International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 06 | June - 2021 ISSN: 2582-3930

Predicting World Happiness using Machine Learning

Sanmati Nikam
Dept. Of Computer Engineering,
BRACT VIT, Pune, India

Abstract

The main aim of this project is to analyze the dataset, pull meaningful insights, and predict happiness scores. As this is a regression project, various algorithms are used, a comparison between R Square and RMSE is used to decide which model is better. Happiness is linked with better decision-making, It is the key to success. World Happiness Report is analyzed by United Nations, it is part of the UN's efforts to emphasize world happiness. The rankings of national happiness are based on the Cantril ladder score, questions asked in the survey regarding social support, freedom to make life choices, generosity, and perception of corruption. The other two factors included for calculating national happiness are GDP Per Capita and healthy life expectancy, data for these two factors are publicly available.

Keywords: World happiness, Regression, RMSE

I. INTRODUCTION

The first world happiness report was published in 2012 for a UN High-level meeting. The second report was issued in 2013 and the third in 2015. Since 2015 annual report of world happiness is published by the UN. The features used in the report reflect what has been broadly found within the investigative writing to be imperative in clarifying national-level contrasts in life assessments. A few vital factors, such as unemployment or inequality, don't show up since comparable international data are not yet accessible for the complete test of countries. The factors are expecting to demonstrate imperative lines of relationship instead of reflecting clean causal estimates since a few of the information are drawn from the same study sources, a few are connected with each other, and on a few occasions, there are likely to be two-way relations between life assessments and the chosen factors.According to the analysts behind the report, released by the Sustainable Development Solutions Network for the UN, this is a critical time to think about the happiness of nations—because they can teach us what's important in a pandemic and which nations are likely to handle and weather it well.

All the countries are compared against a hypothetical nation called Dystopia. It is an imaginary country, that has the world's least happy citizens. The purpose of establishing such a country is to have a benchmark with which all countries can be fairly compared. No country can perform poorly than Dystopia. Since life would be very unpleasant and unhappy with the least GDP per capita, the lowest income, and high corruption, it is referred to as "Dystopia", opposite of Utopia.

A) Matplotlib

Matplotlib is an interactive cross-platform, data visualization and graphical plotting library for Python and its numerical expansion NumPy.Matplotlib is open source and can be used freely. Matplotlib is generally written in python, a couple of sections are composed in C, Objective-C, and Javascript for Platform compatibility.

B) Seaborn

Seaborn is a library in python used for data visualization. It is based on matplotlib. It provides a higher-level interface for making Statistical and graphical plots, while matplotlib was used to make basic plots.

C)Plotly

The Plotly Python library is an interactive, open-source plotting library that underpins over 40 special chart sorts covering a wide extend of factual, financial, geographic, scientific, and 3-dimensional use-cases. only a few lines of code are essential to make aesthetically satisfying, interactive plots. Plotly library was used to make interactive plots.

II.Literature Review

[1] have predicted a model of human happiness in the aspect of the work field. They have used a dataset obtained by a survey in the office to predict the happiness of office employees. KNN, Decision tree, Naive Bayes, and Multilayer perceptron models are used. They have used two major techniques namely, oversampling and under-sampling. The experimental results showed that when an imbalanced data issue was unraveled, the prediction accuracy was higher than the results from the first data. The prediction accuracy of the proposed model was around 87.66 percent.

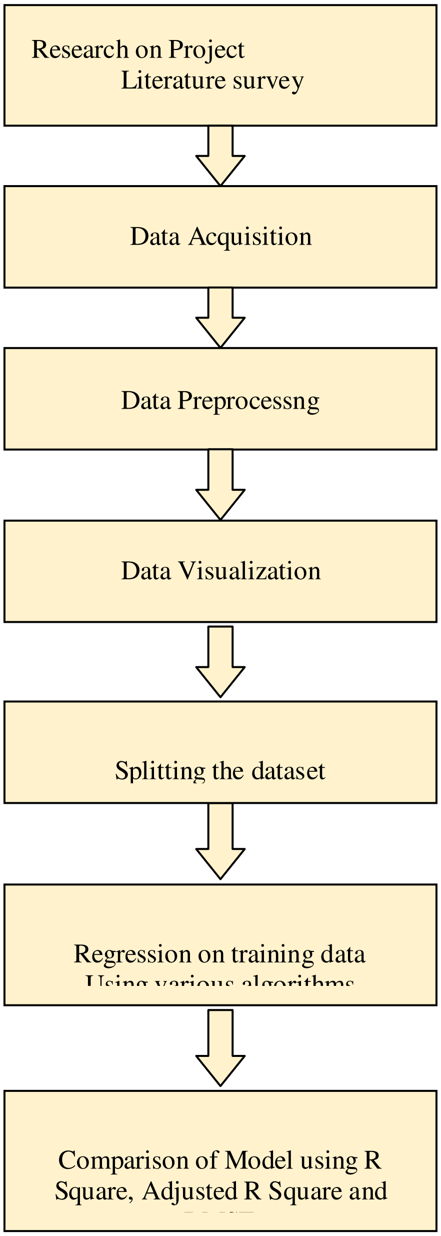
After appropriate preprocessing, resultant data was analyzed by [2] by utilizing 2 sorts of computational strategies viz. Predictive models: for predicting the happiness index(‘Life Ladder’) of a nation and Bayesian Systems: for exploring causal relationships among factors. Over 30 distinctive Machine Learning and Deep Learning models were trained on happiness data from 2016-18 and their execution was evaluated while determining Happiness Index for 2019.

a supervised two-tier ensemble approach for predicting a country’s BLI score was proposed in [3]. The work presented a cost-effective method of BLI prediction with a high degree of efficiency using a recursive elimination method with the 10-fold cross-validation model was built utilizing an ensemble approach and was evaluated using r, R, RMSE, and exactness performance evaluators. The model was about 90% accurate for predicting the life satisfaction score of a country.

In the research paper by [4], Of the 6 explanatory variables measured, GDP, Social Support, and Healthy Life Expectancy showed up most significant. Where instability exists, social supports such as healthcare and education suffer. As a result, life expectancy can be unfavorably influenced. Trade and hence GDP will too be influenced. They have used multiple linear regression. RMSE of 0.67 and an MAE of 0.50 were obtained.

[5]attempted using K-Means clustering to predict Happiness for a given test data-set after “training” a K-Means cluster with a bigger training set of data from the World Happiness Data. Repeated testing with different random seeds showed that the prediction is far from absolute, however - there was some error suggested by the histogram distribution of Happiness in the clusters.

III. PROPOSED METHODOLOGY



IV. PROJECT METHODOLOGY

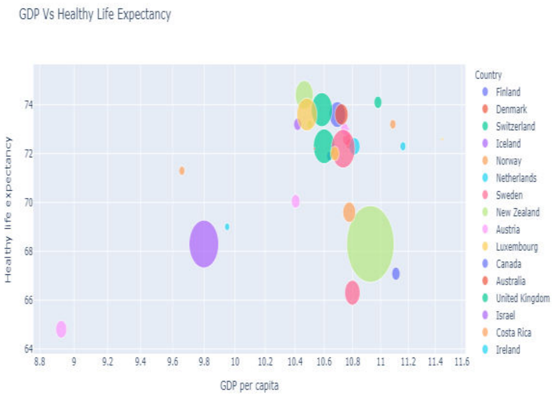
A) Data Description

Dataset consists of 7 CSV files of world happiness report from 2015-2. Almost every CSV file consists of a different number of columns as the UN keeps updating its criteria for the evaluation of world happiness. The total number of columns is 95.

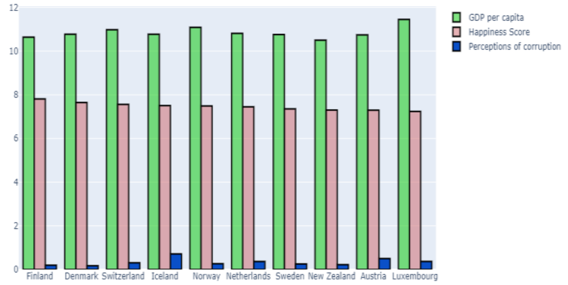
B) Data Preprocessing and Visualization

The next step was to check whether the dataset contained missing values, there were none. 7 data frames were created to store CSV files of respective years. Column names differed in each report, they were replaced with a comfortable feature name.

Correlation between the features indicated that there is a strong relationship between GDP per Capita and Healthy Life expectancy, which shows that more income is better. To make this more clear plot between these features for 2020 is plotted.



Another plot that shows the relation of perceptions of corruption, happiness score, and GDP of top 10 countries.



This graph indicates that the lower the perception of corruption more is the happiness score of the country.

C) Splitting the dataset

Splitting of the dataset is important to avoid a bias prediction on training data. While fitting ML algorithms can cause overfitting. Hence, the dataset was split into a 70:30 ratio.

D) Linear Regression

Linear regression is used to fit the model. It is the most basic and commonly used regression model. After checking the correlation between all the features and dependent variable happiness score, 3 features namely happiness ranking, Freedom of life choices, and healthy life expectancy were used as independent variables. The equation for multiple linear regression is given below.

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

R square, RMSE, and adjusted R Square value are obtained for all the algorithms discussed ahead. These three evaluating metrics are discussed in depth later in the paper.

E) KNN

Before fitting the KNN regressor, data is normalized using MinMaxScaler.KNN is a lazy learning, non-parametric algorithm. It uses data with several classes to predict the classification of the new test point. KNN is non-parametric since it doesn't make any assumptions on the data being considered, i.e., the model is distributed from the data. It checks the neighboring data point and decides the class for the test point.

F) Decision Tree

The decision tree determines the statistical probability. It is used to clarify and get insights on the answer i.e, predicting the happiness score. Each branch of the decision tree depicts a possible decision and the outcomes. It uses tree representation for solving the problem, leaf nodes correspond to the class label, and attributes are represented as internal nodes of the tree.

G) Random Forest

Every decision tree has a high variance, in the random forest, several trees are combined together in parallel creating a low variance. It uses the bagging technique to get a better result. The basic idea is to combine many decision trees and determine a final output instead of relying on a single decision tree. The model is fitted using a random forest regressor on training data and predictions are made on testing data.

H) R Square, Adjusted R Square and RMSE

R Square is a statistical measure that indicates how close the data are fitted to the regression line. For example, the R-Squared value of 0.8 would indicate that 80% of the variance of the dependent variable being studied is explained by the variance of the independent variable. when features, on which target variable is not much relied on are included in the model it results in the high value of R Square. Hence, Adjusted R Square can be used to avoid this problem it is a modified version of R Square, the equation is given below.

$$\mathbf{R^2\ adj = 1 - [\ (1 - R^2\ )*(n\text{-}1) / (n\text{-}k\text{-}1)\ ]}$$

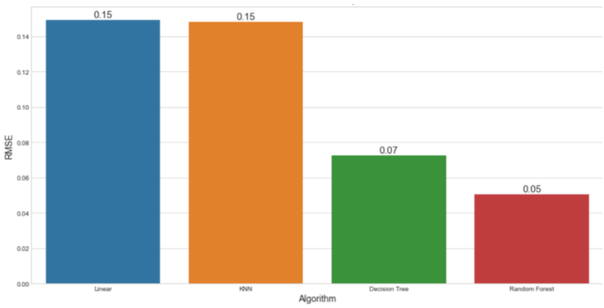
n is the number of data samples and k is the number of variables in the model. But R Square does not indicate whether a regression model is adequate. Hence, RMSE is used to calculate the performance of a model. It is the standard deviation of prediction errors, it tells how concentrated is the data around the best fit line. It is the square root of the value obtained from mean squared error(MSE), MSE determines the average difference between actual and predicted values of a feature.

V. CONCLUSION

In this advanced Python project, visualization and predictions of the World happiness report from 2015-21 are achieved. R Square and RMSE results of the algorithms are shown in the table below.

	R Square	RMSE
Linear Regression	0.9813	0.15
KNN	0.9958	0.15
Decision Tree	0.9956	0.07
Random Forest	0.9979	0.05

Comparison of Different algorithms is shown in the figure



Random forest showed the best RMSE result of 0.05. Lower the RMSE score, the model is better at predicting values therefore Random forest is concluded as the best model. Other models are showing RMSE values closer to 0.1 which indicates that the regression line fits the data well and the model performance is better

VI. REFERENCES

[1] P. Chaipornkaew and T. Prexawanprasut, "A Prediction Model for Human Happiness Using Machine

Learning Techniques," 2019 5th International Conference on Science in Information Technology (ICSITech), 2019, pp. 33-37, doi: 10.1109/ICSITech46713.2019.8987513.

[2] Dixit, Siddharth & Chaudhary, Meghna &Sahni, Niteesh. (2020). Network Learning Approaches to study World Happiness.

[3] Prashanthi, B., and R. Ponnusamy. "Future Prediction of World Countries Emotions Status to Understand Economic Status using Happiness Index and SVM Kernel." *Future* 6.11 (2019).

[4] Moore, Lisa. "Exploring trends and factors in the world happiness report." (2020).

[5] Ed Bullen, 12 August 2016Simple Machine Learning Prediction with the UN World Happiness Data-Set

http://rstudio-pubs-static.s3.amazonaws.com/201826_cab699be72ca47f99debadf16ee54c95.html

[6] R. A. Nugrahaeni and K. Mutijarsa, "Comparative analysis of machine learning KNN, SVM, and random forests algorithm for facial expression classification," 2016 International Seminar on Application for Technology of Information and Communication (ISemantic), 2016, pp. 163-168, doi: 10.1109/ISEMANTIC.2016.7873831.

[7] Min-Ling Zhang and Zhi-Hua Zhou, "A k-nearest neighbor based algorithm for multi-label classification," 2005 IEEE International Conference on Granular Computing, 2005, pp. 718-721 Vol. 2, doi: 10.1109/GRC.2005.1547385.

[8]M. R. Al-Hadidi, A. Alarabeyyat and M. Alhanahnah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm," 2016 9th International Conference on Developments in eSystems Engineering (DeSE), 2016, pp. 35-39, doi: 10.1109/DeSE.2016.8.

[9] Y. Benmahamed, M. Teguar and A. Boubakeur, "Application of SVM and KNN to Duval Pentagon 1 for transformer oil diagnosis," in IEEE Transactions on Dielectrics and Electrical Insulation, vol. 24, no. 6, pp. 3443-3451, Dec. 2017, doi: 10.1109/TDEI.2017.006841.

[10] M. Gashler, C. Giraud-Carrier and T. Martinez, "Decision Tree Ensemble: Small Heterogeneous Is Better Than Large Homogeneous," 2008 Seventh International Conference on Machine Learning and Applications, 2008, pp. 900-905, doi: 10.1109/ICMLA.2008.154.

[11] R. E. Banfield, L. O. Hall, K. W. Bowyer and W. P. Kegelmeyer, "A Comparison of Decision Tree Ensemble Creation Techniques," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 1, pp. 173-180, Jan. 2007, doi: 10.1109/TPAMI.2007.250609.

[12] Tin Kam Ho, "The random subspace method for constructing decision forests," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 832-844, Aug. 1998, doi: 10.1109/34.709601.

[13]K. P. Bennett and J. A. Blue, "A support vector machine approach to decision trees," 1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227), 1998, pp. 2396-2401 vol.3, doi: 10.1109/IJCNN.1998.687237.

[14] R. W. Selby and A. A. Porter, "Learning from examples: generation and evaluation of decision trees for software resource analysis," in IEEE Transactions on Software Engineering, vol. 14, no. 12, pp. 1743-1757, Dec. 1988, doi: 10.1109/32.9061.

[15]Shichao Zhang, Z. Qin, C. X. Ling and S. Sheng, ""Missing is useful": missing values in cost-sensitive decision trees," in IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 12, pp. 1689-1693, Dec. 2005, doi: 10.1109/TKDE.2005.188.

[16] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660-674, May-June 1991, doi: 10.1109/21.97458.