DOI: 10.15849/IJASCA.220328.02

Int. J. Advance Soft Compu. Appl, Vol. 14, No. 1, March 2022
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Analysis of World Happiness Report Dataset Using Machine Learning Approaches

Moaiad Ahmad Khder[1,*] Mohammad Adnan Sayfi[1] and Samah Wael Fujo[3]

[1*] Department of Computer Science – Applied Science University-Bahrain.
[2] Data Engineer – Umniah Jordan.
[3] Data Scientist – Nasser Artificial Intelligence Research and Development Centre - Nasser Vocational Training Center-Bahrain.
[1*]e-mail: moaiad.khder@asu.edu.bh

Abstract

happiness is a dream goal to be achieved by governments and individuals and it can be considered as a proper measure of social development progress. The purpose of this paper is to conduct a study on World happiness report dataset, to classify the most critical variables regarding the life happiness score. The strong evidence of the identified main features classified from the outcomes of applying the supervised machine learning approaches using the Neural Network training model and the OneR models in classifications and feature selection. The trained model used in predictions revealed the insights derived from applying the data analysis, where the study found out that the GDP per capita is the critical indicator of life happiness score as well as the health life expectancy is the second primary feature. Findings from study evaluated using different performance metrics such as accuracy and confusion matrix to prove the insights gained from the data.

Keywords: *world happiness, machine learning, Neural Network.*

1 Introduction

An increasing corpus of scientific data links happiness to physical health and overall well-being. Most individuals define happiness as purpose and well-being. This cheerful mindset has a number of physical and mental health benefits, including:

- Happiness enhancing your health by lowering your blood pressure, improving your sleep, improving your diet, allowing you to exercise regularly, and reducing stress.

Received 6 October 2021; Accepted 22 December 2021

- Optimism and energy are crucial to a person's well-being.
- enhancing problem-solving skills, positive thinkers feel they “can” and want to attain their goals.
- Building physical, mental, and social resources helps people learn better because they seek like-minded individuals.

World happiness report [1] shows that the proper measure of social development progress is Happiness and the aim of public policy, national average life equation taking into account six critical variables: GDP per capita, social support, healthy life expectancy, freedom from corruption. Happiness indicators relying on World Values Survey applied to 65 countries and Gallup World Poll covering 155 countries [1]. The Organisation for Economic Co-operation and Development(OECD) is making essential proposals for globally standard measures could help the world happiness analytics. This research helps us understand the features and conditions that lead to happiness relying on the world happiness report such as: (Economy .GDP per. Capita, Health. life. Expectancy, Family, Freedom, Destopia. Residual, Trust. Government. Corruption, and Generosity).

2 Related Work

A decent 2020 paper [2] included a descriptive insight for the analysis of happiness and wellbeing relationships and the distinction between happiness and various items such as(competence, emotional stability, positive emotions, and engagement) [2]. critical previous studies, where positivity items defined as the level of satisfaction of life without distinction between them in a sense. The study indicates that GDP and similar labels are links to happiness, but it is far from complete correlation.

A 2009 study by [3] showed that there is a difference in the correlation between the self-assessment happiness score and the values of happiness provided by the family and friends. The study found that there is a discrepancy between the three evaluations that must take into account. In 2011, the scope of Machine Learning on the analysis of global happiness by [4] using different approaches. The principal component analysis(PCA) used to analyse gender equality and satisfaction of life, and in feature selection decision trees were used as well as for life satisfaction predictions. Finings of this study the key features life expectancy, Income distribution, and freedom by using permutation testing which summarized its results in the form of visualizations map for life happiness. An analysis report by [5] used machine learning approaches to split the UN World Happiness dataset into training and test data-set and used K-Means cluster reached imperfect due to use of one approach due not considering other metrics while clustering.

The accuracy of studies and analyses are considered limitations. Because of the limited data available, such as Gallup World Poll (GWP)survey data are not freely available for use [4], as well as variables that are not available or difficult to measure, may be better than current variables. The timing of the studies varies as the happiness of the people differs from one season to another. The answers to the different questionnaires differ which provide some biases with happiness variables. The time also affects the processing of the studies and the variables for each year, with the start date of the annual analysis to produce data sets thus made data analysts and researchers to use time series for some variables [6].

Many studies explored different areas of the world happiness, such as the list of the studies shown in table 1:

Table 1. List of Happiness research studies

Study	Year	Study Focus
[7]	2022	freedom of citizens and COVID-19
[8]	2021	Quality of Life
[9]	2021	Causes of Suicides and identification of Vulnerable Categories
[10]	2021	Prediction of world happiness scenario effective in the period of COVID-19 pandemic
[11]	2021	Clustering countries according to the world happiness report
[12]	2020	the important features affecting happiness which would be useful in policy making
[13]	2020	Predict Happiness
[14]	2020	Exploring trends and factors in the world happiness report
[15]	2020	Clustering Countries According to The World Happiness Report
[16]	2020	Southeast Asia Happiness Report Analysis
[17]	2019	Emotions Status to Understand Economic Status
[18]	2019	national quality of life scoring
[19]	2018	Global Happiness
[20]	2018	Exploring Subjective Well-Being Factors
[21]	2015	cross-national differences in Happiness

Upon reviewing different studies, the research gap is to find the degree of importance of variables that lead to the happiness and have a high impact on it, in addition to identify which one of the machine learning approaches can derive the highest accuracy in happiness prediction.

3 The proposed Machine learning approaches

Kaggle World happiness report dataset [22] is a landmark survey of the state of global happiness, which ranks 155 countries among their happiness levels. Happiness scores rated on a scale from 0 to 10 depending on survey respondents to think a ladder with the best possible life for them. The dataset provides several insights from variables affecting the world happiness scores as critical features and supporting features.

The data set features rely on getting answers from the respondents based on their day to day life experiences taking into consideration the highest persuasive life and the most extremely bad life being. The following describes the features in detail.

- Sample of the first 10 rows includes eight main features is shown in table 2, data statistics is shown in table 3, and variable types is shown in table 4:

Table 2. data sample (10 rows)

Country name	Regional indicator	GDP per capita	Social support	Healthy life expectancy	Freedom	Generosity	corruption
Finland	Western Europe	10.6393	0.95433	71.9008	0.94917	-0.0595	0.19545
Denmark	Western Europe	10.774	0.95599	72.4025	0.95144	0.0662	0.16849
Switzerland	Western Europe	10.9799	0.94285	74.1024	0.92134	0.10591	0.30373
Iceland	Western Europe	10.7726	0.97467	73	0.94889	0.24694	0.71171
Norway	Western Europe	11.0878	0.95249	73.2008	0.95575	0.13453	0.26322
Netherlands	Western Europe	10.8127	0.93914	72.3009	0.90855	0.20761	0.36472
Sweden	Western Europe	10.7588	0.92631	72.6008	0.93914	0.11162	0.25088
New Zealand	North America and ANZ	10.5009	0.94912	73.2026	0.93622	0.1916	0.22114
Austria	Western Europe	10.7428	0.92805	73.0025	0.89999	0.08543	0.49996
Luxembourg	Western Europe	11.4507	0.90691	72.6	0.90564	-0.0046	0.36708

Table 3. Dataset Statistics

Number of variables	Number of observations	Missing cells	Missing cells (%)	Duplicate rows	Duplicate rows (%)
20	153	0	0.00%	0	0.00%

Table 4. Variable Types

Number of variables	Number of observations
20	153

The Supervised Machine learning approaches used to reveal the insights from the data, where the classification approach used to define the variable importance by using Neural Networks Classifier. Neural networks facilitate to find which label from the data fits best to the second supervised learning approach used in this study which is a prediction. This report proposed two vital inquiries derived from the data:

Does the world happiness score rely on the economy GDP per capita?

Does the Health life expectancy lead to happy life?

[23] Witnessing the growing availability of vast quantities of data and developments in the fields of artificial intelligence (AI), machine learning (ML), and optimisation. Statistical breakthroughs, discrete new algorithms, and applied mathematics create a new interdisciplinary domain called data science

ML is used to make machines ready to learn and can handle a big amount of data efficiently and to reveal patterns from a sea of data that was hidden without the use of ML [24].

3.1 Supervised Learning Approaches

First of all, the dataset divided into two parts world happiness data 2019 and 2020 consists of ten similar variables (Country, Region, Happiness Rank, Happiness Score Economy GDP per capita, family, Health life expectancy, Freedom, Generosity, trust government corruption, dystopia residual)and has no NA’s observations. So the two data sets merged into one dataset to ease the progress and expand the training data, which increase the accuracy while building machine learning classifiers and feature selection process.

3.1.1 Classification:

Classification :is arranging the data into predefined groups or classes. Common algorithms include nearest neighbor, Naive Bayes classifier and neural network [25].

For classifying the data three intervals made to deal with unbalanced data regarding happiness score for each observation as follows (Low, Medium, High) which each interval indicates values of the happiness score variable as shown in density plot in figure 1:

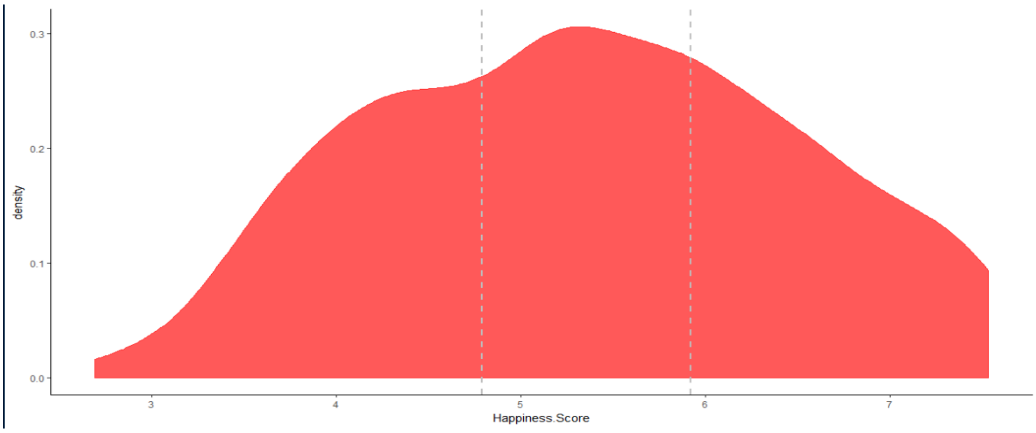


Figure 1. shows the Happiness score intervals(bins)

The happiness score intervals is shown in table 5

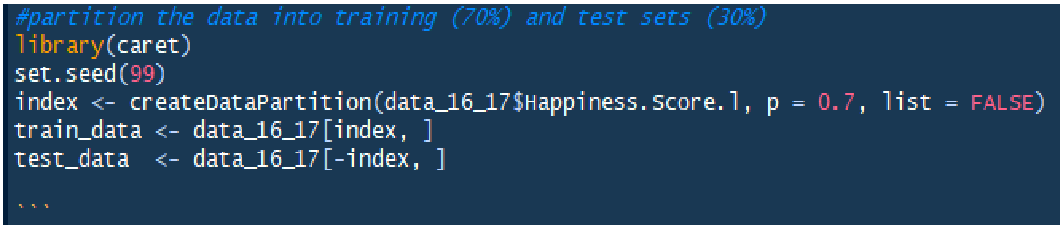
Table 5. Happiness score intervals

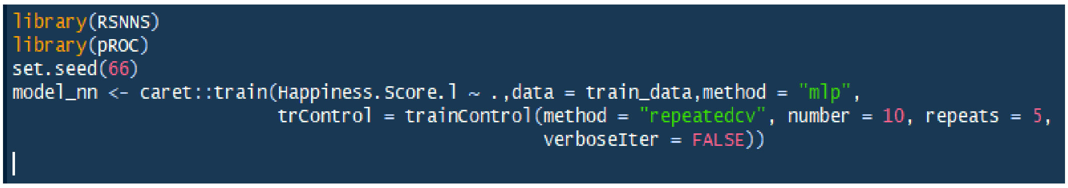
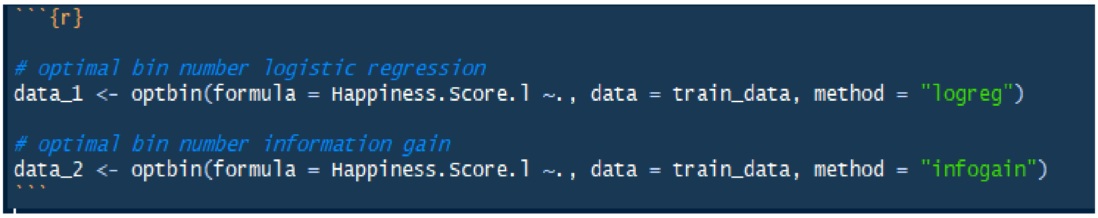
Interval	Happiness Score
Low	[2.69,4.79]
Medium	[4.79,5.92]
High	[5.92,7.54]

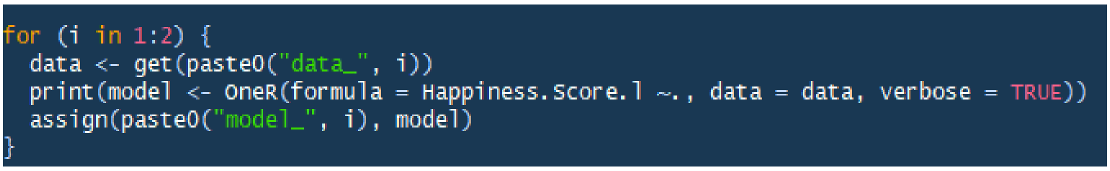
Neural Network Classifier

Data partitioned into two datasets the training (70%) and test sets (30%) regarding Happiness score variable to define the rest of important labels and select the best-fit feature using the multilayer perceptrons as a model training method which used the backpropagation for training. Building model with repeated cross-validation for five repeats and dividing the training data set randomly into ten parts by using the R caret library.

The figure2 shows the code segment that divides the dataset into training and test sets



```
#partition the data into training (70%) and test sets (30%)
library(caret)
set.seed(99)
index <- createDataPartition(data_16_17$Happiness.Score.1, p = 0.7, list = FALSE)
train_data <- data_16_17[index, ]
test_data  <- data_16_17[-index, ]

```
```

Figure 2. data partition code segment

The figure 3 shows neural network model building



```
library(RSNNS)
library(pROC)
set.seed(66)
model_nn <- caret::train(Happiness.Score.1 ~ .,data = train_data,method = "mlp",
                         trControl = trainControl(method = "repeatedcv", number = 10, repeats = 5,
                                                  verboseIter = FALSE))
```

Figure 3. Neural Networks training model building

OneR Classification Algorithm

OneR is an algorithm method used to create machine learning models with sufficient accuracy and faster training time comparing with other models, while there is a library on R cran based on the one algorithm method called OneR created excellently to handle the NA values. Regarding oneR using only categorical features so splitting the variables into bins needed by using built-in function in the OneR package to prepare the data for model building. The methods used for building the OneR models are logistic regression and the info again method. The figure 4 shows code segment to split the data using the optimal bin numbers



```
```{r}

# optimal bin number logistic regression
data_1 <- optbin(formula = Happiness.Score.1 ~., data = train_data, method = "logreg")

# optimal bin number information gain
data_2 <- optbin(formula = Happiness.Score.1 ~., data = train_data, method = "infogain")
```
```

**Figure 4.** splitting data into bins (converting values into factor intervals)

OneR model building using the split data above using OneR package to produce the highest features accuracy among two methods (log reg, info again) as shown in figure 5:
```



```
for (i in 1:2) {
  data <- get(paste0("data_", i))
  print(model <- OneR(formula = Happiness.Score.l ~., data = data, verbose = TRUE))
  assign(paste0("model_", i), model)
}
```

Figure 5. OneR Models building

3.1.2 Prediction:

After building the NN classifier trained model now, the turn is for the prediction on the test data set based on the feature selected by the model as shown in figure 6, where it shows the most critical variable across the three classes is the GDP per capita which the model used it as the critical feature and continue the progress of the prediction based on it.

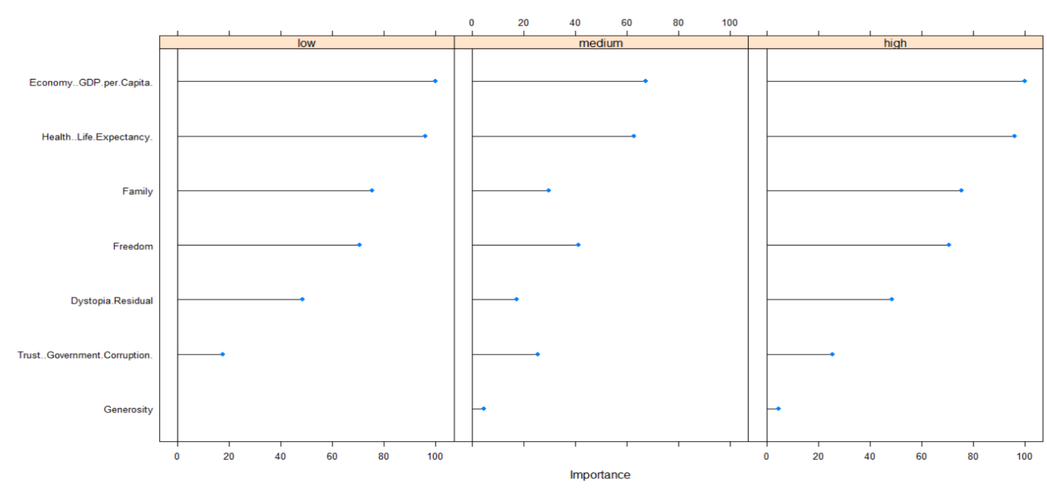
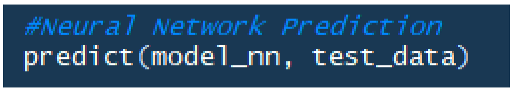


Figure 6. variable importance using the NN model across three happiness score classes

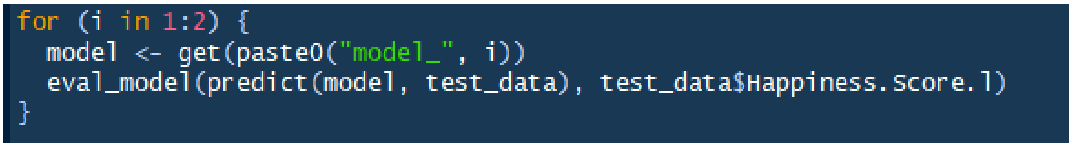
To validate the accuracy of the NN model using the code segment shown in figure 7:



```
#Neural Network Prediction
predict(model_nn, test_data)
```

Figure 7. Neural Network Prediction

Which give the prediction on the test dataset using the trained model above to produce the confusion matrix which facilitates to measure the accuracy and the error rate in the following section. Also, OneR prediction model using for loop to use each trained model and assign them into data frames which can be evaluated in the following section as shown in figure 8:



```
for (i in 1:2) {
  model <- get(paste0("model_", i))
  eval_model(predict(model, test_data), test_data$Happiness.Score.1)
}
```

Figure 8. Generating prediction models from OneR trained models

Using the eval_model function to generate the confusion matrix absolute and relative for each model produced above to evaluate the models' accuracy.

4 The proposed evaluation methods

This section is covering the proposed evaluation method. Which we will consider the classification accuracy.

4.1 Classification Accuracy

This section is covering the different models' classification accuracies: Neural Network, OneR, and finalize that by the confusion matrix.

4.1.1 Neural Network Part

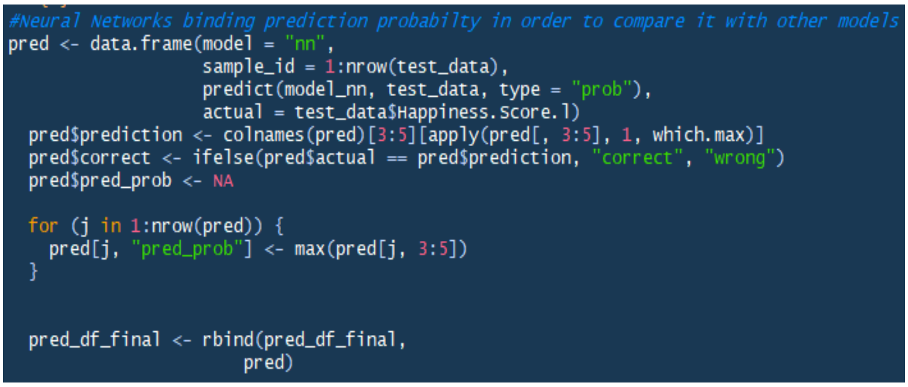
This evaluation metric used to measure the accuracy of the neural network classification model using the formula (1):

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made} \qquad (1)$$

Accuracy generated automatically by code and results as discussed in the following section. Also to confirm the model accuracy by comparing it with other machine learning methods: Random forest model and Extreme gradient boosting trees (xgb) applied on the same training data set to confirm the accuracy of the Neural Network model. Thus comparison made manually by finding the prediction probability for each class of happiness score(Low, Medium, High) and then mutate the results into one data frame to visualise it by using the equation(2):

$$Prediction\_Probability = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made} \times 100 \qquad (2)$$

Prediction probability generated by code by binding the prediction model values predicted with type "prob" to calculate the formula above for each class on the test data set by using the segment of code shown in figure 9:



```
#Neural Networks binding prediction probabilty in order to compare it with other models
pred <- data.frame(model = "nn",
                   sample_id = 1:nrow(test_data),
                   predict(model_nn, test_data, type = "prob"),
                   actual = test_data$Happiness.Score.1)
  pred$prediction <- colnames(pred)[3:5][apply(pred[, 3:5], 1, which.max)]
  pred$correct <- ifelse(pred$actual == pred$prediction, "correct", "wrong")
  pred$pred_prob <- NA

  for (j in 1:nrow(pred)) {
    pred[j, "pred_prob"] <- max(pred[j, 3:5])
  }


  pred_df_final <- rbind(pred_df_final,
                         pred)
```

Figure 9. the predicted   probability for the neural network model

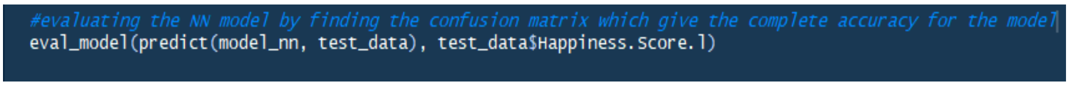
The code above defining a data frame contains the samples from the test data set and using the predict function on the test data by applying it on the NN trained model also the actual values for each class. Generating a variable for the correct predicted values by comparing the predicted value with the actual sample value .and finally the for loop assign the maximum values between each predicted class value and consider it as probability of the particular sample in order to compare it with the rest of model used in this performance measure the outcome of this measurement discussed in the following section.

4.1.2 OneR part

Using accuracy metric evaluation measure to find the performance for the OneR models used (LogReg, Infogain) by finding the accuracy for each model mainly using the OneR model structure data, as well as compare its prediction probability with decision tree method as described in the following section.

4.2 Confusion Matrix

This metric used to evaluate the complete performance of the Neural Network model and for other methods used to confirm the accuracy of the model used. Generating confusion matrices for the models by using built-in R function which takes the prediction model and the test data as input and provide the absolute and the relative confusion matrix which gives the complete accuracy for the model as shown in figure 10 :



```
  #evaluating the NN model by finding the confusion matrix which give the complete accuracy for the model
  eval_model(predict(model_nn, test_data), test_data$Happiness.Score.1)
```

Figure 10. shows code used to produce the confusion matrix

The output of the code shown in figure 10 is mentioned in the following section as confusion matrices for each model used for the evaluation, where the same code above used to evaluate the OneR models to generate their confusion matrix for each model.

5 The evaluation results and the proposed method for improvement

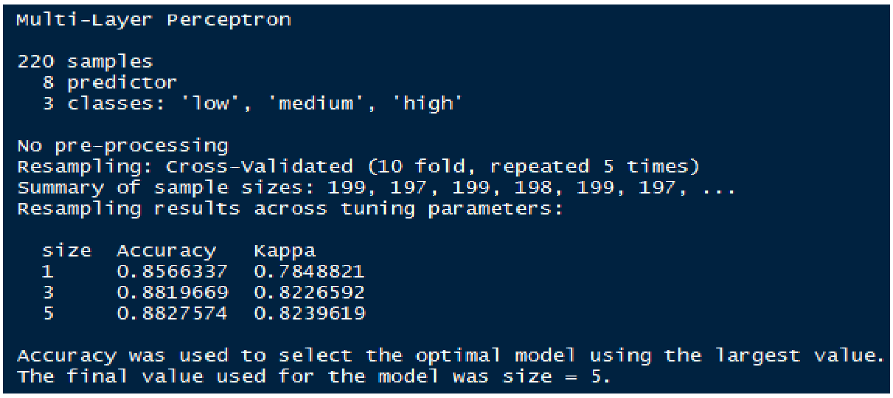
The discussion of the classification accuracy evaluation is covered in the following sections:

5.1 Classification accuracy evaluation results:

The discussion of the classification accuracy evaluation for (: Neural Network, OneR, and finalize that by the confusion matrix) is covered in the following sections:

5.1.1 Neural network part

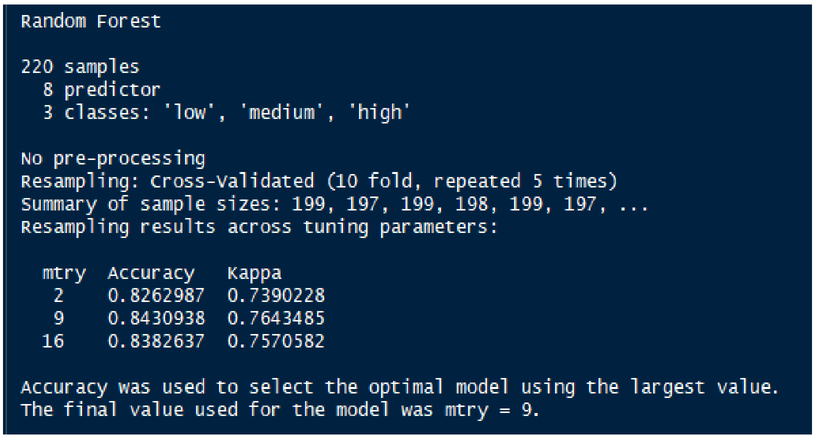
The Neural Network model accuracy is shown in figure 11, while selected the highest model accuracy as the final model.



```
Multi-Layer Perceptron

220 samples
  8 predictor
  3 classes: 'low', 'medium', 'high'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 199, 197, 199, 198, 199, 197, ...
Resampling results across tuning parameters:

  size  Accuracy   Kappa
  1     0.8566337  0.7848821
  3     0.8819669  0.8226592
  5     0.8827574  0.8239619

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was size = 5.
```

Figure 11. Neural Network trained models accuracy

The model considered as a final model with value size =5 used as the final neural network model because it has the highest accuracy(0.882) and kappa (0.823)among the other models while training the data. Comparing the NN model with the random forest model accuracy in the same manner used to describe its accuracy as shown in the figure 12 :



```
Random Forest

220 samples
  8 predictor
  3 classes: 'low', 'medium', 'high'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 199, 197, 199, 198, 199, 197, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.8262987  0.7390228
   9    0.8430938  0.7643485
  16    0.8382637  0.7570582

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 9.
```

Figure 12. Random Forest trained models accuracy

The model considered as final after building the random forest method is the model with the accuracy value of (0.843) and kappa (0.764).

The prediction probability formula discussed in the above evaluation section between the three methods used (NN, RF, XGB) and generated by earlier code is visualised for the Neural Network model as shown in figure 13 and grouped the results for all model as shown in figure 14.

The figure 13 shows that the correct predicted value using the NN model and compare it with the wrong predictions, where the box plot indicates that most of the correct values incident between the high and the low class of Happiness score.

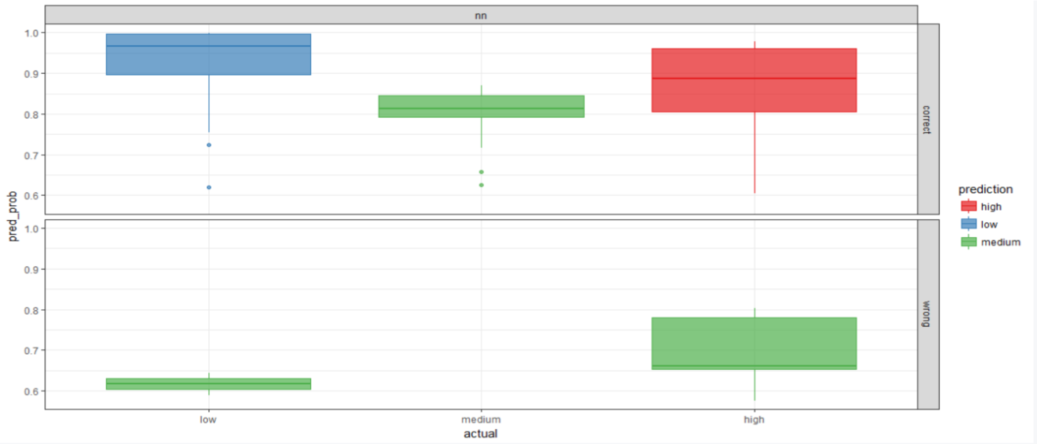


Figure 13. Prediction probability for NN model

The figure 14 discussed the probability of prediction using each model and the nearest to the one has the highest accuracy among the models which is the Neural Network model used in this report.

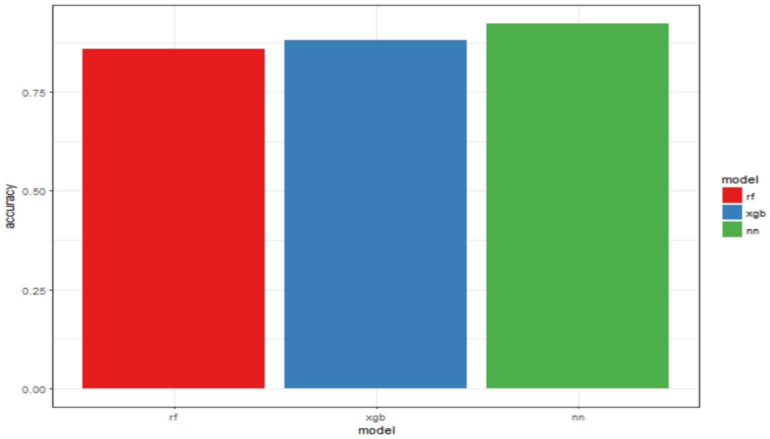
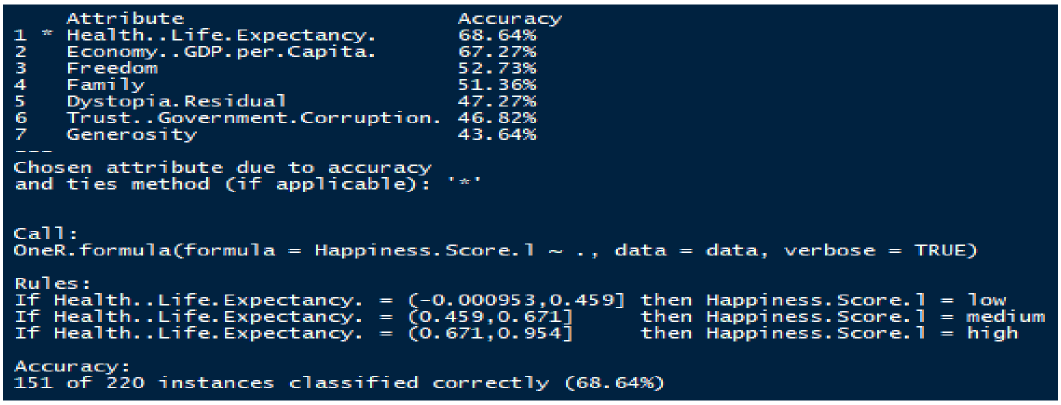


Figure 14. rf, xgb, NN models evaluation

5.1.2 OneR part:

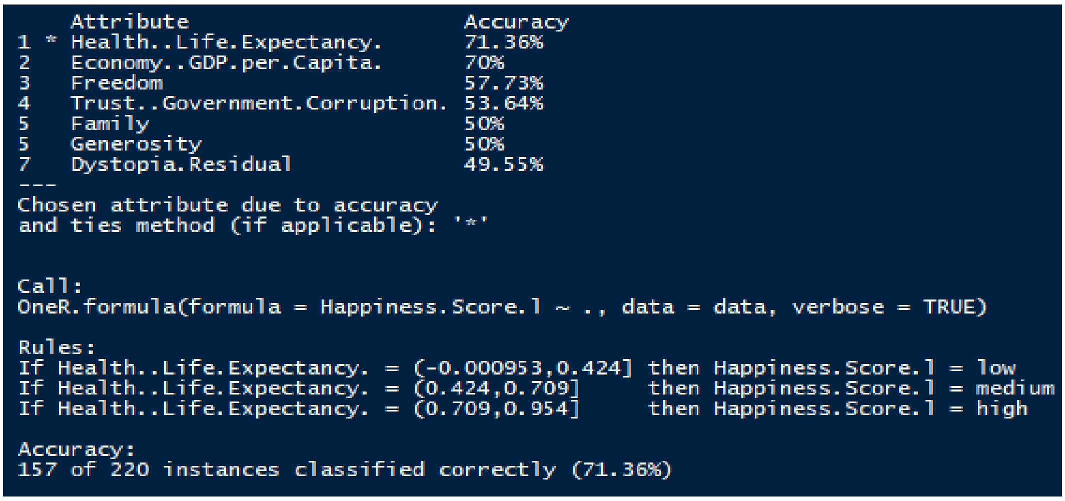
In this evaluating metric, the OneR models evaluated by its accuracy as shown in the figures 15:



```
    Attribute                         Accuracy
1 * Health..Life.Expectancy.          68.64%
2   Economy..GDP.per.Capita.          67.27%
3   Freedom                           52.73%
4   Family                            51.36%
5   Dystopia.Residual                 47.27%
6   Trust..Government.Corruption.     46.82%
7   Generosity                        43.64%
---
Chosen attribute due to accuracy
and ties method (if applicable): '*'


Call:
OneR.formula(formula = Happiness.Score.1 ~ ., data = data, verbose = TRUE)

Rules:
If Health..Life.Expectancy. = (-0.000953,0.459] then Happiness.Score.1 = low
If Health..Life.Expectancy. = (0.459,0.671]     then Happiness.Score.1 = medium
If Health..Life.Expectancy. = (0.671,0.954]     then Happiness.Score.1 = high

Accuracy:
151 of 220 instances classified correctly (68.64%)
```

Figure15. Logreg model accuracy and variable importance

The OneR algorithm shows that the highest selected feature in the model is the Health life expectancy thus leads to the answer to the second research question and discussed in the results section. The best fit model gives a sufficient accuracy and has (68.64) among other features. OneR indicates three rules as described in the figure 15 where each rule identified for each happiness score class. The figure 16 identified the accuracy of the classification using the (infogain) method.



```
    Attribute                     Accuracy
1 * Health..Life.Expectancy.      71.36%
2   Economy..GDP.per.Capita.      70%
3   Freedom                       57.73%
4   Trust..Government.Corruption. 53.64%
5   Family                        50%
5   Generosity                    50%
7   Dystopia.Residual             49.55%
---
Chosen attribute due to accuracy
and ties method (if applicable): '*'


Call:
OneR.formula(formula = Happiness.Score.l ~ ., data = data, verbose = TRUE)

Rules:
If Health..Life.Expectancy. = (-0.000953,0.424] then Happiness.Score.l = low
If Health..Life.Expectancy. = (0.424,0.709]     then Happiness.Score.l = medium
If Health..Life.Expectancy. = (0.709,0.954]     then Happiness.Score.l = high

Accuracy:
157 of 220 instances classified correctly (71.36%)
```

Figure16. Infogain model accuracy and variable importance

OneR indicates that highest model gains the highest accuracy which is, in this case, the model trained by the infogain method as shown above that got (71.36%) and the selected feature is the Health life expectancy. Comparing the OneR models accuracy with the decision tree method using the prediction probability to define its effectiveness while producing models as shown in the figure 17:

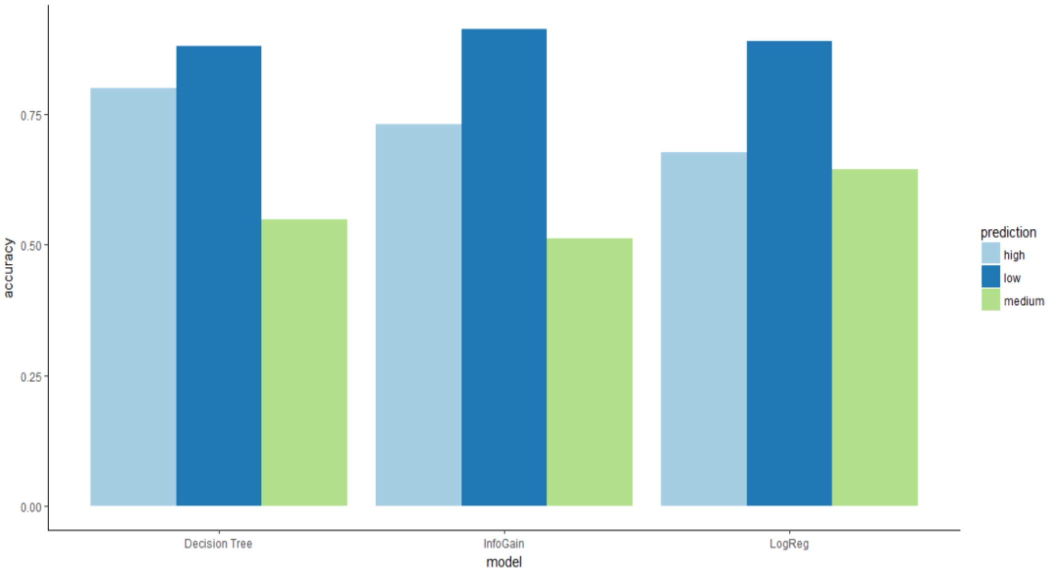


Figure17. OneR models accuracy with decision tree model's accuracy

The figure 17 identified the OneR models accuracy and comparing it with the decision tree model by using the prediction probability for each Happiness score class which shows a sufficient accuracy with faster building model runtime.

5.2 Confusion Matrix evaluation results:

5.2.1 Neural network part

In this metric evaluation, the neural network confusion matrix produced to reveal the complete performance of the trained model as shown in figure 18(a) .

Outcomes of the figure 18(a) that the accuracy for the complete model calculated using the formula (3):

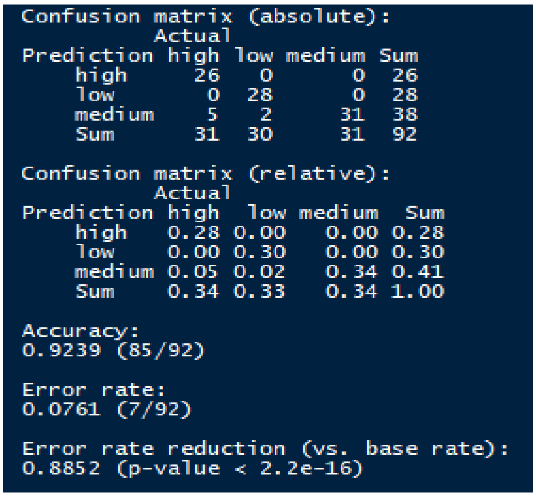
$$Accuracy = \frac{TruePositives + FalseNegatives}{Total\ number\ of\ Samples} \qquad (3)$$

For the NN model the accuracy calculated as the equation(4):

$$Accuracy = \frac{26 + 28 + 31}{92} \qquad (4)$$

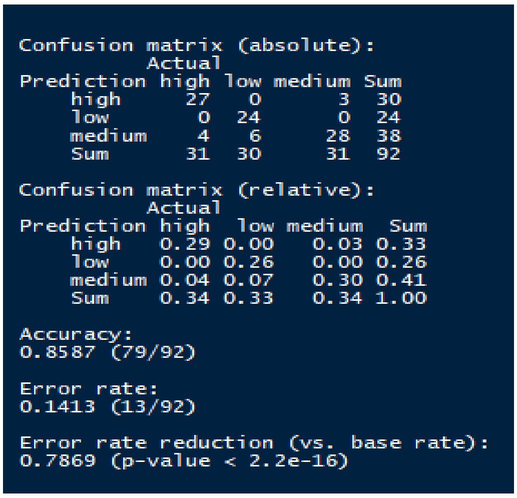
Where each number indicates the value for each happiness score class (High, Medium, Low) and results to 0.9239 as the complete model accuracy.

For comparing the NN model with random forest model in term of the complete performance using the confusion matrix for the rf model described in the figure 18(b):



```
Confusion matrix (absolute):
          Actual
Prediction high low medium Sum
    high     26   0      0  26
    low       0  28      0  28
    medium    5   2     31  38
    Sum      31  30     31  92

Confusion matrix (relative):
          Actual
Prediction high  low medium  Sum
    high   0.28 0.00   0.00 0.28
    low    0.00 0.30   0.00 0.30
    medium 0.05 0.02   0.34 0.41
    Sum    0.34 0.33   0.34 1.00

Accuracy:
0.9239 (85/92)

Error rate:
0.0761 (7/92)

Error rate reduction (vs. base rate):
0.8852 (p-value < 2.2e-16)
```

Figure 18(a). Confusion matrix for the RF trained model



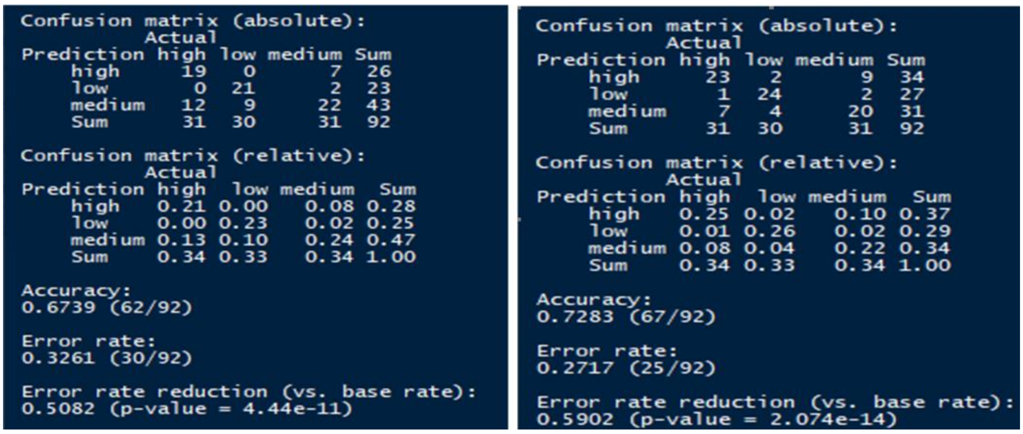
```
Confusion matrix (absolute):
          Actual
Prediction high low medium Sum
    high     27   0      3  30
    low       0  24      0  24
    medium    4   6     28  38
    Sum      31  30     31  92

Confusion matrix (relative):
          Actual
Prediction high  low medium  Sum
    high   0.29 0.00   0.03 0.33
    low    0.00 0.26   0.00 0.26
    medium 0.04 0.07   0.30 0.41
    Sum    0.34 0.33   0.34 1.00

Accuracy:
0.8587 (79/92)

Error rate:
0.1413 (13/92)

Error rate reduction (vs. base rate):
0.7869 (p-value < 2.2e-16)
```

Figure 18(b). Confusion matrix for the NN trained model

The matrix shows the accuracy of the RF model is 0.924 with 79 positive predicted values so comparing to the NN model the trained nn model provides higher accuracy which indicates the massive performance of the model used.

5.2.2 OneR part:

Confusion matrix used to identify the accuracy of two models generated by the OneR method as described in the figure 19:



```
Confusion matrix (absolute):
          Actual
Prediction high low medium Sum
    high     19   0      7  26
    low       0  21      2  23
    medium   12   9     22  43
    Sum      31  30     31  92

Confusion matrix (relative):
          Actual
Prediction high  low medium  Sum
    high   0.21 0.00   0.08 0.28
    low    0.00 0.23   0.02 0.25
    medium 0.13 0.10   0.24 0.47
    Sum    0.34 0.33   0.34 1.00

Accuracy:
0.6739 (62/92)

Error rate:
0.3261 (30/92)

Error rate reduction (vs. base rate):
0.5082 (p-value = 4.44e-11)
```

```
Confusion matrix (absolute):
          Actual
Prediction high low medium Sum
    high     23   2      9  34
    low       1  24      2  27
    medium    7   4     20  31
    Sum      31  30     31  92

Confusion matrix (relative):
          Actual
Prediction high  low medium  Sum
    high   0.25 0.02   0.10 0.37
    low    0.01 0.26   0.02 0.29
    medium 0.08 0.04   0.22 0.34
    Sum    0.34 0.33   0.34 1.00

Accuracy:
0.7283 (67/92)

Error rate:
0.2717 (25/92)

Error rate reduction (vs. base rate):
0.5902 (p-value = 2.074e-14)
```

Figure 19. Confusion matrices for the infogain(Left)and Logreg(Right)

Outcomes from the matrices above-identified performance of the Logreg based model as the highest best fit model for the OneR method which has higher accuracy(0.7283) than the trained model using the infogain accuracy (0.6739).

5.3 The proposed method for improvement

Some improvements for the proposed methods are introduced in the following sections:

5.3.1 Improvement for the Neural Network approach:

The existence of an ensemble of randomised regression trees can facilitate to restructure the neural networks with the random forest.

This ensemble is to provide a collection of multilayered neural networks with particular connection weights called neural random forests by using different hybrid procedures identified as random neural forests which this approach proved by real experiment and give an excellent performance [26], thus can enhance the feature selection using this approach

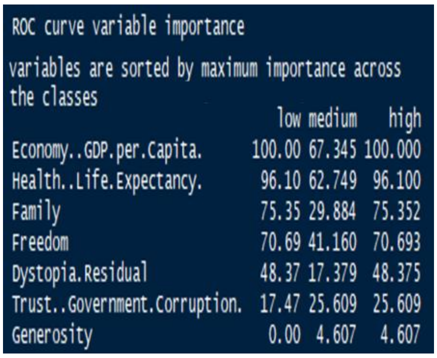
5.3.2 Improvement for OneR approach:

The OneR package enhanced from the one rule algorithm [27] by discretisation of numeric data which the new OneR optimises the numeric data into cut points to deal with them. The enhanced OneR is treating the missing values by separating them into the specific level called (Level "NA") or omitting NA’s by default. Finally, the tie-breaking issue while the old one rule algorithm takes the first attribute while the enhanced version taking all best rules by merging them into the contingency tables of all best rules [28].

6 The final results and the insights gained from the study

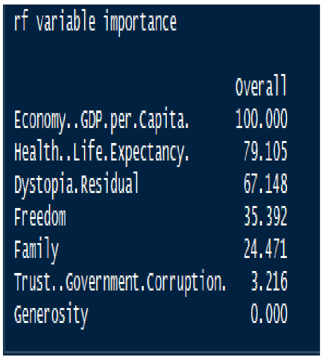
The first results indicate the GDP per capita as the most important variable which is the most feature affecting the world happiness score and these results proved by using the neural network trained model as shown in the figures 20(a) :

Economy GDP per Capita proved as the most important variable by using the ROC curve. Also confirmed by the RF model as shown in figure 20(b) and the xgb model as shown in figure 20(c).



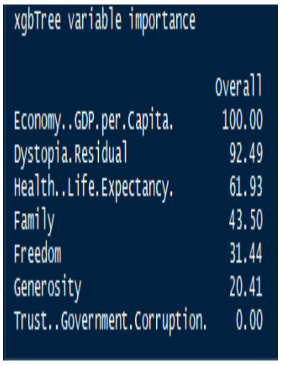
```
ROC curve variable importance

variables are sorted by maximum importance across
the classes
                                 low medium    high
Economy..GDP.per.Capita.       100.00 67.345 100.000
Health..Life.Expectancy.        96.10 62.749  96.100
Family                          75.35 29.884  75.352
Freedom                         70.69 41.160  70.693
Dystopia.Residual               48.37 17.379  48.375
Trust..Government.Corruption.   17.47 25.609  25.609
Generosity                       0.00  4.607   4.607
```

Figure20(a). NN model variable importance



```
rf variable importance

                              Overall
Economy..GDP.per.Capita.      100.000
Health..Life.Expectancy.       79.105
Dystopia.Residual              67.148
Freedom                        35.392
Family                         24.471
Trust..Government.Corruption.   3.216
Generosity                      0.000
```

Figure22(b). RF model variable importance

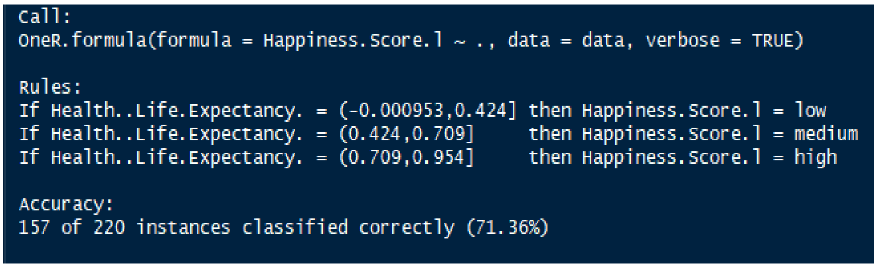


```
xgbTree variable importance

                              Overall
Economy..GDP.per.Capita.       100.00
Dystopia.Residual               92.49
Health..Life.Expectancy.        61.93
Family                          43.50
Freedom                         31.44
Generosity                      20.41
Trust..Government.Corruption.    0.00
```

Figure22(c). xgb model variable importance

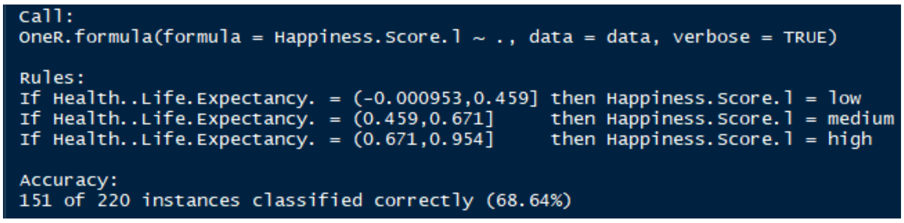
Finally, by the evidence of the evaluation of the machine learning methods used and its high accuracy in classifying and prediction thus leads to the insights gained from the data to answer the first research question as coding the GDP per Capita as one of the critical elements in the happiness life score.

The second result indicates that health life expectancy attribute as the first rule in both logistic regression model and the information gain model, thus reveal the insights of that high life expectancy may lead to an excellent happiness score. The figure 21,22 show that how OneR considered the life expectancy attribute as the first rule among the three different classes life happiness score.



```
Call:
OneR.formula(formula = Happiness.Score.1 ~ ., data = data, verbose = TRUE)

Rules:
If Health..Life.Expectancy. = (-0.000953,0.424] then Happiness.Score.1 = low
If Health..Life.Expectancy. = (0.424,0.709]     then Happiness.Score.1 = medium
If Health..Life.Expectancy. = (0.709,0.954]     then Happiness.Score.1 = high

Accuracy:
157 of 220 instances classified correctly (71.36%)
```

Figure21. OneR rules using (Logistic Regression method)



```
Call:
OneR.formula(formula = Happiness.Score.1 ~ ., data = data, verbose = TRUE)

Rules:
If Health..Life.Expectancy. = (-0.000953,0.459] then Happiness.Score.1 = low
If Health..Life.Expectancy. = (0.459,0.671]     then Happiness.Score.1 = medium
If Health..Life.Expectancy. = (0.671,0.954]     then Happiness.Score.1 = high

Accuracy:
151 of 220 instances classified correctly (68.64%)
```

Figure22. OneR rules using (information gain method)

These rules proved by applying the same training data using the decision tree approach to find out if it gives the same derived rules from the models above, the decision tree on the training dataset is shown in figure 23.

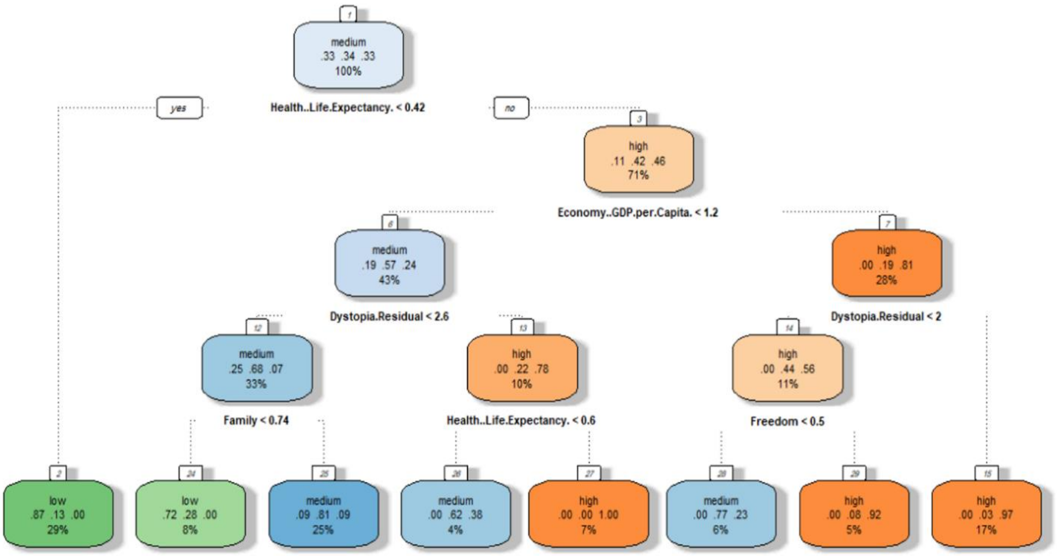


Figure23. Decision tree approach

The figure 23 indicates that the health life expectancy attribute used as the first rule while producing the fit model thus leads to the answer of the second research question which discovering the insights by using the OneR method to identify the health life expectancy as one of the critical attributes leading to high life happiness score.

The table 6. Summarize the gained insight after applying the three models:

Table 6. gained insights after applying models: nn, rf and xgb

Model	NN	RF	xgb
Economy .GDP per. Capita	100%	100%	100%
Health. life. Expectancy	96.1%	79.11%	61.93%
Family	75.35%	24.47%	43.50%
Freedom	70.69%	35.40%	31.44%
Destopia. Residual	48.38%	67.15%	92.49%
Trust. Government. Corruption	25.61%	3.22%	0.00%
Generosity	4.61%	0.00%	20.41%

The figure 24 shows the summary for importance of the happiness variables, where its too clear that the most importance variables are*: economy. Gdp.per.capital* and *health.life.expectancy*

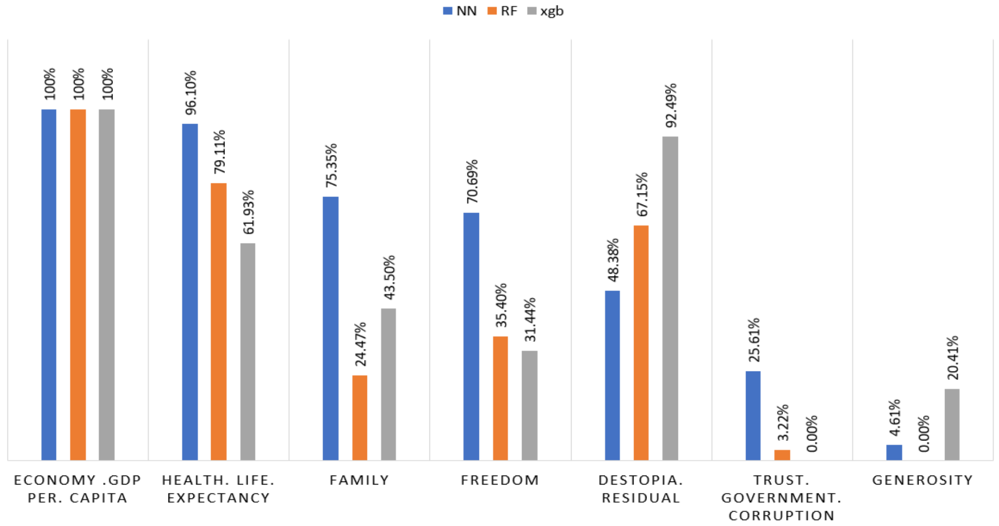


Figure24. summary for importance of the happiness variables

7 conclusion and future work

This research paper identified vital issues across used data starting with the importance of variables included in the dataset. The used machine learning approaches classified the GDP per Capita as the most crucial feature affecting the life happiness score and the health life expectancy the second. Due to the outcomes from the Neural Network approach its recognised the GDP is one of the primary indicators of the life happiness score, furthermore the findings evaluated by using different approaches to increase the accuracy of the conducted work. The second gained insight from the study that is the high life expectancy might lead into a good life happiness score by classified it as the first rule while using the OneR classification method and its outcome evaluated by different performance metrics which increased the findings strengthen. The future work of this research is to apply more machine learning approaches to more dataset and covering longer range of years, in addition its planned to apply deep learning methods to explore more insights of the happiness drivers.

Acknowledgement

The authors are grateful to the Applied Science University, Bahrain for financial support granted to this research paper.

References

[1] J. HELLIWELL, R. Layard and J. Sachs, "WORLD HAPPINESS REPORT 2016," Sustainable Development Solutions Network., New York, 2016.

[2] K. Ruggeri, E. Garcia-Garzon, Á. Maguire, S. Matz and F. A. Huppert, "Well-being is more than happiness and life satisfaction: a multidimensional analysis of 21 countries," Health and Quality of Life Outcomes, vol. 18, no. 192, 2020.

[3] E. Sandvic, E. Diener and L. Seidlitz, "Subjective well-being: The convergence and stability of self-report and non-self-report measures," Assessing WellBeing, vol. 39, no. Social Indicators Research Series, pp. 119-138, 2009.

[4] L. Millard, "Data Mining and Analysis of Global Happiness: A Machine Learning Approach," University of Bristol:DEPARTMENT OF COMPUTER SCIENCE, Bristol, 2011.

[5] E. Bullen, "Ed Bullen," 12 8 2016. [Online]. Available: http://rstudio-pubs-static.s3.amazonaws.com/201826_cab699be72ca47f99debadf16ee54c95.html. [Accessed 20 09 2021].

[6] J. F. Helliwell, H. Huang and S. Wang, "Statistical Appendix for \The social foundations of world happiness," Sustainable Development Solutions Network, Newyork, 2017.

[7] Farooq, S. A., & Shanmugam, S. K. . A Performance Analysis of Supervised Machine Learning Techniques for COVID-19 and Happiness Report Dataset. In Sentimental Analysis and Deep Learning (pp. 591-601). Springer, Singapore. 2022

[8] Jannani, A., Sael, N., & Benabbou, F., Predicting Quality of Life using Machine Learning: case of World Happiness Index. In 2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT) (pp. 1-6). IEEE. 2021

[9] Shreemali, J., Chakrabarti, P., Chakrabarti, T., Poddar, S., Sipple, D., Kateb, B., & Nami, M., A Machine Learning Perspective on Causes of Suicides and identification of Vulnerable Categories using Multiple Algorithms. medRxiv. 2021

[10] Khemraj, S., Thepa, P. C. A., Chi, H., Wu, W. Y., Samanta, S., & Prakash, J., Prediction of world happiness scenario effective in the period of COVID-19 pandemic, by artificial neuron network (ANN), support vector machine (SVM), and regression tree (RT). NVEO-NATURAL VOLATILES & ESSENTIAL OILS Journal| NVEO, 13944-13959. 2021

[11] Ulkhaq, M. M., & Adyatama, A., Clustering countries according to the world happiness report 2019. Engineering and Applied Science Research, 48(2), 137-150. 2021

[12] Dixit, S., Chaudhary, M., & Sahni, N., Network Learning Approaches to study World Happiness. arXiv preprint arXiv:2007.09181. 2021

[13] Tan, Y., Singhapreecha, C., & Yamaka, W., Applying Machine Learning to Predict Happiness: A case study of 20 Countries. Mukht Shabd Journal, 9(6), 3433-3437. 2020

[14] Moore, L. , Exploring trends and factors in the world happiness report. 2020

[15] Ulkhaq, M. M. , Clustering Countries According to The World Happiness Report. Statistica & Applicazioni, 18(2). 2020

[16] Riyantoko, P. A., Southeast Asia Happiness Report in 2020 Using Exploratory Data Analysis. IJCONSIST JOURNALS, 2(1), 16-21. 2020

[17] Prashanthi, B., & Ponnusamy, R., Future Prediction of World Countries Emotions Status to Understand Economic Status using Happiness Index and SVM Kernel. Future, 6(11). 2019

[18] Kaur, M., Dhalaria, M., Sharma, P. K., & Park, J. H., Supervised machine-learning predictive analytics for national quality of life scoring. Applied Sciences, 9(8), 1613. 2019

[19] Bond, R. R., Zhang, S., & Marshall, F. (2018, June). Measuring and Visualising Global Happiness. In British HCI Conference 2018.

[20] Du Ni, M. K., Xiao, Z., & Feng, X., Exploring Subjective Well-Being Factors with Support Vector Machine. International Journal of Engineering & Technology, 7(3.36), 164-166. 2018

[21] Saputri, T. R. D., & Lee, S. W., A study of cross-national differences in Happiness factors using machine learning approach. International Journal of Software Engineering and Knowledge Engineering, 25(09n10), 1699-1702. 2015

[22] "World Happiness Report 2015-2021," Kaggle, March 2021. [Online]. Available: https://www.kaggle.com/mathurinache/world-happiness-report-20152021. [Accessed 5 10 2021].

[23] Moscato, P. and Jane, N, ‘Memetic algorithms for business analytics and data science: a brief survey’, in Business and Consumer Analytics: New Ideas, Springer International Publishing. 2019

[24] Khder, M. A., Fujo, S. W., Sayfi, M. A. (2021). A roadmap to data science: background, future, and trends. International

[25] Journal of Intelligent Information and Database Systems.14(3), 277-293, 2021.

[26] Abazeed, A., & Khder, M. (2017). A Classification and Prediction Model for Student's Performance in University Level. J. Comput. Sci., 13(7), 228-233.

[27] G. Biau, E. Scornet and J. Welbl, "Neural Random Forests," arXiv.org, p. arXiv:1604.07143v2 , 2018.

[28] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," Machine Learning, vol. 11, no. 1, pp. 63-91, 1993.

[29] H. v. Jouanne-Diedrich, "OneR - Establishing a New Baseline for Machine

Learning Classification Models"," 2017. [Online]. Available: https://rdrr.io/cran/OneR/f/vignettes/OneR.Rmd. [Accessed 4 10 2021].

Notes on contributors

Moaiad Ahmad Khder is (Senior Member, IEEE) received the Ph.D. degree from the Faculty of Information Science and Technology, The National University of Malaysia, in 2015. He is currently an Assistant Professor with the Computer Science Department, Applied Science University, Bahrain. He has been working on the area of mobile environment, mobile database, data science, big data and cloud computing.

Mohammad Adnan Sayfi Mohammad Adnan Sayfi received his Master’s degree in Data Science on 2019 from University of Sunderland, UK. He is currently working as data engineer at Uminiah Jordan.

Samah Wael Fujo, received her Master in Computer Science and information technology from College of Information Technology, Ahlia University- Bahrain in 2021, and her Bachelor degree in Computer Science from Applied Science University – Bahrain in 2019. She is currently working in Nasser Artificial Intelligence Research & Development Centre (NAIRDC) - Nasser Vocational Training Centre (NVTC) - Bahrain. She has been working on the area of data science, machine and deep learning.