Analyzing Determinants of Happiness Score: A Comparison Based on Machine Learning Approaches

Yuxuan Xiong

Department of Economics, Brandeis University, Waltham, 02453, USA
yuxuanxiong@brandeis.edu

Abstract. In this research, the determinants of happiness scores across countries are explored using a data-driven, machine learning-based approach. The study employs a dataset comprising variables such as GDP per capita, social support, healthy life expectancy, freedom to make life choices, etc. to predict the Happiness Index Score for the years 2018 and 2019. Three distinct machine learning models - K-Nearest Neighbors (KNN), Random Forest (RF), and Linear Regression (LR) - are implemented individually and as an ensemble to ascertain the most accurate predictor. Model performance is evaluated via three key metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Findings indicate that while each individual model offers valuable insights, the ensemble model outperforms them with an MAE, MSE, and RMSE respectively. Feature importance, derived from the RF model, revealed 'Social support', 'GDP per capita', and 'Healthy life expectancy' as the most influential parameters. This study underscores the utility of machine learning techniques and ensemble modeling in exploring the multifaceted nature of societal well-being.

Keywords: Happiness Score, Machine Learning, Ensemble model

1 Introduction

The examination of happiness, commonly referred to as subjective well-being, has emerged as a thriving field of scholarly investigation over the past several decades. It focuses on understanding the multifaceted nature of happiness from an individual to a national level. The advent of the first World Happiness Report in 2012 [1], published in the context of a UN High-Level Meeting, has played a pivotal role in elevating the prominence of such inquiries on a global scale. The reports employ a variety of indicators, including but not limited to GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption to establish happiness scores for each country. More recently, the integration of machine learning models has been introduced to analyze and anticipate data pertaining to individual well-being. This fusion of data science and happiness studies provides fresh, computational perspectives on this abstract, subjective concept,

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_32

refining the tools and methodologies used to quantify and comprehend happiness globally.

Numerous studies have focused on employing Machine Learning to identify the correlation between happiness and a person’s life, but some research gap still exists. The article by Fabina Ibnat thoroughly analyzes the World Happiness Report of 2019 using various machine learning models [2]. One of the main research gaps identified in this article pertains to the machine learning models used in the analysis. As mentioned in the conclusion, future research plans involve the integration of significant machine learning algorithms, including but not limited to linear regression, logistic regression, and Support Vector Machines (SVM) [2]. It suggests that the current analysis was limited to Decision Tables, Random Forest, and SMOreg algorithms. Applying and comparing other machine learning techniques, like linear or logistic regression and SVM, could further expand the insights and perhaps yield more precise predictions or analyses. Another potential research gap arises from the dataset used in the study. The article states, "The recent and future happiness datasets may have different results and attributes." [3]. This implies that the generalizability of findings derived from the 2019 World Happiness Report to subsequent years' datasets is potentially limited, as these datasets may possess distinct attributes and yield dissimilar outcomes. Therefore, continuous and updated analyses with more recent datasets are needed to understand how happiness and its associated variables have evolved over time.

Besides, in the study [3], You et al., exploited the strengths of various models - namely Linear Regression, Decision Tree, Random Forest, and Gradient Boosting - to predict the happiness index [3]. Although these models possess distinct advantages, they are not impervious to shortcomings, which may result in inconsistencies in predictions across varying conditions. Particularly noteworthy is the absence of an ensemble approach within the study, which could harness the combined predictive capabilities of diverse models. This absence may result in suboptimal predictive performance due to individual model shortcomings. As the authors rightly observed, "model accuracy varies across different socioeconomic contexts"[3], an issue that ensemble techniques could potentially mitigate by balancing the biases of individual models. The study's limited exploration of this diversity is a key research gap, suggesting future work could benefit from using the Random Forest to explore the major property for improved robustness and accuracy in predicting the happiness index, potentially leading to a more comprehensive understanding of the factors influencing happiness across various contexts.

This study will leverage the complementary strengths of K-Nearest Neighbors (KNN), Random Forest (RF), and Linear Regression (LR)algorithms to optimize predictive performance. By amalgamating these diverse algorithms, the ensemble method is expected to reconcile individual model biases, thereby enhancing robustness and accuracy. This approach reflects the realization that the complexity of happiness as a construct necessitates a more sophisticated method, capable of capturing the myriad socio-economic factors influencing its variance across different contexts. Thus, this study aims to enhance the predictive model's performance in line with the broader objective of developing more reliable and nuanced tools for policy decision-making in relation to societal happiness.

2 Methodology

2.1 Data Preparation

The dataset used in this research consists of annual Happiness Index Scores for the years 2018 and 2019, coupled with several associated independent variables [4]. The dataset includes a comprehensive listing of 156 countries, with each country assigned a rank based on its Happiness Index Score. The dataset used in this research consists of annual Happiness Index Scores for the years 2018 and 2019, coupled with several associated independent variables. The dataset includes a comprehensive listing of 156 countries, with each country assigned a rank based on its Happiness Index Score. The independent variables considered in this research include 'GDP per capita', 'Social support', etc. These parameters provide a well-rounded representation of the various socio-economic and ethical factors that potentially influence the overall happiness of individuals within these nations. Each variable is a continuous numerical value, representing various facets of a country's societal and economic health.

In preparation for the analysis, the data was imported utilizing Pandas, allowing efficient manipulation and analysis. The independent and dependent variables were separated into X and y respectively for model development purposes. To facilitate model training and testing, the dataset was split into an 80% training set and a 20% testing set.

2.2 Developed Ensemble Model

KNN Model. The KNN algorithm is a widely adopted and intuitive model used for both regression and classification tasks [5, 6]. The core principle it utilizes is that instances within a dataset will generally resemble other instances that are close to them in the feature space. This algorithm is advantageous due to its simplicity and flexibility to fit complex data patterns. However, selecting the right number of neighbors (k) to consider is critical in achieving optimal performance. Furthermore, its computational cost can be relatively high on larger datasets, given its need to calculate distances between instances.

Random Forest. RF is a powerful and versatile machine-learning model used for both regression and classification tasks [7]. As an ensemble learning method, it operates by constructing a multitude of decision trees at training time and outputting the average prediction (in the case of regression) or the class that is the mode of the classes (classification) of the individual trees. The 'Random' in Random Forest comes from two random aspects incorporated in the model: each tree in the ensemble is trained on a different bootstrap sample of the original training data, and at each node, a random subset of features is considered for splitting. This randomness helps to improve model generalization and robustness to overfitting. Despite its effectiveness, one downside of Random.

Linear Regression. LR is a fundamental algorithm in the machine learning toolkit [8]. The model assumes a linear relationship between the input variables (X) and a single output variable (Y). When there is a single input variable, the method is referred to as simple LR. For more than one input variable, it is called multiple LR. Y can be calculated from a linear combination of the input variables (X). If the linear relationship is correctly identified, the algorithm is capable of predicting the output for any arbitrary input. Despite its simplicity, LR can be incredibly powerful when applied to real-world problems such as economic forecasting, financial analysis, and healthcare applications. It is also worth noting that understanding LR paves the way for understanding more complex algorithms that often use concepts from this fundamental technique.

Ensemble Model. The ensemble methodology applied in this study involves combining the individual predictions from KNN, RF, and LR models. The individual models were trained on the training dataset and subsequently used to make predictions on both the training and testing sets. The predictions from the individual models were then consolidated into new features, essentially creating a new, higher-level training and testing dataset.

The ensemble model shown in Fig. 1, a KNN model in this case, was then trained on the consolidated training meta-features and used to predict the final output based on the consolidated testing meta-features. The objective of this ensemble approach is to leverage the strengths of each model, thus resulting in a more robust and accurate prediction of the Happiness Index Score. This strategy is believed to alleviate the limitations associated with using a single modeling approach, potentially leading to a more comprehensive understanding of the relationships within the data.

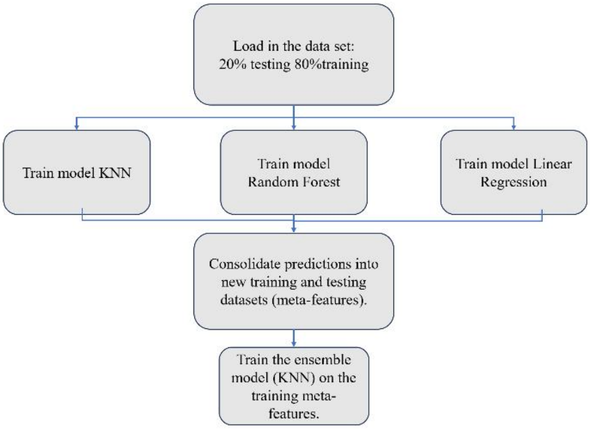
Load in the data set:
20% testing 80%training
Train model KNN
Train model Random Forest
Train model Linear Regression
Consolidate predictions into new training and testing datasets (meta-features).
Train the ensemble model (KNN) on the training meta-features.

Fig. 1. The structure of the proposed ensemble model (Photo/Picture credit: Original).

2.3 Implementation Details

The implementation of the KNN model, setting the “n_neighbors” value of 1 that the model considers the single nearest neighbor to a new data point and assigns the target output based on this nearest neighbor. The RF model, an ensemble learning method constructing multiple decision trees at training time, is applied using the default parameters as offered by the Scikit-learn library. The output of the model is derived as the mode of the classes (classification) or mean prediction (regression) of the individual trees. The application of the LR model, a fundamental statistical and machine learning method, is also executed utilizing the Scikit-learn library's default parameters.

The performance of all employed models is evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics offer a holistic perspective on the prediction errors, accounting for the average error, the impact of larger errors, and providing an interpretable measurement scale.

Upon completion of model training, predictions are generated for both training and testing datasets. These predictions are then consolidated and serve as new meta-features to train an additional KNN model. For this step, the “n_neighbors” parameter is explored within a range of 1 to 300 to establish an optimal configuration.

Finally, the importance of features is ascertained using the feature importance property of the trained RF model. This provides a score for each feature, with higher scores indicating greater relevance to the output variable. This measure is computed as the normalized total reduction of the criterion brought by that feature.

3 Results and Discussion

3.1 The Performance of Ensemble Model

In the research, the performance of the KNN, RF, LR, and ensemble models was assessed and compared using three primary metrics shown in Table 1, namely MAE, MSE, and RMSE. Each of these metrics offers a distinct evaluation perspective, encapsulating prediction accuracy and error distribution.

The individual KNN model demonstrated a moderate average deviation from the actual values, as indicated by an MAE of 0.3688. The substantial penalty for larger errors was highlighted by an RMSE of 0.5019, with an MSE of 0.2519 indicating a considerable difference between predictions and actual values. On the other hand, the RF model exhibited improved prediction accuracy with a reduced MAE, MSE, and RMSE of 0.3464, 0.2429, and 0.4929 respectively. These figures suggest a lower deviation from true values and better handling of larger errors by the RF model. The LR model showed a comparatively higher deviation, evidenced by an MAE of 0.4209, an MSE of 0.2787, and an RMSE of 0.5279.

By contrast, the ensemble model, integrating the predictive capabilities of the KNN, RF, and LR models, displayed the highest accuracy with the lowest MAE (0.3256), MSE (0.2180), and RMSE (0.4669). Compared to individual models, the ensemble model decreased the MAE by approximately 11.7% (KNN), 6.0% (RF), and

22.6% (LR); the MSE by around 13.5% (KNN), 10.2% (RF), and 21.7% (LR); and the RMSE by about 7.0% (KNN), 5.3% (RF), and 11.6% (LR). These results substantiate the superiority of the ensemble model in achieving high prediction accuracy by leveraging the strengths of the individual models and alleviating their weaknesses.

The ensemble model used in this research implements the unique strengths of individual models – KNN, RF, and LR models providing robust predictions that capture both local and global trends as well as linear and non-linear relationships. This ensemble model effectively balances the variance-bias trade-off, reducing the likelihood of overfitting or underfitting, which are common pitfalls in individual models. Consequently, its superior performance can be attributed to the integration of diverse model strengths and efficient handling of the variance-bias trade-off, enabling a more accurate understanding of the complex relationships within the data. In the future, the neural network may be considered as a component of the developed ensemble model due to its excellent performance in various tasks [9, 10].

Table 1. Performance for each model.

Model	MAE	MSE	RMSE
KNN	0.3688	0.2519	0.5019
LR	0.4209	0.2787	0.5279
RF	0.3464	0.2429	0.4930
Ensemble	0.3256	0.2180	0.4669

3.2 Feature Importance

The investigation of feature importance offers insights into the relative influence of individual variables on the prediction of the Happiness Index Score shown in Table 2. 'Social Support' emerged as the most influential, holding approximately 50.81% of the decision-making power within the RF model's prediction process. 'GDP per Capita' followed, contributing 20.03%, indicating the economic influence on happiness levels. 'Healthy Life Expectancy' was third, accounting for 14.61% of the model's decision-making process, reflecting the health and longevity aspect. 'Freedom to Make Life Choices', 'Perceptions of Corruption', and 'Generosity' held less decision-making power at 6.27%, 5.03%, and 3.26% respectively, however, their contributions should not be overlooked.

Table 2. Importance level for each feature

Feature	Importance
Social support	0.517546
GDP per capita	0.217676
Healthy life expectancy	0.123826
Freedom to make life choices	0.064358

Perceptions of corruption	0.042753
Generosity	0.033841

The findings of the significance of 'Social Support' in influencing a nation's happiness index. In real-world scenarios, 'Social Support' refers to the networks of relationships and community ties that individuals can lean on during challenging times. It contains a spectrum of supportive interactions, from emotional care and companionship to practical help in times of need. The prevalence of robust social support systems in a country can potentially indicate a culture of empathy, collective responsibility, and mutual aid. Such systems might manifest through strong family bonds, close-knit neighborhoods, extensive friend networks, and effective social services provided by governments or community organizations. In times of personal crises such as illness, job loss, or emotional distress, individuals with strong social support are likely to experience a buffer against the negative impact of these stressors, leading to improved mental well-being.

4 Conclusion

In conclusion, predicting a country's happiness score using features like social support, GDP per capita, and others has emerged as a significant area of study. The research highlights the efficacy of ensemble learning techniques in enhancing prediction accuracy. Specifically, an ensemble model integrating the strengths of KNN, RF, and LR consistently surpassed the accuracy of individual models. In terms of performance metrics, there was an approximate 11.7%, 6.0%, and 22.6% reduction in the MAE for KNN, RF, and LR respectively; 13.5%, 10.2%, and 21.7% decrease in the MSE; and 7.0%, 5.3%, and 11.6% drop in the RMSE. This remarkable performance can be attributed to the ensemble model's unique capability to harness the strengths of each individual model, meanwhile concurrently addressing, and neutralizing their respective weaknesses. Such a cohesive strategy lends the model its high robustness and adaptability in prediction. Besides, there are boundless opportunities for enhancement. Future research could contemplate diversifying the ensemble by integrating more varied predictive models. Additionally, there's potential in expanding the feature set by incorporating newer, perhaps less traditional, indicators that might further refine and bolster the accuracy of the happiness score prediction. Such advancements could pave the way for a deeper understanding of the intricate factors that contribute to a nation's well-being and happiness.

References

1. John, F. H.: The 2012 World Happiness Report. UN High-Level Meeting (2012).
2. Fabih, I.: Understanding World Happiness using Machine Learning Techniques, IEEE (2021).
3. Lexin, Y.: Utilizing Machine Learning to Predict Happiness Index, IEEE (2021).

4. Kaggle.: Happiness Index 2018-2019, https://www.kaggle.com/datasets/sougatapramanick/happiness-index-2018-2019?resource=download (2023).
5. Zhang, S., Li, X., Zong, M., et al.: Learning k for knn classification. ACM Transactions on Intelligent Systems and Technology (TIST), 8(3): 1-19 (2017).
6. Deng, Z., Zhu, X., Cheng, D., et al.: Efficient kNN classification algorithm for big data. Neurocomputing, 195: 143-148 (2016).
7. Rigatti, S. J.: Random forest. Journal of Insurance Medicine, 47(1): 31-39 (2017).
8. Weisberg, S.: Applied linear regression. John Wiley & Sons (2005).
9. Yu, Q., Yang, Y., Lin, Z., et al.: Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV. China Communications, 17(3): 46-57 (2020).
10. Brause, R. W.: Medical analysis and diagnosis by neural networks, Medical Data Analysis: Second International Symposium, ISMDA 2001 Madrid, Spain, October 8–9, 2001 Proceedings 2. Springer Berlin Heidelberg, 1-13 (2001).