Bhartiya Vidya Bhavan's

Sardar Patel Institute of Technology

(Autonomous Institute Affiliated to University of Mumbai)

Department of Computer Engineering

Exploring Global Happiness Index

By

Oren Coelho(2023300033)
Prem Choithani(2023300032)
Omkar Bhoir(2023300022)
Kalhan Bhat(2023300018)

Guided by

Prof. Nikita Mishra

Course Project

Python for Data Science (S.Y.)

Abstract

This project, titled **"Exploring Global Happiness Index,"** aims to analyze and predict the happiness levels of countries worldwide by examining socio-economic indicators and leveraging data science techniques. Happiness, as measured by the Global Happiness Index, is increasingly recognized as a key factor in assessing the quality of life within nations. By analyzing factors such as **GDP per capita, social support, life expectancy,** and **freedom**, we attempt to uncover patterns and correlations that influence happiness. The dataset, sourced from the world happiness report website, provided valuable insights into these socio-economic variables across various countries.

The project's methodology involved three main phases: **data exploration**, **predictive modeling**, and **clustering analysis**. First, through **Exploratory Data Analysis (EDA)**, we visualized relationships between happiness and different socio-economic factors, identifying trends and outliers. For predictive modeling, we applied machine learning algorithms, specifically **Linear Regression** and **Random Forest**, Polynomial regression etc, to estimate happiness scores based on the key factors. Random Forest outperformed all other models in this scenario, highlighting its ability to capture complex, non-linear interactions in the data. Lastly, clustering techniques like **K-Means** were used to group countries with similar happiness profiles, providing an insightful categorization based on socio-economic characteristics.

The results revealed that factors like GDP, social support, and life expectancy were the most significant in predicting happiness, with Random Forest achieving the best performance metrics. Clustering analysis further illustrated distinctive groupings among countries, suggesting that certain socio-economic patterns are common among happier nations. Although the dataset size and feature scope were limited, this project demonstrates the potential for using machine learning to assess and interpret happiness levels globally. Future improvements could include expanding the dataset, incorporating additional features, and experimenting with advanced algorithms for more accurate predictions.

Introduction

Problem Statement: The project seeks to analyze and predict the Global Happiness Index by examining various socio-economic factors. It aims to identify key contributors to happiness and understand how these factors influence well-being across nations.

Objective: The primary goal is to predict happiness scores based on socio-economic indicators and cluster countries with similar happiness profiles.

Motivation: Understanding the drivers of happiness is crucial for informing policy and supporting well-being on a global scale. This project provides insights into factors that governments and organizations can target to improve happiness.

Outline: The report begins with data preparation and exploration, followed by the methodology for model building and clustering analysis. Finally, it presents results, discussion, and conclusions.

Colab link: https://colab.research.google.com/github/Oreniscool/happiness-index/blob/main/happiness_predictor.ipynb

GitHub Link: https://github.com/Oreniscool/happiness-index

Dataset

Dataset Description:
The dataset utilized in this project, sourced from https://worldhappiness.report/data/ , contains socio-economic indicators from multiple countries, focusing on metrics related to happiness and well-being. Key features include:

- **GDP per Capita**: Reflects the economic wealth of each country.
- **Social Support**: Indicates the level of support individuals feel they have.
- **Life Expectancy**: Represents the average lifespan, which serves as a health measure.
- **Freedom to Make Life Choices**: Reflects the level of autonomy in citizens' lives.
- **Happiness Score**: This is the target variable representing the overall happiness level of each country.

The dataset consists of roughly 150 observations with both numerical and categorical data types. Initial analysis revealed that some data cleaning and transformation steps were necessary to ensure accurate model performance.

Preprocessing Steps:

1. **Handling Missing Values**:
   During initial data examination, missing values were identified in several socio-economic indicators. These were handled as follows:
   - **Numerical Columns**: Missing values in numerical columns were replaced with the median of each feature to avoid skewing the distribution while maintaining data integrity.
   - **Categorical Columns**: Missing values in categorical columns were imputed using the mode. This approach ensured that common categories were preserved without introducing new, potentially outlying data points.
2. **Normalization and Encoding**:
   - **Normalization**: Standardization was applied to numerical features such as GDP, life expectancy, and social support using `StandardScaler`. Standardization scales features to a mean of zero and a standard deviation of one, which prevents features with larger values (like GDP) from disproportionately influencing model training.
   - **Encoding Categorical Variables**: While most variables were numerical, any categorical features were label-encoded to make them compatible with machine learning algorithms. Encoding ensures that the model can interpret categorical data as numerical values for effective training.

3. **Data Splitting**:
   To ensure a robust model evaluation, the dataset was split into training and testing sets using an 80-20 split. The training set was used for model learning, while the testing set provided an unbiased performance evaluation on unseen data. This split enabled an accurate assessment of the model's predictive capabilities.

Data Exploration (EDA):

- **Summary Statistics**: Summary statistics (mean, median, standard deviation) of the key features were calculated to understand the data's central tendency and variability. For example, GDP and life expectancy were analyzed for their typical ranges across countries, highlighting differences that might impact happiness.
- **Data Visualization**: Several visualization techniques were used:
    - **Histograms**: Created for GDP, life expectancy, and social support to observe their distributions and detect any skewness or outliers.
    - **Correlation Heatmap**: A heatmap displayed the correlations between features, identifying that GDP, social support, and life expectancy had the strongest positive correlations with happiness scores. This correlation analysis guided model selection and feature engineering by emphasizing the most impactful predictors.
- **Insights from EDA**:
  EDA helped identify that, social support, Freedom to make life choices, and Generosity were key indicators of happiness. This informed feature selection and justified the inclusion of these predictors in modeling, as well as the choice of machine learning algorithms that could capture both linear and non-linear relationships.

Machine Learning Models and Implementation

1. Machine Learning Models

- Models Used:
- Linear Regression:

  Rationale: A foundational model for its simplicity and interpretability, suitable for identifying linear relationships between happiness and socio-economic factors.

- Polynomial Regression:

  Rationale : Polynomial regression is suitable for the Happiness Index dataset because it can capture non-linear relationships between predictors and happiness scores, which linear regression may miss. By introducing higher-degree terms, it allows for better modeling of curvatures in the data, improving the model's fit and predictive accuracy

- Random Forest:

  Rationale: A powerful ensemble method that can handle complex, non-linear relationships and capture intricate interactions between variables, making it well-suited for socio-economic data.

- **Additional Models**:

  To further enhance our analysis and potentially improve predictive accuracy, we incorporated the following models:

- Decision Trees:

  Rationale: A versatile algorithm capable of modeling both linear and non-linear relationships, providing interpretable decision rules.

- Gradient Boosting:

- Rationale: An ensemble technique that builds on decision trees, iteratively improving predictions by focusing on errors made by previous models.

2. Model Implementation

- **Training and Testing Split**:
  An 80-20 train-test split was applied to the data to evaluate model performance reliably. This split ratio ensured that the model was trained on a majority of the data, while a sufficient portion was reserved for testing its predictive accuracy on unseen data. Cross-validation techniques, such as k-fold validation, were also implemented to further assess model reliability across multiple data folds.
- **Hyperparameter Tuning**:
  - **Linear Regression**: Minimal tuning was necessary due to the simplicity of the model. The focus was on feature selection and preprocessing to maximize linear regression's interpretability.
  - **Random Forest**: Grid search was employed to tune key hyperparameters, such as the number of estimators (`n_estimators`), which controls the number of decision trees, and the maximum depth (`max_depth`), which limits tree complexity. Grid search allowed for systematic exploration of parameter combinations, identifying the optimal settings for high predictive performance.

3. Feature Selection and Extraction

- **Feature Selection**:
  Features were selected based on correlation analysis and domain knowledge. Freedom to make life choices, social support, Generosity and Life ladder were chosen for their strong, positive correlations with happiness scores. Using a correlation target of 0.3 as our dataset wasn't strongly correlated or

have many columns. These features were retained to reduce model complexity and enhance interpretability, focusing the model on the most relevant predictors.

- **Feature Extraction Techniques**:
  Techniques such as Recursive Feature Elimination (RFE) and correlation-based selection were evaluated. Ultimately, correlation-based selection was deemed suitable due to the strength of relationships with happiness scores, simplifying the model and increasing interpretability by emphasizing the primary socio-economic drivers of happiness.
- **Impact on Performance**:
  By selecting the most relevant features, the model's performance improved in both interpretability and accuracy. Removing less impactful features reduced overfitting risk and enhanced the generalization capability of both Linear Regression, Random Forest and other models.

4. Model Performance Evaluation

- **Evaluation Metrics**:
  - **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)** were used to assess prediction accuracy for both Linear Regression and Random Forest. These metrics provided insight into the average magnitude of errors in predictions, with RMSE being more sensitive to large errors.
  - **Silhouette Score** (for K-Means Clustering): This score measured the coherence within clusters, assessing how well countries grouped by happiness profiles. A high silhouette score indicated well-defined clusters with meaningful patterns.

Experimental Setup

- **Tools Used**: Python, with libraries such as Pandas, NumPy, Scikit-learn, and Matplotlib and Seaborn for data analysis and modeling.
- **Hardware/Environment**: Google Colab
- **Evaluation Metrics**: R^2 and RMSE metrics are used to evaluate model accuracy for regression models.

Performance Comparison:

The performance of different models was evaluated using metrics such as **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and **R-squared (R^2)**. These metrics allowed us to compare the accuracy and generalization ability of each model in predicting happiness scores.

As shown, **Random Forest** performed better than **Linear Regression** in terms of both MAE and RMSE, indicating its superior ability to capture complex, non-linear relationships in the data. Additionally, the clustering model (K-Means) achieved a high silhouette score of 0.62, suggesting that it successfully grouped countries with similar socio-economic profiles and happiness levels.

Model Interpretation:

- **Feature Importance**: Random Forest's feature importance analysis highlighted that Freedom to make life choices, **Generosity**, **social support**, and **Life ladder** were the most influential factors, aligning well with findings from the exploratory data analysis (EDA).
- **Model Coefficients**: For Linear Regression, GDP, social support, and life expectancy had the highest positive coefficients, supporting their strong influence on happiness. This interpretability offers valuable insights into how incremental changes in these factors could impact happiness.
- **ROC and Precision-Recall Curves**: Since the task involved regression rather than classification, ROC and precision-recall curves were not directly applicable. Instead, we used the residual distribution plots, which showed that errors were randomly distributed, confirming a good model fit without significant bias.

Error Analysis:

- **High Error Cases**: Analysis of the residuals showed that **outliers** (countries with extreme socio-economic conditions) often contributed to higher prediction errors, particularly for countries with very low GDP or social support scores. These outliers may require a more flexible model or specialized handling, such as **robust regression** techniques.
- **Overfitting and Underfitting**: The Random Forest model displayed slight overfitting due to its complex structure, as evidenced by marginally lower R^2 scores on the test set compared to the training set. Techniques such as **regularization** or reducing tree depth could mitigate this.
- We achieved an RMSE of 0.07 which the scale of predicted data from 0.1-1

Suggested Improvements:

- Implementing robust regression or ensemble methods that account for extreme cases could further improve model stability.
- Additional features, like **education level or mental health indicators**, could add context, potentially reducing prediction errors for countries with unique socio-economic conditions.

Conclusion

Summary:

This project focused on analyzing and predicting the Global Happiness Index using socio-economic indicators such as GDP per capita, social support, life expectancy, and freedom. Through exploratory data analysis (EDA), we visualized relationships between these factors and happiness scores across countries. We then applied machine learning techniques, including Linear Regression and Random Forest, to predict happiness scores, and used K-Means clustering to group countries with similar happiness profiles. This approach provided valuable insights into the drivers of national happiness and allowed us to predict happiness levels based on socio-economic conditions.

Findings:

The Random Forest model emerged as the best-performing model, outperforming Linear Regression in terms of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). This model successfully captured non-linear relationships between happiness scores and socio-economic factors, highlighting the significant influence of GDP per capita, social support, and life expectancy. Clustering analysis revealed distinct country groups, suggesting that these key factors play a crucial role in defining national happiness levels. Overall, the project demonstrated that socio-economic variables can effectively predict and interpret happiness levels, providing useful insights into global well-being.

Limitations:

Several limitations affected the project's scope and results:

1. **Data Size**: The dataset was relatively small, potentially limiting the model's ability to generalize across a broader range of countries.
2. **Data Bias**: Certain countries may be underrepresented in the dataset, leading to biased predictions that favor socio-economic profiles of dominant regions.
3. **Feature Scope**: Only a limited set of socio-economic indicators was available. Happiness is influenced by complex, interwoven factors, and additional data could enhance prediction accuracy.

Future Work:

To build on these results, several improvements and extensions could be explored:

1. **Incorporating Additional Features**: Adding more predictors, such as education level, mental health indicators, and environmental quality, may provide a more comprehensive view of happiness factors.
2. **Using Advanced Algorithms**: Trying other algorithms like Gradient Boosting or neural networks could capture more intricate patterns in the data, further improving prediction accuracy.
3. **Expanding Data Sources**: Increasing the dataset size by combining data from other sources or over multiple years could enhance model robustness and accuracy, improving its applicability across diverse regions.
4. **Enhanced Feature Engineering**: Developing new features or combining existing ones to better capture non-linear relationships could yield more nuanced insights.

Timesheet

Work Done	HOURS
Oren Coelho	
Finding dataset and creating basic template of project	3
Creating basic plots showing trends of every country over time	2
Creating a correlation heatmap and understanding which features to include	3
Took feedback from model and changed the various techniques used in feature selection	4
Created graphs to display machine learning model accuracy comparisons	3
Fixed input issues with input data not being normalized	4
Collaborated with all the member to decide on the project goals	1
Fixed errors with representation of graphs overlapping	2

Prem Choithani

Model Research	5
Model Justification	4
Initial Model Setup	3
Model Implementation	5
Hyperparameter Tuning	2
Selecting The Most Accurate Model	2

Omkar

Finding dataset	3
Data understanding	3
Missing data handling	3
Normalization and encoding	4
Data splitting	2
Checked if data is ready for exploration	1
Collaborated with all members to decide on the project goals	1
Collaborated to ensure data is properly fed into ML models	2
Code commenting	2

Kalhan Bhat

Dataset Finding on World Happiness websites	3
Researched on models which may be used	3
Write initial code for model implementation.	4
Made contributions in I/O final prediction code using random forest	4
Report writing part 1	2
Report writing part 2	3

References

- Predicting Happiness Index Using Machine Learning from IEEE by K. Akanbi 2024
- Predicting student teacher happiness
- Predicting life satisfaction using machine learning
- Predicting happiness using random forest
- Project reference