Predicting Happiness, Migration, and Democracy through Socio-Economic Factors

Student Number 210250792

10/07/2024

Contents

- **Introduction**
  - Background
  - Objectives
  - Research Questions
- **Dataset Description**
  - Dataset Overview
  - Variables
- **Methodology**
  - Regression Analysis
    - Objectives
    - Techniques used
    - Results
    - Interpretation
  - Classification
    - Objectives
    - Techniques used
    - Results
    - Interpretation
  - Unsupervised Learning
    - Objectives
    - Techniques used
    - Results
    - Interpretation
- **Discussion**
  - Summary of findings
  - Comparison with existing literature
  - Implication of findings
- **Conclusion**
  - Key takeaways
  - Limitations
  - Future work

Introduction

Background

This project utilises datasets from various sources to analyse socio-economic factors influencing happiness, migration, and democracy. The primary datasets are sourced from the United Nations and include comprehensive information on GDP, population growth, fertility, and mortality indicators. Additional datasets from other reputable sources provide data on happiness scores, net migration, and democracy indicators.

Objectives

The main objectives are:

1. To use regression analysis to understand the relationship between socio-economic factors and the happiness of a country's population.
2. To use classification techniques to predict net positive migration based on happiness and other socio-economic factors.
3. To perform unsupervised learning to identify patterns in democracy status based on socio-economic indicators.

Research Questions

Each task involves focusing on one of the following questions.

1. How do countries' various economic and population factors interplay in impacting the happiness of their population?
2. To what extent do a country's measured happiness and other socio-economic factors predict its likelihood of experiencing net positive migration?
3. How effectively can we identify and differentiate democratic from non-democratic countries based on socio-economic, net positive migration and happiness indicators?

Dataset Overview

The analysis uses multiple datasets. All data was merged into one comprehensive dataset by matching country names and only including countries with complete data in all databases, resulting in a dataset with 153 countries.

1. **GDP and GDP Per Capita:** This dataset has economic information for UN-recognized countries. For this analysis, we focused on data from 2015. Columns used are total GDP and GDP growth.
2. **Population Growth, Fertility, and Mortality Indicators:** This dataset contains information on population growth, fertility, and mortality for UN-recognized countries. We focused on data from 2015. Columns used are life expectancy, maternal mortality, population annual rate of increase, and total fertility rate.
3. **Happiness and Regions:** Additional columns from a happiness database from OpenML provide average self-reported happiness scores (on a scale of 1-10) and regional information for countries for the year 2015.
4. **Net Migration:** A database from the world bank was used to add a column indicating net migration for all UN-recognized countries. For this analysis, we focused on data from 2015. The data was transformed into a binary variable indicating whether a country has a net positive migration.
5. **Democracy Indicators:** A dataset produced by Professor José Antonio Cheibub from Pittsburgh University was used to add a binary indicator of whether each country is democratic. The latest data was from 2008, so we had to use that information and assume that few countries changed their democratic status between 2008 and 2015.

Variables

- **GDP Per Capita:** GDP per person in a country.
- **GDP Growth:** Yearly GDP growth percent of a country.
- **Life Expectancy:** Average number of years a person is expected to live.
- **Maternal Mortality:** Number of maternal deaths due to births per 100,000 live births.
- **Population Growth Rate:** Annual percentage of population increase.
- **Total Fertility Rate:** Average number of children born per woman.
- **Happiness:** Average self-reported happiness on a scale of 1-10.
- **Region:** Region of the country.
- **Net Migration:** Indicator of whether a country experiences net positive migration.
- **Democracy Indicator:** Indicator of whether a country is democratic or not.

Database application

Regression Analysis

To examine the relationship between socio-economic factors and happiness, we performed a regression analysis. We added columns from the happiness database, which included average happiness scores and regional

information, to the combined dataset of GDP and population indicators. We ensured that the country names matched across datasets for accurate merging.

Classification

For predicting net positive migration, we used classification techniques. The dataset was enhanced with a column from the World Bank database indicating net positive migration. This variable was converted into a binary indicator to classify countries as experiencing net positive migration or not. The country names were standardised to match across datasets.

Unsupervised Learning

To identify patterns in democracy status, we used unsupervised learning methods. We added a binary indicator from Professor José Antonio Cheibub's database, showing whether a country is democratic. This column was merged with the main dataset after ensuring consistent country naming.

Methodology

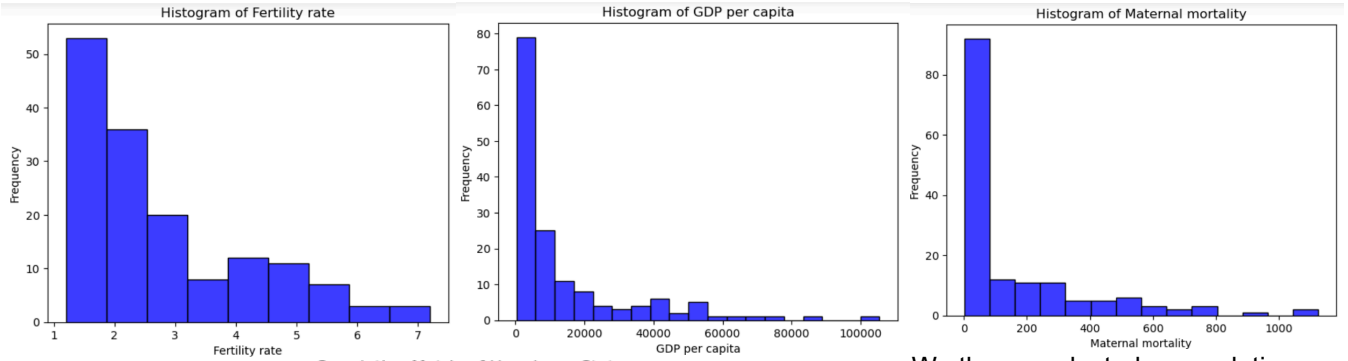
Regression Analysis

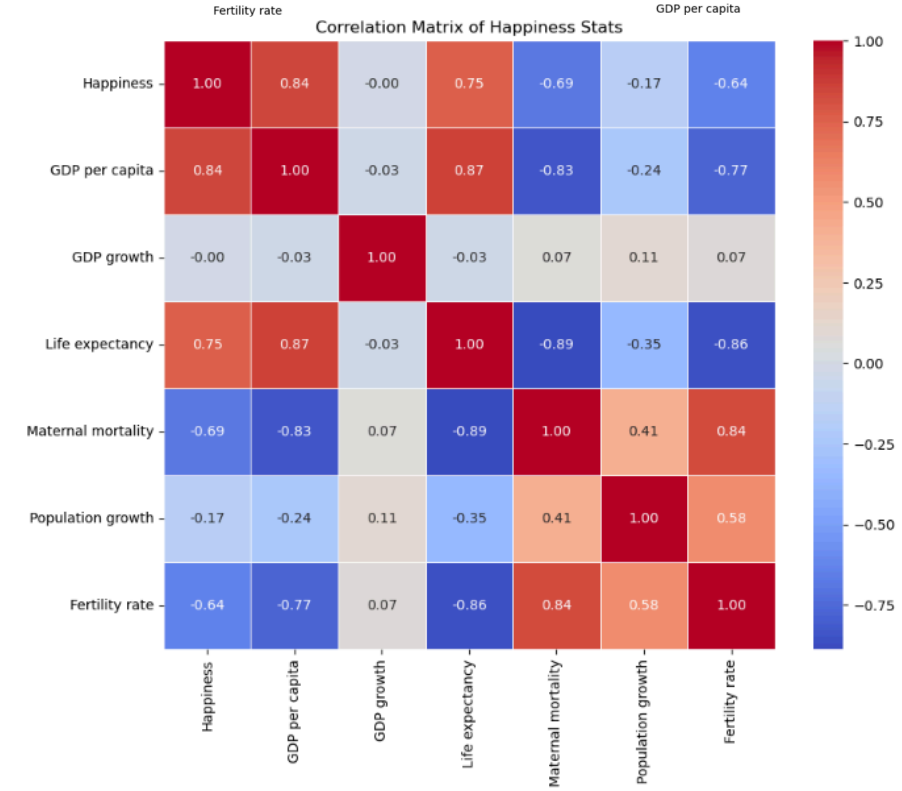
Objective: The objective of this regression analysis is to predict Happiness scores based on socio-economic variables in order to explore how various economic and population factors interact to impact the happiness of a population.

Techniques Used: Initially, Ordinary Least Squares (OLS) regression was used to examine the relationship between Happiness scores and various predictors. Concerns about multicollinearity led us to employ Bayesian Information Criterion (BIC) for model refinement. Additionally, ridge and lasso regression techniques were explored to manage multicollinearity and enhance model interpretability. Finally, random forests using decision trees were employed as an alternative approach. For data preparation, a correlation analysis among all variables in the dataset was conducted to identify relationships and potential multicollinearity issues. Visual representations such as scatter plots and correlation matrices were utilised to explore the data structure.

Results: Before fitting a regression model, it is crucial to inspect the data to identify potential issues such as skewness, which may necessitate further transformations.

The histograms depict the distributions of three variables that exhibit right skewness. To address this skewness and ensure more meaningful regression analysis, we applied a logarithmic transformation to these variables.





We then conducted a correlation analysis among all variables in our dataset and found noteworthy patterns. Happiness shows strong positive correlations with GDP and life expectancy while demonstrating moderate negative correlations with maternal mortality and fertility rate. Notably, these four variables also exhibit high inter-correlation. Multicollinearity, therefore, emerges as a significant concern, as correlation among predictors may inflate standard errors and complicate result interpretation for ordinary linear regression models. Due to this, we will employ several models and model selection techniques that could provide more reliable insights. Interestingly, GDP growth displays

notably low correlation coefficients with all other variables in our analysis.

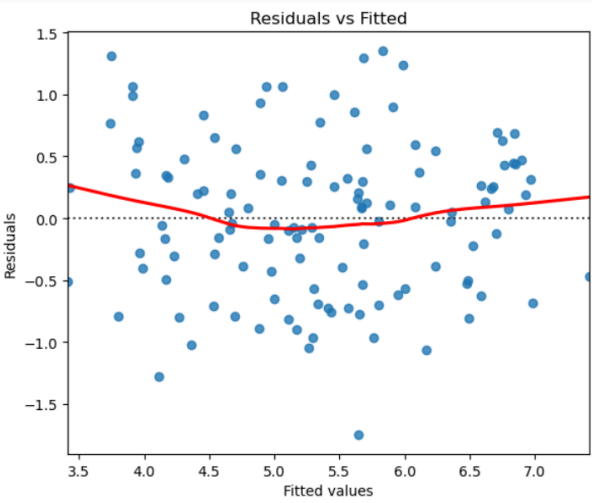
We initiated our analysis by conducting an Ordinary Least Squares (OLS) regression using all available variables. To test the accuracy of our model, we split the data into an 80% train set and a 20% test set. The results indicate a model with a robust R-squared of 0.719 on the training data, suggesting that approximately 71.9% of the variability in Happiness scores is explained by the predictors included in our model. Notably, GDP per capita emerges as highly significant (coef = 0.5972, p < 0.001), indicating a positive association with Happiness scores. Conversely, variables such as GDP growth, life expectancy, maternal mortality, population growth, and fertility rate exhibit coefficients with non-significant p-values, suggesting they do not significantly influence Happiness scores within this model. Turning to the test data, the Mean Squared Error (MSE) of 0.418, considering the Happiness variable ranges from 0 to 10, indicates a high accuracy in predicting Happiness scores, with an average deviation of less than half a unit. However, the model's predictive performance on the test set shows a diminished but still good R-squared of 0.612 compared to the training data. This suggests that while the model performs well within the training dataset, its ability to generalise to new observations is slightly less robust. When we generalised the test of our model with 153 folds, i.e., leave-one-out cross-validation, we obtained an MSE of 0.423, confirming the accuracy of our model.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.8457	1.645	-1.122	0.264	-5.104	1.413
GDP per capita	0.5972	0.087	6.849	0.000	0.425	0.770
GDP growth	0.0006	0.010	0.059	0.953	-0.020	0.021
Life expectancy	0.0268	0.020	1.343	0.182	-0.013	0.066
Maternal mortality	0.0465	0.076	0.613	0.541	-0.104	0.197
Population growth	0.0799	0.065	1.221	0.225	-0.050	0.210
Fertility rate	-0.1020	0.328	-0.310	0.757	-0.752	0.549

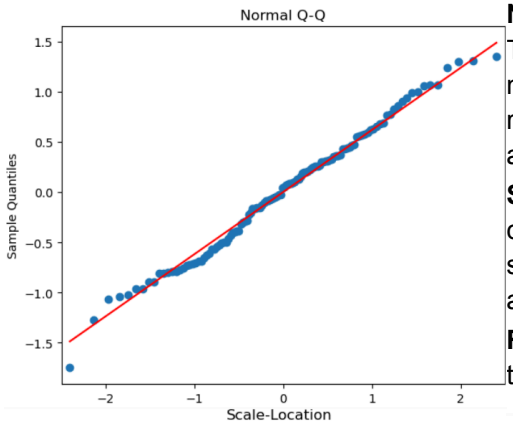
```
R-squared:      0.719
Adj. R-squared:0.704
Mean Squared Error on Test Set: 0.418
R-squared on Test Set: 0.612
```

```
Average MSE for 153 folds ('leave one out'): 0.423
```

Now, I will present some plots that demonstrate the accuracy and predictive capabilities of our regression model, highlighting its adherence to linearity and normality assumptions, and its good enough adherence to the homoscedasticity assumption with only minor heteroscedasticity.



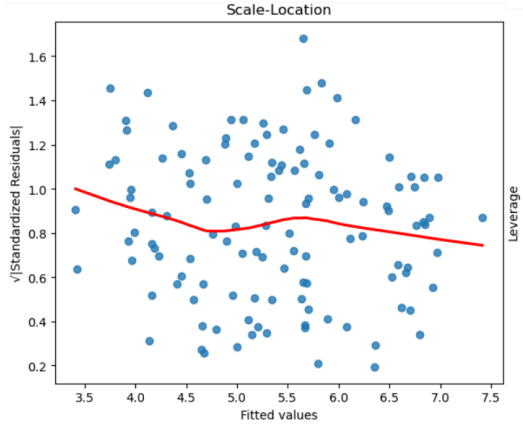
Residual vs Fitted Analysis: The Residual vs. Fitted plot is used to examine non-linearity and heteroscedasticity in our regression model. Residuals are mostly randomly scattered around zero, indicating adherence to the linearity assumption, though with slight curvature suggesting mild non-linearity. The spread of residuals remains relatively constant across fitted values, with a slight reduction towards higher fitted values, indicating minor heteroscedasticity. Given our dataset's size constraints, we will rely on the fact that this heteroscedasticity appears minor, rather than using heteroskedasticity-robust standard errors.

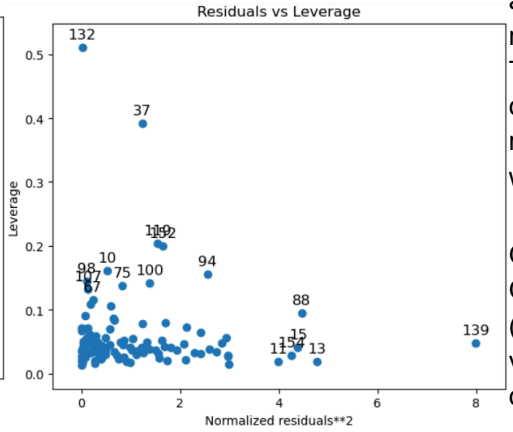


Normal Q-Q Plot: This plot assesses the normality of residuals. The observed points closely follow the 45-degree reference line, even towards the extremes, suggesting that residuals approximate a normal distribution.

Scale-Location Plot: This plot examines homoscedasticity, or the consistency of residual variance across fitted values. Points are scattered around a mostly horizontal line, showing a slight funnel shape at higher fitted values indicative of mild heteroskedasticity.

Residuals vs Leverage Plot: This plot identifies influential data points that significantly affect the regression model. Sixteen points are flagged as outliers due to their large residuals or high leverage. These outliers exert considerable influence on our regression results and may warrant further investigation.





Our initial analysis using Ordinary Least Squares (OLS) regression with all variables revealed GDP per capita as the only statistically

significant predictor of Happiness scores. Concerns about multicollinearity led us to use the Bayesian Information Criterion (BIC) model selection technique to determine which variables to include in our model. BIC suggested that a simpler model including only GDP per capita would be more reliable. On the test set, the Mean Squared Error (MSE) of 0.376 indicates improved prediction accuracy. Using 10-fold cross-validation, we obtained a slightly higher MSE of 0.407, suggesting a very slight improvement in predictive performance compared to our OLS model. This confirms that GDP per capita is a strong predictor of Happiness when considered alone and highlights how multicollinearity affected our initial model, leading to reduced predictive accuracy despite including additional variables.

	coef	std err	t	P>|t|	[0.025	0.975]
GDP per capita	0.6293	0.007	92.652	0.000	0.616	0.643

```
Mean Squared Error on Test Set: 0.376
Average MSE for 10 fold CV: 0.407
```

In addition to our initial Ordinary Least Squares (OLS) regression, we explored ridge and lasso regression techniques to refine our model and address multicollinearity. Ridge regression controls multicollinearity by adding a penalty term to the coefficient estimates, while lasso regression further promotes sparsity and automatically selects important variables. Ridge regression resulted in a slightly worse 10-fold MSE of 0.436. The coefficients from ridge regression show that GDP per capita remains highly significant (coefficient = 0.836), reinforcing its positive association with Happiness. Other variables, such as Life expectancy, also show moderate influences, albeit with smaller coefficients.

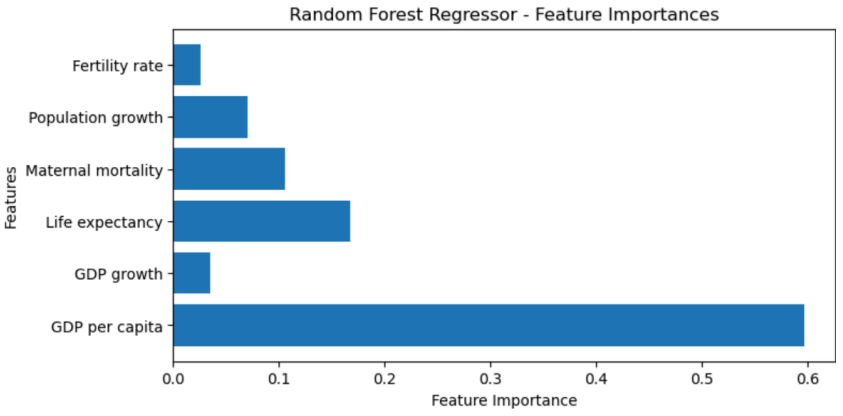
```
Ridge Regression Mean Squared Error: 0.418
Ridge Regression Coefficients:
              Feature  Coefficient
0               const     5.395777
1      GDP per capita     0.836430
2          GDP growth     0.003924
3     Life expectancy     0.224143
4  Maternal mortality     0.065755
5   Population growth     0.127894
6      Fertility rate    -0.055141
Ridge Regression 10-fold CV MSE: 0.436
Lasso Regression 10-fold CV MSE: 0.421
```

```
Lasso Regression Mean Squared Error: 0.371
Lasso Regression Coefficients:
              Feature  Coefficient
0               const     5.382202
1      GDP per capita     0.809693
2          GDP growth     0.000000
3     Life expectancy     0.091374
4  Maternal mortality    -0.000000
5   Population growth     0.000000
6      Fertility rate    -0.000000
```

Lasso regression produced a lower MSE of 0.371. However, upon conducting 10-fold cross-validation, the MSE for lasso regression was 0.421, suggesting that the initially low MSE was likely anomalous. Lasso regression effectively reduced some coefficients to zero, indicating variable selection. GDP per capita remained influential (coefficient = 0.810), while variables such as GDP growth, Maternal mortality, Population growth, and Fertility rate were penalised to zero, demonstrating that they do not significantly contribute to explaining Happiness scores in this model. Life expectancy, however, was not reduced to zero and remains somewhat influential, similar to its role in the ridge regression model.

Next, we employed random forests to predict Happiness. Random forests improve prediction accuracy by averaging the results of multiple decision trees, each built from random subsets of the data and features, thereby reducing overfitting and capturing complex patterns. We will explore decision trees further in the classification task. Utilising 100 estimators, the model achieved a 10-fold MSE of 0.35—our best result so far. Our analysis highlights GDP per capita as the most crucial predictor, with a feature importance score of almost 0.6, while Life expectancy and maternal mortality also emerge as notable predictors.

```
Mean Squared Error (Random Forest - 10-fold CV): 0.35 (± 0.12)
```



Interpretation: It's important to note that while GDP per capita consistently emerged as a robust predictor of Happiness scores across all models, our analysis is limited to explaining correlation rather than establishing causation. Demonstrating causation is inherently more complex and requires fulfilling stringent criteria beyond statistical association. Hypothesising about causation in this context suggests that an increase in GDP could potentially lead to higher Happiness scores not only because individuals enjoy increased purchasing power but also indirectly through factors strongly correlated with GDP per capita, such as fertility rate and life expectancy. The use of advanced techniques like random forests has improved our model's accuracy and revealed more nuanced relationships between predictors and Happiness, underscoring the benefit of sophisticated methods in capturing complex patterns that simpler models might miss. Although GDP per capita consistently proves to be a strong predictor, the slight variations in model performance across different techniques (OLS, Ridge, Lasso, and Random Forests) highlight the importance of careful model selection to achieve the best possible predictive accuracy. This

variability emphasises the need for thorough evaluation to ensure reliable results. These insights collectively suggest that while GDP per capita is a critical factor in explaining Happiness, the pathways through which it influences well-being are complex and multifaceted. Further research, ideally involving experimental designs or rigorous longitudinal studies, would be necessary to definitively establish causal relationships and deepen our understanding of these dynamics.

Classification

Objective: This study aims to evaluate how a country's measured happiness and various socio-economic factors predict its likelihood of experiencing net positive migration. By analysing these factors, the research seeks to understand their relative contributions in shaping migration patterns.

Techniques Used:This study utilises logistic regression and decision tree classifiers to analyse how variables predict net positive migration patterns. Logistic regression initially identified significant predictors such as GDP per capita and fertility rate, while decision trees further explored variable interactions and importance.

Results: We initiated our analysis with a basic logistic regression to determine which socio-economic variables best predict whether a country will experience net positive migration. GDP per capita emerged as highly significant (coef = 1.3637, p < 0.001), indicating a strong positive association with the likelihood of net positive migration. Conversely, variables such as GDP growth, life expectancy, and maternal mortality displayed non-significant coefficients (p > 0.05), suggesting they do not substantially affect migration outcomes in this context. Unexpectedly, the fertility rate also showed significance, with an even higher coefficient than GDP per capita (coef = 3.2778, p = 0.004), suggesting that higher fertility rates are associated with increased odds of net positive migration. This correlation is intriguing, as poorer countries typically have higher fertility rates. To understand this better, we need to explore the underlying causes of fertility rates. It's possible that migrants themselves have higher fertility rates, leading to countries with net positive migration having a greater average fertility rate. The model's accuracy on test data is reported at 71%, with the confusion matrix showing balanced predictions but with higher sensitivity than specificity. Leave-one-out cross-validation also produced an accuracy of 71%, revealing slightly higher specificity than sensitivity.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-19.4217	5.991	-3.242	0.001	-31.164	-7.680
GDP per capita	1.3637	0.374	3.645	0.000	0.630	2.097
GDP growth	0.0237	0.050	0.472	0.637	-0.075	0.122
Life expectancy	0.0680	0.074	0.919	0.358	-0.077	0.213
Maternal mortality	-0.1000	0.278	-0.360	0.719	-0.645	0.445
Fertility rate	3.2778	1.149	2.854	0.004	1.027	5.529

Test Set Accuracy: 0.71
Confusion Matrix:

	Predicted Negative	Predicted Positive
Actual Negative	10	6
Actual Positive	3	12

LOOCV Accuracy: 0.71
Confusion Matrix:

	Predicted Negative	Predicted Positive
Actual Negative	67	18
Actual Positive	26	42

We then used Bayesian Information Criterion (BIC) to select the most effective model for predicting net positive migration. The BIC-selected model, which includes GDP per capita, life expectancy, and maternal mortality as significant predictors, is more parsimonious than the initial model. It highlights the importance of GDP per capita (coef = 1.2070, p < 0.001) in predicting higher odds of net positive migration, while life expectancy and maternal mortality do not show statistically significant effects. Notably, fertility rate, previously a significant predictor, is excluded from the reduced model, suggesting it is not crucial for predicting net migration after accounting for collinearity. This shift might explain why maternal mortality's coefficient changed from negative in the earlier model to positive in the BIC-selected model as it is strongly positively correlated with fertility rate. The 10-fold cross-validated accuracy of 0.69 for the BIC model is only 0.02 lower than the original logistic regression, demonstrating that we can halve the predictors while maintaining comparable accuracy.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-9.8736	4.603	-2.145	0.032	-18.895	-0.852
GDP per capita	1.2070	0.335	3.607	0.000	0.551	1.863
Life expectancy	-0.0216	0.061	-0.353	0.724	-0.142	0.098
Maternal mortality	0.2309	0.249	0.929	0.353	-0.256	0.718

Test Set Accuracy (BIC Model): 0.61
Confusion Matrix (BIC Model):

	Predicted Negative	Predicted Positive
Actual Negative	6	10
Actual Positive	2	13

Average Test Set Accuracy (10-Fold CV): 0.69
Average Confusion Matrix (10-Fold CV):

	Predicted Negative	Predicted Positive
Actual Negative	6.3	2.2
Actual Positive	2.5	4.3

Next, we examined a simple logistic regression model predicting net positive migration based solely on Happiness. The

	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.3436	1.111	-4.812	0.000	-7.520	-3.167
Happiness	0.9492	0.203	4.683	0.000	0.552	1.346

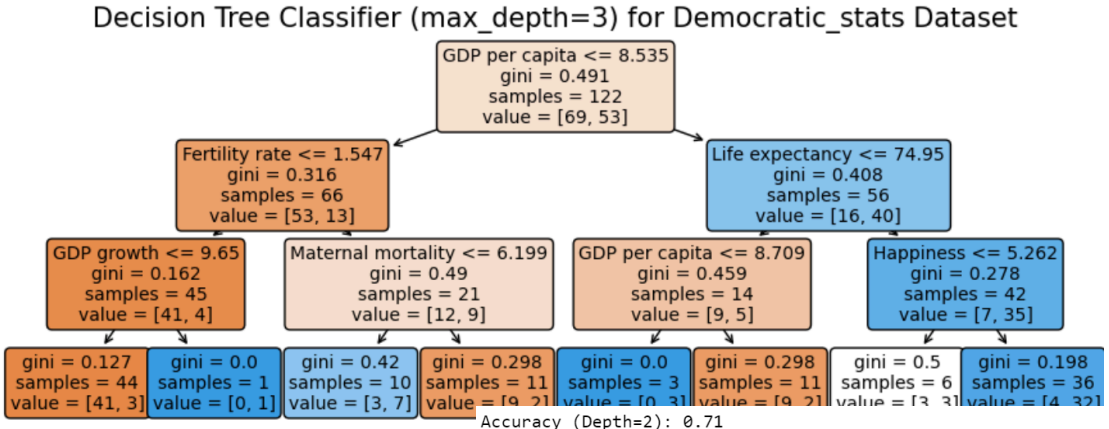
model reveals a coefficient of 0.9492 ($p < 0.001$), indicating a significant positive association with the likelihood of net positive migration. Despite its simplicity, this model achieves an accuracy of 0.71 on both the test set and in 50-fold cross-validation, comparable to the performance of more complex models. Evaluating the predictive accuracy of each variable in isolation using 50-fold cross-validation shows that Happiness can be a strong predictor of net positive migration, on par with GDP per capita and life expectancy. However, it's essential to consider the ultimate drivers of net migration. While Happiness might appear crucial for migrants, other variables like GDP and health (e.g., life expectancy) are also significant drivers, often correlating with Happiness. Demonstrating causation and isolating the effects of GDP or Happiness on net migration under controlled conditions is challenging due to the limited number of countries and incomplete data on migrants.

```
Test Set Accuracy: 0.71
Confusion Matrix:
                 Predicted Negative  Predicted Positive
Actual Negative                  11                   5
Actual Positive                   4                  11
```

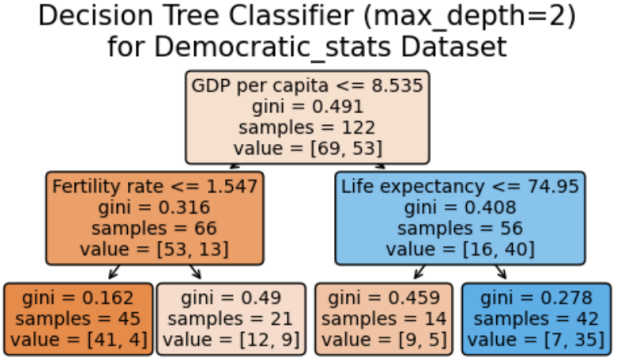
```
Average Test Set Accuracy when predicting
Net Migration using one variable on 50 folds:
GDP per capita 0.71
GDP growth 0.55
Life expectancy 0.70
Maternal mortality 0.68
Fertility rate 0.63
Happiness 0.71
```

We utilised a decision tree classifier to further analyse how variables predict net positive migration patterns. Decision trees simplify complex decision-making by recursively splitting data based on feature values, providing a clear visualisation of the impact of various factors. Using a tree with a maximum depth of 3, we achieved an accuracy of 0.74 and a leave-one-out accuracy of 0.72. GDP emerged as the initial split, with all variables appearing at one level of the tree. This suggests that each variable contributes to predicting net migration, albeit with some variability. Reducing the depth to 2 resulted in an accuracy of 0.71 but a notably lower leave-one-out accuracy of 0.59, indicating that the initial model's performance may have been somewhat fortuitous. Only GDP per capita, fertility rate, and life expectancy remained important predictors. GDP per capita consistently outperformed Happiness in predicting net migration, likely due to their high correlation, which led to the exclusion of Happiness from the model. This suggests GDP per capita plays a more significant role in influencing net migration, though we refrain from implying causation due to data limitations.

```
Accuracy (Depth=3): 0.74
```




```
Accuracy (Depth=2): 0.71
```




```
LOO Accuracy (Depth=3): 0.72
LOO Accuracy (Depth=2): 0.59
```

Interpretation: Across all models, GDP per capita consistently emerged as a robust predictor of net positive migration, highlighting its pivotal role in migration decisions. The BIC-selected model performed almost as well as the basic logistic regression model, indicating multicollinearity among our variables. When tested individually, GDP per capita, happiness, and life expectancy all proved to be strong predictors of net positive migration, with average MSEs better than our BIC model and as good as our logistic model, suggesting potential overfitting and the need for more data to improve model performance. A key question remains: do migrants prioritise happiness, economics, or health when deciding to migrate or stay? Decision tree analysis underscored GDP per capita as the primary predictor, emphasising its significant role in migration patterns. Other factors contributed to varying extents based on model depth, showing the balance between simplicity and capturing nuanced interactions. This suggests GDP per capita may be the most significant factor in migration decisions, more so than happiness and life expectancy, whose significance in logistic models might stem from their correlation with GDP per capita. However, decision trees did not significantly outperform simple one-variable models, making definitive conclusions challenging. Overall, the analysis indicates that richer, healthier, and happier countries are more likely to experience net positive migration, and with these variables, we can predict migration likelihood with over 70% accuracy. This suggests migrants are somewhat rational decision-makers. However, our prediction accuracy being just over 70% indicates that many other factors, such as distance between countries and migration policies, likely influence their decisions

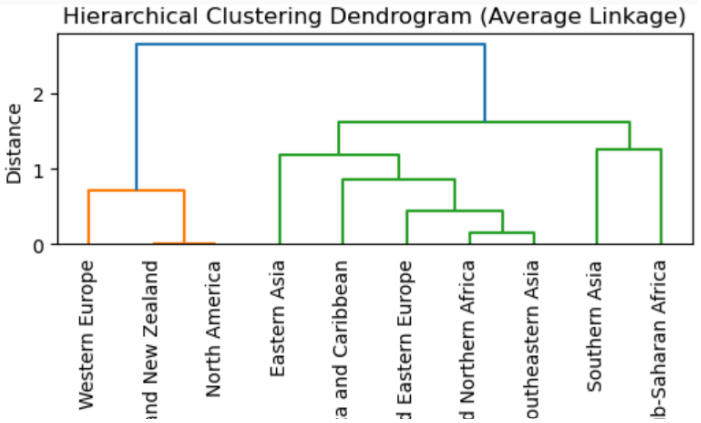
and were not explored in this study.

Unsupervised Learning

Objective: The purpose of this study is to use unsupervised learning to identify homogeneous groups of countries and determine the extent to which happiness, net migration, and other socio-economic variables can predict a country's democratic status. By leveraging clustering, dimensionality reduction, and neural learning techniques, we aim to explore the connection between democracies and their socio-economic variables and assess how effectively we can categorise a country as democratic or non-democratic based on these variables.

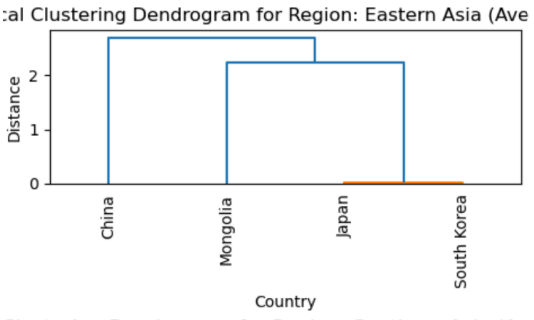
Techniques Used: We started with hierarchical clustering to build a dendrogram using the Happiness and Net Migration variables. We also employed the K-means clustering method to partition our data into two clusters based on all variables except country, region, and democracy, and analysed how these clusters matched the democratic status of countries. Next, we used Principal Component Analysis (PCA) to reduce the dimensionality of our data, focusing on the first few principal components to capture the majority of the variance and understand the relationships between variables. Finally, we built a neural network using Keras to predict democracy based on our variables, testing the model's performance with and without PCA-reduced data to evaluate efficiency and accuracy.

Results: We initiate our analysis by building a dendrogram using hierarchical clustering with average linkage. This method constructs a hierarchy of clusters by iteratively merging the pair of clusters with the smallest average distance between their data points. We use only the Happiness and Net Migration variables to explore how well we can group regions into democratic and non-democratic categories based on these two simple variables. We grouped countries into regions and averaged the variables for each region to analyze them as a whole, allowing us to explore whether regions with similar values for these two variables also have similar levels of democracy.

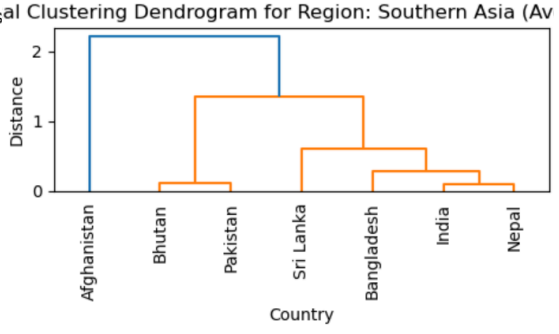


Western Europe, New Zealand, and North America form the first split with a significant drop, showing little variance among these regions. This is notable since all the countries in these regions are democratic, with a democracy score of 1. Latin America and Eastern Asia also split off early and both have high democratic rates of 0.955 and 0.75, respectively. In contrast, the Middle East and Southeastern Asia are very similar, with the lowest democratic rates of 0.1 and 0.3, respectively. We can already see that, without any training method, our two variables seem to sort our regions into more and less democratic ones. Next, we examine individual countries within a few regions. In Eastern Asia, China, the only non-democratic country, splits off from Mongolia, Japan, and South Korea, the latter two being very similar to each other. In Southern Asia, Afghanistan, one of the two non-democratic countries, splits off immediately with a large gap, whereas Bangladesh appears similar to other countries in the region. We chose to analyse these regions as they have fewer countries, allowing for easier visualisation. However, this trend holds across all regions to some extent. It appears that by sorting countries based on happiness and net positive migration, they categorise into democratic and non-democratic groups. Democratic countries tend to group together when analysing these two variables. While we will later hypothesise whether democracy affects these variables or the other way around, it is worth noting for now that democratic states may have more freedom, directly improving happiness and attracting migration. Additionally, democracies may influence happiness and migration through other variables, such as better GDP and healthcare, due to free market and non-corrupt policies.

```
Democratic rate by Region
Australia and New Zealand          1.000
North America                      1.000
Western Europe                     1.000
Latin America and Caribbean        0.955
Eastern Asia                       0.750
Southern Asia                      0.714
Central and Eastern Europe         0.679
Sub-Saharan Africa                 0.350
Southeastern Asia                  0.333
Middle East and Northern Africa    0.105
```

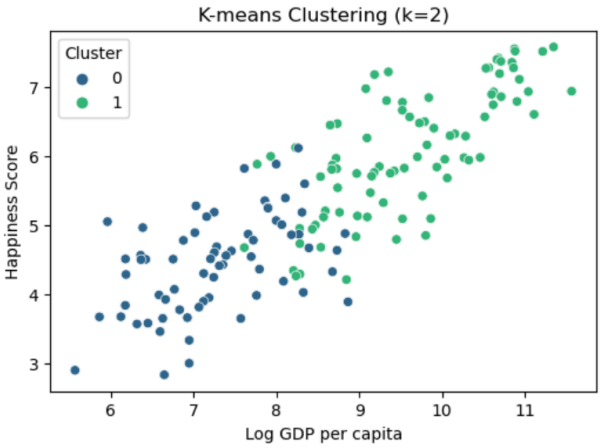


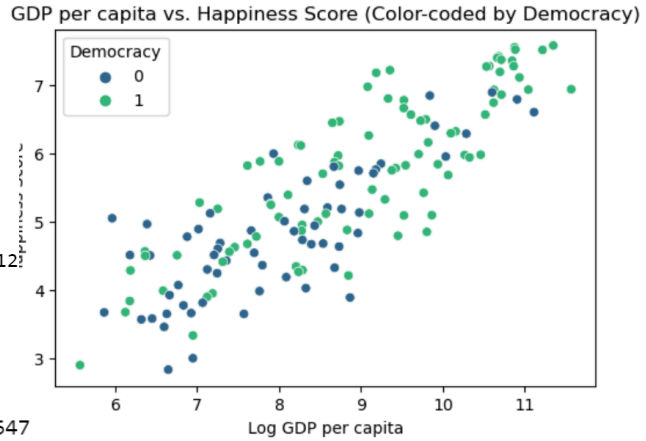

```
Democracies in Eastern Asia
       Country  democracy
44       Japan          1
45  South Korea         1
95     Mongolia          1
```



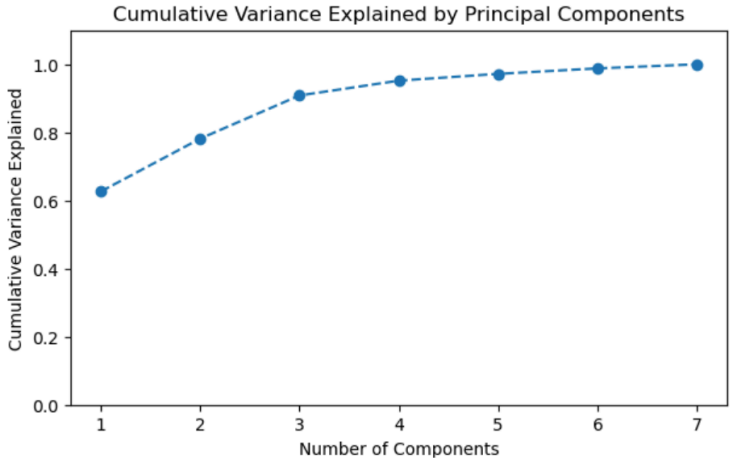

```
Democracies in Southern As
       Country  democracy
74      Bhutan          1
76    Pakistan          1
111      India          1
115      Nepal          1
126  Sri Lanka          1
```

Next, we employed the K-means clustering method to partition our data into two clusters based on all variables except country, region, and democracy. K-means clustering is a method that partitions data into a specified number of clusters by assigning each data point to the cluster with the nearest mean, then recalculating the means and reassigning data points until convergence. The initial plot depicts GDP against happiness for our countries, colour-coded by the two clusters. Notably, the clusters predominantly segregate into groups of countries with low GDP and happiness versus those with high GDP and happiness. Next, we recreated the plot, this time color-coded by whether a country is democratic or not. It's observed that democracies tend to exhibit higher levels of happiness and GDP per capita, although the distinction is less pronounced compared to our clusters. When we examine the matrix of cluster counts against democracy, we find that 71.2% of countries are classified into identical clusters and democratic groups. This represents a significant improvement compared to splitting the data by GDP, even when using identical quantile sizes for democratic and non-democratic countries, which achieved an accuracy of 64.7%. This demonstrates the strength of our correlation and the power of unsupervised learning; without using any training or targeting of the democratic variable, we can group the countries based on our other variables alone with 71.2% accuracy.

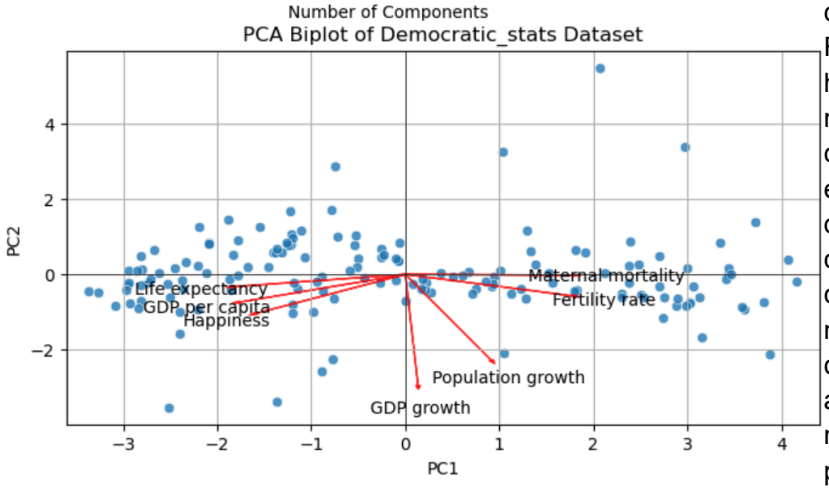





```
Cluster Counts:
democracy   0   1
Cluster
0          41  23
1          21  68
Overall Accuracy: 0.712
GDP split_counts:
democracy   0   1
GDP split
0          35  27
1          27  64
Overall Accuracy: 0.647
```



Before building our neural network, we performed PCA analysis to reduce the number of variables required to effectively represent our data. Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms a large set of correlated variables into a smaller set of uncorrelated variables called principal components, which capture the most variance in the data. The initial plot shows the cumulative variance explained by the principal components. Notably, the first three components capture approximately 91% of the variance, and the fourth captures 95%, indicating substantial data compression with minimal information loss. Following this, the subsequent plot illustrates how our various continuous variables are mapped using the first two principal components. The arrows in the plot, exaggerated for clarity, reveal interesting directions: Life expectancy and GDP per capita, which are strongly positively correlated with happiness, are oriented negatively along the first principal component. In contrast, maternal mortality and fertility rate, which are strongly negatively correlated with happiness, are positively oriented in the same component.



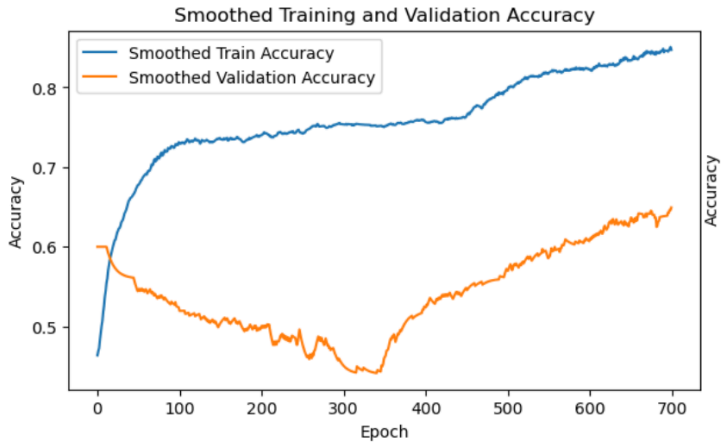
Meanwhile, GDP growth and population growth, which show weaker correlations with happiness, are predominantly aligned with the second principal component. The majority of our arrows run close to the first principal component, explaining why more than 60% of the variance is captured by just the first PCA alone.

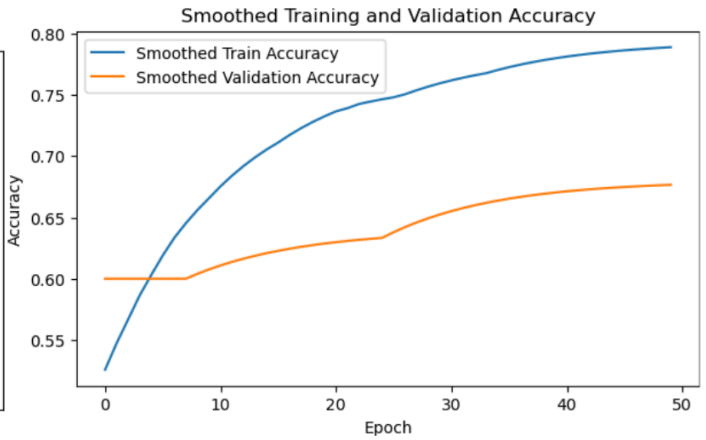
Finally, we built a neural network using Keras, a popular deep learning library, to predict democracy based on our variables. Neural networks are computational models consisting of interconnected layers of nodes (neurons) that learn to recognize patterns in data through iterative adjustments of connection weights during training. In neural

network training, we use three different datasets: training, validation, and test sets. The training set is used to adjust weights and biases to minimise error, while the validation set helps in tuning hyperparameters and evaluating performance to prevent overfitting. The test set provides an unbiased evaluation of the model's ability to generalise to new, unseen data. After 700 epochs, our model achieved an excellent test accuracy of 0.77, significantly outperforming previous clustering methods in predicting democracy. While the training accuracy improved consistently, the validation accuracy initially declined before stabilising after 350 epochs, likely due to initial overfitting with the validation set. Further training beyond 700 epochs offered only marginal improvements and became both time-consuming and computationally intensive. To enhance our model, we can experiment with adjusting the neural network structure, optimising parameters such as the learning rate, increasing training data volume, applying regularisation methods like dropout, and testing different optimization algorithms. Additionally, we explored using the first 4 principal components from PCA to train the model. This approach achieved a test accuracy of 0.74, nearly matching the previous accuracy of 0.77 but required only 50 epochs compared to 700 epochs with the original model. This reduction in epochs and computational complexity highlights PCA's effectiveness in reducing data dimensionality, which translates to significant savings in time and computational resources. Such efficiency enables quicker model training and opens up opportunities for developing more sophisticated models.

```
Prediction Accuracies: Normal Model
Train Accuracy: 0.83
Test Accuracy: 0.77
```

```
Prediction Accuracies: 4 PCA Component Model
rain Accuracy: 0.77
est Accuracy: 0.74
```





Interpretation: Our models illustrate that unsupervised learning techniques effectively group countries based on happiness, migration, and other socio-economic variables, revealing distinct patterns that differentiate democratic from non-democratic states. K-means clustering, in particular, demonstrates strong predictive power for democratic status, outperforming traditional economic indicators like GDP. The neural network achieved an excellent prediction accuracy of 0.77, indicating robust correlations between democratic status and socio-economic variables. While establishing causation is challenging, we can hypothesise in two main ways. First, democracy might drive improvements in happiness, GDP per capita, and population health through increased political freedoms and free market policies. These freedoms could directly enhance happiness and indirectly boost GDP and healthcare quality, creating a positive feedback loop. Second, economic growth might lead to more democratic societies, with GDP serving as a catalyst for political reforms. Economic advancements could empower populations to demand greater political freedoms and democratic governance, linking prosperity to democracy. The truth likely involves elements of both explanations, but historical evidence leans toward the first. For instance, countries that had similar economies after World War II diverged significantly based on their political systems. Nations like Poland and Hungary, with comparable pre-WWII economies to West European countries, lagged under Soviet influence but saw rapid economic growth after transitioning to democracy post-Soviet Union collapse, suggesting that democratic governance played a crucial role in their development. The study also underscores the importance of dimensionality reduction for enhancing model efficiency and interpretability. PCA's ability to retain significant variance while reducing data complexity facilitates quicker and less resource-intensive model training, demonstrating its value in data analysis and machine learning applications.

Discussion

Summary of Findings: Our findings consistently highlight the crucial role of GDP per capita in predicting both happiness and net positive migration. GDP per capita emerges as a robust predictor across all models, demonstrating its strong correlation with happiness, fertility rate, and life expectancy. This suggests that economic prosperity influences broader societal well-being through interconnected mechanisms, emphasizing the need for further research to clarify definitive causal relationships. Decision tree analyses show that economic conditions significantly shape migration decisions, often outweighing subjective well-being indicators like happiness. Additionally, unsupervised learning techniques, including K-means clustering and neural networks, effectively distinguish democratic from non-democratic states based on socio-economic variables. These results imply that the

political freedoms associated with democracy may enhance happiness, GDP per capita, and healthcare quality, potentially creating a positive feedback loop.

Comparison with Existing Literature:

https://greggvanourek.com/what-leads-to-happiness/

In this article, Gregg Vanourek demonstrates how research suggests that while genetics and circumstances play a role, intentional activities and practices, such as regular exercise, acts of kindness, meaningful relationships, and gratitude, are key to boosting our happiness and well-being. Our study, however, demonstrated that these activities alone are not enough; populations in poor and undemocratic countries may engage in these positive practices but remain significantly less happy. I would posit that although money, health, and freedom alone are not sufficient to guarantee happiness, it is very difficult to be happy if they are lacking.

https://journals.sagepub.com/doi/10.1177/0197918320949825?icid=int.sj-abstract.citing-articles.16

This study identifies that prospective migrants from the Middle East and North Africa prioritise liberal democratic governance and employment prospects when choosing destinations, with welfare benefits as a secondary factor, while geographic distance and co-ethnic stock have little impact. This aligns with our findings that democracy and GDP per capita are closely linked to causing net positive migration to a country. Additionally, migrants from the Middle East and North Africa (the least democratic region) likely prioritise political stability and freedoms even more, as they are often fleeing from these issues.

Implications:

- **Migration and Decision-Making:** Our research indicates that GDP, happiness, and democracy are critical factors driving net positive migration, suggesting that migrants are rational decision-makers seeking better opportunities. Future studies should explore potential long-term impacts of migration on GDP and happiness to provide a comprehensive understanding of these dynamics.
- **Democracy and Societal Prosperity:** The correlation between democracy and economic, health, and happiness indicators implies that democratic governance may play a pivotal role in societal prosperity and stability, positively influencing migration patterns. Democratic states tend to exhibit higher levels of GDP and better health outcomes, which in turn attract more migrants.
- **Basic Needs for Happiness:** While many studies emphasise the importance of activities and relationships in achieving happiness, our findings suggest that basic health and wealth are also essential. There are minimum levels of wealth and health necessary for individuals to achieve high happiness through their activities and relationships, indicating that economic and health stability are foundational to well-being.

Conclusion

Key Takeaways:

- GDP per capita is a robust predictor of both happiness and net positive migration, indicating that economic prosperity significantly influences societal well-being.
- Economic conditions, as revealed by decision tree and logistic regression analyses, play a more critical role in shaping migration decisions than subjective well-being indicators like happiness.
- Political freedoms associated with democratic governance appear to improve happiness, GDP per capita, and healthcare quality, potentially creating a positive feedback loop.

Limitations:

- The analysis primarily relies on cross-sectional data, limiting the ability to establish definitive causal relationships between GDP per capita, happiness, and migration.
- Potential confounding variables, such as cultural factors and regional differences, may influence the observed relationships and were not fully accounted for in the study.
- The study's findings are based on available data, which may not capture the full spectrum of factors influencing happiness and migration decisions.
- There may be biases in the self-reported measures of happiness and other subjective well-being indicators used in the analysis.

Future Work: Future research should investigate the long-term effects of net migration on GDP and happiness within individual countries through longitudinal studies and econometric models. This research should analyse the impact of migration on various economic sectors and assess the fiscal contributions of migrants. Additionally, to evaluate isolationist claims, it should explore how migration influences social cohesion and happiness in host populations, with a focus on integration policies. Furthermore, examining the disparity in happiness between poorer and wealthier individuals in the UK could shed light on how GDP per capita affects well-being, particularly given the relatively uniform access to healthcare and infrastructure across income groups. This could reveal the extent to which the impact of GDP per capita is mediated through these factors.