# Happiness Index Predictor

Final Report Paper

Gariel Mahwastu, Garien Rahadi, & Vedant Kanabar

CIS*4780

## Table of Contents

## Introduction

The happiness index is a widely used metric to assess the well-being and the quality of life in countries around the world. Understanding what aspects affect the happiness index of a country can be beneficial for countries to continuously reflect on their policies and conditions to determine how happy people are. This report will document the machine-learning project we created to predict the happiness index of EU countries using data from 2012 to 2021, analyzing features such as economic performance, employment rate, educational indexes, governance, and various relevant metrics.

## Background/Motivation

There are several reasons why we decided to study and predict the happiness index. Firstly, the happiness index can help policymakers in their decision-making process for new policies. Predicting how policy changes affect the happiness index can be useful in helping governments and policymakers make the right decisions and create policies that can benefit everyone (“Report Calls”, 2013). Secondly, the happiness index can guide the government’s investment decisions in social programs by assessing which social programs are the most effective for the improvement of the development and well-being of the country, which can also benefit governments in allocating funds effectively (Cotofan, 2023). For example, if a country’s poor educational programs are significantly hurting the country’s happiness index, it will allow them to focus more on the improvement of educational programs. Thirdly, the happiness index can give us a more complete picture of assessing a country by also focusing on the citizen’s quality of life rather than purely looking at the economic measures (Stanley, 2024). This can be helpful so that the citizen’s quality of life can also be considered for growth rather than only on economic factors.

## Literature

There have been several studies in the past that we came across in predicting happiness index using machine learning. A project by Azad et al. (2023) and a project by Akanbi et al. (2024) are similar to what we are trying to achieve by looking at various indexes like economics, freedom, etc. However, the main difference between our project and theirs is that we only focus on countries in the European Union instead of the whole world. This is because we realize that countries in the European Union have closer characteristics in their economy, policies, and culture compared to countries outside of the European Union. Focusing only on European Union countries can help us study the pattern of features more easily.

Another difference is that they also seem to only take data from 2 to 3 years, while we are taking data from 2012 to 2021. This can help us in using more data to train our model to be more accurate with its prediction.

## Problem Statement

In this project, there are two key questions that we want to address:

- Can we predict a country's happiness index based on socio-economic data?
- Which factors have the most significant impact on happiness?

This project is aimed at predicting a country's happiness index based on a given socioeconomic data using machine learning. We also would like to find out the most significant features of our socioeconomic data that have a greater impact on the happiness index.

## Dataset Description

We gathered our data from several sources, which include:

- World development indicators from World Bank Group (World Bank Group, n.d.).
    - Includes data like employee compensation, educational attainment, unemployment rate, GDP growth, etc.
- World Happiness Report (World Happiness Report, 2024).
    - Includes data like the happiness index, social support, perception of corruption, freedom to make life choices, etc.
- Fraser Institute (Fraser Institute, n.d.).
    - Includes data like economic freedom, size of government, tax policy, reliability of police, tariffs, etc.

In total, our dataset consists of data from countries in the European Union from 2012 to 2021, except for the United Kingdom. We decided not to include the United Kingdom since they have their own currency and policies and recently have left the European Union. This brings our total dataset to consist of 267 rows representing each country and year from 2012-2021 and 240 columns (features) for each row consisting of various features mentioned above. However, we also have some missing data that we will discuss further in the next section. Our data was collected by downloading files in XLSX or CSV format from the datasets mentioned above.

# Design/Methodology

## Data Preprocessing:

For our data preprocessing, our main objective is to connect all of the three datasets above into a single CSV file. We used Pandas as a tool to help us preprocess our data more easily.

Data Preprocessing Steps:

1. Matched the names of countries that are written differently from our dataset.
2. Changed the dataset that is in XLSX format to CSV.
3. Changed the structure of each CSV file to be consistent with one another.
4. Combine all of the CSV files into a single CSV file.
5. Drop any columns that have missing data of more than 40%.
6. With data missing less than 40%, we put the missing data as 0.
7. Convert any ranges data (e.g. 50-55) to the midpoint of the range (e.g. 52.5)
8. Drop data with missing happiness index.
9. Convert the master CSV to a JSON file to help us work with our data easier.

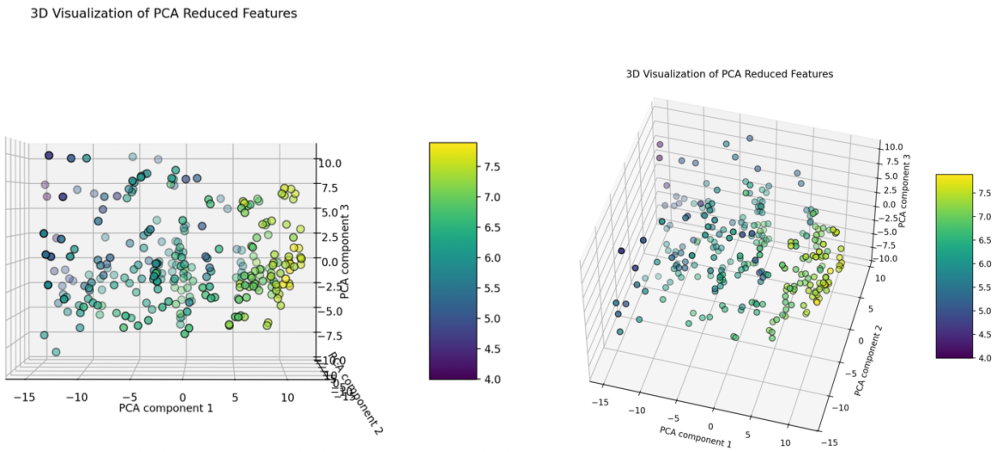
Some of our missing data are from the World Development Indicators dataset that we decided to drop if more than 40% of the data are missing. We also dropped data from some countries and years, like the Czech Republic in 2019 and Luxembourg in 2020 and 2021, since they are missing the happiness index.

Additionally, combining the data and creating the JSON files with labels and features was a challenge for us as we had never worked with large amounts of data and columns at such. Through research and trial and error, we were able to achieve our goals with the data wrangling into JSON format.

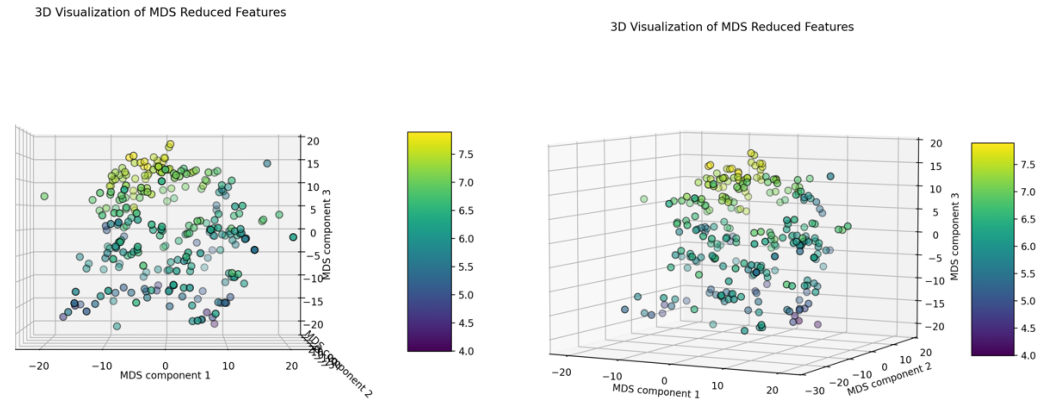
## Data Visualization:

For our data visualization, we utilized different dimensionality reduction methods such as PCA (Principal Component Analysis, which aims to preserve variance in data most), MDS (Multidimensional Scaling, which aims to preserve relative distance between observations), and t-SNE (t-distributed Socratic Neighbours Embedding which aims to preserve local neighbourhoods in the data). The results are as follows:

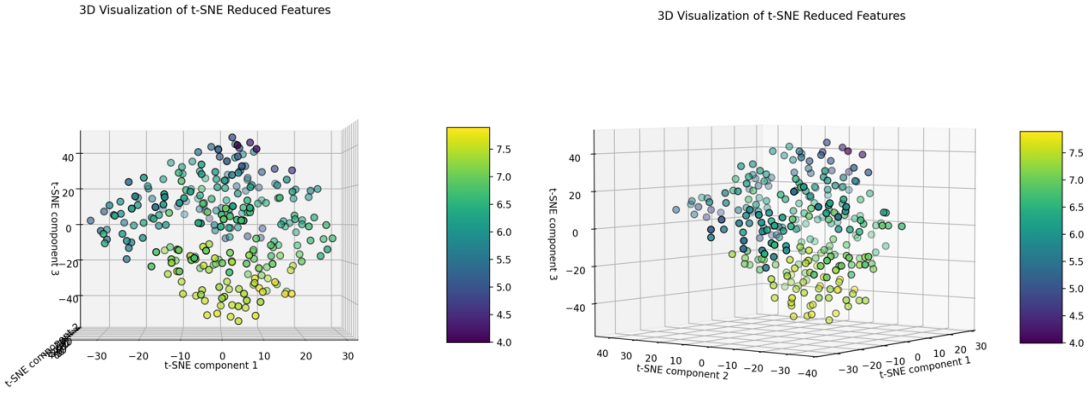
Result for PCA:




Result for MDS:




Result for t-SNE:




From these data visualizations, the shades represent the happiness index, and each dot represents a specific feature in our dataset. We can see that there is a pattern between the features and the happiness index for each of the countries, which highlights the potential correlation between the two.

## Model Evaluation:

Our next step is model evaluation. We have a Python script that will evaluate several different regression models to help us determine the best model. For this step, we evaluated Linear Regression, Lasso Regression, Random Forest, Decision Tree, and Support Vector Regression (SVR) from Scikit-learn.

Model Evaluation Steps:

1. Read the data from the JSON file, extracting the labels for each data point as well as the features and corresponding feature names.
2. Normalize our data using the StandardScaler function from Scikit-learn, which calculates the z-score.
3. Run 10-fold cross-validation for each model to determine the $R^2$ Scores and Mean Squared Error.
4. Select the best model based on the average $R^2$ scores and average MSE (Mean Squared Error) score.

To normalize the data, we use a function called StandardScaler, which applies normalization by finding the z-score of the data using the following formula:

$$z = \frac{x - \mu_{feature}}{\sigma_{feature}}$$

Where $\mu_{feature}$ is the mean of the feature, $\sigma_{feature}$ is the standard variance of the feature, and $x$ is the feature value. We chose these since most of our features are indexes, which, in theory, are normally distributed, hence fitting our data well. It is important to note, however, that some of our features weren't best suited for this, e.g., the Rank of country in Police Reliability. We still chose to use a standard scaler for this to have the normalization be the same, but ideally, we would want to differentiate the normalization for such features in our dataset

For evaluation, we used two metrics, $R^2$ and MSE (Mean Squared Error). $R^2$ is the Correlation of determination, which is a statistical measure that tells us how much of the variance in the dependent variable (happiness index) depends on the model (features). It is a value between 0 and 1, and $R^2$ value of 1 indicates that all the variance in the dependent variable is based on the model, while $R^2$ value of 0 indicates none of the variance in the dependent variable depends on the model. $R^2$ value closer to 1 indicates a strong model and is what we are looking for.

MSE (Mean Squared Error), on the other hand, is the average of the squares of the differences between the estimated values (predicted by the model) of the dependent variable (happiness index) and the actual values (From our datasets) of the dependent variable. A value of MSE closer to 0 indicates a strong model, while a very high MSE would indicate issues in the model.

This is the result of our 10-fold cross-validation:

| Model | Average R2 Score | Standard Deviation R2 | Min R2 | Max R2 | Average MSE | Standard Deviation MSE | Min MSE | Max MSE |
|---|---|---|---|---|---|---|---|---|
| **Linear Regression** | -2.718546e+19 | 8.155638e+19 | -2.718546e+20 | 0.729543 | 2.286630e+19 | 6.859890e+19 | 0.119516 | 2.286630e+20 |
| **Lasso Regression** | -9.349858e-02 | 8.404257e-02 | -2.901783e-01 | -0.008302 | 6.084468e-01 | 2.032107e-01 | 0.352297 | 1.017455e+00 |
| **Random Forest** | 9.022545e-01 | 6.996983e-02 | 7.607446e-01 | 0.962070 | 5.070453e-02 | 2.404718e-02 | 0.018996 | 8.841463e-02 |
| **Decision Tree** | 8.170855e-01 | 8.437978e-02 | 6.305710e-01 | 0.935452 | 9.662081e-02 | 2.870384e-02 | 0.061784 | 1.658099e-01 |
| **SVR** | 9.039415e-01 | 4.724066e-02 | 8.011841e-01 | 0.958329 | 5.097689e-02 | 2.498583e-02 | 0.020975 | 1.030664e-01 |

Green = worse
Blue = better

According to our evaluation, the best model based on the average $R^2$ score is SVR, and based on the average MSE score, it is Random Forest. This evaluation result prompted us to train both models to see which one gave us a more accurate result after conducting hyperparameter optimization.

## Hyperparameter Optimization:

Our next step is to optimize the hyperparameter for both SVR and Random Forest. To do this, we utilized the RandomizedSearchCV function from Scikit-learn. The RandomizedSearchCV function is a “fit” and “score” function that fits the models to a sample of defined model hyperparameters and tries to find the best parameter setting for your model through random sampling. This helps with the tradeoff of quality and runtime when trying to optimize the hyperparameters of a model.

Hyperparameters for Random Forest:

| Hyperparameter | Value |
|---|---|
| Number of trees (n_estimators) | Choose a number between 50 and 200. |
| Maximum depth for each tree (max_depth) | Choose between None, 10, 20, and 30. |
| Minimum number of samples required to create a new branch (min_samples_split) | Choose a number between 2 and 10. |
| Minimum number of samples required to be at a leaf node (min_samples_leaf) | Choose a number between 1 and 5. |
| Whether to use bootstrap or not (bootstrap) | Choose between True or False. |

Best Hyperparameters for Random Forest:

| Hyperparameter | Value |
|---|---|
| Number of trees (n_estimators) | 196 |
| Maximum depth for each tree (max_depth) | 30 |
| Minimum number of samples required to create a new branch (min_samples_split) | 2 |
| Minimum number of samples required to be at a leaf node (min_samples_leaf) | 1 |
| Whether to use bootstrap or not (bootstrap) | True |

Hyperparameters for SVR:

| Hyperparameter | Value |
|---|---|
| Regularization parameter (c) | Choose a number between 0.01 to 10. |
| Degree of polynomial kernel (degree) | Choose between 2, 3, 4, and 5. |
| Epsilon (epsilon) | Choose a number between 0.01 to 1. |
| Types of kernel (kernel) | Choose between linear, poly, rbf, and sigmoid |
| Gamma (gamma) | Choose between scale or auto |

Best Hyperparameters for SVR:

| Hyperparameter | Value |
|---|---|
| Regularization parameter (c) | np.float64(2.472234061341474) |
| Degree of polynomial kernel (degree) | 3 |
| Epsilon (epsilon) | np.float64(0.07110400554102801) |
| Types of kernel (kernel) | poly |
| Gamma (gamma) | auto |

## Model Training:

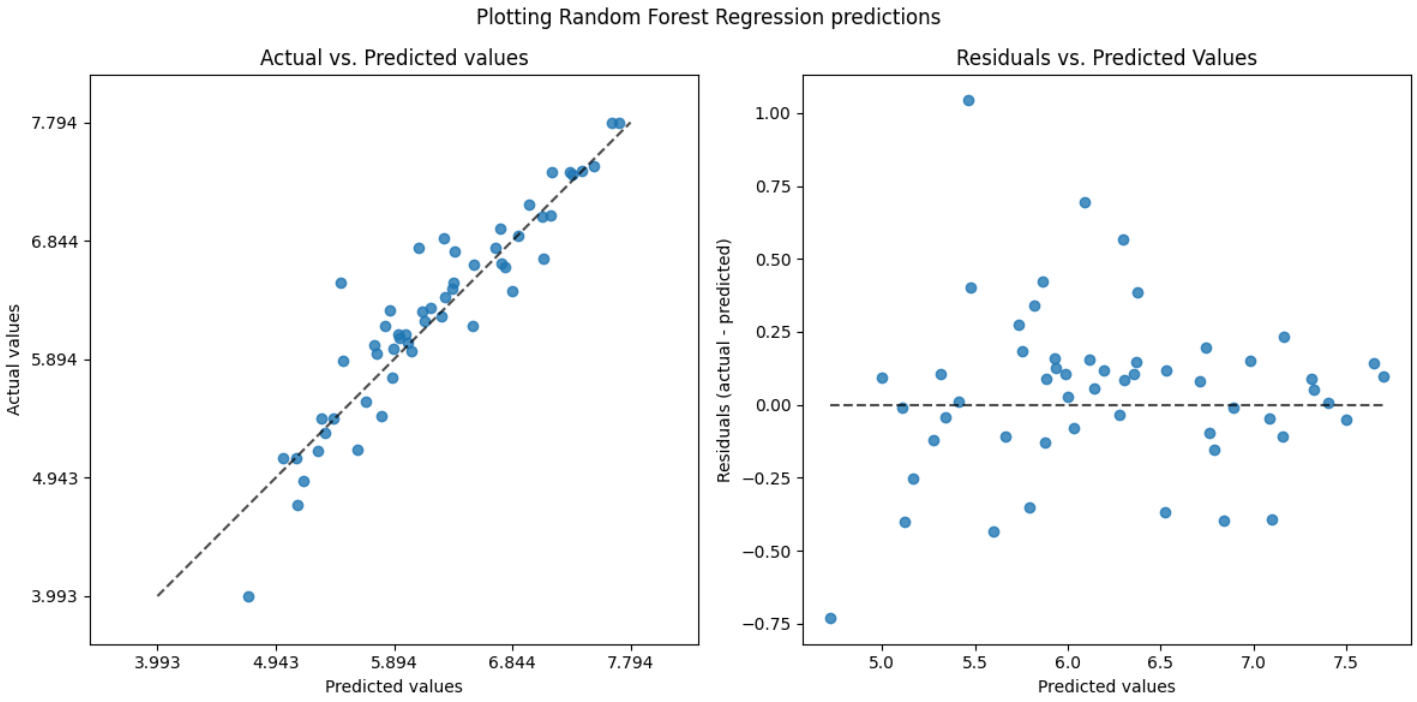
Our next step is model training. In this step, we trained both the SVR and Random Forest models. For our training, we opted not to use any dimensionality reduction given that our dataset is relatively small, and we want to preserve as much detail as possible. We also utilized the result of the hyperparameter optimization above on our models and analyzed the results of both models. To split our data into training and testing data, we utilized the train_test_split function from Scikit-learn by using 80% of our data for training and 20% for testing. The train_test_split randomly splits the data into training and testing based on defined percentages. To evaluate the model, we used $R^2$(As explained above), Mean Squared Error (MSE, as explained above), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). RMSE is the square root of the MSE value, while MAE is like MSE, where instead of squaring the errors, we just take their absolute values

## Random Forest Model:

| Testing Evaluation (Using Testing Data) | Result |
|---|---|
| $R^2$ | 0.8692900547197634 |
| Mean Squared Error (MSE) | 0.08466197865876188 |
| Root Mean Squared Error (RMSE) | 0.29096731544756343 |
| Mean Absolute Error (MAE) | 0.20713802917564086 |

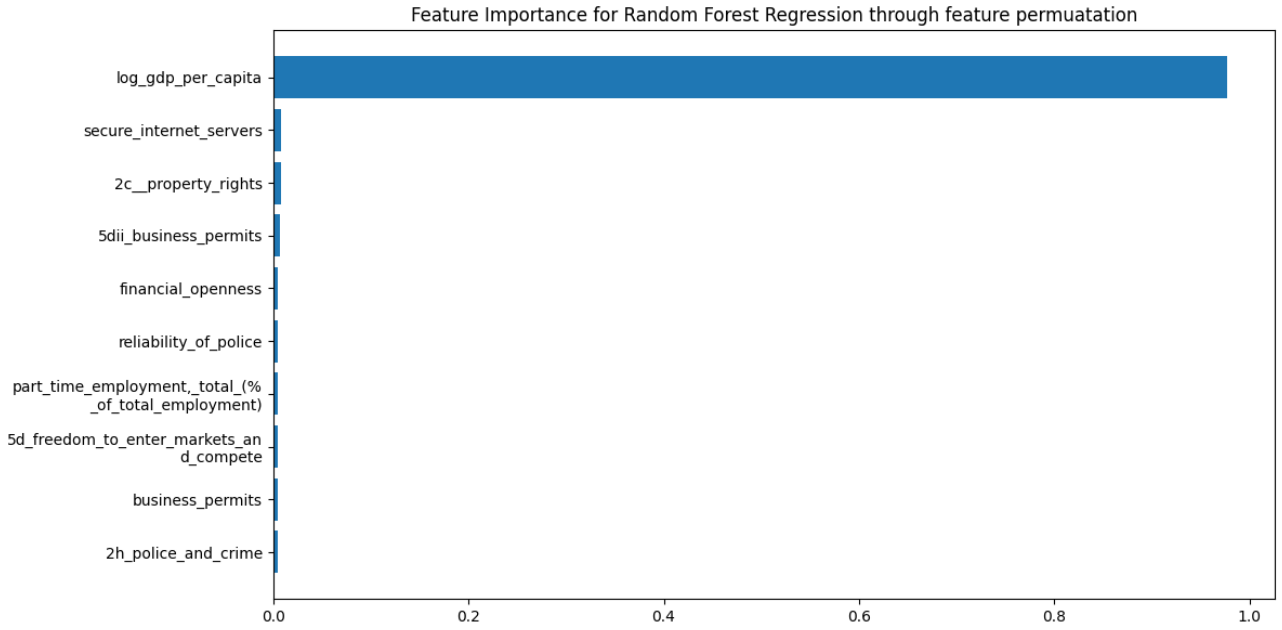
The $R^2$ value of 0.869 indicates that 86.9% of the variability we predicted was due to the model indicating a strong model. The MSE of 0.084, RMSE of 0.290 and MAE of 0.207 all indicate very low errors. As the happiness index is from 0-10, an error of 0.207 is about ± 2.7%, which is still relatively low.

We also computed the residual plots for the data:




As we can see from the left chart above, our Random Forest Model predicts a pretty accurate result on the testing data. The chart on the right can tell us whether our model is overfitting or underfitting. Looking at the plots, the residuals appear to be random with no underlying patterns, suggesting an appropriate model for the data and it's a good fit.

We evaluate the most important features using the permutation feature importance method. To calculate the importance of a feature, we calculate the increase in model prediction error when we permute the feature (i.e. move it around). A feature is "important" if shuffling its value increases the prediction error, while it is said to be unimportant if shuffling its value causes no change in model prediction error (Molnar, 2024). This concept was introduced in the Random Forest paper (Breiman, 2001), which introduced the Random Forests model. The results are below:




| Features | Importance Value |
|---|---|
| log_gdp_per_capita | 0.977179 |
| secure_internet_servers | 0.008349 |
| 2c__property_rights | 0.007729 |
| 5dii_business_permits | 0.006594 |
| financial_openness | 0.005140 |
| reliability_of_police | 0.005078 |
| part_time_employment,_total_(%_of_total_employment) | 0.004902 |
| 5d_freedom_to_enter_markets_and_compete | 0.004681 |
| business_permits | 0.004470 |
| 2h_police_and_crime | 0.004098 |

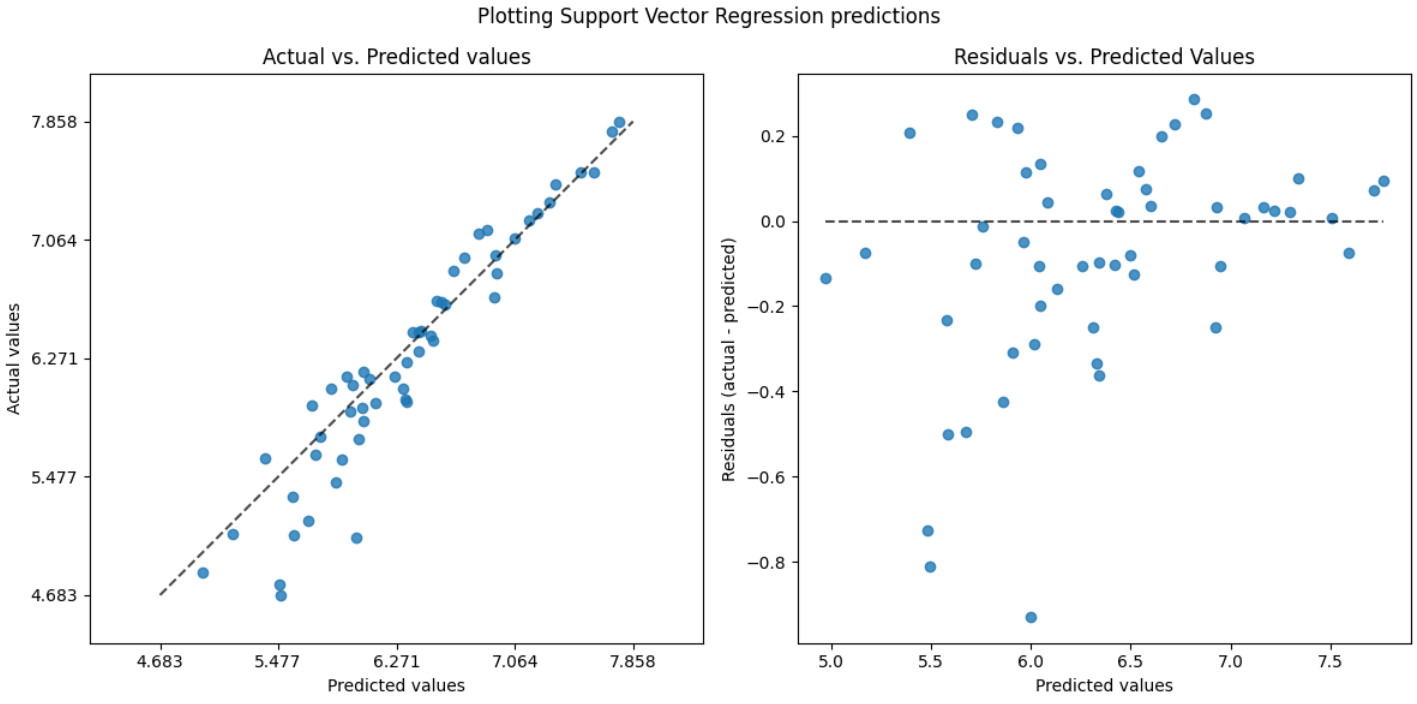
From the feature importance graph above, we can see that the feature that seems to heavily influence the prediction is GDP per capita by a significant margin. We thought this was because of the strong correlation

between the two variables, which led the Random Forest Model to pick up on it and be heavily influenced by it as it is a feature that easily discriminates the data. This is because, through our Randomized hyperparameter search, we found a value for min_samples_leaf to be 1. This hyperparameter controls the minimum number of samples to the right and left branches of any tree. If we had the time to go through all possibilities and do a Grid search instead of a Randomized search, we might have been able to find a model that is less reliant on GDP per capita. The fact that we were still able to make accurate predictions shows how important a feature of GDP per capita is in our data.

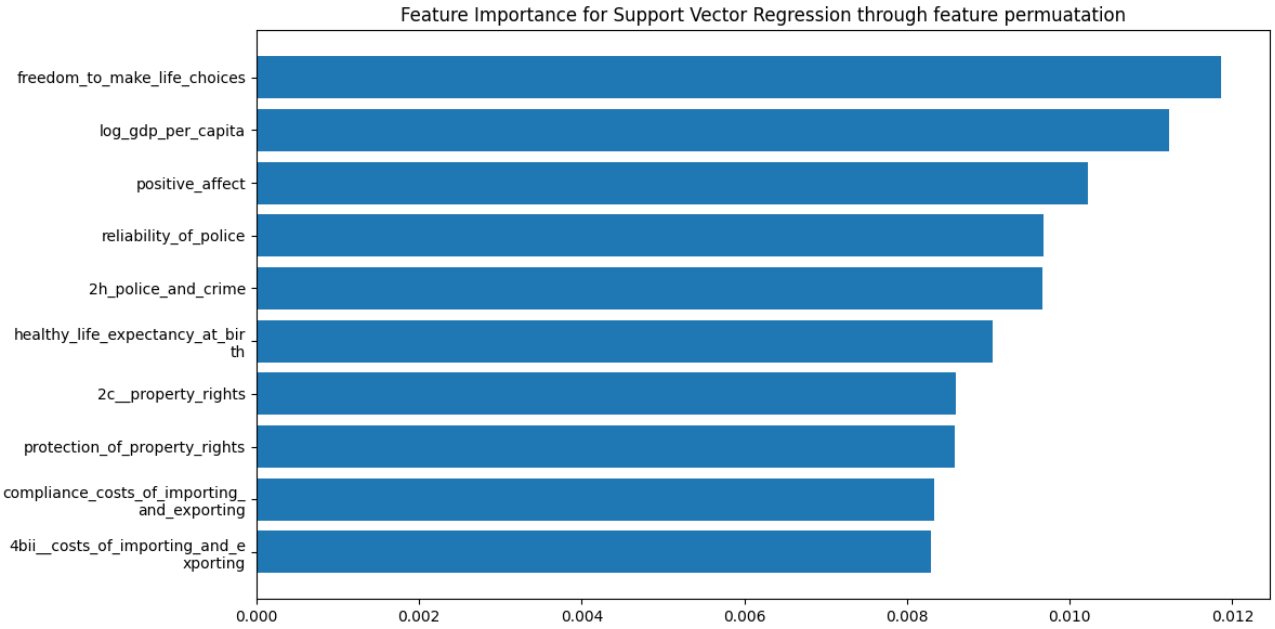
## Support Vector Regression:

| Testing Evaluation (Using Testing Data) | Result |
|---|---|
| $R^2$ | 0.8784683475008479 |
| Mean Squared Error | 0.07488697701878194 |
| Root Mean Squared Error | 0.27365485016491475 |
| Mean Absolute Error | 0.19150189358417094 |

The $R^2$ value of 0.878 indicates that 87.8% of the variability we predicted was due to the model indicating a strong model. The MSE of 0.074, RMSE of 0.273 and MAE of 0.191 all indicate very low errors. As the happiness index is from 0-10, an error of 0.191 is about ± 1.91%, which is still relatively low.



As we can see from the chart above, the SVR model seems to be performing well, just like our Random Forest Model. From the left chart, we can see how our SVR model seems to be predicting the testing value well. The residual vs predicted values chart also seems to show randomly scattered dots, which indicates that our model has a good fit.




| Features | Importance Value |
|---|---|
| freedom_to_make_life_choices | 0.011867 |
| log_gdp_per_capita | 0.011222 |
| positive_affect | 0.010228 |
| reliability_of_police | 0.009680 |
| 2h_police_and_crime | 0.009662 |
| healthy_life_expectancy_at_birth | 0.009051 |
| 2c__property_rights | 0.008602 |
| protection_of_property_rights | 0.008592 |
| compliance_costs_of_importing_and_exporting | 0.008331 |
| 4bii__costs_of_importing_and_exporting | 0.008295 |

This table shows a unique feature importance compared to the one from the Random Forest model. This shows a more balanced weight distribution of influence towards the happiness index. This indicates that the SVR is more sensitive to other features, unlike our random forest model, which only seems to be

heavily influenced by the GDP per capita feature. Hence why, we think SVR is more representative of the important features of the same-year model. Comparing both models, we can see some overlap in important features such as GDP per capita, property rights, reliability of police, and police and crime.

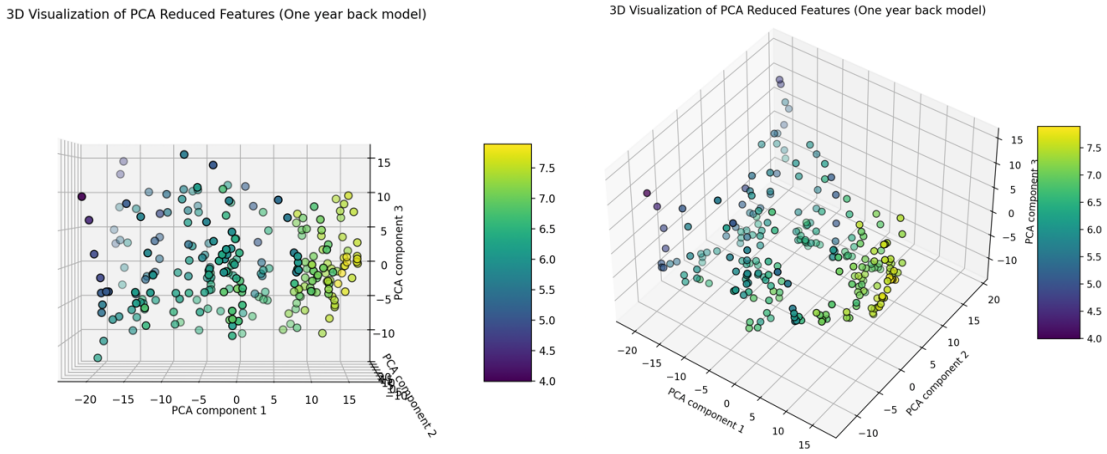
We can also see from the result of the testing evaluation with testing (unseen) data between both models the SVR model performs slightly better than the Random Forest model. It can be shown that the SVR model has a slightly higher $R^2$ value and lower MSE, RMSE, and MAE.

## Furthering the Model

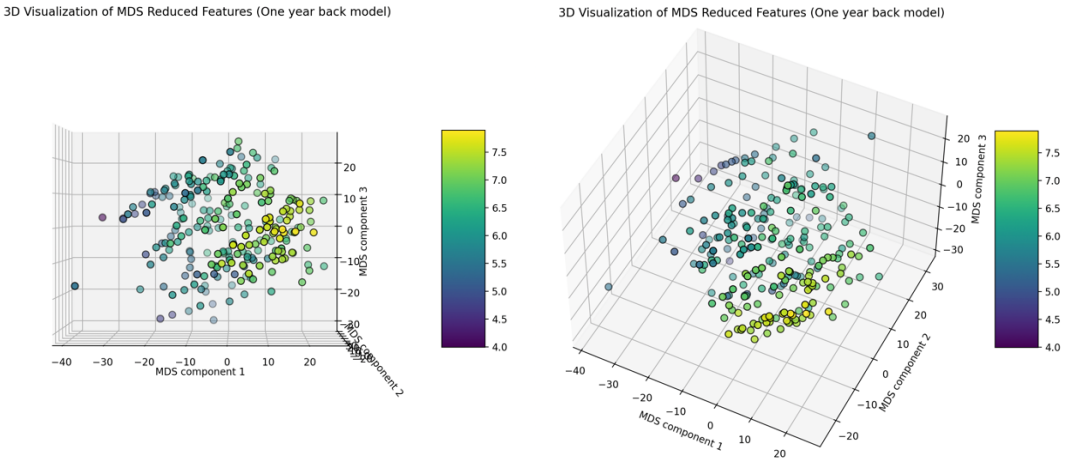
We decided to conduct experimentation by combining 2 years of features to make a prediction to see if our models can improve their prediction accuracy. For example, in making a prediction for a specific year, e.g. 2013, we will give our model features of data from that year (2013) and the year prior to that (2012). This was to study if we could come up with a stronger and more accurate model. This is also to study what the model can predict with long-term data and which features can stand out in the long term.

### Data Visualization:

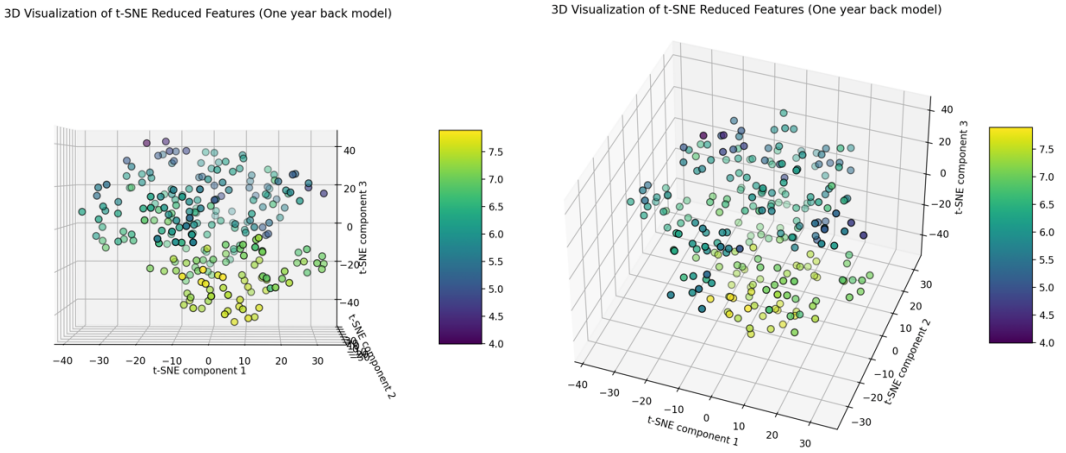
Result for PCA:



Result for MDS:




Result for t-SNE:




From the plots above, we can see they are similar to the previous model. The separation between higher happiness index data points and lower happiness index data points has increased but not by a greater margin. Since the data is not being discriminated more, we don't expect a great increase in model accuracy.

## Model Evaluation:

This is the result of our 10-fold cross-validation:

| Model | Average R2 Score | Standard Deviation R2 | Min R2 | Max R2 | Average MSE | Standard Deviation MSE | Min MSE | Max MSE |
|---|---|---|---|---|---|---|---|---|
| **Linear Regression** | 0.422308 | 0.421684 | -0.702614 | 0.756053 | 0.291031 | 0.227691 | 0.107502 | 0.928661 |
| **Lasso Regression** | -0.091939 | 0.094368 | -0.342002 | -0.000236 | 0.577358 | 0.224058 | 0.311951 | 1.110297 |
| **Random Forest** | 0.876354 | 0.106729 | 0.599731 | 0.962583 | 0.055981 | 0.044010 | 0.022575 | 0.177335 |
| **Decision Tree** | 0.782689 | 0.112373 | 0.568977 | 0.908899 | 0.099573 | 0.056461 | 0.044140 | 0.249276 |
| **SVR** | 0.874204 | 0.082743 | 0.707828 | 0.971719 | 0.066816 | 0.064754 | 0.013356 | 0.256366 |

Blue = best

Based on the average $R^2$ score and the average MSE score above, we can determine that the best model in this case is the Random Forest model. However, we still want to check using the SVR model so that we can also make a comparison with our previous SVR model as well.

## Hyperparameter Optimization:

Similar to the previous model, we ran a hyperparameter optimization script. Below are the results

Best Hyperparameters for Random Forest:

| **Hyperparameter** | **Value** |
|---|---|
| Number of trees (n_estimators) | 151 |
| Maximum depth for each tree (max_depth) | 30 |
| Minimum number of samples required to create a new branch (min_samples_split) | 4 |
| Minimum number of samples required to be at a leaf node (min_samples_leaf) | 2 |
| Whether to use bootstrap or not (bootstrap) | True |

Best Hyperparameters for SVR:

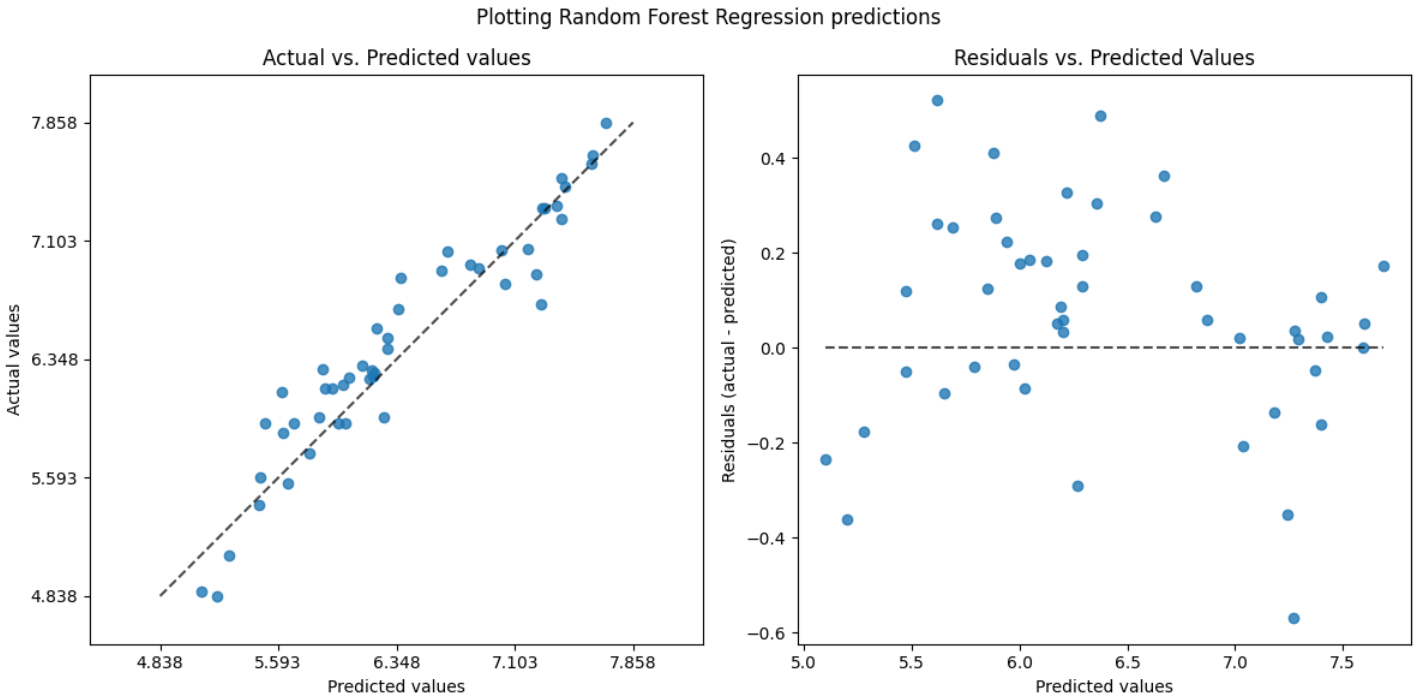
| Hyperparameter | Value |
|---|---|
| Regularization parameter (c) | 4.081446113734713 |
| Degree of polynomial kernel (degree) | 2 |
| Epsilon (epsilon) | 0.16245405416477765 |
| Types of kernel (kernel) | linear |
| Gamma (gamma) | scale |

## Random Forest Model (With features for 2 years combined):

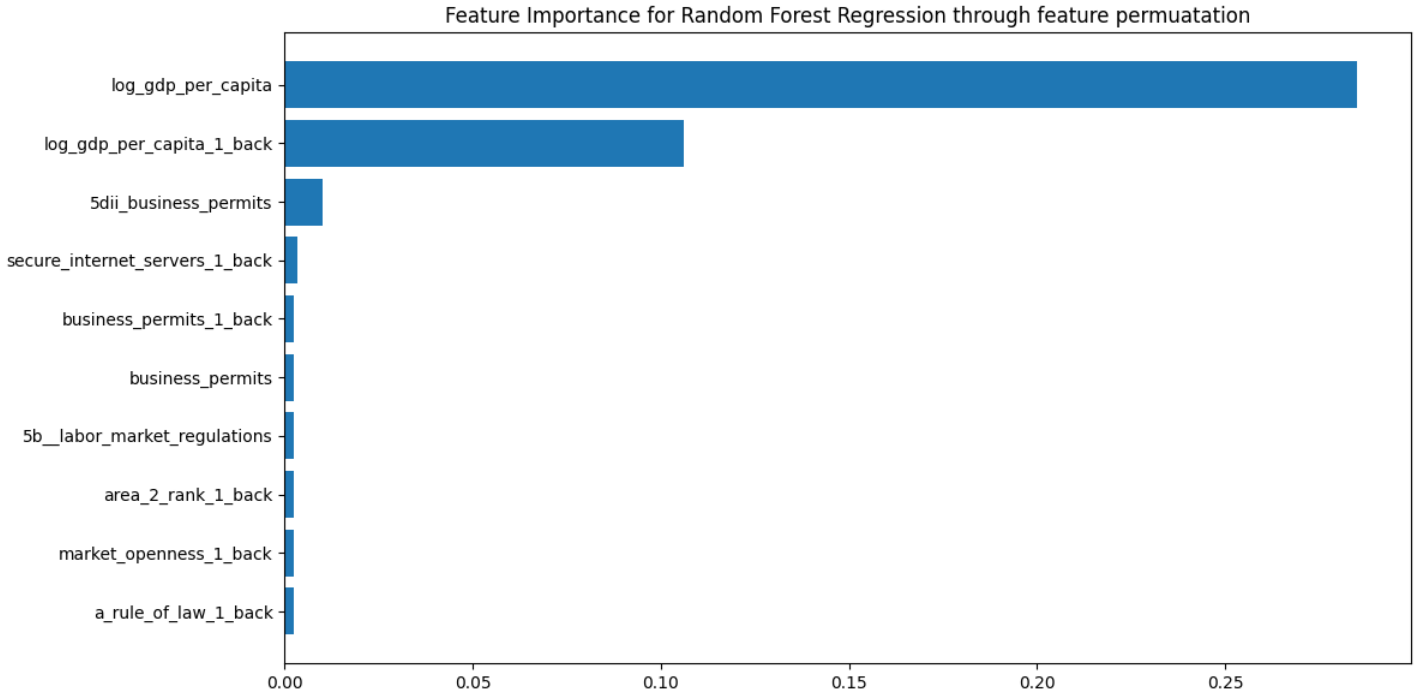
| Testing Evaluation (Using Testing Data) | Result |
|---|---|
| $R^2$ | 0.8918640350093442 |
| Mean Squared Error (MSE) | 0.05521379053329457 |
| Root Mean Squared Error (RMSE) | 0.23497614886046322 |
| Mean Absolute Error (MAE) | 0.18611439150764783 |

The $R^2$ value of 0.891 indicates that 89.1% of the variability we predicted was due to the model indicating a strong model. The MSE of 0.055, RMSE of 0.234 and MAE of 0.186 all indicate very low errors. As the happiness index is from 0-10, an error of 0.186 is about ± 1.86%, which is still relatively low. This is an improvement from the previous Random Forest model, but not by a great margin. As we can see and compare, the testing evaluation of our experimentation of combining 2 years did not result in a significant change compared to our initial model. This is expected from our plot's discussion earlier.

The residual plot we generated:




As we can see, compared to our initial Random Forest model, there don't seem to be any changes. Looking at the residual plot, it seems that our experimental model is also a good fit since the dots appear to be randomly scattered.



| Features | Importance Value |
|---|---|
| log_gdp_per_capita | 0.285110 |
| log_gdp_per_capita_1_back | 0.106132 |
| 5dii_business_permits | 0.010217 |
| secure_internet_servers_1_back | 0.003335 |
| business_permits_1_back | 0.002553 |
| business_permits | 0.002541 |
| 5b__labor_market_regulations | 0.002518 |
| area_2_rank_1_back | 0.002428 |
| market_openness_1_back | 0.002407 |
| a_rule_of_law_1_back | 0.002268 |

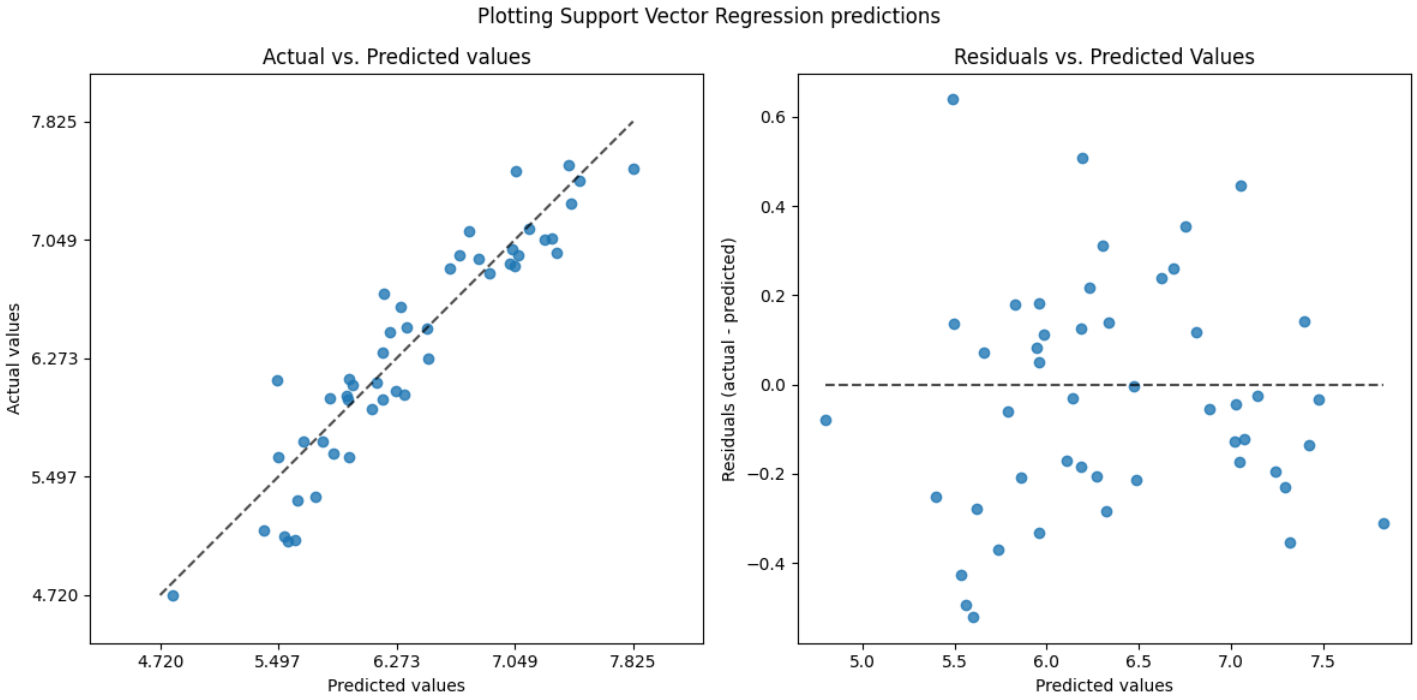
Comparing the feature importance graph above with the previous one, we can still see that both GDPs per capita from the two combined years are still heavily influencing its prediction, but the influence has reduced in value. However, importantly, we see that there are repeat significant features over time, e.g. log_gdp_per_capita & log_gdp_per_capita_1_back (the log GDP per capita) and business_permits & business_permits_1_back (Rating 0-10 time taken to get a new business permit). Also, features that are important from one year back, e.g. market_openness_1_back (Economy Open market index), secure_internet_servers_1_back (Internet security), area_2_rank_1_back (Legal Systems and Property Rights Ranking). These show the long-term features that are important in the happiness of an economy and guide the policy makers for what to focus on in improving the economy. Overall, in comparison to the first model, this model seems less reliant on one feature

## Support Vector Regression (With features for 2 years combined):

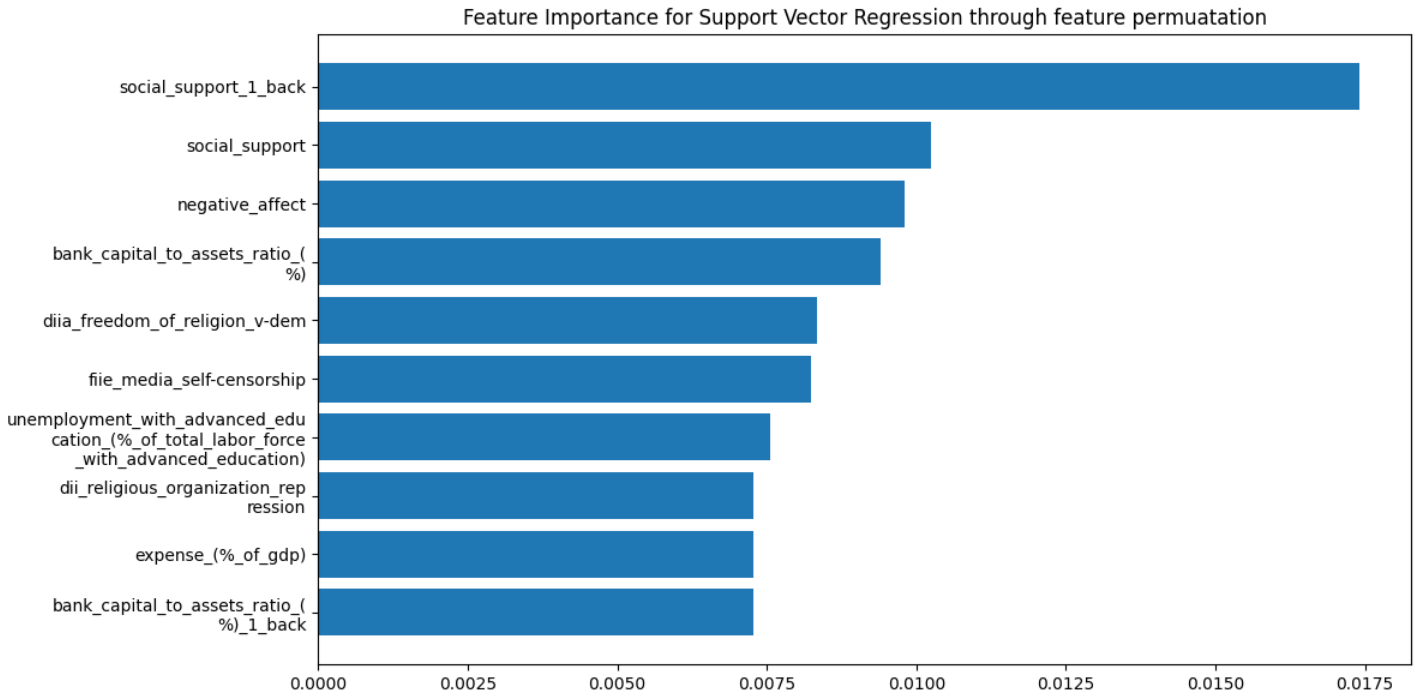
| Testing Evaluation (Using Testing Data) | Result |
|---|---|
| $R^2$ | 0.8749984923648766 |
| Mean Squared Error (MSE) | 0.06675894180221652 |
| Root Mean Squared Error (RMSE) | 0.25837751798911707 |
| Mean Absolute Error (MAE) | 0.21302912188003922 |

The $R^2$ value of 0.874 indicates that 87.4% of the variability we predicted was due to the model indicating a strong model. The MSE of 0.066, RMSE of 0.258 and MAE of 0.213 all indicate very low errors. As the happiness index is from 0-10, an error of 0.213 is about ± 2.13%, which is still relatively low. Compared to the previous SVR model, the errors are lower but the $R^2$ variable is similar. Overall, there are no significant improvements from the previous model. As seen from the plot's discussion, this was expected.

The residual plot that we generated:




As we can see from this graph, our model is still performing well, residuals are still random, and it shows a good fit as well, with no significant changes compared to our initial SVR model.




| Features | Importance Value |
|---|---|
| social_support_1_back | 0.017398 |
| social_support | 0.010241 |
| negative_affect | 0.009813 |
| bank_capital_to_assets_ratio_(%) | 0.009401 |
| diia_freedom_of_religion_v-dem | 0.008349 |
| fiie_media_self-censorship | 0.008233 |
| unemployment_with_advanced_education_(%_of_total_labor_force_with_advanced_education) | 0.007552 |
| dii_religious_organization_repression | 0.007278 |
| expense_(%_of_gdp) | 0.007278 |
| bank_capital_to_assets_ratio_(%)_1_back | 0.007274 |

The feature importance, like the previous SVR model, is more spread out. As we noticed in the Random Forest model, there are repeated significant features over time, e.g. social_support & social_support_1_back (National average of the binary response on question about social support), bank_capital_to_assets_ratio_(%) & bank_capital_to_assets_ratio_(%)_1_back (the ratio of bank capital and reserves to total assets) which show important features over time. Apart from those, compared to the initial model, a lot of important features have changed, and some more have been added, e.g. Religion and

Unemployment with Advanced Education indexes, which could be more important in the long term. Unlike in the previous model, there were no common features between the two models.

## Results and Discussion

### Best Model

Based on our research above, we can conclude that our best overall model is the Random Forest model with features from 2 years combined. It can be seen from the testing evaluation using unseen data we conducted for each of the chosen models that it has the highest $R^2$ value of 0.891. This indicates that this model fits our data the best and will be able to make better predictions. It also has the lowest MSE, RMSE, and MAE score of 0.055, 0.234, and 0.186, respectively. This indicates that this model can predict a happiness index value that is the closest to what the actual value is compared to the other models. Compared to the initial one-year model, this model has a lower reliance on only GDP per Capita with a bit more importance on other features. However, when only looking at our two initial models, we can see that SVR performs slightly better than the Random Forest model, as discussed previously.

### Top features for same-year model

Looking at the top 10 most important feature list for both models in the same year, we can see common features: log_gdp_per_capita (GDP per capita on logs scale), 2c__property_rights (index for level of protection of people and their property rights by law), reliability_of_police (index for poll asking for police reliability on a scale 1-7) and 2h_police_and_crime (index based on police reliability and impact of crime). These features being common in both models indicate that these are the most important features over a year to predict the happiness index. GDP per capita had the strongest importance in both models, outlining the importance of a strong and growing economy. The other two features related to police, security and property rights secured by the law point out how protection of oneself and one's assets stands out in our models over one year. Overall, this could help policymakers and the government see these key areas where policy change (e.g. personal property laws), better service (e.g. bettering police force) or economic growth are needed for an economy's overall growth.

## Top features for one year back model

Looking at both models, we see that there were no overlapping features between the two models. However, there were a few important features in these categories.

Firstly, we have important features that had their repeated counterparts from one year back also included in the top 10 feature list. These were log_gdp_per_capita (GDP per capita on logs scale) and business_permits (Rating 0-10 time taken to get new business permit) for the Random Forest model and social_support (National average of the binary response on question about social support) and bank_capital_to_assets_ratio_(%) (the ratio of bank capital and reserves to total assets) for the SVR model. This is interesting and insightful as we can see over the long term, these are the features that impact the happiness of a country the most. The fact that their one-year-back counterparts also made it to the top 10 list shows that these are core features that help define the base level of happiness in a country. The GDP per capita is important for an economy as it's a measure of the average earnings. The bank capital to assets ratio is important as it shows how well a bank can deal with unexpected losses and, hence, how safe one's savings or investments in a bank are. The time to get a business permit is also important as it shows how innovation and enterprise spirit are active in a country. Having good social support is a great indicator of one's happiness index as well. We can see logically why these features made it to the important features.

In the Random Forest model, we also had a few important features which were for 1 year back: market_openness_1_back (Economy Open market index), secure_internet_servers_1_back (Internet security) and area_2_rank_1_back (Legal Systems and Property Rights Ranking). These again deal with the ability to support innovation (market openness) and security (Internet security and property rights), which, as we explained above, logically make sense in the context.

Overall, these features give a good idea of the core features that make the happiness index, as these will have long-term effects on the happiness level of a country.

## Conclusion and Future Work

Overall, our project has been able to answer the main questions we had above. We can see how our models are able to make great predictions of the happiness index given its features. We can also see that from the results of our model, some important features, such as GDP, reliability of police, property

rights, and social support, are some of the most important aspects that help generate a greater happiness index. We can also see that most of our models show that the GDP has the most power out of all our features in influencing the happiness index, making it into the top 10 features for 3 of the four models we created. This explains why it is such an important and widely used indicator for a country. We were also able to get some core features that affect happiness in the long term with our two-year model, which is a great helper in understanding what really makes up the happiness index and how one can predict the overall happiness of the population.

For future improvement, we would like to focus more on making the model stronger and have better hyperparameter optimization compared to just random search. Another area of improvement is data processing. Although standard scaling works for most of our features, there are some features which are ranked and applying different normalizations there would be more beneficial.

Furthermore, it would be interesting to create a model with data from 2 years or 3 years back to further study and get more insights into the long-term important features of the happiness index. One could also divide our current model into 2-3 different models with different features focusing on economics, human freedom and other indexes, each having its own model.

## References

*Report Calls on Policy Makers to Make Happiness a Key Measure and Target of Development*. (2013). World Happiness Report. https://worldhappiness.report/news/report-calls-on-policy-makers-to-make-happiness-a-key-measure-and-target-of-development/

Cotofan, M. (2023). What is the role of government in creating a happier world? *Economics Observatory*. https://www.economicsobservatory.com/what-is-the-role-of-government-in-creating-a-happier-world

Stanley, A. (2024). Looking Beyond GDP. *International Monetary Fund*. https://www.imf.org/en/Publications/fandd/issues/2024/03/Picture-this-looking-beyond-GDP

Azad, A., Talwar, B., Shah, S., Prasad Shendre, H., Jyoti, A., Pandit, G., & Tiwari, V. (2023). Prediction model for world happiness index using machine learning technique. *SSRN*. https://doi.org/10.2139/ssrn.4485326

Akanbi, K., Jones, Y., Oluwadare, S., & Nti, I. K. (2024). Predicting Happiness Index Using Machine Learning. *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*, 1–5. https://doi.org/10.1109/ICMI60790.2024.10586193

*World Development Indicators*. (n.d.). World Bank Group. https://databank.worldbank.org/source/world-development-indicators

*World Happiness Report 2024*. (2024). World Happiness Report. https://worldhappiness.report/ed/2024/#appendices-and-datahttps://worldhappiness.report/ed/2024/#appendices-and-data

*Economic Freedom Dataset*. (n.d.). Fraser institute. https://efotw.org/economic-freedom/dataset?geozone=world&year=2022&page=dataset&min-year=2&max-year=0&filter=0

Molnar, C. (2024). 8.5 Permutation Feature Importance. In *Interpretable Machine Learning*. Essay. Retrieved November 28, 2024, from https://christophm.github.io/interpretable-ml-book/feature-importance.html.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324