SSRG International Journal of Economics and Management Studies
ISSN: 2393 – 9125 / https://doi.org/10.14445/23939125/IJEMS-V8I10P101

Volume 8 Issue 10, 1-7, Oct, 2021

Original Article

Classification of Countries' HDI Through Development Indicators

Zaheer Abbas[1], Abdul Hakim H M Mohamed[2], Abdur Rehman[3], Dr Faris Omar [4]

[1]*Department of Statistics, University of Gujrat, Pakistan*
[2]*Department of Management Information Systems, A'Sharqiyah University, Oman*
[3]*Department of Computer Science, Bahria University Islamabad, Pakistan*
[4]*International College of Engineering and Management, Oman.*

Received Date: 26 August 2021
Revised Date: 29 September 2021
Accepted Date: 10 October 2021

Abstract *- United Nations Development Program (UNDP) annually publishes a report for the Human development index based on Health, Educational, Social and Economic development factors. This study proposed a method for classifying and predicting the human development index (HDI) using important development indicators. Kernel principal component analysis (KPCA) and k-nearest neighbour (KNN) classifiers are used for dimensionality reduction and classification. A sample of 757 Omani students was selected, of which 81.2% were female. Sixty per cent of the data used in this study was extracted from the United Nations Development Program and World Bank databases. Dimensionality reduction technique was applied to the data to overcome the over-fitting and extract the important information with a minimum loss. To address the violation of linearity assumption, Kernel principal component analysis was used for classification purposes. Correlation Matrix for development indicators, classification Report, Confusion Matrix and Classification Boundaries are constructed. The results of KNN showed 77 per cent classification accuracy in predicting any country HDI.*

Keywords *- HDI, Development Indicators, Dimensions, UNDP, Supervised Learning, Unsupervised Learning, KPCA, KNN, Dimensionality Reduction.*

I. INTRODUCTION

Humans are the actual treasure of any Nation. There is no doubt that the growth and development of people are actually the development of Nations. Basically, the development is the process of social and eco-nomic conversion caused by environmental and complicated cultural factors and their interactions. Development means "betterment in country's social and economic situation. It is pointed out that advancement in a such way of controlling the areas of human and natural resources so that it can improve the lives of human beings. Every human wants to live a good life, education, health and nice standard of living are the three dimensions from which we can judge the goodness of human's life. These three dimensions are combined to make the term "Human development".

Human development is a measurement of attainments by human beings through improvement of knowledge, natural changes, habit formations or other benchmark that displays variation over time. The knowledge of human development can help a country to handle international trade efficiently. In literature development of countries is a popular topic which had been argued almost in every discipline. Economists, Educationists, Public Administrators, Sociologists and Medical Experts have their own but different viewpoints for country development. So, the definition of development is so controversial. Human development is defined as "The process of waxing the decisions of human beings that grant them to lead a long and healthy lives, to get good education and decent and honorable standard of living. The human development is about giving people more freedom and space to live their lives according to their choices."

HDI is the compound statistic used to classify countries by human development levels. The HDI is amplitude (measure) of income, education and health. It measures the overall achievements in a country according to the above three dimensions (Health, Education and Income).

HDI was introduced as a tool to assess the economic and social improvement levels of countries. There are three major key points to examine the rank of countries in the dimensions of Education, health and standard of living. The key points or indicators of these three dimensions are "Mean year of schooling", "expected years of schooling" (educational dimensions), "life expectancy at the time of birth" (Health dimension), "gross national income per capita" (standard of living dimension). This index (HDI) carry out feasible to follow variation which exist in development levels of different countries with the passage of time. The HDI was produced to put great attentions that people and their potential should be the main criteria for determining the growth of a

country, not only economic development. It is also raised question to national policies, that how two countries with the same GNI per capita can show different human development results? This diversity can arouse dispute about government policies and its priorities. The HDI is the summary of statistical measure to judge the attainments in key dimensions of human development: such a healthy and long life, to be educated and have a nice standard of living. The geometric mean of normalized indicators of these three dimensions becomes the HDI.

The HDI was produced as a substitute to assess about the attainments of any country. It is a quantitative statistical measure that is easy to understand and made up of that what majority of people think about the very basic factors of people's well-being, education, health and income. The first HDI report was published in 1990 by the United Nation Development Program (UNDP). HDI provide the country's development at macro level in which the development of every human being was counted. HDI has become one of the universally used indices of welfare in the World and has flourished in advancing the assessment, and discussion of prosperity over the important. HDI used widely very successfully but yet it is narrow approach because it gives limited view economically. Currently, the role of HDI become very strong at government level being an official statistic. Its yearly publications commence deliberate (serious) political discussion and renewed struggles, to improve lives at national and regional levels.

As Human development index (HDI) is based on major three dimensions i.e. Health, Knowledge and standard of living. United Nation Development Program (UNDP) has classified countries in accordance with these dimensions along with their four indicators (Life expectancy at birth, Expected year of schooling, Mean year of schooling and Gross national income (GNI) per capita). The present study has these three underlying dimensions, but with different indicators of development.

II. DIMENSIONS OF DEVELOPMENT

Development of any country based on the following three dimensions:

A. Health

Development of any country based on social, political, economic, health and educational indicators. Development and health are historically associated; health is defined as the healthy population not only the absence of diseases, and also necessary for social and economic development. When the development is stable, it contributes to the health of population (Buss, 2016). Health is the very important dimension for the development of any country.

B. Education

Education is the key to development, education is the human right that play a vital role in the development of any country. Anyone can judge the country's development by observing the literacy rate of that country. In the report of (UNESCO, 2006) literacy is more important to grow and compete with the other nations. It is obvious that competition is a way towards development. Since the education is also the important dimension human development.

C. Standard of Living

Standard of living dimension refer to the level of comfort, wealth, material goods and necessities accessible to a certain area in a country. The standard of living combines the strength of factors such that political, social and economic stability, national economic growth, gross domestic product and income.

Beneath these dimensions there are some development indicators that are used in this study. The indicators are given in the figure1 with their corresponding dimension.

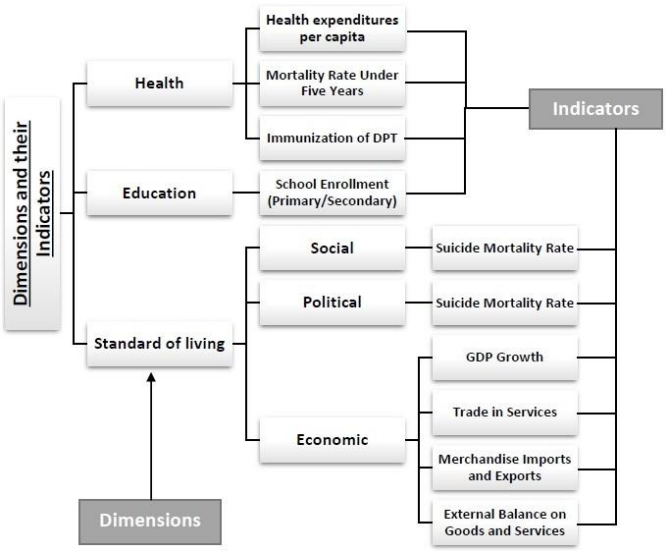

Fig 1. Dimensions and their corresponding Indicators

III. OBJECTIVE OF THE STUDY

The objective of this study is to classify the country type using health, economic, social, Educational and political indicators, which presumably helps in filling the gap of classifying and predicting human development index (HDI) of any country without applying traditional methods.

IV. LITERATURE REVIEW

Ozturk, S. G. (2007) classify and predict the country types based on develop- ment indicators by using discriminant analysis as a statistical technique. Countries are pre-classified according to their HDI (Human development index) value published each year by the United Nations. Factors regarding the Health are the good indicators to discriminate the country type, on the other hand Economic growth, trade on goods and services, involvement of women in national and government parliament are the good indicators for the distinction between developed and developing countries. The results from the discriminant analysis gives us the opportunity to

determine the level of development based on some factor related to education, health, women's involvement in National parliament and trade. Immunization of Hepatitis B3, Mortality rate can easily be discriminating the developed and developing countries. Imports, Exports, School enrollment, GDP growth, ratio of women's in parliament can categorize the countries. Factors regarding health are powerful indicators for discrimination between developed and undeveloped countries.

The data reduction and classification are very important aspects, as they improve the performance of models through dimensionality reduction (Ozturk, S. G., 2007). Principal component analysis is a well-known technique to reduce the dimensions of the data to get the meaningful patterns of the data. In this study PCA is used to reduce the dimensions of medical data through feature extraction. After feature extractions, multiple classifiers are used to classify the data. The results from the Cardiac Arrthymia data set 280 actual features are convert into 12 meaningful features after applying the PCA as a dimensionality reduction technique. With classifiers PCA shows better results than other data reduction techniques.

Sasikala, S., & Balamurugan, S. A. A. (2013) proposed a unique analysis method for the clas- sification of gene expression data. This procedure contains dimensionality reduction using kernel principal component analysis (KPCA) and logistic regression classifier is used for classification (discrimination). KPCA is a generalization and nonlinear extension of principal component analysis. The proposed method was applied to five different gene expression datasets involving human tumor samples. This method of classification is more effective and powerful than other classification methods such as neural networks and support vector machines in classifying gene expression data.

Liu, Z., Chen, D., & Bensmail, H. (2005) used the principal component analysis (PCA), an un- supervised technique to reduce the dimensionality of the high dimension's data and classify it into different groups. It is the transformation of the high dimensional data into a meaning patterns that have low dimensions. Native based and K-Nearest neighbor classifier are used in this study and the K-Nearest neighbor classifier shows the better results from the Native based in data classification.

The Kernel Principal Component Analysis for texture classification. KPCA is the non-linear extension of Principal Component Analysis. It is the powerful tool to extract features by making a new space with the help of the product space of input pixels for making the texture patterns effectively. The principal components are efficiently measure by using the input space of pixels. Neural Network classifier is used to classify these features. The experimental results indicate that the Kernel PCA is an amazing and powerful tool for texture classification (Liu, Z., Chen, D., & Bensmail, H., 2005).

V. METHODOLOGY

A. Introduction

Kernel principal component analysis is an unsupervised technique which reduce the dimensionality of independent variables without losing much information. In this study supervised and unsupervised learnings are used for data analysis. The former is used for dimensionality reduction of data, while the latter is used for classification of data. The methodology adopted by this research is illustared in figure 2.

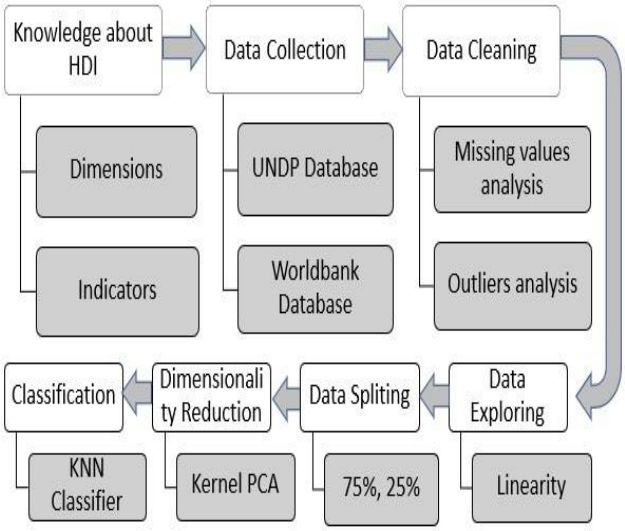

Fig 2. The Methodology

B. Data Collection

The data used in this study is taken from the United Nation Development Program and World Bank databases. The data consists information about Human Development Index of 188 countries across the world. After initial review of the data, one hundred and fifteen (115) countries were selected that has maximum information on the different development indicators. The Human development index' (HDI) score indicate the country type and this score was provided every year for each country by the United Nations. The range of the HDI value is zero to one. The countries those have the HDI value 0.800 or greater go to the very high HD's category, and the countries those have the HDI value between 0.700 to 0.800 go to the High HD's category, and the countries those have the HDI value between 0.555 to 0.700 go to medium HD's category. And the countries those have the HDI value less than 0.555 go to the low HD's category.

The dependent variable is country type (CT) that is cate- gorical variable with four categories defined above. and all the indicators "Suicide mortality rate (SMR), Proportion of seats held by women in national parliaments (WNP), GDP growth (GDP), Trade in services (TS), Merchandise imports (IMP), Merchandise exports (EXP), External balance on goods and services (EB-GS), Primary School enrollment (PSE), Secondary School enrollment (SSE) (Health Expendi- tures/capita (HE/c), Immunization DPT (IM-DPT), Mortality rate under-5 (MR¡5) are independent variables.

C. Data Analysis Technique

Kernel Principal Component Analysis (KPCA) is used to extract the features and reduce the dimensionality of the nonlinear data while K-Nearest Neighbor classifier is used for countries classification. The mathematical background of these techniques are given below:

a) Kernel Principal Component Analysis

In the field of classification KPCA is an astonishing technique that captured the complex structure of data by mapping the data from lower dimensional space to higher dimensional space through Kernel trick. Kernel principal component analysis (KPCA) is an unsupervised technique and a nonlinear extension of principal component analysis (PCA). If the data has complicated structure which cannot be visualize on linear subspace, standard PCA is not suitable technique for data reduction. In the situations of non-linearity of the data Kernel PCA will be very helpful for dimensionality reduction and classification. According to Vapnik-Chervonenkis theory higher dimensional space provide the greater classification power than the lower dimensional space (Goldberg, P. W., & Jerrum, M. R., 1995).

Same as the basic idea behind the Kernel PCA is to map the original d- dimensional data into a new D-dimensional feature space, (where d is less than D), each data point is projected to a new feature space . Kernel PCA use the Kernel trick to compute the principal components in higher dimensional feature space. KPCA is not a classifier itself but it improves the classification results by knowing the pattern of the entire data. Before applying the classifier, it is necessary to reduce the dimensionality through feature extraction method.

b) K-Nearest Neighbour Classifier

K Nearest Neighbor classifier use the Euclidian distance to classify the unknown unit into a certain group after training the data set. The algorithm based on the value of K that varies from 1 to N, here N are the number of units in the data set. For example, there are two categories (0 and 1) in training data set and the value of K = 5, calculate the five nearest neighbor from the unknown unit that is to be classified, if the majority of the neighbors are from the category 0 then the unknown unit classified into 0 category. The drawback in this algorithm is the different value of K may not give the same results all the time. It is very effective when the training data set is large enough. Euclidian Distance. KNN classifier has no model so it is called lazy learning. KNN do classification and prediction by exploring or investigating the classified pattern of data in training session. Predictions with KNN is based on voting scheme in which the winner category takes the new unit (Imandoust, S. B., & Bolandraftar, M.,

VI. RESULTS AND DISCUSSION

A. Assumptions

Before performing the kernel principal component analysis there are some assumptions that must be fulfilled, the assumptions are given below:

- Require multiple independent variables that are on continuous scale.
- There should be adequate non-linear relationship between independent variables.
- There should be no outliers.

All the variables of this study is on continuous scale and there are no adequate relationship between variables and also free from outliers with no missing value. Pearson correlation test is applied to check the linearity. From the pairwise correlation scatter matrix and correlation matrix (fig. 4) , it is clear that there is no adequate linear relationship between the pairs of most variables. Significance of the relationship can be checked by the p-value of the test. The variables that have the correlation value between 0 to 0.2 or -0.2 to 0 does not have the significant relationship because p-value is greater than 0.05 which provide the reason to accept the null hypothesis that there is no linear relationship between variables. Correlation value between +/-0.2 to +/-0.5 have the significant linear relationship but this relationship is not adequate for the use of linear model in classification [9]. Nonlinearity exist in data, cleared from the correlation matrix and scatter correlation matrix. So a nonlinear extension of principal component analysis KPCA is used for dimensionality reduction.

B. Splitting the Data

In machine learning techniques data is divided into two subsets, one is training dataset and other is testing dataset. Machine train a model using training dataset and understand the entire pattern and extract the features of the data. And model will test the dataset for the prediction and accuracy. For this study data is divided in the ratio of 3 and 1. Total units (after removing the outliers) in the dataset are 101 in which 75 units (75% of the whole data) are in the training dataset and 26 units (25% of the whole dataset) are in the test dataset.

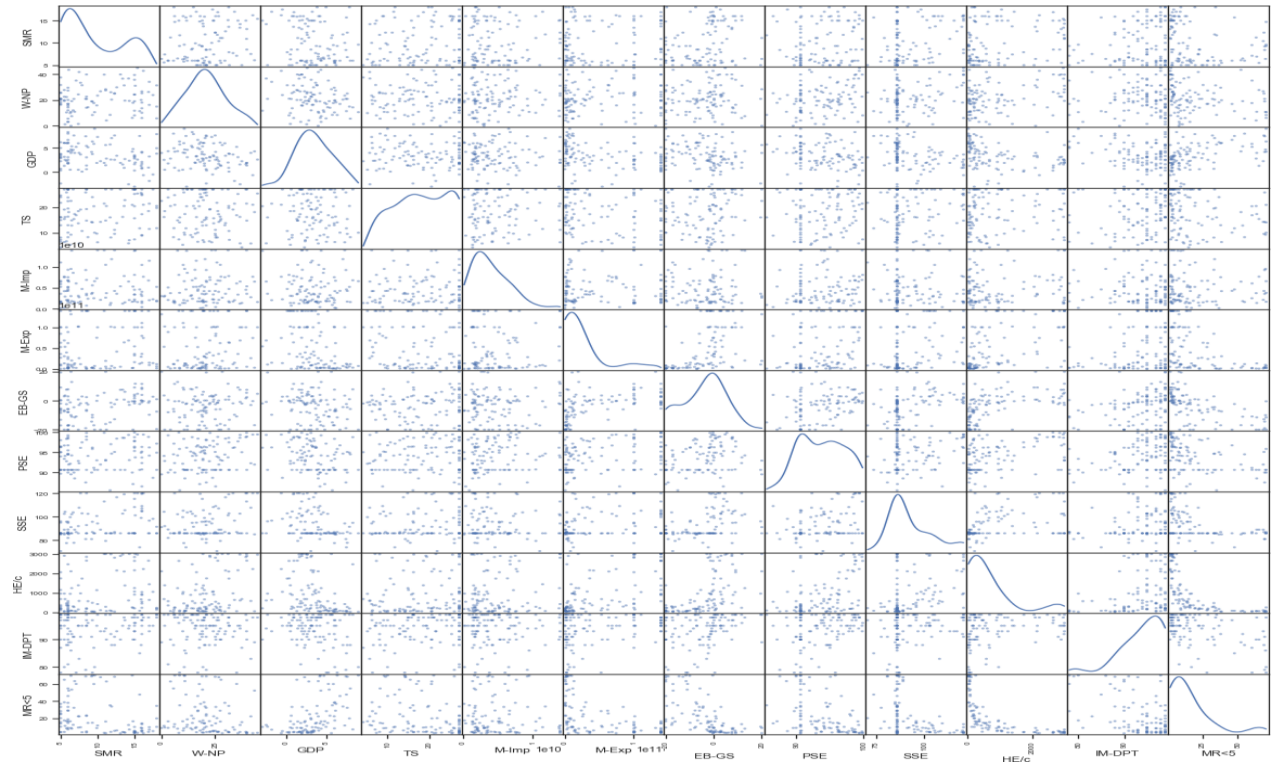

Fig. 3 Scatter Correlation Matrix

C. Dimensionality Reduction and Classification

After splitting the data Kernel PCA is applied to both training and test dataset to extract the features and classify the countries using K-nearest neighbor classifier. As discussed above that the value of K is important that how many neighbors are considered to classify the new county. To overcome this problem a graph is plotted between K-value and Mean error (Figure 5). The point/value where the mean error is minimum will be the best value of K. When the value of K is 5 then mean error is 0.31 and the accuracy is 96%. By changing the value of K from 5 to 6, error rate decreased and end at 0.225. The error rate and accuracy will be the same if value of K is 6, 8, 11, 13 and 14. Because at these values the error rate is minimum and accuracy is maximum. The maximum accuracy for this classification is 77%.

Table 1. Classification Report

Category	Precision	Recall	F1-Score	Support
First (0)	0.92	1.00	0.96	11
Second (1)	0.67	0.86	0.75	7
Third (2)	0.50	0.40	0.44	5
Fourth (3)	1.00	0.33	0.50	3
Average	0.77	0.65	0.66	**n=26**

From the table 1 precision of first category is 0.92. which means, from all the actual countries of first category 92% correct classification belongs to the first category and 8% classification belongs to the other category. Similarly, precision of second category is 0.67, which means that from all the actual countries of second category 67% correct classification belongs to the second category and 33%

classification belongs to the other category. The precision of third and fourth category is 0.50 and 1.00 respectively (Table 1).

The recall of first category is 1.00 which means that from all the predicted countries of first category 100% correct classification belongs to the first category. similarly, the recall of second category is 0.86, which means that from all the predicted countries of second category, 86% correct classification belongs to the second category and 14% wrong classification belongs to the other category. The recall of third and fourth category is 0.40 and 0.33 respectively.

	SMR	W-NP	GDP	TS	M-Imp	M-Exp	EB-GS	PSE	SSE	HE/c	IM-DPT	MR<5
SMR	1.000000	0.153206	-0.233208	0.173323	0.027807	0.127018	0.219314	0.121565	0.312823	0.260338	0.200043	-0.279292
W-NP	0.153206	1.000000	-0.239807	-0.036107	0.074387	0.105792	0.062885	0.261128	0.280624	0.353703	-0.013238	-0.194921
GDP	-0.233208	-0.239807	1.000000	-0.051994	-0.062626	-0.329511	-0.228851	-0.292737	-0.333413	-0.336372	-0.268720	0.480918
TS	0.173323	-0.036107	-0.051994	1.000000	0.024631	-0.234042	0.100409	0.091169	0.163192	0.142627	0.164971	-0.221624
M-Imp	0.027807	0.074387	-0.062626	0.024631	1.000000	0.054384	0.042518	0.101298	-0.055217	0.152002	0.000354	-0.074590
M-Exp	0.127018	0.105792	-0.329511	-0.234042	0.054384	1.000000	0.500520	0.319227	0.304656	0.570799	0.294223	-0.531848
EB-GS	0.219314	0.062885	-0.228851	0.100409	0.042518	0.500520	1.000000	0.278124	0.408901	0.576356	0.319431	-0.621755
PSE	0.121565	0.261128	-0.292737	0.091169	0.101298	0.319227	0.278124	1.000000	0.317993	0.464865	0.314545	-0.445796
SSE	0.312823	0.280624	-0.333413	0.163192	-0.055217	0.304656	0.408901	0.317993	1.000000	0.390584	0.256611	-0.462304
HE/c	0.260338	0.353703	-0.336372	0.142627	0.152002	0.570799	0.576356	0.464865	0.390584	1.000000	0.288030	-0.645622
IM-DPT	0.200043	-0.013238	-0.268720	0.164971	0.000354	0.294223	0.319431	0.314545	0.256611	0.288030	1.000000	-0.471430
MR<5	-0.279292	-0.194921	0.480918	-0.221624	-0.074590	-0.531848	-0.621755	-0.445796	-0.462304	-0.645622	-0.471430	1.000000

Fig. 4 Correlation Matrix

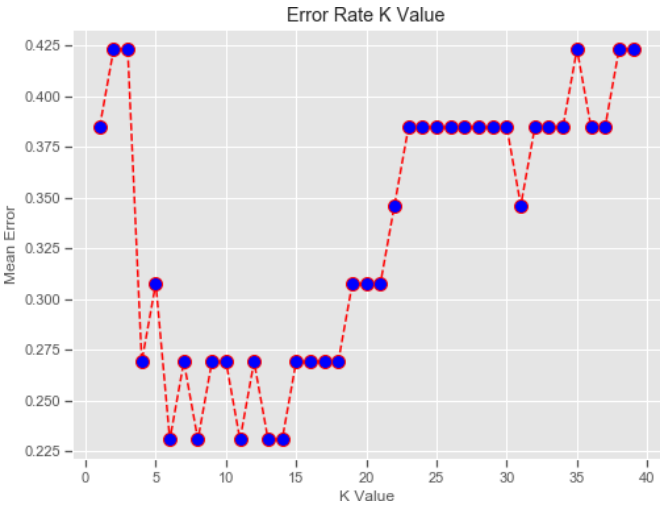

Fig. 5 Error Rate K-Value

F1-Score is the geometric mean of precision and recall. The last column of classification report represent that, in the scenario of classification, out of 26 countries, eleven countries supports the first category,seven countries supports the second category, five and three countries supports the third and fourth category respectively. All results in the classification report are calculated from the confusion matrix given below.

Classification Boundaries

Figure 6. gives the graphical view of classification boundaries of KNN classifier, from the figure it is clear that there are four categories presented by a unique color. Dots in the dark purple area represents that from the 12 countries of first category, 11 are correctly classified but 1 country is wrongly predicted to the second category by the classifier. Dots in the light purple area represents that from the 9 countries of second category, 6 are correctly classified but 3 countries are wrongly predicted to the third category.

Similarly, Dots in the grey area represents that from the 4 countries of third category, 2 are correctly classified but 2 are wrongly predicted to the fourth category. Dot in the green area represent that there are only one country in the fourth category, and also correctly predicted by the classifier.

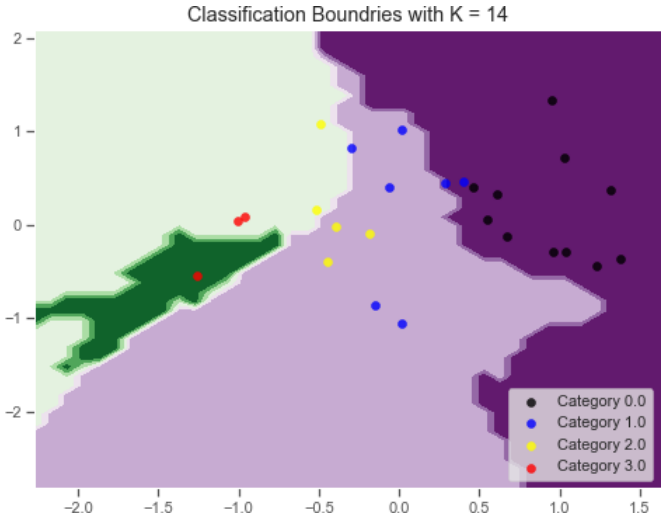

Fig. 6 Classification Boundaries of KNN Classifier

VII. CONCLUSION

The technique that is used in this study is a nonlinear extension of PCA. It is an effective tool that is sed to reduce the dimensionality of nonlinear data and make a new subspace where the data patterns become linear. KPCA is not a classifier itself but it can improve the classification results. KNN classifier done the 77% classification correctly after applying KPCA on countries data.

REFERENCES

[1] United Nations Development Programme (UNDP). The Real Wealth of Nations - Pathways to Human Development. New York (2010). http://hdr.undp.org/en/content/human-development-report-2010.

[2] Ozturk, S. G. Classifying and predicting country types through development factors that influence economic, social, educational and health environments of countries. SWDI Proceedings papers , 759 (2007)

[3] 665-674.

[4] Sasikala, S., & Balamurugan, S. A. A. Data classification using PCA based on Effective Variance Coverage (EVC). In IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN) IEEE (2013) 727-732.

[5] Liu, Z., Chen, D., & Bensmail, H. Gene expression data classification with kernel principal component analysis. Journal of Biomedicine and Biotechnology, 2 (2005) 155.

[6] Telgaonkar Archana, H., & Sachin, D. Dimensionality Reduction and Classification through PCA and LDA. International Journal of Computer Applications, 975 (2015) 8887.

[7] Kim, K. I., Park, S. H., & Kim, H. J. Kernel principal component analysis for texture classification. IEEE Signal Processing Letters, 8(2) (2001) 39-41.

[8] Goldberg, P. W., & Jerrum, M. R. Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers. Machine Learning, 18(2-3) (1995) 131-148.

[9] Imandoust, S. B., & Bolandraftar, M. Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. International Journal of Engineering Research and Applications, 3(5) (2013) 605-610.

[10] Mukaka, M. M. A guide to appropriate use of correlation coefficient in medical research. Malawi medical journal, 24(3) (2012) 69-71.

[11] Klugman, J., Rodríguez, F., & Choi, H. J. The HDI 2010: new controversies, old critiques. The Journal of Economic Inequality, 9(2) (2011) 249-288.

[12] Junior, P. N. A., Mariano, E. B., & do Nascimento Rebelatto, D. A. Using data envelopment analysis to construct human development index. In Emerging Trends in the Development and Application of Composite Indicators . IGI Global (2017) 298-323.

[13] Smit Shah, Determinants of Human Development Index: A Cross-Country Empirical Analysis. SSRG International Journal of Economics and Management Studies 3(7) (2019) 40-43.

[14] Ahmad, Muhammad Syarif, Fajar Saranani, Wali Aya Rumbia, The Impact of Human Development Index on Poverty in Southeast Sulawesi. SSRG International Journal of Economics and Management Studies 6(12) (2019) 30-36.