**Happiness Prediction**

**Comparison of models**

Kumbham Nuthan Manideep

National Institute of Technology, Silchar

Data Mining Methods

Dr. Ripon Patigiri

**Abstract**

The World Happiness Report 2015 dataset contains data on the happiness levels of people in 159 countries around the world. It includes information on a variety of factors that are thought to contribute to happiness, such as economic factors (GDP per capita), social factors (family, health, and freedom), and trust in government. The data also includes a happiness score and a happiness rank for each country. The happiness score and happiness rank are based on a poll in which people were asked to rate their overall happiness on a scale of 0 to 10. The economy, family, health, freedom, trust, and generosity columns contain various measures of those factors. Overall, the World Happiness Report 2015 dataset provides valuable insights into the factors that contribute to happiness and well-being at the national level, and can be used by researchers, policymakers, and others to better understand and promote happiness around the world.

**Overview**

This report compares the performance of various machine learning models on the World Happiness Report 2015 dataset. The models that have been compared are Logistic Regression, Neural Network, Classification Tree, Random Forest, Bagging and Boosting. The dataset contains information on the happiness levels of people in 159 countries around the world and includes a variety of factors that are thought to contribute to happiness. The models have been evaluated based on their accuracy and performance in predicting the happiness scores of countries in the dataset. The report aims to identify the most effective machine learning model for predicting happiness levels in different countries and provide insights into the factors that contribute to happiness and well-being at the national level. The results of the analysis can be used by researchers, policymakers, and others to promote happiness and well-being around the world.

**Data**

The dataset can be found on Kaggle, and it contains data on happiness levels of people in 159 countries around the world. The data is collected in the year 2015 and it includes information on a variety of factors that are thought to contribute to happiness, such as:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Country or Region | Name of the country |
| Happiness Score | A composite score of overall well-being |
| Economy (GDP per Capita) | Measure of the economic production of a country |
| Family | Measure of social support |
| Health (Life Expectancy) | Measure of the health of citizens |
| Freedom | Measure of freedom to make life choices |
| Generosity | Measure of generosity of citizens |
| Trust (Government Corruption) | Measure of trust in government |

In this dataset, the response variable is 'Happiness Score', and predictor variables are 'Economy (GDP per Capita)', 'Family', 'Health (Life Expectancy)', 'Freedom', 'Generosity', and 'Trust (Government Corruption)'.

**Research Questions**

1. How do the different machine learning models (Logistic Regression, Neural Network, Classification Tree, Random Forest, Bagging, and Boosting) compare in their accuracy and performance in predicting happiness levels in different countries?
2. Which features or factors have the strongest impact on happiness levels across different countries, as identified by the different machine learning models?
3. Can the machine learning models be used to predict the happiness levels of countries that were not included in the original dataset, based on their economic, social, and political factors?
4. How do the machine learning models perform when predicting happiness levels in countries with very different cultural and socio-economic backgrounds?
5. How do the results of the machine learning analysis compare with other studies on the determinants of happiness at the national level, such as those based on survey data or qualitative research methods?

**Analysis**

**Summary of Logistic Regression analysis**

A full Logistic Regression model is fitted by using all predictors. The coefficients of the full model are given in the table.

| **Variable** | **Estimate** | **P-value** |
| --- | --- | --- |
| (Intercept) | -19.3208119 | 0.0009845987 |
| Economy (GDP per Capita) | 7.6631705 | 0.0337889054 |
| Family | 6.4534031 | 0.0807727018 |
| Health (Life Expectancy) | 5.9366848 | 0.1557630723 |
| Freedom | 3.1143209 | 0.4418690026 |
| Trust (Government Corruption) | 2.0524514 | 0.7336194356 |
| Generosity | 0.6044032 | 0.8784020918 |

From the table the Economy is statically significant because the p-value is less than 0.05.

The confusion matrix of full model training data is given below.

|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 47 | 6 |
| Actual Positive (1) | 2 | 39 |

The model's accuracy is 87.5%, with a specificity of 88.6% and a sensitivity of 86.4%.

The confusion matrix of full model validation data is given below.

|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 30 | 7 |
| Actual Positive (1) | 5 | 22 |

The model's accuracy is 81%, with a specificity of 81% and a sensitivity of 81%.

The best model in building models is the one that utilizes Economy, Family, Health, Freedom, and Trust (Government Corruption) as its fitting parameters.

The coefficients of the model-1 are given in the table.

| **Variable** | **Estimate** | **P-value** |
| --- | --- | --- |
| (Intercept) | -19.142768 | 0.0007052961 |
| Economy (GDP per Capita) | 7.477746 | 0.0253183052 |
| Family | 6.566475 | 0.0710079997 |
| Health (Life Expectancy) | 5.910844 | 0.1481226606 |
| Freedom | 3.238972 | 0.4162510765 |
| Trust (Government Corruption) | 1.735307 | 0.7592808345 |

From the table the Economy and Family are statically significant because the p-value is less than 0.05.  
The confusion matrix of best model training data is given below.

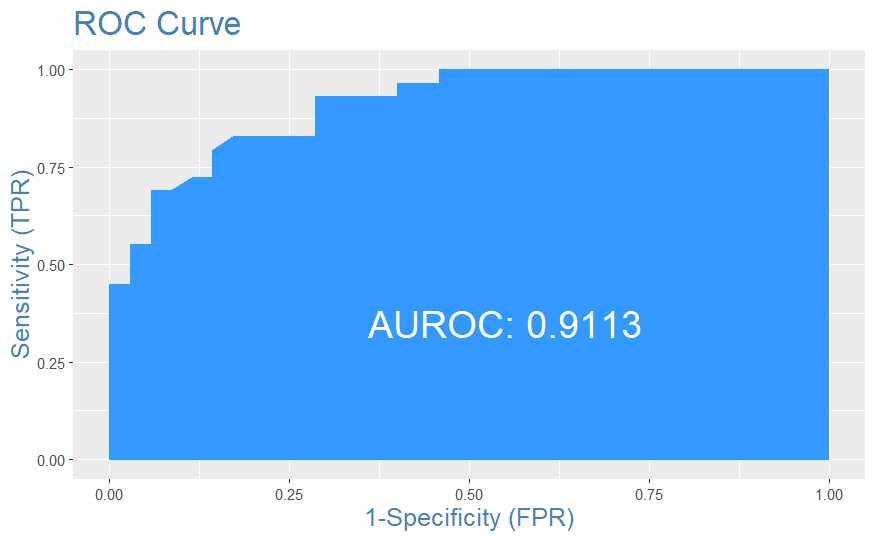
|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 47 | 6 |
| Actual Positive (1) | 2 | 39 |

The model's accuracy is 87.5%, with a specificity of 88.6% and a sensitivity of 86.4%.

The confusion matrix of best model validation data is given below.

|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 30 | 7 |
| Actual Positive (1) | 5 | 22 |

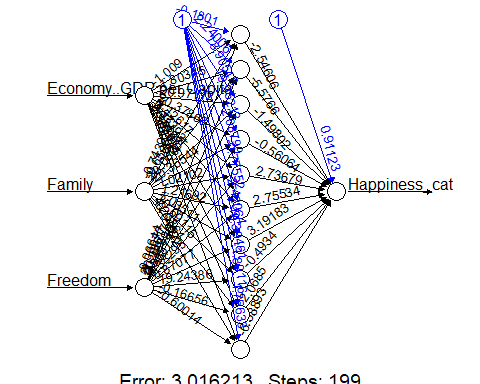
The model's accuracy is 81.25%, with a specificity of 81% and a sensitivity of 81%.



The plot shows that the model has a high AUC-ROC value of 0.9113, indicating high accuracy in correctly classifying the data with low false positives and false negatives.

**Summary of Neural Network analysis**

The neural network model with 10 neurons and a feature set consisting of Economy(GDP per Capita), Family, and Freedom has been identified as the best model.



The confusion matrix of best model training data is given below.

|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 46 | 7 |
| Actual Positive (1) | 3 | 38 |

The model's accuracy is 88%, with a specificity of 87% and a sensitivity of 93%.

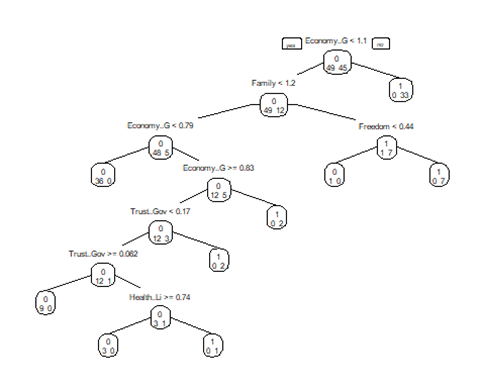
The confusion matrix of best model validation data is given below.

|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 30 | 6 |
| Actual Positive (1) | 5 | 23 |

The model's accuracy is 81%, with a specificity of 82.14% and a sensitivity of 83.33%.

**Summary of Classification Tree, bagging, boosting, random forest analysis**

The tree was built with all variables. Variables actually used in tree construction are “Economy..GDP.per.Capita.”, “Family”, “Freedom”, “Health..Life.Expectancy.”, “Trust..Government.Corruption”. The classification tree is shown below



The confusion matrix of classification tree training data is given below.

|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 49 | 0 |
| Actual Positive (1) | 0 | 45 |

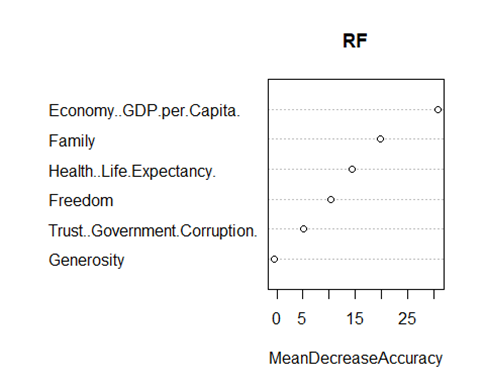
The model's accuracy is 100%, with a specificity of 100% and a sensitivity of 100%.

The confusion matrix of classification tree validation data is given below.

|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 29 | 9 |
| Actual Positive (1) | 6 | 20 |

The model's accuracy is 76.56%, with a specificity of 68.8% and a sensitivity of 82.86%.

The random forest plot is shown below



The most important variables in the plot are “GDP”, “Family”, “Life Expectancy”, “Freedom”.

The confusion matrix for the training data of random forest model is shown below

|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 49 | 3 |
| Actual Positive (1) | 0 | 42 |

The model's accuracy is 97.87%, with a specificity of 93.33% and a sensitivity of 100%.

The confusion matrix for the validation data of random forest model is shown below

|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 29 | 9 |
| Actual Positive (1) | 6 | 20 |

The model's accuracy is 76.96%, with a specificity of 68.97% and a sensitivity of 82.86%.

Bagging, short for bootstrap aggregating, is ensemble learning method that combines multiple models to improve the accuracy and stability of the predictions. The basic idea is to generate multiple bootstrap samples of the training data, and then train a separate model on each sample.

The confusion matrix for the training data of bagging model is shown below

|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 47 | 5 |
| Actual Positive (1) | 2 | 40 |

The model's accuracy is 92.55%, with a specificity of 88.89% and a sensitivity of 95.92%.

The confusion matrix for the validation data of bagging model is shown below

|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 28 | 9 |
| Actual Positive (1) | 7 | 20 |

The model's accuracy is 76.56%, with a specificity of 68.97% and a sensitivity of 80%.

Boosting is ensemble learning method that combines multiple models to improve the accuracy of the predictions. The basic idea is to train multiple weak models sequentially, and each model is trained on the residuals of the previous model.

The confusion matrix for the training data of boosting model is shown below

|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 46 | 4 |
| Actual Positive (1) | 3 | 41 |

The model's accuracy is 92.55%, with a specificity of 91.11% and a sensitivity of 93.88%.

The confusion matrix for the validation data of boosting model is shown below

|  | **Predicted Negative (0)** | **Predicted Positive (1)** |
| --- | --- | --- |
| Actual Negative (0) | 27 | 9 |
| Actual Positive (1) | 8 | 20 |

The model's accuracy is 73.43%, with a specificity of 68.97% and a sensitivity of 77.14%.

The examination shows that out of all the models developed, the random forest model has emerged as the top performer. When used on the validation data, this model shows the highest accuracy, reaching 76.56%.

**Comparison of LR, NN and Tree methods**

|  |  |  |
| --- | --- | --- |
| Model Name | Model Description | Validation Accuracy |
| Logistic Regression Model | Model with 5 predictors Economy, Family, Health, Freedom, and Trust (Government Corruption) | 81.25% |
| Neural network model | Model with 1 hidden and 10 neurons having of Economy(GDP per Capita), Family, and Freedom. | 81% |
| Best Classification Tree | classification tree model with full predictors | 76.56% |
| Random Forest | Random forest model with all predictors and 500 decision trees. | 76.56% |
| Bagging | Bagging method with all predictors | 76.56% |
| Boosting | Boosting method is used to fit all predictors | 73.43% |

**Conclusion**

In this analysis, four different models were fitted to predict the happiness score of different

countries. The models used were logistic regression, neural network, classification tree, bagging,

boosting, and random forest analysis. Overall the best model is Logistic model with Economy, Family, Health, Freedom, and Trust (Government Corruption) as predictors. The model's accuracy for validation data is 81.25%, with a specificity of 81% and a sensitivity of 81%. Overall, the model demonstrates good performance in predicting the outcome of interest, suggesting that the factors included in the model are important predictors.

**References**

*World Happiness Report*. (2019, November 27). Kaggle. <https://www.kaggle.com/unsdsn/world-happiness>

Menard, S. (2010). *Logistic Regression: From Introductory to Advanced Concepts and Applications*. SAGE Publications.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, *4*(1), 1–58. <https://doi.org/10.1162/neco.1992.4.1.1>

Alfaro, E., Gámez, M., & García, N. (2013). **adabag**: An*R*Package for Classification with Boosting and Bagging. *Journal of Statistical Software*, *54*(2). <https://doi.org/10.18637/jss.v054.i02>

Hofner, B., Mayr, A., Robinzonov, N., & Schmid, M. (2014). Model-based boosting in R: a hands-on tutorial using the R package mboost. *Computational Statistics*, *29*(1–2), 3–35. <https://doi.org/10.1007/s00180-012-0382-5>

Wright, M. N., & Ziegler, A. (2017). **ranger**: A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R*. *Journal of Statistical Software*, *77*(1). <https://doi.org/10.18637/jss.v077.i01>