International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 11 | Nov 2019 www.irjet.net p-ISSN: 2395-0072

Future Prediction of World Countries Emotions Status to Understand Economic Status using Happiness Index and SVM Kernel

B. Prashanthi[1], Dr. R. Ponnusamy[2]

[1]PG Scholar in CSE Department, CVR College of Engineering, Telangana, Hyderabad, India.
[2]Professor in CSE Department, CVR College of Engineering, Telangana, Hyderabad, India

---

Abstract - For many years there has been a focus on individual welfare and societal advancement. In addition to the economic system, diverse experiences and the habitats of people are crucial factors that contribute to the well-being and progress of the nation. The predictor of quality of life called the Better Life Index (BLI) visualizes and compares key elements—environment, jobs, health, civic engagement, governance, education, access to services, housing, community, and income—that contribute to well-being in different countries. This paper presents a supervised machine-learning analytical model that predicts the life satisfaction score of any specific country based on these given parameters. This work is a stacked generalization based on a novel approach that combines different machine-learning approaches to generate a meta-machine-learning model that further aids in maximizing prediction accuracy. The work utilized an Organization for Economic Cooperation and Development (OECD) regional statistics dataset with four years of data, from 2015 to 2019. Using the data of 187 countries from the UN Development Project, this work is able to identify which factor needed to be improved by a certain country to increase the happiness of their citizens. The novel model achieved a high root mean squared error (RMSE) value of 0.3 with 10-fold cross-validation on the balanced class data. Compared to base models, the ensemble model based on the stacked generalization framework was a significantly better predictor of the life satisfaction of a nation. It is clear from the results that the ensemble model presents more precise and consistent predictions in comparison to the base learners.

Keywords- data mining; classification; feature selection; principal component analysis; support vector machine.

1. INTRODUCTION

World happiness has been actively studied throughout the last ten years. The work in [8] argues that the government of a country is usually driven by the happiness of their citizens. Some factors that are controlled or authorized by the government positively correlate with the happiness level. That work shows that the key role to determine the citizen happiness is the improvement of public policy. Understanding happiness factors will help governments to make a better policy and legislation. However, the factors that influence happiness could be different due to different human perspectives. We cannot just simply say that The United State is happier than Indonesia country because The United State has higher GDP. Peggy in [1] stated that happiness is correlated with national economic and cultural living conditions. The work in [2] determined Happiness using three factors which are life expectancy, experienced well-being and Ecological Footprint. Other work in [9] shows a new measurement to improve the happiness of a country. Unlike the previous work, this work studies that happiness is not only related to physical but also mental needs. Therefore, they also consider mental health, which includes stress, depression, and emotional problems. As a result of the increase of human social complexity, the factors proposed by [2] and [9] may not be reliable anymore. Additional factors such as health and human development index should be examined carefully. However, analyzing the factor to determine happiness of a particular country is not a trivial problem. A single factor can have a bigger impact than another. NEF organization in [2] proposes an equation to calculate the happiness index. However, this equation does not consider the economical aspects. Therefore, this work proposes an approach by extending the factors and adopting machine learning techniques to learn about those factors. Due to numerous size of the features, it is unwise to rely on the prediction of a world happiness done by manual analysis. That process will result in high cost of analysis. Therefore, this work also proposes the use of machine learning to predict the world happiness. Machine learning is a widely known technique to learn about patterns in data. There are several machine learning techniques which can be used to perform a prediction task [3]. One of the remarkable techniques is the support vector machine. This work uses support vector machine because its outstanding ability to perform a classification task.

2. RELATED WORK

This section briefly explains the related work in this project. Firstly, the national happiness analysis is described. It will discuss the importance of happiness analysis. Secondly, the used machine learning is introduced. Lastly, the proposed factor analysis is discussed.

A. World Happiness Analysis

The work in [4] mentioned that happiness could be a good indicator for how well a society is doing. This becomes important because Betham [5] said that the best society is the one where the citizens are happiest. Several researches have been conducted on positive aspects and the matters of

happiness in policy making [6][7]. As mentioned in the previous section, happiness can be determined based on various factors. Unfortunately, these factors were analyzed manually[8]. The complexity of the factor leads to the expensive cost of analysis. Therefore, the automatic analysis is needed.

B. Support Vector Machine

Due to its capability to learn from the past experiences, machine learning has been used in various areas. Support vector machine is one of the powerful machine learning algorithm. Support Vector Machine (SVM) is a learning technique which is used for classifying unseen data correctly. It is a learning technique which usually used for classifying the unseen data correctly. This technique has been used in various research field due to its remarkable performance. In order to perform the classification task, support vector machine builds a hyper plane which separates the data into different categories [9]. One of the important advantages of support vector machine is its ability to handle the scarcity of the data. Moreover, support vector machine is able to learn about the complex decision boundaries in the high dimensional feature space efficiently. Due to the complex features used to predict the national happiness, it is important to apply the technique with ability to handle the complex features.

C. Factor Analysis

As mentioned in the previous section, there are a large number of features used to predict the world happiness. However, some of the features may have no significant contribution to the prediction. Therefore, it is unwise to use the entire features to analyze the happiness. This work uses factor analysis to analyze the related features. Factor analysis aims to determine the contribution of a certain feature. This technique does not focus on dimension reduction. Therefore, there will be no features removed. The works in [10] and [11] have introduced the advantages of factor analysis. The first advantages mentioned is the ability to identify latent Dimensions or constructs that cannot be done using direct analysis. Moreover, this approach is easy to run and inexpensive in term of resources.

3. PROPOSED APPROACH

The aim of this project is to predict the world happiness of a particular country using machine learning techniques. The proposed approach contains four main steps in the data mining process which are data collection, data preprocessing, data analysis, and classification process as seen in Figure 1. The first process in the process approach is collecting the data. The data used in this project are gathered from the UN Human Development Project. The data contains of the human development index, GDI, healthy index of each country in the world. However, these data are quite dirty. It cannot be used directly as the input data for the learning process. Therefore, the second process is data preprocessing. Data preprocessing is used to increase the data quality. By increasing the quality of data results to the increasing number of prediction accuracy and consistency. The processes included in this process are data cleaning and data integration. The routine processes that should be done are filling the missing values, reduce the noise and identify the outliers.

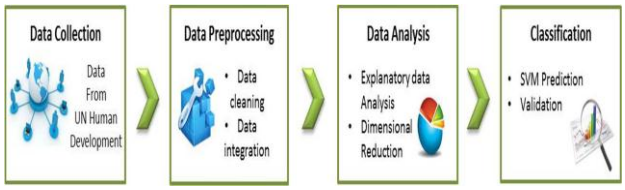

Figure 1. The proposed Approach

The third process in the proposed approach is data analysis. This explanatory data analysis is used for finding the relationship among the attributes of the features. This analysis is done by visualizing the data. Dimensional reduction also be done in this step. Using the information gain technique, the features can be reduced. Information gain technique is used because it can explore the interrelationships among a set of variables. The last part is this work is the classification process. In this classification process, SVM technique is used to predict the happiness of the data based on the important features. The validation process using k-fold cross validation technique is used to measure the performance of the data based on the accuracy, sensitivity and specificity values

4. RESULT AND ANALYSIS

This section discusses about the result of each step in the proposed approach. Moreover, this section also presents the analysis of significant factors to determine the happiness of a particular country.

A. Data Collection

As mentioned in the proposed approach section, the data used for this work are gathered from the UN Development Project. In total there are 187 countries listed in the data. Different types of factors are also mentioned in this data, such as human development index, education, environment, health care. The data consists of 105 types of features from 14 different factors. The dataset contained no missing values.

B. Data Preprocessing

In order to increase the data quality, the data preprocessing is needed. The collected data are scattered in various tables. Therefore, creating an integrated data is needed. The single integrated table consists of the entire features and sample that we are going to use in the next process. The dataset contained no missing values. The values within the dataset were present on a mixed scale, in the form of ratios,

percentage and average scores. The data was normalized using the min.max() normalization function in R. We used a realistic dataset so we did not have to consider outliers of the dataset and simulated our work using the original dataset [11] .

C. Feature Selection

Selecting the related features is important in order to improve the performance of the classifier. In order to perform this process, WEKA package for attribute selection is used [12]. The evaluator used in this work is information gain. This technique is chosen due to its ability to measure the amount of information in bits about the class prediction [13]. Therefore, it measures the expected reduction in entropy.

Figure 3. Selected features for classification

ATTRIBUTE NAME	ATTRIBUTE NAME
1.Ladder 2.SD Ladder 3.Positive affect 4.Negative affect 5.freedom 6.Corruption 7.Socila support	11.Generosity 12.LogofGDPpercapita 13.Health life expectancy 14.Happiness score 15.Happiness rank 16.Family
8.Dystopia Residual 9.Trust(Government corruption) 10.Lower confidence Interval	17.Higher confidence Interval 18.Whisker low 19.Whisker high

There are two main properties in the ranker evaluator. The first property is numToSelect property, which defines the number of attributes to keep, an Integer number that is -1 (all) by default. The next property is the threshold which defines the minimum value that an attribute has to get in the evaluator in order to be kept. In this case, the threshold is set to 0. After running this process, the number of remaining attributes is 19 includes class attribute. Those attributes are listed in Figure 3. The selected attributes such as inequality in life expectancy (inequality in the distribution of the expected length of life based on data from life tables estimated using the Atkinson inequality index), the number of homeless people, and the mortality rate is used to classify the happiness of a certain country.

D. Classification

In order to evaluate the selected attribute, this work also runs the classification using the entire gathered attributes. The same parameter used to compare both scenarios. The kernel for SVM is chosen based on cross validation. Table 1 shows the result for comparing different types of kernel. We can see that the normalized poly kernel gave an outperform result. Therefore, this kernel is chosen for this work.

TABLE I.

Kernel Type	Accuracy Rate
Normalize Poly Kernel	68.456 %
Poly Kernel	60.402 %
RBF Kernel	38.926 %
String Kernel	43.624 %

COMPARISON RESULT OF THE KERNEL

As mentioned before, this work also runs the classification using the entire attributes in order to validate the performance of selected attributes. Using the entire attributes we can see that the classification process result in 56.25% of accuracy. This result shows the improvement of accuracy rate and true positive rate. It also shows that using the selected attribute is able to reduce the mean square error. It means that the selected attributes have strong correlation with the class attribute that can be used to predict the class which is the happiness of the country.

E. Analysis

Instead of classifying the data into happy and unhappy countries, this work classified the data into three categories which are happy, mid, and unhappy. Based on the classification results, it shows that most countries in the world are not in a happy state. As seen in Figure 4, 39 %, 38%, 23% of the countries are happy, mid-happy, and unhappy, respectively.

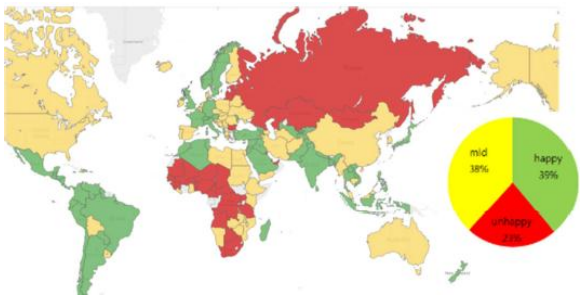

Figure 4. Distribution of Country Happiness

In order to analyze the happiness factor, this work also shows the distribution of the happiness based on the country. A country with red shade is a country in an unhappy state such as Russia and Nigeria. Moreover, the country with yellow shade is mid-state country such as the United State and Australia. Lastly, the country with green shade has happy state such as Brazil and Indonesia. This figure showed one surprising fact that even though a country is developed, it does not mean that it has a happy state.

5. CONCLUSIONS AND FUTURE WORK

In the current work, a supervised two-tier ensemble approach for predicting a country's BLI score was proposed. The work presented a cost-effective method of BLI prediction with a high degree of effciency. The dataset consisted of four different files for 2015 to 2019. The capability of the model to predict life satisfaction relied on the proper training features, chosen using a recursive elimination method with 10-fold cross-validation. The work combined three of the top four models, with simple averaging, to enhance the performance of the regression. The model was built using an ensemble approach and was evaluated using r, R, RMSE, and accuracy performance evaluators. The empirical relevance of diversity estimates were assessed with regard to combining the regression models by stacking. The model was about 90% accurate for predicting the life satisfaction score of a country. The present study was the first step towards forecasting the BLI score using machine learning based regression model that can influence the survival of future generations and further aid the immigration process. The work can be extended by tuning the parameters of the base models using meta-heuristic approaches to improve the prediction accuracy.

6. REFERENCES

[1] Schyns, Peggy "Cross national differences in happiness: Economic and Cultural factors explored." Social Indicators Research 43.1-2, pp. 3-26, 1998.

[2]http://www.happyplanetindex.org/assets/happy-planet-index-report.pdf

[3] Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[4] Gudmunds dottir, Dora Gudrun. "The impact of economic crisis on happiness." Social indicators research 110.3. pp: 1083-1101, 2013.

[5] Bentham, J. (1789/1996).An Introduction of the principles of morals and Legislation .Oxford: Clarendon Press. (Originally from 1789)

[6] Diener, E., Lucas, R. E., Schimmack, U., & Helliwell, J. Well-being for public policy. New York: Oxford University Press, 2009.

[7] Dolan, Paul, and Mathew P. White. "How can measures of subjective well-being be used to infor public policy?." Perspectives on Psychological Science 2, no.1 pp .71-85.2007.

[8] Viinamäki, H., Kontula, O., Niskanen, L., & Koskela, K. "The association between economic and social factors and mental health in Finland." Acta Psychiatrica Scandinavica 92, no. 3, pp.208-213, 1995.

[9] Malhotra, R., & Jain, A. "Software Effort Prediction using Statistical and Machine Learning Methods." International Journal of Advanced Computer Science and Applications 2., pp. 1451-1521, 2011.

[10] Garson, G. David, "Factor Analysis," from Statnotes: Topics in Multivariate Analysis.

[11] Tucker, L. R., & MacCallum, R. C.. "Exploratory factor analysis." Unpublished manuscript, Ohio State University, Columbus, 1997.

[12]http://weka.sourceforge.net/doc.dev/weka/filters/supervised/attribute/AttributeSelection.html

[13] Roobaert, D., Karakoulas, G., & Chawla, N.V. "Information gain, correlation and support vector machines." In Feature Extraction, pp. 463 - 470. Springer Berlin Heidelberg, 2006.