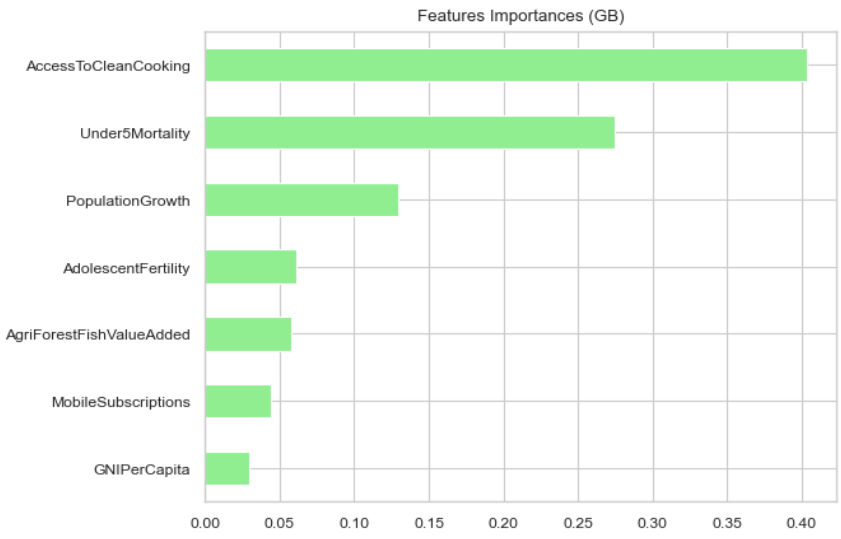
between the predicted and actual values have increased. However, the model is still making relatively accurate predictions.

5 Fold cross-validation metrics for refined random forest model:

```
5-fold X-Validation for refined random forest model
Cross-validation R^2 scores:
0.441  0.474  0.4887  0.4885  0.4537
Mean R^2: 0.4692
```

The revised random forest model is not doing as well as the original model, which has an R-squared value of 0.5119, according to the cross-validation mean R-squared value of 0.4692. This shows that the improved random forest model might be overfitting the training set and not generalizing effectively to fresh data. Additional analysis may call for more tweaking or the use of a different machine learning method.

3.3. Gradient Boosting



Gradient Boosting model evaluation metrics:

```
Gradient Boosting model evaluation metrics:
R^2: 0.6732064429357187 (67.3%)
MSE: 0.01954543394590838 (2.0%)
MAE: 0.11385609315007199 (11.4%)
MSLE: 0.010659071558157138 (1.1%)
MedAE: 0.09521817836561305 (9.5%)
```

The gradient boosting model is a predictive model that can explain 67.3% of the changes in the target variable. This model's predictions are very close to the actual values, with an average difference of only 0.1138, and half of the absolute errors being smaller than 0.0952. The MSLE value of 0.0106 indicates that the model is making accurate predictions across the entire range of target values.

5-fold CVmetrics for gradient boosting model:

```
5-fold X-Validation for Gradient Boosting model:
Cross-validation R^2 scores:
0.6499  0.6606  0.6356  0.6901  0.6527
Mean R^2: 0.6578
```

The gradient boosting model has a R-squared value of 0.6732, which means that it can explain 67.3% of the variance in the target variable. However, when the model was evaluated using cross-validation, the R-squared value was found to be 0.657, which is lower than the original R-squared value. This indicates that the model is overfitting to the training data, and may not perform well on new data. Cross-validation gives an estimate of how well the model is likely to perform on new data.

3.4. Models Summary Evaluation

	Model	R-Squared	RMSE	MAE	MSLE	MedAE
0	Decision Tree	0.49	0.17	0.14	0.02	0.12
1	Random forest	0.93	0.06	0.05	0.00	0.03
2	Refined Random forest	0.51	0.17	0.14	0.02	0.12
3	Gradient boosting	0.67	0.14	0.11	0.01	0.10

Comparing 4 machine learning models on dataset using below metrics:

- Random forest model performs the best across all metrics with 93% variance.
- The Decision Tree model performs worst across all metrics.

- Refined random forest model performs similar to the Decision Tree model.
- Gradient boosting model has intermediate performance with R-squared value of 67%.

4. Result

According to the feature importance of modeling and the correction matrix, democratic institutions are unaffected by education. Further, the Economic welfare of citizens has a slight impact on democratic institutions, while Healthcare does impact democratic institutions.

5. Discussion

5.1. Limitations

This research project has several limitations that should be considered when interpreting the results. One of the limitations is that this analysis is focused on macro-level factors, which may overlook important micro-level factors that could impact democracy. Thus, the results should be interpreted cautiously and not used as the sole basis for decision-making. Another limitation of this study is the relatively small sample size of 3740 entries, which may not be representative of all countries and may need to be more sufficient to capture the complex dynamics of the relationship between the variables of interest. Additionally, part of the analysis uses 20-year averages instead of the time series data, which may mask critical short-term changes and fluctuations in the variables. It is also important to note that while the analysis controls for certain factors, there may still be unobserved confounding variables that could affect the relationship between the independent and dependent variables. Finally, it is essential to consider the potential for selection bias, as countries may self-select into specific categories or policies that are not necessarily representative of the general population. As with any research project, these limitations should be considered when interpreting the results and drawing conclusions.

5.2. Further Research

This research could be expanded by incorporating variables impacting the democracy index, such as topography, natural resources, and geopolitical factors. For example, countries with abundant natural resources may be more susceptible to corruption and political instability, which could negatively impact democracy. Similarly, countries with strategic geopolitical positions may face more significant external pressures and interference in their political systems. Incorporating these variables into the analysis could provide a more comprehensive understanding of the relationship between various factors and the democracy index. Furthermore, the analysis could also benefit from using data at more minor geographic scales, such as city or state-level data. This would allow for a more nuanced understanding of how local factors impact democracy.

Additionally, using data at a closer interval of time series, such as quarterly or monthly data, could provide a more detailed understanding of how democracy changes over time and may reveal short-term trends and patterns not captured in the current analysis. Finally, future research could also explore the use of other models, such as Bayesian or neural network models, to analyze the relationship between various factors and the democracy index. These models may provide additional insights and more accurate predictions than the current model used in the analysis.

6. Conclusion

Based on the data that was collected and analyzed throughout this process, we now return to the three SMART questions that were asked at the beginning of this analysis:

Does education have an impact on a country's democratic institutions?

Does the economic welfare of citizens have an impact on a country's democratic institutions?

Does healthcare have an impact on a country's democratic institutions?

As pointed out above in the results section, the findings from the modeling that was conducted, as well as the correlation matrix, suggest that education does not impact education. At the same time, the economic welfare of citizens has a slight impact, and healthcare has a more noticeable impact. Unfortunately, these conclusions were obtained from variables selected to represent these broader categories and had severe limitations. More specifically, while our analysis found that primary school enrollment does not appear to impact the democracy index of countries, it might be the case that secondary school enrollment could have a much more substantial impact. The same applies to economic welfare and healthcare when looking at other variables. Unfortunately, due to the constraints of our research, we cannot look at a significant amount of variables. Therefore, we must limit our analysis to a more simplified version so that, at the very least, some basic understanding of the data can be obtained. We are confident with the results as they reflect the variables used.

References

Alemán, E., & Kim, Y. (2015). The democratizing effect of education. Research & Politics, 2(4).

Barceló, J., & Rosas, G. (2021). Endogenous democracy: Causal evidence from the potato productivity shock in the old world. Political Science Research and Methods, 9(3), 650-657.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, Agnes Cornell, M. Steven Fish, Lisa Gastaldi, Haakon Gjerløw, Adam Glynn, Sandra Grahn, Allen Hicken, Katrin Kinzelbach, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Anja Neundorf, Pamela Paxton, Daniel Pemstein, Oskar Rydén, Johannes von Römer, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Aksel Sundström, Eitan Tzelgov, Luca Uberti, Yi-ting Wang, Tore Wig, and Daniel Ziblatt. 2023. "V-Dem Codebook v13" Varieties of Democracy (V-Dem) Project.

Dahl, R. A., & Shapiro, I. (1998). Why Market-Capitalism Favors Democracy. In On Democracy (pp. 166–172).

Franco A, Alvarez-Dardet C, Ruiz MT. Effect of democracy on health: ecological study. BMJ. 2004 Dec 18;329(7480):1421-3.

Lipset, S. (1959). Some Social Requisites of Democracy: Economic Development and Political Legitimacy. American Political Science Review, 53(1), 69-105.

Narita, Yusuke and Sudo, Ayumi, "Curse of Democracy: Evidence from the 21st Century" (2021). Cowles Foundation Discussion Papers. 2632.

Tipps, D. C. (1973). Modernization Theory and the Comparative Study of Societies: A Critical Perspective. *Comparative Studies in Society and History*, *15*(2), 199–226.