Classifying Countries based on World Happiness Index data using Principal Component Analysis & Machine Learning

Mian Basit Mahmood (40185188)

https://github.com/Basit1994/Basit_INSE-6220-Proj

***Abstract*—Principal Component Analysis (PCA) is a technique for reducing data dimensionality and converting correlated to uncorrelated properties. The goal is to conduct a statistical study of data from the World Happiness Report and how it relates to nations' economic income, Gross Domestic Product (GDP), as well as to discover additional variables that impact happiness. The method employs research methods and Machine Learning techniques to create models that describe the function used to forecast a country's happiness score. Two different machine learning methods will be tested and compared.**

Keywords — Principal Component Analysis (PCA), Classification, Dimensional Reduction, World Happiness Index features, Machine Learning.

I. INTRODUCTION

Happiness, typically corresponds to love, a feeling of happiness, an experience of delight, amazement, mixed with the belief that one's life is significant and good, is becoming a more popular subject in cross-national research. Now being recognised as an acceptable metric to evaluate social progress and achieving public policy objectives. It has been studied from several angles throughout the years, including links to wellness and positive psychological factors. It also has a connection to "the economics of happiness," which reports empirical relationships between happiness and other factors.

Wealth Economics was the first to investigate happiness from an economic standpoint after WWII. Prior to that, economics was concerned with economic growth, development, poverty alleviation, and inequality reduction. A unique interest was created at that time, centred around the quality of life. This resulted in development of many questionnaires to track people's quality of life.

Several contemporary approaches have relied on mathematical modelling with tools that measure happiness or a particular component with high internal consistency. It appears to be an essential causative component; therefore, one must acknowledge that it impacts many facets of pleasure, despite the superficial assumption that economic indicators bring enjoyment.

As data grows more intricate with an increasing number of feature vectors, time and effort have become major problems for academics when it comes to data analysis. Science has demonstrated that data with high dimensionality may be condensed and clarified by focusing on some of the important aspects and excluding the insignificant or non-contributing ones. For dimensionality reduction, this study will employ approaches such as feature selection and feature extraction. Feature selection is the procedure of picking a subset of suitable characteristics or variables from a larger dataset. While separating features is a process of reducing the dimensionality of the initial data set to ease processing. The type of the dataset, as well as the machine learning approach to be utilised, dictate the dimensionality reduction strategy.

In this report, Principal Component Analysis (PCA) technique will be used to reduce the dimensionality of World Happiness Index Dataset to distinguish between Happy countries and Not Happy countries. This separation is based on classification of data by Happiness score. Happiness score above ‘5’ is considered Happy, while below ‘5’ score is considered Not Happy. Moreover, two machine learning techniques will be applied on the same dataset to be assessed and compared. The classification techniques are Logistic Regression (LR), and Gaussian Naïve Bayes (GNB).

II. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is a multivariable approach that analyses a dataset with associated quantitative variables that represent measurements. PCA is used to minimise the dimensionality of data while retaining as much information as feasible. PCA reduces the complexity of big data sets by reducing many variables into a lesser number that retains most of the original dataset's information.

The first principal component is the combination of the variable with the greatest variation (when compared to all other combinations). It accommodates for as much data volatility as feasible. While the second principal component is the combination of variables which accounts for the majority of the variation that remains as possible, the correlation between the first and second components must be zero. This technique minimises the complexity of large amounts of data while keeping patterns. This is accomplished by compressing the data into fewer dimensions that function as feature summaries.

Given a data matrix X, the PCA algorithm consists of four main steps:

Step 1: Compute the centered data matrix Y = HX by subtracting off-column means.

Step 2: Compute the p × p covariance matrix S of the centered data matrix as follows:

$$S = (1/n - 1)\,(Y'Y)$$

Step 3: Compute the eigenvectors and eigenvalues of S using eigen-decomposition, Where:

$$S = A\Lambda A' = \sum_{j=1}^{p} \lambda_j a'_j a_j$$

Step 4: Compute the transformed data matrix Z = YA of size n × p

III. Classification Algortihms

In this report, two classifiers will be applied and eventually compared in terms of their accuracy and time. The classifiers to be used are Ridge regression and Agnes.

Ridge regression is an algorithm optimisation technique used to examine data with convergence. L2 regularisation is performed via this approach. When there is a problem with convergence, least-squares are unbiased, and variances are enormous, resulting in projected values that are distant from the actual values. It works in the following manner for the binary classification problems by making use of Ridge regression algorithm:

a. Transforms the target variable to +1 and -1 as needed.

b. Develop a Ridge model with a mean square loss loss function and L2 regularisation (ridge) as the penalty term.

c. If the expected value is less than zero, the predicted class label is -1; otherwise, the predicted class label is +1.

The first step in ridge regression is to standardise the variables (both dependent and independent) by removing their means and dividing them by their normal deviations. This is a notational problem since we must declare whether or not the variables in a specific formula are standardised.

When it comes to creating ridge regression models using actual datasets, the trade-off between bias and variance is often tricky. However, following the general trend which one needs to remember is the bias increases as λ increases and the variance decreases as λ increases.

AGNES falls within the category of hierarchical clustering. It begins by treating each observation as a separate cluster. Pairs of clusters are mixed successively until all clusters are merged into one "big cluster" containing all observations. This approach produces a dendrogram, which is a tree-based representation of the whole set of data. The method takes into account the (dis)similarity of each pair of observations in the full data set.

The linkage function is then used to integrate data that are near together to generate the dendrogram. The end point of the tree's hierarchy should be defined if clusters are to be created.

IV. Data set Description

World Happiness Index data was chosen to explore the dimensionality reduction abilities of PCA. The dataset consists of 156 subjects. The number of the subjects were grouped as 'Happy' is 99 throughout the study and 57 for 'Not Happy'.

The Box-plots shown in Fig. were constructed after getting the data centered. As we can see, there are no outliers in pelvic radius, however, all other variables have at least one outlier. Overall, the data can be considered normally distributed, thus; the data is ready for analysis.

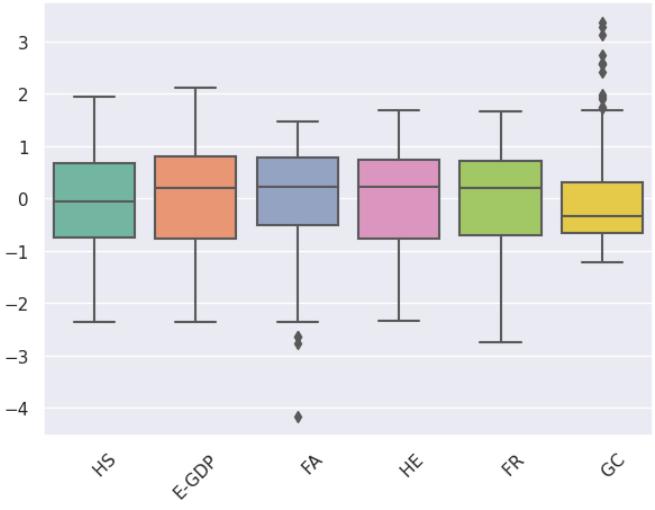

Fig 1. Box plots

The correlation between parameters can be shown by the covariance matrix in Fig.2. There is a high correlation between Happiness Score (HS) and Economic GDP (E-GDP). In addition, we see a positive correlation between degree Economic GDP (E-GDP) and Health (HE). It is also clear that there is a negative correlation between Economic GDP (E-GDP) and all other five variables.

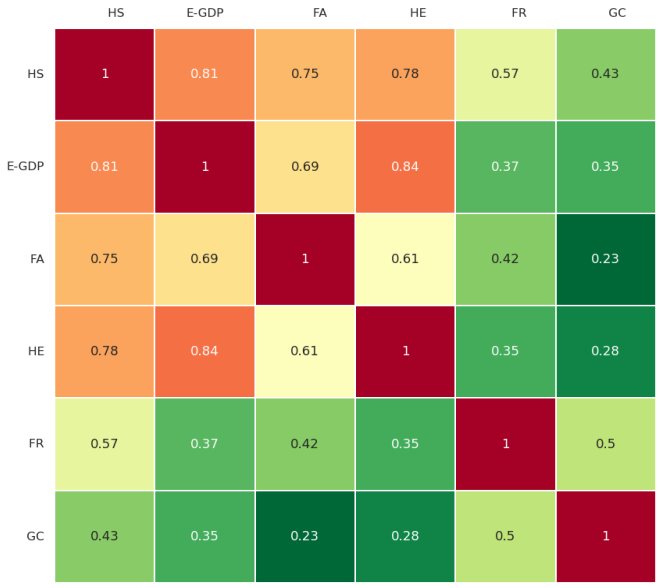

Fig 2. Covariance Matrix

When interpreting correlations, it is important to visualize the relationships between all variables. The Pairplot is shown in Fig. and the relationship is present between all the six variables. we see a positive correlation between degree Economic GDP (E-GDP) and Health (HE). It is also clear that there is a negative correlation between Economic GDP (E-GDP) and all other five variables.

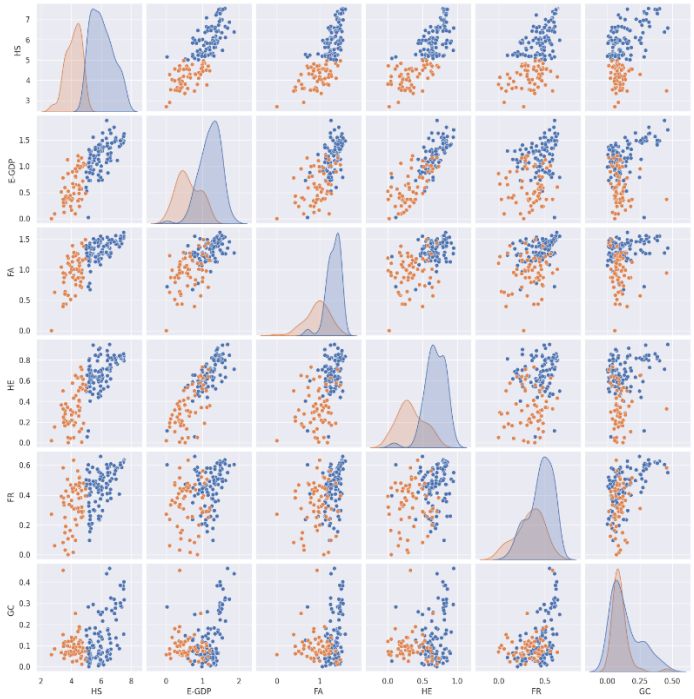

Fig 3. Pair plot

In the next step, we are going to use PCA to reduce World Happiness Index features data dimensionality. It is always advisable to reduce the data into two or three dimensions. The reason behind it is to ease the visualization of the components. In our case, the Dataset originally has six dimensions, by using PCA, we will reduce dimensions into 2 or 3 dimensions only.

In order to know the optimal dimensions, the data should be reduced to, we use Scree plot (Fig. 4) and Pareto Diagram (Fig. 5). In Fig. 4, the graph has an elbow shape at (r = 2) where the explained variance percentage is almost equal to 80%.

Similarly, in Fig. 5, the Pareto Diagram also shows that the first two components have a cumulative explained variance value of approximately 79%. From the two graphs, we can conclude that the optimal number of dimensions to be reduced for the World Happiness Index Dataset is (r = 2).

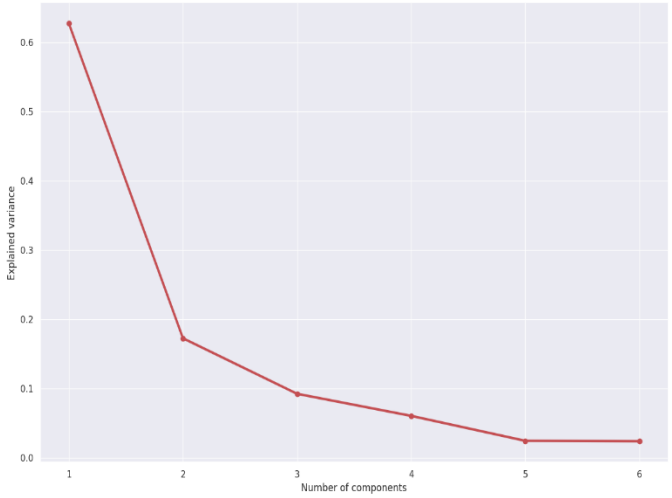

Fig 4. Scree Plot

In other words, two out of six components contribute the most to our data, and dismissing the other two will not significantly affect the data analysis but will simplify it to make it easier to study and visualize.

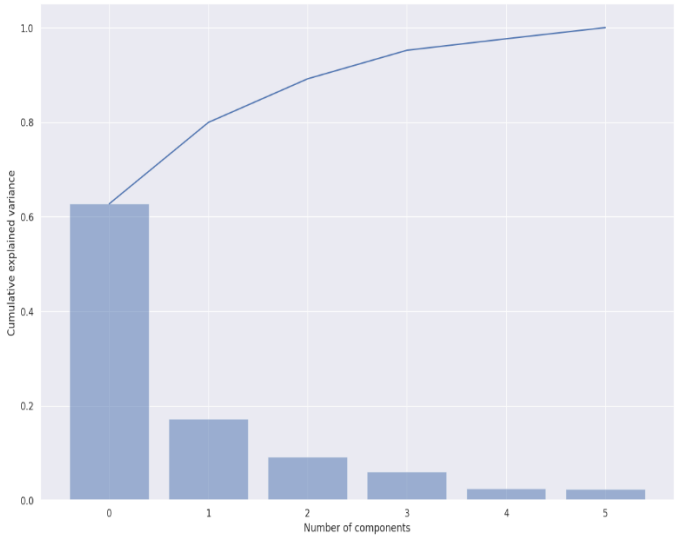

Fig 5. Pareto Diagram

The eigenvectors matrix is given by:

−0.48	−0.05	−0.06	−0.01	−0.80	0.33
−0.46	−0.25	0.26	−0.10	0.52	0.60
−0.41	−0.22	−0.40	0.71	0.17	−0.26
−0.44	−0.30	0.27	−0.44	−0.003	−0.66
−0.33	0.55	−0.60	−0.41	0.21	−0.01
−0.27	0.69	0.56	0.32	0.01	−0.11

And the first two Principal Components are:

Eq 1:

Z1 = -0.48 X1 - 0.46 X2 - 0.41 X3 - 0.44 X4 - 0.33 X5 - 0.27 X6

Eq 2:

Z2 = -0.25 X1 - 0.22 X2 - 0.30 X3 + 0.55 X4 + 0.69 X5

In the first Principal Component, we can see how $x1$, $x2$, x3 and x4 have a very similar contribution. For the second Principal Component, $x5$ has the highest contribution while one component (-0.05) have minimal or no contribution and thus; neglected. The following vector indicates eigenvalues:

3.78

1.04

0.55

0.36

0.14

0.143

As we can see in the Scatter Plot below in Fig. 6 shows that the data for Government Corruption (GC) located top right of the plot is exceptionally different from the others. It is also illustrated which features have similar contributions within the same component; Like Economic GDP (E-GDP), Happiness Expression (HE) and Family (FA), they are responsible by about 72% for the first principal component in PC1. Happiness Score (HS) and Family (FA) also have the most involvement for the second component in PC2.

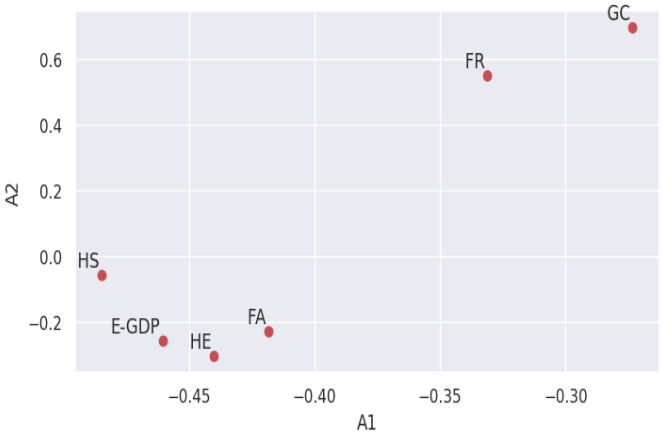

Fig 6. Scatter plot

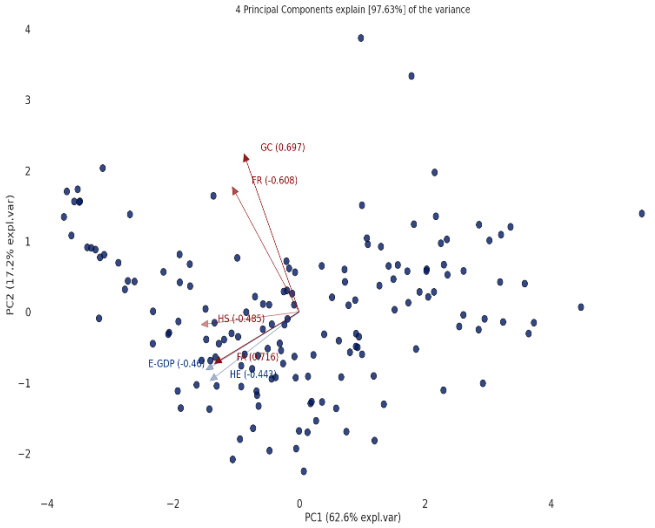

Fig 7. Biplot

A biplot is a graphical representation of the information of the variables. In Fig. 7, the Bi-plot has been constructed showing both Principal Components coefficients for each variable and scores for each measurement. The horizontal and vertical axes in the graph are the Principal Components and the variables are the six vectors presented in the graph. While the points present the measurements. It is clear from the biplot that the vectors that tend to point in the same direction have a positive correlation.

The first component has a positive coefficient for Economics GDP (E-GDP), Family (FA), and Happiness Expression (HE), with almost the same value as we noticed that in Eq.(1) and they are having narrow angles between them because they are almost equally in PC1. While Government Corruption (GC) has the longest arrow as it is the most contributed variable in the second component. Moreover, the points in the Bi-plot represent the 159 countries, where the location of the points represents the score of each observation for both PC1 and PC2.

Another interpretation that can be concluded from the Biplot is that Economics GDP (E-GDP), Family (FA), and Happiness Expression (HE) more than Government Corruption (GC).

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
dt	Decision Tree Classifier	0.9900	0.9875	1.0000	0.9857	0.9923	0.9783	0.9802	0.0830
ada	Ada Boost Classifier	0.9900	0.9875	1.0000	0.9857	0.9923	0.9783	0.9802	0.0750
gbc	Gradient Boosting Classifier	0.9900	0.9875	1.0000	0.9857	0.9923	0.9783	0.9802	0.2750
xgboost	Extreme Gradient Boosting	0.9900	1.0000	1.0000	0.9857	0.9923	0.9783	0.9802	0.0970
lr	Logistic Regression	0.9689	1.0000	0.9833	0.9714	0.9755	0.9334	0.9394	0.7920
rf	Random Forest Classifier	0.9678	0.9917	0.9667	0.9857	0.9723	0.9354	0.9434	0.4120
knn	K Neighbors Classifier	0.9589	1.0000	0.9833	0.9607	0.9689	0.9097	0.9205	0.2640
lightgbm	Light Gradient Boosting Machine	0.9578	1.0000	0.9667	0.9750	0.9657	0.9117	0.9245	0.3440
et	Extra Trees Classifier	0.9478	0.9917	0.9500	0.9690	0.9556	0.8937	0.9018	0.4680
svm	SVM - Linear Kernel	0.9289	0.0000	0.9690	0.9315	0.9470	0.8376	0.8517	0.0560
ridge	Ridge Classifier	0.9267	0.0000	0.9190	0.9690	0.9346	0.8549	0.8687	0.0570
lda	Linear Discriminant Analysis	0.9167	0.9917	0.9190	0.9571	0.9282	0.8313	0.8460	0.1350
nb	Naive Bayes	0.9156	0.9764	0.9000	0.9732	0.9232	0.8304	0.8523	0.2120
qda	Quadratic Discriminant Analysis	0.8956	0.9778	0.9000	0.9407	0.9083	0.7897	0.8113	0.0720
dummy	Dummy Classifier	0.6433	0.5000	1.0000	0.6433	0.7821	0.0000	0.0000	0.1570

Fig 8. Comparison among classification models before applying PCA

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ridge	Ridge Classifier	0.9455	0.0000	0.9571	0.9583	0.9559	0.8839	0.8897	0.0680
lda	Linear Discriminant Analysis	0.9455	0.9643	0.9571	0.9583	0.9559	0.8839	0.8897	0.0900
nb	Naive Bayes	0.9364	0.9571	0.9571	0.9486	0.9501	0.8607	0.8710	0.0850
gbc	Gradient Boosting Classifier	0.9364	0.9625	0.9429	0.9583	0.9492	0.8632	0.8662	0.2490
xgboost	Extreme Gradient Boosting	0.9355	0.9625	0.9571	0.9440	0.9482	0.8622	0.8699	0.1150
lr	Logistic Regression	0.9273	0.9714	0.9286	0.9583	0.9363	0.8518	0.8640	0.5180
dt	Decision Tree Classifier	0.9273	0.9214	0.9429	0.9500	0.9423	0.8421	0.8524	0.0840
rf	Random Forest Classifier	0.9264	0.9679	0.9262	0.9583	0.9378	0.8471	0.8575	0.4000
qda	Quadratic Discriminant Analysis	0.9264	0.9286	0.9405	0.9486	0.9410	0.8407	0.8527	0.1490
knn	K Neighbors Classifier	0.9173	0.9304	0.9119	0.9583	0.9272	0.8318	0.8457	0.0900
et	Extra Trees Classifier	0.9173	0.9679	0.9405	0.9389	0.9351	0.8174	0.8340	0.3710
lightgbm	Light Gradient Boosting Machine	0.9173	0.9679	0.9119	0.9583	0.9272	0.8318	0.8457	0.3460
ada	Ada Boost Classifier	0.8800	0.9565	0.8976	0.9167	0.8997	0.7479	0.7609	0.3760
svm	SVM - Linear Kernel	0.8627	0.0000	0.8976	0.8911	0.8887	0.7053	0.7205	0.0650
dummy	Dummy Classifier	0.6291	0.5000	1.0000	0.6291	0.7722	0.0000	0.0000	0.1350

Fig 9. Comparison among classification models after applying PCA

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	0.9091	0.8750	1.0000	0.8750	0.9333	0.7925	0.8101
2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	0.9091	0.9286	0.8571	1.0000	0.9231	0.8136	0.8281
4	0.7273	0.7321	0.7143	0.8333	0.7692	0.4407	0.4485
5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	0.9091	0.8750	1.0000	0.8750	0.9333	0.7925	0.8101
8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Mean	0.9455	0.9411	0.9571	0.9583	0.9559	0.8839	0.8897
Std	0.0833	0.0857	0.0915	0.0645	0.0697	0.1729	0.1685

Fig 10. Metrics score after hyperparameter tuning

V. CLASSIFICATION

In this project, two Machine Learning Classification algorithms were applied on the World Happiness Index Dataset. As described earlier, we are going to use Ridge Classifier and Agnes.

In this section, the performance of two classification algorithms on the World Happiness dataset is discussed. In order to observe the effects of PCA on the World Happiness dataset, the classification algorithms are applied on the original dataset as well as the PCA applied dataset with three PCA components.

The classification is performed using PyCaret library of Python. The original dataset is split into train and test set with the proportion of 70% and 30%, respectively. For the sake of reproducibility, the session id is set with 123. Using PyCaret, it is possible to create a performance comparison table among all available classification algorithms on the target dataset and find the best model with the highest accuracy.

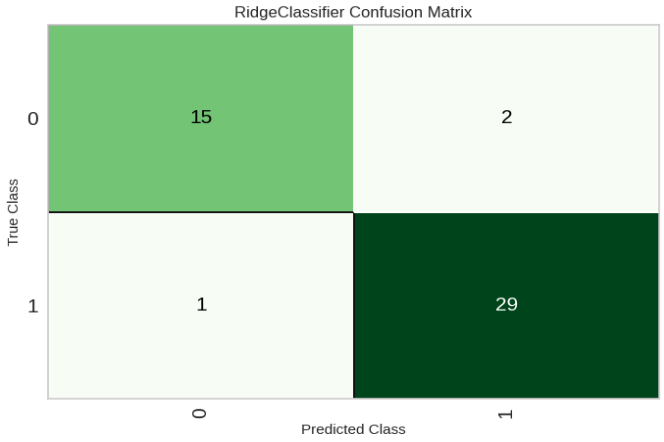

Fig 11. Confusion Matrix

Ridge classification is a machine learning approach for analyzing linear discriminant models. It is a type of regularization in which model coefficients are penalized to prevent overfitting. Ridge classification functions by introducing a penalty element into the cost functional that discourage complication. Typically, the penalty term is the sum of the square coefficients of the model's features.

This causes the coefficients to stay modest, preventing overfitting. The penalty period can be changed to adjust the amount of regularisation. A higher penalty leads to more regularisation and lower coefficient values. This can be useful when there is a scarcity of training data. Nevertheless, if the repercussions term is too long, underfitting can occur.

Once applied on World Happiness Index data set it yielded the following results:

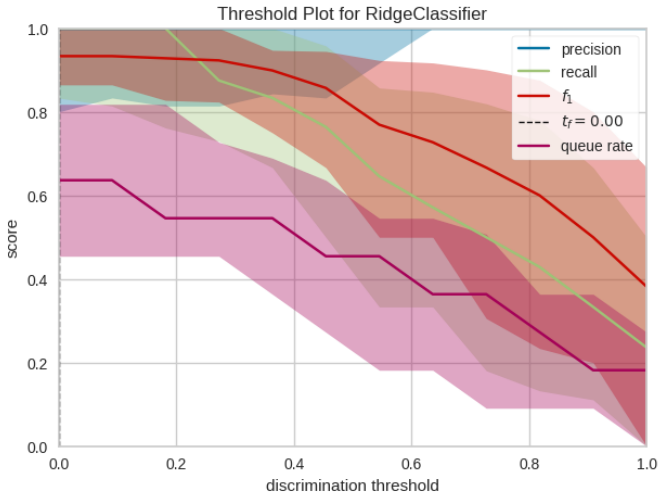

Fig 12. Threshold plot

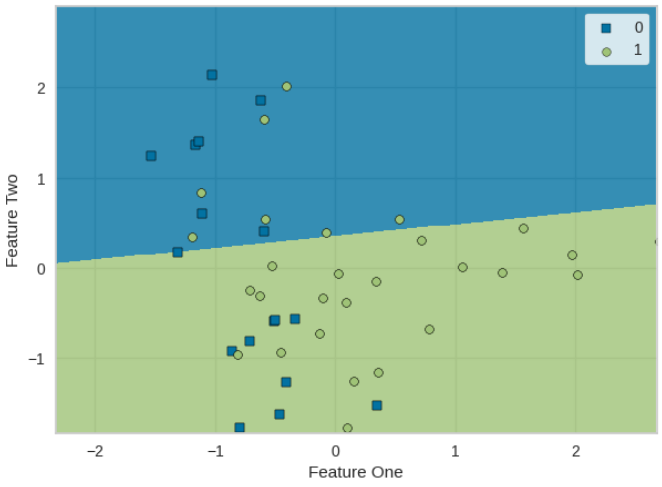

Fig 13. Decision Boundaries of the Ridge classifier algorithm applied on transformed dataset

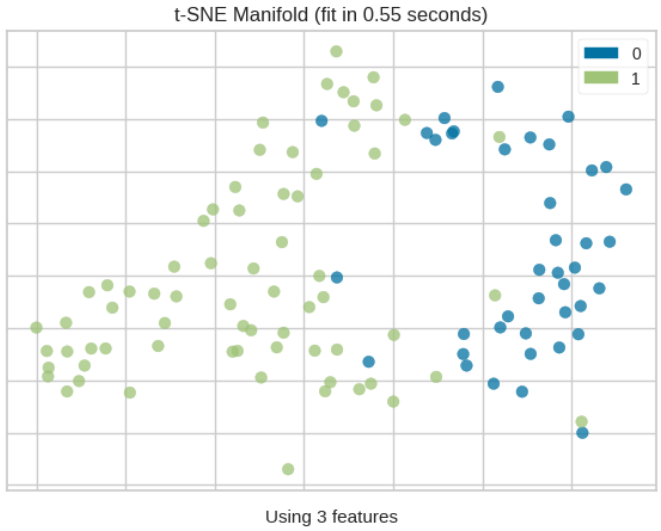

Fig 14. T-SNE Manifold

AGNES (AGglomerative NESting) is a bottom-up approach to network analysis. In other words, each observation is originally treated as a separate-element cluster (leaf). Two clusters which are more comparable are joined into a new larger cluster (node) at each phase of the procedure. This method is repeated till every point are members of a single large cluster (root).

Once applied, It yielded following result:

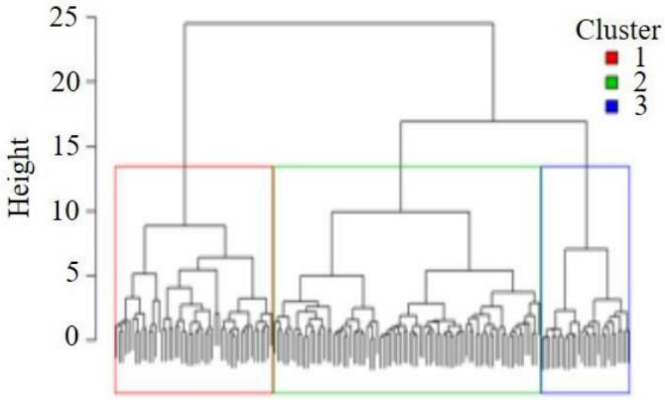

Fig 15. Agnes Cluster

VI. Conclusion

In this project, we applied PCA on six World Happiness Index features dataset to distinguish between Happy countries and Not Happy countries. Using both scree plots and explained variance we found that the first two components contain 79% of the whole variation, in addition to that, a covariance matrix was built and examined to identify the correlation between the six variables.

Furthermore, the relationship between both components and their directions were analyzed by Bi-plot. We also applied two classification algorithms on the original data, all components, and the first two components. And compared the results for the two algorithms.

REFERENCES

[1] United Nations. Happiness should have greater role in development policy, https://news.un.org/en/story/2011/07/382052.

[2] E. Diener, M. Tamir and C. Scollon, “Happiness, life satisfaction, and fulfillment: The social psychology of subjective well-being.”, 2006

[3] J. Ott, “Beyond Economics, happiness as a standard in our personal life and politics” pp71, 2020

[4] J.Helliwell, L Aknin, “Expanding the social science of happiness”, Feb 2018

[5] The Worldwide Governance Indicators Project [Internet]. 2020 [cited 2020Apr]. Available from: https://info.worldbank.org/governance/wgi/.

[6] https://towardsdatascience.com/understanding-happiness-dynamics-with-machine-learning-part-2-4df36e52486

[7] United Nations. Happiness should have greater role in development policy, https://news.un.org/en/story/2011/07/382052

[8] World Happiness Report https://worldhappiness.report/

[9] https://www.kaggle.com/code/josignaciofernandez/happiness-report-exploratory-data-analysis