2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) | 978-1-6654-1460-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICT4SD50815.2021.9396950

# Clusterization of Different Vulnerable Countries for Immigrants due to COVID-19 using Mean Probabilistic Likelihood Score and Unsupervised Mining Algorithms

Shovan Bhowmik, Saha Reno*, Sharmin Sultana, Mamun Ahmed
Dept. of Computer Science and Engineering
Bangladesh Army International University of Science and Technology
Cumilla, Bangladesh
bhowmik.sshovon5795@gmail.com, reno.saha39@gmail.com, sharminsultana977@gmail.com, mamun57@gmail.com

***Abstract*—In many developing and poor countries, people often migrate to suitable countries to earn their livelihood and support their families. Due to the ongoing pecuniary disaster that ensued because of COVID-19, many immigrants are coming back to their ancestry from different migrant-friendly countries for several reasons. In this paper, a novel approach has been proposed to segregate these countries into five vulnerability labels based on probabilistic likelihood score (LHS) and unsupervised clustering algorithms (CA). A survey dataset of returnee people including various information has been collected and leveraged as attributes in this study. Depending on the dissemination of attributes, LHS has been generated using Bayes' Theorem for each vulnerable country and three unsupervised mining algorithms (KMeans++, Agglomerative and BIRCH) have been applied to the LHS for categorization. Output labels obtained from CA are then evaluated appropriating the average LHS. Multiple performance measurement metrics (Adjusted Rand Index, Mutual Information based Score) have been consolidated to get an incisive comparison of vulnerability labels resulting from CA and expected LHS. The highest value of 0.74 has been attained as Normalized Mutual Information based Score for BIRCH clustering accompanying ample results for the remaining algorithms. The result has shown that the combined application of probabilistic LHS and unsupervised CA can be a reliable method to identify the vulnerability of different countries generally chosen by migrant people.**

***Index Terms*—COVID-19, Migrant People, Probabilistic Likelihood, Unsupervised Machine Learning, Discretization, Silhouette Coefficient**

## I. Introduction

Coronavirus pandemic has severely affected all classes of people around the globe, both socially and economically. In Bangladesh, approximately 15 billion USD is contributed to the country's GDP by the immigrants, resulting in a noticeable socio-economic development each year [1]. Due to COVID-19, lockdown and economic recession have imposed continuous pressure on the migrant workers to return to their own country. Bangladeshi migrant workers in other countries are acutely struggling over the issues like unemployment, tremendous psychological pressure, substandard lifestyle, poor wages and isolation. The adverse effects of this pandemic are mostly inflicted on the impoverished class of the society.

Various researches and surveys have been conducted for a better analysis of the current global pandemic situation with respect to migrant workers across the globe. In [2], Abdul Azeez et al. talked over the negative impacts of COVID-19 on migrant women workers in India and proposed policy interventions to tackle the sufferings and economical crisis. The policy intervention and strategies which can mitigate the negative impacts of COVID-19 on the migrant workforce in context to Bangladeshi migrant workers are discussed in [3]. However, no research has been conducted to analyze the vulnerability of the countries that generally provide working opportunities to the migrants but currently sending them back because of the ongoing crisis. Finding a different level of vulnerability can deliver a clear perception about the hosts' inefficiency to render a robust employment scope for the migrants and help people to choose suitable countries in the near future.

In this perspective, clustering can be an effective technique for categorizing different countries into disparate vulnerability labels by evaluating attributes related to migrant data. Thirty hierarchical and density based CA have been considered in different categorical real-world applications by Sami [4]. Also, instead of label encoding, probabilistic Bayes framework and Bayesian Inference based clustering has been proposed in [5] and [6].

In this study, we have applied Bayesian rule based probabilistic LHS following an improved density based CA (K-Means++) and two hierarchical CA (Agglomerative and BIRCH) to the attributes of different unsafe host countries in our collected dataset for partitioning those countries into multilevel vulnerability. We have also attempted to validate our research outcome juxtaposing results generated from mean LHS and CA estimating different performance evaluation metrics and made our model more vigorous in this case.

The contents of the article are arranged as follows: section 2 contains some works related to our study. Section 3

978-1-6654-1460-9/21/$31.00 ©2021 IEEE

provides the detailed experimental study with the proposed methodology. Section 4 shows our obtained result with a brief discussion and section 5 concludes our work by mentioning some future works associated with this research.

## II. Background

So far very few works have taken place regarding the migrant workers and their current problems in this coronavirus pandemic. Strict lockdown enforcement caused an intense blow to the livelihood of the disadvantaged and poor class of the society, especially the migrant workers. After doing a telephone survey, Abdul Azeez found that 92% of migrant workers inside North-India have lost their job [2]. Assessment of the vulnerability effects of migrant workers can be the first initiative for policy interventions, according to the authors of [7]. Fixed term contracts, working under non-standard contracts, economic downturns, temporary visas are some of the important reasons why the migrants have to lose their jobs and return to their origins.

Instead of only survey related data, the unsupervised mining technique can be a helpful tool to identify vulnerability labels of different countries from many standpoints. Hans discussed density and hierarchical based cluster sequence analysis using probabilistic models in [8]. A maximum hierarchical likelihood clustering technique has been proposed by Sharma in [9] where biological data that belongs to several groups can be classified effectively. In [10], Khan et al. focused mainly on utilizing Latent Gaussian Models (LTMs) and proposed a likelihood function for categorical LTMs which efficiently captures the correlations in discrete data and well suited for categorical data analysis.

A unique labeled clustering method has been introduced using an unsupervised K-Means algorithm along with correlation analysis and frequent member function in [11]. Here, 40 countries have been labeled into different energy sectors by measuring 27 sustainable energy indicators as attributes with little deviations from original results. K-Means++, Mini Batch, Agglomerative, BIRCH and GMM algorithms have been applied on 40000 fashion products by examining PCA and Singular Value Decomposition (SVD) to create a recommendation system by observing images in [12]. In [13], a modified data labeling technique based on the changes in intra and inter class dissimilarities is proposed by the authors to tackle the segregation problem of a large database, which can be well applied in categorical data.

After studying the analysis of the above mentioned literature, we have focused on establishing vulnerability labels of immigrant countries computing LHS and applying unsupervised CA(K-Means++, Agglomerative, BIRCH).

## III. Experimental Studies

In this section, we have described the working procedure of our experiment in detail. For our experimental study, we have utilized $Spyder3$ as the IDE, $Python3.6$ as the language and scikit-learn tools for cluster analysis and performance measurement. The overall workflow of our proposed methodology for vulnerability label assignment is illustrated in "Fig. 1".

The following subsections highlight step-by-step procedure for determining vulnerability cluster labels for host countries:

### A. Dataset Description

In this paper, the survey dataset is provided by $YPSA$ (Young Power in Social Action), which is an organization intended for sustainable development in Bangladesh. It contains 2576 instances of 12 countries from where many Bangladeshi migrants reverted for multiple reasons during this COVID-19 pandemic. The features inside the dataset correspond to the information regarding the following questions:

Q1: What is the age of the migrant? Q2: What is the sex of the migrant? Q3: What is the educational background of the migrant? Q4: What kind of job was served by the migrant abroad? Q5: What kind of working skill did the migrant show before going abroad? Q6: How long did the migrant stay abroad before coming back? Q7: What is the particular cause of return for the migrant during COVID-19 pandemic? Q8: What was the salary of the migrant Abroad? Q9: Is the migrant the only earning person of the family?

The surveyed information is then placed into several categories. Country-wise data for the above 9 questions and category-wise data for different attributes is provided in Table 1 and Table 2 respectively.

Table I: COUNTRY-WISE SUMMARY DATA FOR SURVEY QUESTIONS

| Country Wise Attribute Data | Bahrain | Iran | Iraq | KSA | Kuwait | Lebanon | Malaysia | Maldives | Oman | Qatar | South Africa | UAE | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Q1-Q9*** | 70 | 81 | 98 | 599 | 32 | 221 | 141 | 46 | 465 | 213 | 57 | 553 | 2576 |

### B. Data Wrangling

In our dataset, some attribute values have been missed for different countries along with some data incompatibilities. We have utilized $WEKA$ [14] to fill up the missing data and for cleansing purposes.

### C. Likelihood Measurement for Each Vulnerable Country

There are multiple attributes for the same country in our study. Only label encoding for each category cannot represent the actual dependency of attributes for measuring vulnerability. Various research previously handled these dependencies in several ways as described in [9][11][13]. In our research, we have exploited Bayes' Theorem [5] in this case to generate likelihood values of each country with respect to each of our surveyed attribute categories. After pre-processing the dataset, the likelihood of each vulnerable country is measured based on the distribution of all the attributes. The equation of the likelihood function for our work is stated in (1).

$$P(C|A) = \frac{P(A|C).P(C)}{P(A)} \tag{1}$$

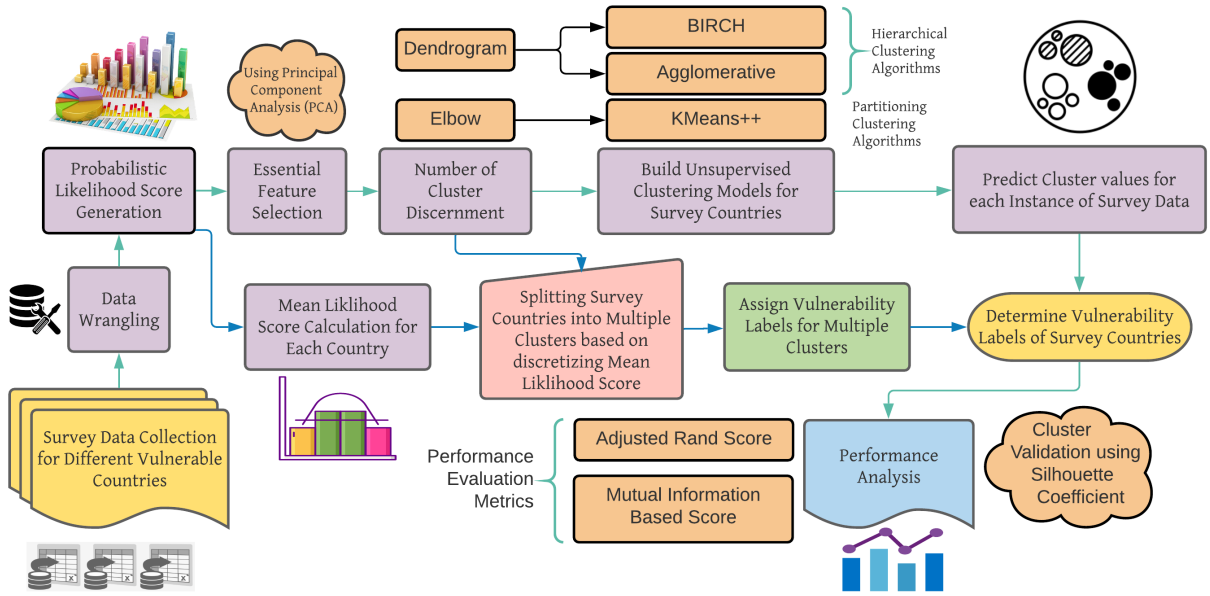



Figure 1: Proposed Methodology to assign Vulnerability Label using mean Likelihood Score and unsupervised Clustering Algorithms

Table II: CATEGORY-WISE SUMMARY DATA FOR VARIOUS ATTRIBUTES

| Age | | | | | | | Sex | | Education Qualification | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *18 to 25* | *26 to 30* | *31 to 35* | *36 to 40* | *41 to 45* | *46 to 50* | *Above 50* | *F* | *M* | *Illiterate* | *Below SSC* | *Graduate* | *HSC Pass* | *SSC Pass* |
| 181 | 699 | 842 | 533 | 167 | 88 | 66 | 355 | 2221 | 134 | 1351 | 30 | 393 | 668 |

| Job in Abroad | | | | | | | | | | | | | Only Earning Member | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Agricultural Work* | *Business* | *Cleaner* | *Construction Work* | *Cook* | *Driver* | *Electrician* | *Labor* | *Machine Operator* | *Mechanic* | *Private Service* | *Sales Person* | *Tailor* | *N* | *Y* |
| 231 | 210 | 163 | 181 | 217 | 172 | 27 | 664 | 153 | 27 | 289 | 213 | 29 | 704 | 1872 |

| Area of Skill | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Agriculture* | *Cleaner* | *Computer* | *Construction* | *Cooking* | *Driving* | *Electronic* | *Hotel Management* | *Hotel Work* | *Mechanic* | *Pathy* | *Plumber* | *Tailor* | *Welding Machine* | *Workshop Activities* |
| 153 | 172 | 153 | 354 | 182 | 264 | 135 | 139 | 133 | 282 | 178 | 99 | 120 | 89 | 123 |

| Years in Abroad | | | | Cause to Return | | | Salary in Abroad | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *2 to 3 Years* | *4 to 5 Years* | *Less Than 1 Year* | *More Than 5 Years* | *Forced Leave* | *No Job* | *Short Leave or Vacation* | *Visa Expired* | *Voluntarily Leave* | *20 Thousand to 30 Thousand Taka* | *30 Thousand to 40 Thousand Taka* | *40 Thousand to 50 Thousand Taka* | *50 Thousand to 80 Thousand Taka* | *80 Thousand to 1 Lac Taka* | *Less Than 20 Thousands Taka* |
| 668 | 464 | 217 | 1227 | 157 | 141 | 1581 | 315 | 382 | 920 | 734 | 339 | 97 | 77 | 409 |

Here, $P(C|A)$ = Likelihood value of a country with respect to an attribute category and $P(A|C)$ = The probability of an attribute category for a given country. $P(C)$= Probability of each vulnerable country and $P(A)$ = Probability of each attribute category.

Following the likelihood value of each endangered country, we have generated Likelihood Score (LHS) for those countries simply multiplying it by 100. These LHS are then assigned to instances of each country.

### *D. Essential Feature Selection*

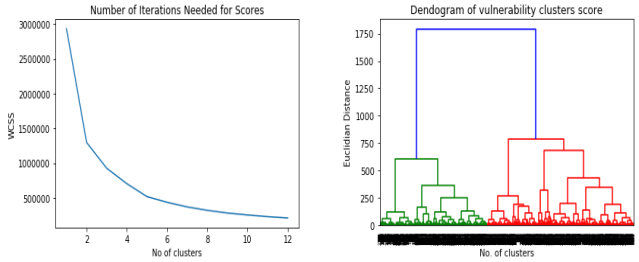
The overall size of our dataset for 2576 instances with 9 attributes is a $2576 \times 9$ matrix. We have reduced it's dimension into 2 principal components which convey the majority of the explained part of all the features by using Principal Component Analysis (PCA) [15]. Consequently, the final feature matrix has been converted into $2576 \times 2$ matrix.

### *E. Determining Number of Clusters*

In our experiment, we have fixed the number of clusters that can contrive all the countries into some hierarchical vulnerable groups. *ElbowMethod* is used for the K-Means++ algorithm observing the total Within-Cluster Sum of Square (WCSS) as the function of the number of clusters and considering the knee in the plot as the appropriate number of clusters [16]. *Dendrogram* representation is observed in case of Agglomerative and BIRCH algorithms to discern the suitable number of groups that can handle all the instances comprising *Ward* as a linkage method [17]. Five clusters have been ascertained for grouping of risky countries. The graphical representation of determining the number of clusters are shown in "Fig. 2".

### *F. Applying Unsupervised Clustering Algorithms*

We have exerted three unsupervised mining algorithms (K-Means++, Agglomerative and BIRCH) to the LHS of each data with respect to all the inspected countries to allocate survey instances of immigrant people into a desired number of




$(a)$ $(b)$

Figure 2: $(a)$ $Elbow\ Method$ and $(b)$ $Dendrogram$ Representation for Number of Clusters Determination

clusters. Different clustering algorithms are applied combining different probabilistic models [8]. In our study, K-Means++ is used instead of the original K-Means which intelligently handles the sensitivity related to the centroid initialization and thus improving the quality of clustering [18]. Agglomerative Clustering using Ward's hierarchical clustering is also applied in our work to minimize the cost associated with each partition [19]. Because the dataset of our research is quite large with lots of attributes, BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) has been used too which summarizes a large dataset into a small one that can hold as much information as it can to partition correctly [20]. Clustering Feature (CF) and CF tree estimating the appropriate threshold (the number of entries in the leaf node) has been rendered in this work after reviewing [21]. Each algorithm has sub-divided our 2576 data of 12 vulnerable countries into five clusters by generating a cluster label for every sample.

### G. Mean Likelihood Score Generation for Each Country

Since the labels of the vulnerability are unknown in the dataset, our primary focus is to label them into five clusters of vulnerability (Very Low, Low, Medium, High and Very High). For this purpose, we have calculated the mean LHS of all the instances of each individual vulnerable country.

### H. Discretizing LHS into Different Vulnerability Labels

For dispensing the countries into our desired cluster labels, equal width distribution is availed to discretize the average LHS for each of the countries. We have computed the class interval of each bin by using the formula in (2).

$$ClassInterval = \frac{MaximumValue - MinimumValue}{N} \quad (2)$$

We have assigned the label for each country from "Very Low" to "Very High" in an ascending order considering the average LHS.

### I. Denoting the Labels of CA output

Discretization of the countries into different clusters has helped us denote the cluster labels of the data obtained after applying mining techniques. The labels are then symbolized as Very Low, Low, Medium, High and Very High categories based on the assignment using expected LHS of the unsafe countries.

### J. Performance evaluation of the vulnerability clusters

To make sure of whether our labeling of the class values considering LHS is pertinent or not, we have compared the results derived from mean LHS with the results achieved from CA. In this paper, Adjusted Rand Index (ARI) and Mutual Information (MI) based Score have been enumerated to evaluate the labeling of our dataset. Also to validate the selection of clusters [17], we have measured Silhouette Coefficient (SC) for each of the CA applied in our task by testing it for clusters from two to seven. The overall result analysis and performance evaluation of our work has been mentioned in the following section.

## IV. Result and Discussion

In this paper, we have focused to identify the risk labels of different hazardous countries for migrant workers of Bangladesh which have been affected largely because of the COVID-19 pandemic by degeneration of job opportunities. We have collected data of those countries which are most commonly selected by Bangladeshi working-class people and accomplished classifying them into five clusters. These are not the only countries that have been suffering from immigrant service. But our proposed model can label other vulnerable countries like this if we get real data of those countries' victim people resonating the attributes like this research.

We have labeled the vulnerability of our twelve countries using mean LHS as disclosed in Table 3.

Table III: PARTITION OF COUNTRIES INTO FIVE VULNERABILITY LABELS USING MEAN LHS

| Vulnerability Label | Country Name |
|---|---|
| ***Very Low*** | Bahrain, Lebanon, Iran, Maldives, South Africa |
| ***Low*** | Qatar |
| ***Medium*** | Iraq, UAE, Malaysia |
| ***High*** | Oman |
| ***Very High*** | KSA, Kuwait |

To determine whether the no. of clusters selected are decisive or not, we validated the clustering labels by measuring SC for different CA. Multiple performance gauging metrics have been employed too for comparative analysis of our work in the following two subsections.

### A. Silhouette Coefficient

It calculates the average SC for each instance of a dataset simply by measuring average within-cluster distance and also computing average outside cluster distance for each sample data. The equation of this validation technique is written in (3).

$$SC = \frac{OC - WC}{max(OC, WC)} \quad (3)$$

Here, $WC$ = our practised CA are shown in Table 4. Mean Within Class Distance and $OC$ = Mean Nearest Class Distance.

The SC range is [-1, 1] which states that the more positive

value gives proper separation of data into desired clusters. Also, the negative SC value refers to the misclassification of data. The SC value for our practiced CA are shown in Table 4.

Table IV: SILHOUETTE COEFFICIENT VALUES FOR THREE UNSUPERVISED MINING ALGORITHMS

| Number of Clusters | SC Values of K-Means++ | SC Value of Agglomerative Clustering | SC Value of BIRCH | | |
|---|---|---|---|---|---|
| | | | Threshold=0.5 | Threshold=1.0 | Threshold=1.5 |
| 2 | 0.37 | 0.52 | 0.29 | 0.47 | 0.40 |
| 3 | 0.43 | 0.60 | 0.39 | 0.59 | 0.44 |
| 4 | 0.42 | 0.59 | 0.33 | 0.55 | 0.44 |
| 5 | 0.51 | 0.61 | 0.39 | 0.65 | 0.44 |
| 6 | 0.33 | 0.49 | 0.32 | 0.44 | 0.41 |
| 7 | 0.40 | 0.51 | 0.34 | 0.55 | 0.42 |

From the above table, we can demystify that the selection of clustering our data into five classes has convincingly streamlined our next steps. In the case of the BIRCH mining algorithm, we have also measured SC values for different thresholds of CF and Branching Factor. In our experiment, we have taken Branching Factor = 25 as it has provided better results. Other Branching Factors are occluded from mentioning in the above table since the results are not satisfying. BIRCH clustering algorithm has shown the highest SC value and each of the mining algorithms have provided the highest SC value for when the number of class labels = 5.

The partitioning outcomes using three CA can be shown by scatter-plot diagram as displayed in "Fig. 3" where variegated colors represent various labels of vulnerability.

The figure shows the distribution of instances into different clusters. In few cases, the output of the CA has been altered because the labeling procedure is not similar for all the CA. We have maintained the same labeling for each CA and exchanged using codes where needed. For example, Very Low (Label 0), Low (Label 1), Medium (Label 2), High (Label 3) and Very High (Label 5) have been considered for depicting vulnerability. If an algorithm's output is Label 3 which is Label 0 for the other two algorithms, we have deliberately swapped the labels in this case. The swapping did not act perversely and it has been a felicitous step by which we have procured a salient outcome.

### *B. Performance Evaluation Metrics*

To evaluate the results obtained from CA and to compare it with the output acquired by mean LHS, we have chosen two frequently applicable performance metrics.

*1) Adjusted Rand Score (ARS):* It has been elicited by measuring the resemblance of cluster outputs generated by the applied CA and the mean LHS generated output. ARS is calculated using (4).

$$ARS = \frac{CARI - LHSRI}{max(CARI) - LHSRI} \tag{4}$$

Here, $CARI$ = Rand Score of CA Output and $LHSRI$ = Rand Score of LHS Output.

*2) Mutual Information Based Score (MI):* It is calculated by drawing out the likeness of two labels of the same identical data following (5) where $|U|$ = the number of instances in $U_i$ cluster, $|V|$ = the number of instances in $V_i$ cluster and $N$ = The total number of instances.

$$MI(U,V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|}\frac{|U_i \cap V_j|}{N}\log\frac{N|U_i \cap V_j|}{|U_i||V_j|} \tag{5}$$

- **Adjusted MI Score (AMI):** It is measured by assimilating the MI Score as described in (6).

$$AMI(U,V) = \frac{MI(U,V) - E(MI(U,V))}{avg(H(U),H(V)) - E(MI(U,V))} \tag{6}$$

Here, $H(U)$ and $H(V)$ are the entropies associated with the clusters, $E(MI(U,V))$ is the expected MI Score.

- **Normalized MI Score (NMI)**: It is measured by balancing the MI score for achieving the correlation between the labels of different algorithm outputs.

Each of these performance metrics provides a score between -1 to +1 where +1 means perfect matching of labeled data and 0 means random labeling of data. A negative value represents mislabeling of data. The evaluation metrics of our research for labeling unsafety states to different countries can be delineated in Table 5.

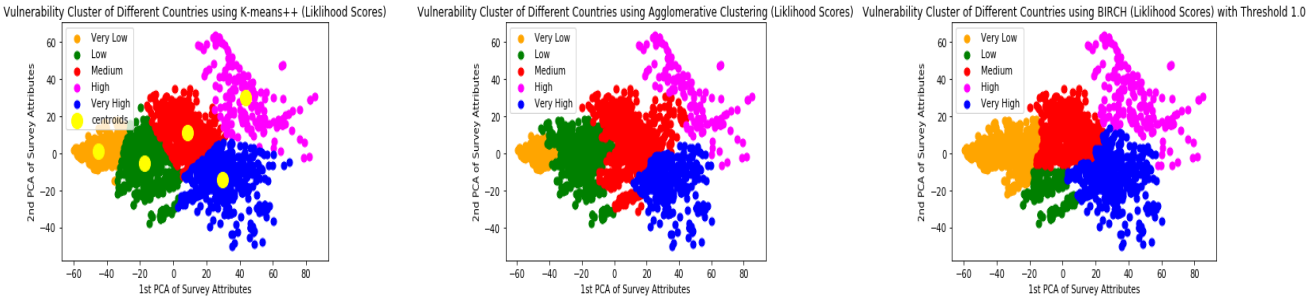
Table V: PERFORMANCE EVALUATION METRICS SCORE FOR THREE UNSUPERVISED MINING ALGORITHMS

| Performance Evaluation Metrics | K-Means++ | Agglomerative | BIRCH |
|---|---|---|---|
| ***Adjusted Rand Score*** | 0.57 | 0.62 | 0.59 |
| ***Adjusted MI*** | 0.51 | 0.63 | 0.70 |
| ***Normalized MI*** | 0.53 | 0.66 | 0.74 |

From this table, we can interpret that the labeling assigned by the average LHS and the CA outputs are significantly similar with the highest NMI score of 0.74. BIRCH Algorithm has achieved a momentous result along with Agglomerative Clustering. K-Means++ clustering didn't partition well compared to the other two algorithms. But still it is far better than random labeling of samples.

The result of not getting a perfect match might be because there were some instances of multiple countries whose likelihood values were so close to the decision boundary of Very Low, Low and Medium clusters. In those cases, different algorithms labeled those data dissimilarly. Moreover, in the step of discretization, a sharp cutting of Average LHS has been performed. Therefore, some "Very Low" label data have been classified as "Low" in Agglomerative Clustering. Similarly, some "Low" label instances have been moved to the "Medium" label in BIRCH Clustering. A very minor part of data has been mismatched arbitrarily between "Medium" and "Very High" classes owing to some faulty data or exceptional data gathered from surveys of the immigrants.

Apart from the aforementioned minor discrepancy, the overall result of our proposed method has worked evenly in determining the label of endangered countries for immigrant people.




$(a)$ $(b)$ $(c)$

Figure 3: Scatter-plot diagram of all the instances for three Unsupervised CA ($(a)$ K-Means++, $(b)$ Agglomerative and $(c)$ BIRCH)

## V. Conclusion

In this paper, we have presented a unique framework for categorizing different unlabeled vulnerable countries without subjective analysis of migration study experts employing probabilistic LHS and unsupervised CA. Clustering validation based on the similarity of labels has been occupied by computing SC. Different performance evaluation metrics have been used to get a good comparison between mean LHS and CA results. Among the data of twelve countries, many countries have fallen in the "Very Low" vulnerability label which provides a good interpretation that when the COVID-19 situation will improve, those countries would be a good choice for migration after the turnaround. Soon, we will try to manage more real data of countries other than our gathered ones for more authenticity of our work as well as to get the overall condition of the whole world in this prospect. We have also planned to sit with experts in this field to explicitly judge our model outcome. Other unsupervised CA and fuzzy-based clustering [12] [4] might be availed later considering maximum likelihood estimation [9] for more robustness of our work. People who are making a choice list of countries for migrating as workers or other groups can depend on our model that can predict the vulnerability of the desired country.

## Acknowledgment

The authors would like to thank $YPSA$ (Young Power in Social Action), a non-profitable social development organization for providing field level dataset in carrying out this work.

## References

[1] T. Siddiqui, "International labour migration from Bangladesh: a decent work perspective," Policy Integration Department Working Paper 66, International Labour Organization, 2005.

[2] A. Azeez E P, D. P. Negi, A. Rani, and S. Kumar A P, "The impact of COVID-19 on migrant women workers in India," *Journal of Eurasian Geography and Economics*, vol. 62, no. 1, pp. 93-112, 2021.

[3] M. R. Karim, M. T. Islam, and B. Talukder, "COVID-19's impacts on migrant workers from Bangladesh: In search of policy intervention," *World Development*, vol. 136, pp. 105123, 2020.

[4] S. Naouali, S. B. Salem, and Z. Chtourou, "Clustering Categorical Data: A Survey," *International Journal of Information Technology & Decision Making*, vol. 19, no. 1, pp. 49-96, 2020.

[5] T. Rigon, A. H. Herring, and D. B. Dunson, "A generalized Bayes framework for probabilistic clustering," 2020. [Online]. Available: arXiv:2006.05451.

[6] A. Agresti, and D. B. Hitchcock. "Bayesian inference for categorical data analysis," *Statistical Methods and Applications*, vol. 14, no. 3, pp. 297–330, 2005.

[7] F. Fasani and J. Mazza, Luxembourg: Publications Office of the European Union, A Vulnerable Workforce: Migrant Workers in the COVID-19 Pandemic, 2020.

[8] H. H. Bock. "Probabilistic Models in Cluster Analysis," *Computational Statistics & Data Analysis*, vol. 23, no. 1, pp. 5-28, 1996.

[9] A. Sharma, K. A. Boroevich, D. Shigemizu, Y. Kamatani, M. Kubo, and T. Tsunoda, "Hierarchical Maximum Likelihood Clustering Approach," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 1, 2016.

[10] M. E. Khan, S. Mohamed, B. M. Marlin, and K. P. Murphy, "A Stick-Breaking Likelihood for Categorical Data Analysis with Latent Gaussian Models," *In Proc. Fifteenth International Conference on Artificial Intelligence and Statistics*, pp. 610-618, PMLR, 2012.

[11] M. Shaheen, S. Iqbal, and F. Basit. "Labeled Clustering: A Unique Method to Label Unsupervised Classes," *8th International Conference for Internet Technology and Secured Transactions (ICITST-2013)*, pp. 210-214, 2013.

[12] S. K. Addagarla, and A. Amalanathan, "Probabilistic Unsupervised Machine Learning Approach for a Similar Image Recommender System for E-Commerce," *Symmetry 2020*, vol. 12, no. 11, pp: 1783, 2020.

[13] H. V. Reddy, B. S. Kumar, and S. Viswanadharaju, "A Data Labeling Method for Categorical Data Clustering Using Cluster Entropies in Rough Sets," *2014 Fourth International Conference on Communication Systems and Network Technologies*, pp. 444-449, 2014.

[14] WEKA: The workbench for machine learning. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/

[15] M. R. Mahmoudi, M. H. Heydari, S. N. Qasem, A. Mosavi, and S. S.Band, "Principal component analysis to study the relations between the spread rates of COVID-19 in high risks countries," *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 457-464, 2021.

[16] Determining The Optimal Number Of Clusters: 3 Must Know Methods. [Online]. Available: https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/

[17] A. Vysala, and D. Gomes. "Evaluating and Validating Cluster Results," 2020. [Online]. Available: arXiv:2007.08034.

[18] D. Arthur, and S. Vassilvitskii, "k-means++: The Advantages of Careful Seeding," Stanford, Tech. Report, ID Code 778, 2006.

[19] V. Marinova–Boncheva, "Using the Agglomerative Method of Hierarchical Clustering as a Data Mining Tool in Capital Market," *International Journal "Information Theories & Applications"*, vol. 15, no. 4, pp. 382-386, 2008.

[20] sklearn.cluster.Birch. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.Birch.html

[21] B. Lorbeer, A. Kosareva, B. Deva, D. Softić, P. Ruppel, and A. Küpper. "A-BIRCH: Automatic Threshold Estimation for the BIRCH Clustering Algorithm," *INNS 2016: Advances in Big Data*, pp. 169-178, 2016.