Tyrone Kent Pastoril – 300356781

**Title: Country Classification Based on World Governance & Development Indicators by Income Level Analysis**

The world development indicators dataset is focused on government factors. This consists of six dimensions of how well the government has run the country for the past years: Control of Corruption, Government Effectiveness, Political Stability and Absence of Violence/Terrorism, Regulatory Quality, Rule of Law, and Voice and Accountability. The dataset is the result of collection on perceptions of governance from multiple sources. The World Bank went through extensive data processing to ensure minimal margin of errors on each factor.

The objective of this analysis is to utilize the six indicators to predict the income group level of a country based on its Gross National Income (GNI) per Capita and relate other dimensions such as unemployment rate, demonstrate the relationship of each factor to another, illustrate the indicators distribution, and fit the best model for predicting the income group level. Multiple classification models are done to fine tune the best and most accurate model for forecasting the future welfare of a country.

The table below serves as the guide for classifying which income group level a country belongs to based on its GNI/ capita.

Table . World Bank Country Classification by Income Level (Hamadeh et al., 2022)

|  |  |
| --- | --- |
| **Group** | **GNI/ capita** |
| Low income | < $1,085 |
| Lower-middle income | $1,086 - $4,255 |
| Upper-middle income | $4,256 - $13,205 |
| High income | > $13,205 |

# Data Exploration

Every measure is correlated with each other to figure out which pairs are highly associated with another. Figure 1 shows that there is a high correlation between the six indicators and GNI/ Capita. This implies that these dimensions could be useable in predicting the income group. All other parameters are dropped due to insignificant relationship with GNI/ Capita.

Other data exploration methods are applied on Tableau and R script. On the Tableau file, the dashboard, World’s Richest Countries filtered in year 2021, illustrates the following: World’s GDP/ Capita Map shows what countries belong to which income group; GNI/ Capita Income Group Distribution indicates that 33.51% or 65 countries are high income level while 11.86% or 23 countries are low income; Switzerland is the leading country in Top 10 GNI/ Capita Countries with a GNI/ Capita of $88.91K. In the R script, a feature plot is used to describe the data distribution of the income group levels among the six dimensions.

Figure . Correlation Heatmap

# Data Preprocessing

There are many null data in the dataset. As only the 6 indicators and GNI/ Capita are utilized, all other fields are dropped. Furthermore, all rows that have null on the remaining fields are also removed.

An additional field is added to the dataset that is going to be the target categorical variable, INCOMEGRP. The field’s value is based on Table 1.

Feature scaling has been applied using the standard normalization. This assumes that the variables come from a normal distribution whose mean is 0 and standard deviation equals to 1.

# Classification Modeling

Multiple classification models are employed to fit the optimal model with the maximum accuracy. The models are the following: Linear Classifier, Linear Discriminant Analysis, Naïve Bayesian, Decision Tree, Random Forest, K-Nearest Neighbors Algorithm (KNN), and Support Vector Machines (SVM). Only KNN and SVM are trained with normalized data as this influences the performance of the said models, while others are trained with the unnormalized dataset since normalization would not make any difference. Nonetheless, all these models have been fine-tuned to provide the highest level of accuracy. Clustering method is also applied on Tableau to categorize the countries based on the average World Government Indicators (WGI) and GNI per Capita.

# Evaluation Results

A train-test split of 0.80 is used to evaluate and train the models – 80% of the preprocessed data for training the models, while the remaining 20% is for testing.

The following table is the result from model evaluation.

Table . Model Evaluation - Accuracy and Kappa

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Method** | **Test Accuracy** | **Kappa** | **Kappa2** |
| 1 | Linear Classification | 0.4170 | 0.2199 | 0.0484 |
| 2 | Linear Discriminant Analysis | 0.5941 | 0.4509 | 0.2033 |
| 3 | Naive Bayes | 0.5649 | 0.4184 | 0.1751 |
| 4 | Decision Tree | 0.6025 | 0.4629 | 0.2143 |
| 5 | Decision Tree with 0.001 Complexity | 0.7183 | 0.6209 | 0.3855 |
| 6 | Random Forest | 0.8285 | 0.7693 | 0.5919 |
| **7** | KNN (k=3) with Standard Normalization Scaling | 0.8605 | 0.8123 | 0.6598 |
| 8 | Linear SVM with Standard Normalization Scaling, c=50 | 0.6151 | 0.4798 | 0.2302 |
| 9 | SVM with RBF and Standard Normalization Scaling, c=50 | 0.7922 | 0.7200 | 0.5185 |

KNN with a k of 3 generates the highest accuracy of 86.05%. Kappa is also taken into consideration as this is frequently used to test interrater reliability that represents the extent to which the study's data are accurate representations of the variables measured (McHugh, 2012). By using kappa, this ensures the percent agreement across multiple data. The KNN model has a kappa of 0.8123, which according to McHugh, signifies a strong level of agreement and a high percentage of data are reliable. Since kappa is some form of correlation coefficient, the coefficient of determination, squared kappa, could be defined. This describes how much variability there is in the model. The squared kappa of the best model is 0.6598 which means 65.98% of the variation in classification can be explained by the independent variables and exhibits a very strong relationship (Ayenew, 2023) with the dependent variable.

Table . Best Fit Model

|  |  |
| --- | --- |
| Model | KNN(k=3) with Standard Normalization Scaling |
| Accuracy % | 86.05% |
| Kappa | 0.8123 |
| Coefficient of Determination | 0.6598 |

In the clustering method, Cluster 2 has the least number of countries with 24 items. This cluster also has the highest GNI/ capita, Control of Corruption, Government Effectiveness, Political Stability and Absence of Violence/Terrorism, Regulatory Quality, Rule of Law, and Voice and Accountability.

# Conclusion

In the study, the 6 dimensions are proven to have high correlation with GNI/ capita and can be used as the predictors for the income group level of a country. Countries with high average value in governance indicators generate more GNI/ capita, thus high economic income level. It is plausible to anticipate that a country that demonstrates the following features are one of the first world or developed countries: an excellent control of corruption; highly efficient governance; stable politics and minimal to zero terrorism; prominent regulatory quality; strong and strict implementations of law; and the citizens have freedom of expression and association as well as democracy in selecting their government. However, these factors cannot be used in the determination of unemployment rate of a country since the parameters are not correlated with another. Time series is also not employed in this study because the analysis would be restricted to a specific country and trends differ from each nation.

# References

Ayenew, D. (2023, February 9). *Coefficient of Determination (R²) — Definition & Interpretation.* Retrieved from Bachelor Print: https://www.bachelorprint.ca/statistics/coefficient-of-determination/

Hamadeh, N., Rompaey, C., Metreau, E., & Eapen, S. (2022, July 1). *New World Bank country classifications by income level: 2022-2023.* Retrieved from World Bank Blogs: https://blogs.worldbank.org/opendata/new-world-bank-country-classifications-income-level-2022-2023

McHugh, M. L. (2012, October). Interrater reliability: the kappa statistic. *Biochem Med, 22*(3), 276-282. https://doi.org/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

*World Development Indicator [data file].* (2023). Retrieved August 3, 2023, from The World Bank: https://databank.worldbank.org/source/world-development-indicators