HAPPINESS INDEX

DATA MINING PROJECT

PALASH RASTOGI

101213037

VARTUL MITTAL

101213055

**HAPPINESS INDEX**

**INTRODUCTION**

The assessment of **gross national happiness** was designed in an attempt to define an indicator and concept that measures quality of life or social progress in more holistic and psychological terms than only the economic indicator of gross domestic product (GDP).

This project aims at using the various measures used to estimate the National Happiness Index of a country in such a way that the data can help us to generate the required conditions which can be directly used in the future to predict the happiness index of a region. The generated conditions can be directly used on the data collected through a survey conducted in a particular region to predict how happy (psychologically and socially satisfied) the people are in that region.

**GROSS NATIONAL HAPPINESS**

GNH has only been officially used in Bhutan, where a Gross National Happiness Commission is charged with reviewing policy decisions and allocation of resources. In 2013, with a new administration, the country shifted the focus from spreading GNH globally to the well-being of people within Bhutan. This shift has been interpreted by some as an abandonment of GNH in favor of more standard development initiatives.

Although there were few adhoc and independent surveys that attempted to measure the happiness or life satisfaction as a subjective score, up to 2005 there was no exact quantitative definition of GNH, but elements that contribute to GNH are subject to quantitative measurement. Low rates of infant mortality, for instance, correlate positively with subjective expressions of well-being or happiness within a country. The practice of social science has long been directed toward transforming subjective expression of large numbers of people into meaningful quantitative data; there is no major difference between asking people "how confident are you in the economy?" and "how satisfied are you with your job?"

In 2005, a second-generation GNH concept (GNH 2.0), treating happiness as a socioeconomic development metric (GNH Index), was proposed by Med Jones, the President of International Institute of Management. The metric measures socioeconomic development by tracking seven development areas including the nation's mental and emotional health. GNH value is proposed to be an index function of the total average per capita of the following measures:

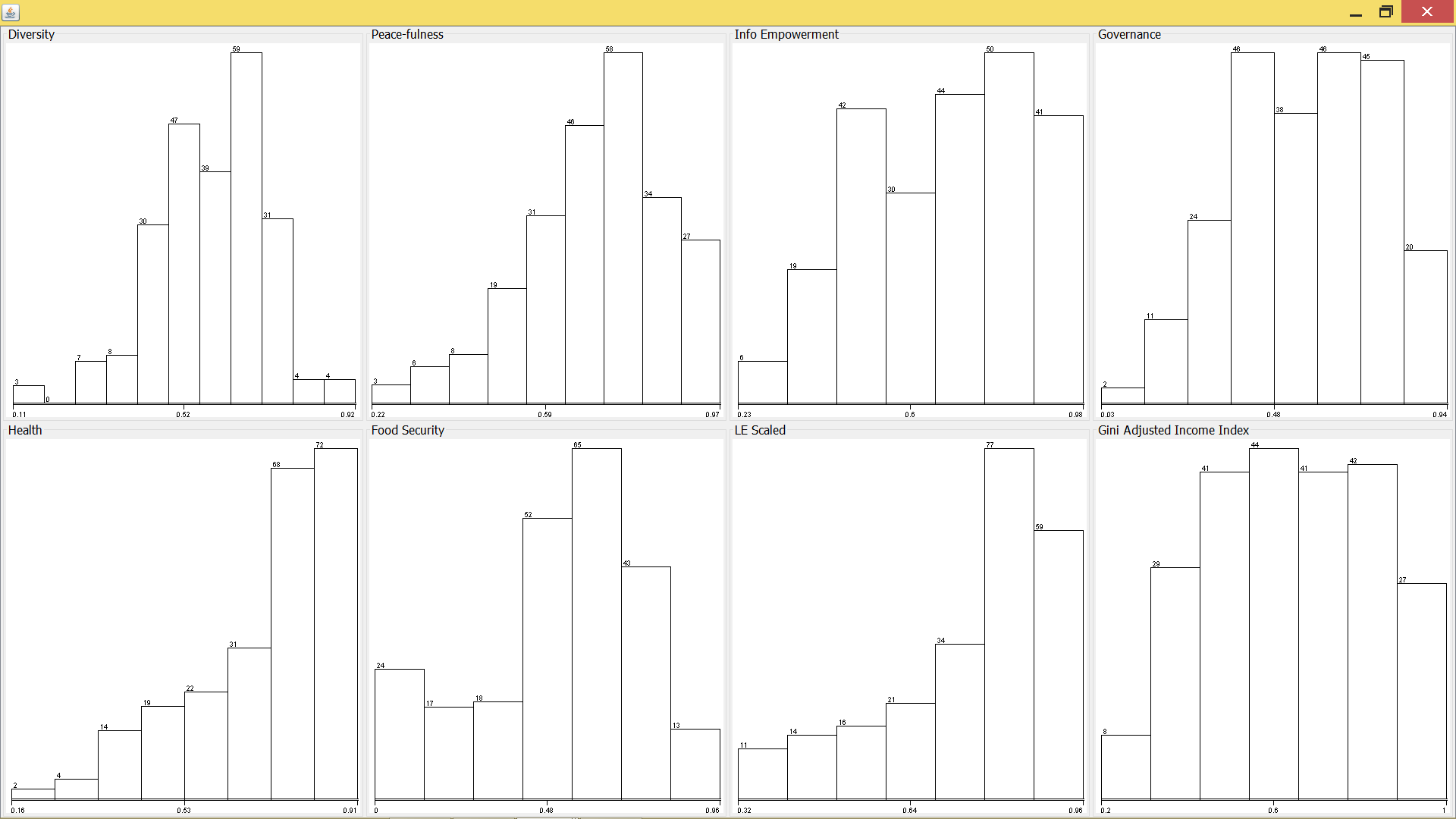
1. Economic Wellness: Indicated via direct survey and statistical measurement of economic metrics such as consumer debt, average income to consumer price index ratio and income distribution
2. Environmental Wellness: Indicated via direct survey and statistical measurement of environmental metrics such as pollution, noise and traffic
3. Physical Wellness: Indicated via statistical measurement of physical health metrics such as disabilities, obesity, severe and chronic illnesses
4. Mental Wellness: Indicated via direct survey and statistical measurement of mental health metrics such as usage of antidepressants and rise or decline of the number of psychotherapy patients
5. Workplace Wellness: Indicated via direct survey and statistical measurement of labor metrics such as jobless claims, job change, workplace complaints and lawsuits
6. Social Wellness: Indicated via direct survey and statistical measurement of social metrics such as discrimination, safety, divorce rates, complaints of domestic conflicts and family lawsuits, public lawsuits, and crime rates
7. Political Wellness: Indicated via direct survey and statistical measurement of political metrics such as the quality of local democracy, individual freedom, domestic and foreign conflicts.

The above seven metrics were incorporated into the first Global GNH Survey.

**ATTRIBUTES**

The following input attributes of 232 countries have been used for generating the conditions favourable for a high happiness index in a region:

* Governance
* Life Expectancy
* Gini adjusted Income Index
* Health
* Food Security
* Peacefulness
* Information Empowerment
* Diversity



**MINING TECHNIQUE**

The present dataset does not contains any labelled data, hence, we need to apply unsupervised learning. Clustering is the needed technique that can be applied on our data.

Clustering can be defined as partitioning of data into subsets or clusters, such that, data in each subset share some common trait. Our aim is to divide the available data into different clusters having maximum intra-cluster similarity and inter-cluster dissimilarity, which can be easily achieved using cluster analysis.

After the data has been clustered into different groups, we can apply Classification (Supervised learning) on the clustered data so as to generate the necessary conditions, required for labelling any unlabelled dataset.

1. **K-MEANS CLUSTERING ALGORITHM**

There are two main attributes which we need to specify before applying clustering:

* numClusters : to specify the number of clusters we want to divide the data into
* displayStdDevs : it is set to true so that we can see standard deviations for the instance cluster assignments

The number of clusters the data should be divided into can be determined by changing the value of the attribute ‘numCluster’ and checking ‘Within cluster sum of squared errors’ as we go. The clusters are incremented until the diminishing returns are achieved. Following are the values of the clusters and the corresponding error values:

|  |  |
| --- | --- |
| Cluster | Error |
| 1 | 140.9 |
| 2 | 44.7 |
| 3 | 31.2 |
| 4 | 27.7 |
| 5 | 25.39 |

With diminishing returns setting in at about fifth cluster, four clusters appear to be good choice for clustering of the data.

Number of iterations: 17

Within cluster sum of squared errors: 27.803451274561795

Missing values globally replaced with mean/mode

Cluster centroids:

Cluster#

Attribute Full Data 0 1 2 3

(232) (67) (66) (43) (56)

==================================================================

Diversity 0.5851 0.6815 0.5854 0.4699 0.5577

+/-0.1286 +/-0.0959 +/-0.0931 +/-0.1434 +/-0.0999

Peace-fulness 0.6977 0.8279 0.7118 0.6058 0.5958

+/-0.1489 +/-0.0761 +/-0.1251 +/-0.1281 +/-0.1308

Info Empowerment 0.6849 0.8763 0.7373 0.4504 0.5743

+/-0.1777 +/-0.0563 +/-0.0878 +/-0.0862 +/-0.1034

Governance 0.5637 0.7637 0.6163 0.3618 0.4175

+/-0.1965 +/-0.1017 +/-0.116 +/-0.1422 +/-0.1129

Health 0.7076 0.8562 0.7639 0.4399 0.6691

+/-0.1609 +/-0.0339 +/-0.059 +/-0.1019 +/-0.0843

Food Security 0.516 0.7314 0.6041 0.1532 0.4329

+/-0.2391 +/-0.1046 +/-0.1089 +/-0.1512 +/-0.1405

LE Scaled 0.7573 0.8995 0.8026 0.4944 0.7356

+/-0.1583 +/-0.044 +/-0.0705 +/-0.0962 +/-0.0918

Gini Adjusted Income Index 0.6363 0.8608 0.6493 0.3784 0.5505

+/-0.1916 +/-0.0696 +/-0.1094 +/-0.0827 +/-0.089

Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

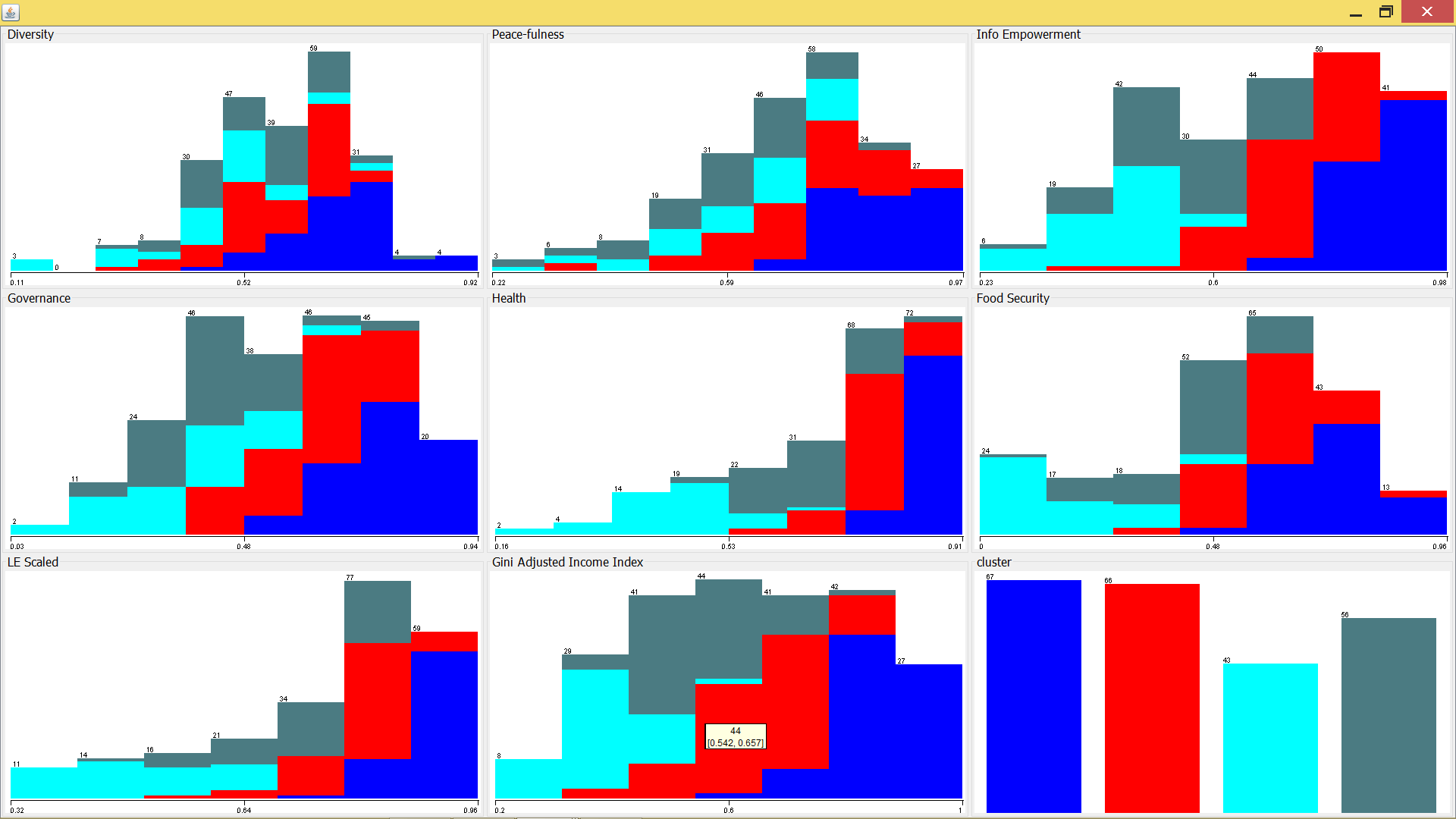
0 67 (29%)

1 66 (28%)

2 43 (19%)

3 56 (24%)

After adding a ‘cluster’ CLASS in the dataset and applying K-Means algorithm, the dataset is automatically updated with the output attribute ‘Cluster’, specifying the cluster in which the given instance is present.



Now, applying clustering with ‘classes to cluster option gives the following result:

Classes to Clusters:

0 1 2 3 <-- assigned to cluster

67 0 0 0 | cluster1

0 66 0 0 | cluster2

0 0 43 0 | cluster3

0 0 0 56 | cluster4

Cluster 0 <-- cluster1

Cluster 1 <-- cluster2

Cluster 2 <-- cluster3

Cluster 3 <-- cluster4

Incorrectly clustered instances : 0.0 0 %

1. **J-48 CLASSIFICATION**

On applying J48 classification on the above dataset, we get the following results:

Correctly Classified Instances 196 84.4828 %

Incorrectly Classified Instances 36 15.5172 %

=== Confusion Matrix ===

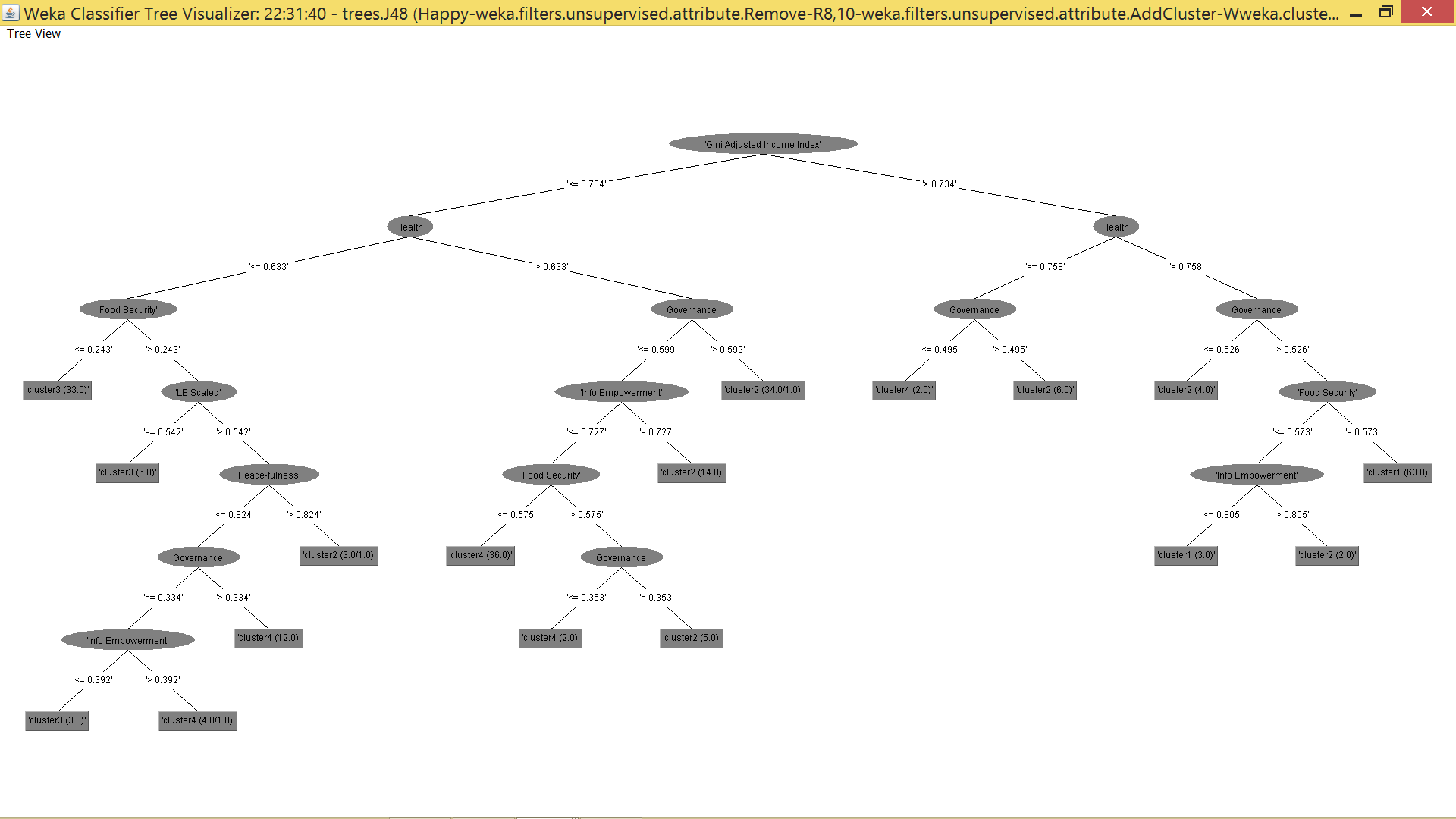
a b c d <-- classified as

61 6 0 0 | a = cluster1

7 55 0 4 | b = cluster2

0 0 37 6 | c = cluster3

1 9 3 43 | d = cluster4



On pruning, we get

J48 pruned tree

------------------

Gini Adjusted Income Index <= 0.734

| Health <= 0.633

| | Food Security <= 0.243: cluster3 (33.0)

| | Food Security > 0.243

| | | LE Scaled <= 0.542: cluster3 (6.0)

| | | LE Scaled > 0.542: cluster4 (22.0/6.0)

| Health > 0.633

| | Governance <= 0.599

| | | Info Empowerment <= 0.727

| | | | Food Security <= 0.575: cluster4 (36.0)

| | | | Food Security > 0.575: cluster2 (7.0/2.0)

| | | Info Empowerment > 0.727: cluster2 (14.0)

| | Governance > 0.599: cluster2 (34.0/1.0)

Gini Adjusted Income Index > 0.734

| Health <= 0.758: cluster2 (8.0/2.0)

| Health > 0.758

| | Peace-fulness <= 0.722: cluster2 (8.0/3.0)

| | Peace-fulness > 0.722: cluster1 (64.0/1.0)

Number of Leaves : 10

Size of the tree : 19

=== Stratified cross-validation ===

Correctly Classified Instances 191 82.3276 %

Incorrectly Classified Instances 41 17.6724 %

=== Confusion Matrix ===

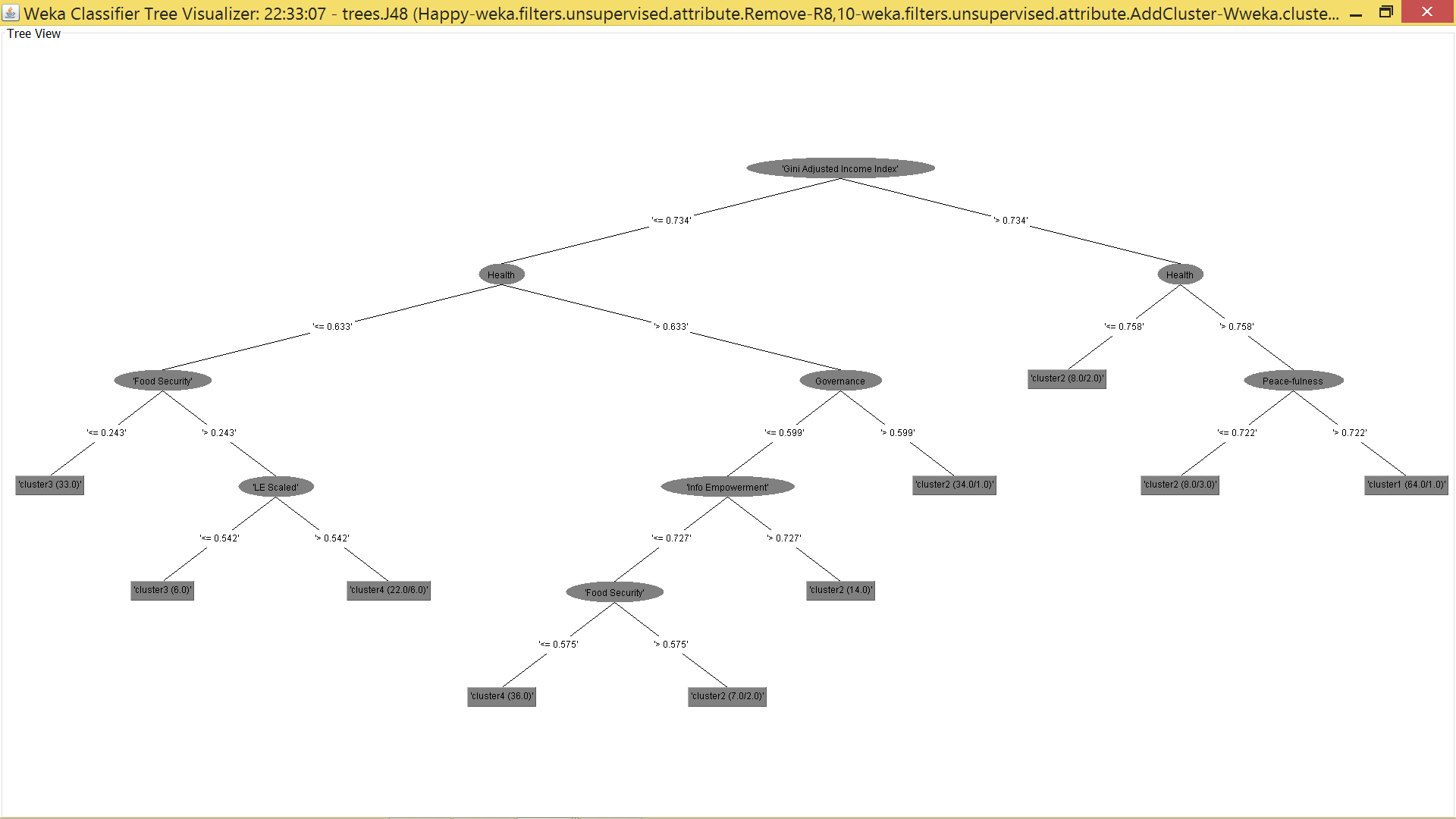
a b c d <-- classified as

58 9 0 0 | a = cluster1

8 54 1 3 | b = cluster2

0 0 36 7 | c = cluster3

1 10 2 43 | d = cluster4

****

Since there is not much increase in the inaccuracy, we can use these conditions created after the pruning (increasing the minNumObj attribute) for classifying any unlabelled dataset in the future.

**CONCLUSION**

* The main attributes deciding the happiness index of any country are the ‘Gini adjusted Income index’ and the ‘Health’ status
* A country having high values for the above two mentioned attributes, along with high level of ‘Peaceful’ experiences have been clustered as the most happy countries
* The different clusters have the following representations:
  + Cluster 1: Happy and highly satisfied
  + Cluster 2: moderate level of happiness and satisfaction
  + Cluster 4: Less happy and satisfaction level
  + Cluster 3: Rarely happy