Analysing Mental Health Trends Among Canadian International Students

Fardin Ahsan Imran
Department of Computer Science
Lakehead University
Thunder Bay, Canada
ID: 1189337

***Abstract*—For international students studying overseas, maintaining good mental health can be especially difficult. Mental health is an essential part of overall well-being. Stress and other mental health issues might result from adjusting to a new culture, being far from home, and dealing with academic demands. In this report, I aimed to emphasize the factors causing impact on International Students mental health(employment, cultural conflicts, living expenses, weather, homesickness, accommodation, funding and other stress etc.) across Canada. Importantly, this paper represents the implementation of an unsupervised machine learning model to analyze a survey participated by students of a specific region and measures the performance of the model using classification.**

I. Introduction

In recent years, scholars such as Alden, Ryder, Paulhus, and Dere (2013), Li, Chen, and Duanmu (2010), and Ruble and Zhang (2013) have extensively explored the myriad challenges faced by international students pursuing education in the United States, the United Kingdom, and Australia [1] [2]. The existing body of research consistently highlights the pervasive difficulties encountered by these students as they navigate academic and social landscapes in foreign countries. From social and cultural adjustments to financial constraints and academic pressures, international students grapple with multifaceted challenges that significantly impact their educational performance. Recognizing the gravity of these issues, there is a collective call for higher educational institutions in the U.S., the U.K., and Australia to proactively address the diverse academic and social needs of international students. This imperative aligns with the assertion of Stoll, Bolam, McMahon, Wallace, and Thomas (2006), emphasizing the necessity for collaborative efforts to enhance learning experiences [3]. The forthcoming discussion delves into the responsibility of host universities and international students alike in fostering an environment conducive to academic success and social integration [4].

The transitional period from high school to college is widely acknowledged as a stress-inducing time for young adults . This phase involves the adoption of new adult roles, coupled with increased academic and economic responsibilities. International students, a distinct subgroup pursuing higher education abroad, face additional unique stressors. Defined by having a student visa and lacking permanent residency in the host country, international students, particularly in Canada, experience escalating enrollment trends in higher education institutions (CBIE, 2017), adding complexity to their transitional challenges [5].

This paper aims to comprehensively address key aspects of clustering methods and how it is causing impact on international students across Canada. In this project we are using both supervised and unsupervised machine learning. It delineates the philosophy, design, and implementation of clustering techniques and classification, emphasizing their relevance in this specific domain. Furthermore, it measures the model accuracy to judge its performance against clustering and furnishes a structured framework for the clustering workflow. Notably, it acknowledges the burgeoning utilization of *k-means* clustering, a method of clustering, in mental health research and provides valuable insights for international students facing various problems in Canada. While primarily catering to mental health researchers, the analytical insights offered in this paper possess broader applicability across related fields such as public health and social sciences. Thus, it serves as a comprehensive guide, equipping researchers with the necessary tools and understanding to effectively utilize clustering methodologies in student's mental health.

II. Dataset Preprocessing

Dataset preprocessing for clustering is essential to prepare data for analysis, ensuring that data is appropriately formatted and scaled to yield meaningful insights. The following steps outline a comprehensive approach to dataset preprocessing for clustering

1. *Dataset:* For the dataset, we are doing a survey and taking responses from all international students across Canada. Nearly 100 participants have submitted their responses and we are hoping to reach 200 students for our dataset to get a better accuracy. The parameters we are considering in this survey are age, degree, difficulties regarding accommodation, employment, housing, cultural barriers, homesickness, anxiety, loneliness and weather change etc.
2. *Data Cleaning:* Begin by examining the dataset's column names. As the dataset consists of 22 columns each representing one question, the descriptive long column names have been renamed to "Q1", "Q2", "Q3" and so on respectively till "Q22" which enhance clarity and

ensure consistency throughout the dataset. For categorical variables, all unique labels are identified across the dataset. This step helps in understanding the range of values and categories within each categorical variable. It is crucial for determining the appropriate encoding method. Common representations of missing data include NaN (Not a Number) for numerical data and blank cells for categorical data. In this dataset, some rows were identified with multiple blank cells. As the proportion of missing values were high and if they did not significantly contribute to the analysis, those rows were deleted.

3. *Encoding Categorical Features:* One critical step of data transformation involves converting categorical data into numerical format using label encoding. This technique assigns a unique numerical value to each category, enabling clustering algorithms to process the data effectively [6]. For instance, this dataset has categorical variables like "Agree", "Disagree", "Strongly Agree" and "Strongly Disagree", label encoding assigns 0, 1, and 2 respectively. This transformation allows algorithms to interpret categorical information while maintaining the integrity of the dataset. By encoding categorical variables into numerical labels, clustering algorithms can identify patterns and similarities within the data more accurately, leading to meaningful insights.
4. *Feature Engineering:* Feature scaling is a preprocessing step of feature engineering in machine learning where numerical features are transformed to a consistent scale. This ensures that all features contribute equally to the analysis, preventing those with larger scales from dominating. Feature scaling is crucial in clustering as it's dependent on similarity metrics. This dataset has been scaled aiming to standardize the values and to ascertain their similarity.
5. *Data Reduction:* Dimensionality has been reduced using Principal Component Analysis(PCA) to capture the variance in the data while reducing the number of features. PCA identifies the most significant components of the dataset, allowing for a reduction in dimensionality without significant loss of information. By reducing the number of features, PCA simplifies the clustering process and helps in visualizing high-dimensional data. Using PCA the dimensions have been reduced to 2 for better visualization of the clusters.

III. Motivation

The challenges faced by international students are numerous and include academic pressures, cultural adjustment, and being far from support networks back home. These problems require creative solutions that may be tailored to the particular requirements of every individual due to their complexity and uniqueness. Utilizing the machine learning models both supervised and unsupervised we will be able to figure out the most important factors that can have an influence on international students' mental health.

IV. Existing Works

This study investigates the relationship between physical activity (PA) and mental health among international students (IntS) within an acculturation framework. Two studies have been done to address the issues. In the first study, analyzing data from both domestic and IntS revealed that PA mediated the connection between stress and mental health, although it didn't mediate the link between stress and mental illness. The second study involved interviews with 12 IntS, uncovering cross-cultural, adaptation, and organizational challenges they face. These challenges suggest the need for higher institutions to delve into the needs and barriers of IntS to provide effective support. The findings emphasize the crucial role of physical activity in mitigating stress and maintaining mental health among international students, urging institutions to address these issues for the well-being of their diverse student populations [5].

This study presents critical implications for universities in addressing the mental health needs of international students (IntS). First and foremost, focus needs to be on male and undergraduate IntS, as they have lower levels of perceived social support and use more avoidant coping mechanisms. It is suggested that integrated support programs be established in which health centers and services for overseas students work together to address the unique mental health and well-being needs of this population. Tailored mental health programs, encompassing symptom recognition, resource awareness, and coping techniques, could be designed for male and undergraduate IntS. To combat self-stigma and encourage help-seeking behaviors, universities may initiate special peer support groups for male students. Secondly, university health services should facilitate interactions among IntS with similar mental health concerns, providing valuable information about available resources. Incorporating mental health education into the curriculum, fostering awareness across all students, and establishing support groups for undergraduates are suggested faculty initiatives. Thirdly, universities are encouraged to promote approach coping strategies through workshops and skill training groups, emphasizing the positive impact on mental health. Lastly, the study advocates for systematic periodic mental health screenings by health professionals, including occupational therapists, social workers, and counselors, to identify at-risk students and implement preventive support programs. This comprehensive approach aims to enhance the mental well-being of IntS and equip them with effective coping skills [7].

This study investigated mental health, mental health literacy (MHL), and help-seeking attitudes in domestic and international students in Australia. Anticipated findings included lower MHL and poorer help-seeking attitudes among international students, aligning with prior research expectations. Additionally, similar levels of psychological distress were expected based on previous examinations of Australian students. Comparing domestic and overseas students in Australia, the study discovered that while psychological distress levels were similar, there were disparities in mental health literacy (MHL) and attitudes toward getting treatment, particularly with relation to suicide ideation. The results point to the necessity for customized interventions that focus on MHL and attitudes among international students; these interventions might make use of low-intensity ones like internet resources. These interventions could enhance engagement with mental health services, offering both student benefits and cost-effective solutions. Future research should explore optimal timing and dosage of interventions during the acculturation process, addressing the growing importance of mental health support for the increasing international student population [8].

This article underscores the importance of recognizing the diverse and growing population of international students on American college campuses and addresses the heightened risk of psychological challenges they face due to cultural adjustments. Despite this, the underutilization of existing counseling services by international students is noted. The article advocates for the establishment of accessible and culturally sensitive mental health services tailored to the unique concerns of international students. It emphasizes the need for a proactive effort to modify counseling facilities and approaches, encouraging counselors to acknowledge and commit to the challenges of working with this population. Such changes could lead to more fulfilling experiences for both international students and counselors, fostering awareness of cultural assumptions and personal values. In an increasingly pluralistic American society, culturally sensitive counseling services are deemed diverse and indispensable for the well-being of this vulnerable student demographic [9].

V. Model Description

Clustering, an unsupervised learning method, groups objects based on similarity, aiming for intra-cluster similarity and inter-cluster dissimilarity. Distance and similarity serve as the foundation for clustering algorithms, with distance being preferred for quantitative data and similarity for qualitative data.

A. Clustering Distance

Commonly used similarity measures include Euclidean distance, Manhattan distance, Minkowski distance, and Cosine similarity, facilitating the grouping of similar data objects in clusters. Here three distance measurement techniques are described to evaluate the most effective one.

1. *Euclidean Distance:* Euclidean distance is one of them which is a common measure in clustering, calculates the ordinary distance between two points by determining the square root of the sum of squared coordinate differences.

$$Dist_{XY} = max_k |X_{i_k} - X_{j_k}| \tag{1}$$

B. The k-means clustering

The *k-means* algorithm is a popular unsupervised machine learning method used for clustering that partitions a dataset into k clusters by minimizing the sum of squared distances within each cluster. As cluster centroids, k sample locations are first chosen at random. Next, using Euclidean distance, data points are assigned to the closest centroid. The algorithm iterates until convergence or a certain number of iterations is attained [10]. The centroids are computed as the mean of data points inside each cluster. However, the algorithm's results are sensitive to the initial centroid selection, leading to potential instability. Choosing the optimal k value is critical as it directly impacts clustering quality, influencing local and global optimality. Therefore, determining the appropriate K value is a key focus of the algorithm, affecting the overall effectiveness and stability of the clustering outcome. [11]

C. Identify Optimum Number of Clusters

For efficient data segmentation, figuring out the ideal number of clusters in clustering is essential. The silhouette index and the elbow method are two popular approaches for this.

1. *Elbow method:* The elbow rule serves as a method to determine the optimal number of clusters in *k-means* clustering. By calculating the Within clusters sum of squares (WCSS) for different K values, where each value represents the number of clusters, the elbow point signifies a significant drop in WCSS. Initially, as K increases, WCSS decreases rapidly, indicating improved convergence. However, beyond the true number of clusters, WCSS continues to decline but at a slower rate [11]. The elbow point represents the optimal K value where the rate of WCSS reduction diminishes, indicating the appropriate number of clusters to capture the data's underlying structure effectively. [12]
2. *Silhouette index:* The silhouette index is used to assess how well clusters created by clustering algorithms like k-means are constructed. It measures a cluster's similarity on how close it is to its own cluster. As opposed to separation, how similar it is to other clusters. The range of the silhouette index is -1 to 1, where +1 means, it matches its own cluster well and its adjacent clusters badly. An object is at the decision boundary between two nearby clusters if its score is around zero. If the object's score is near to -1, it might have been placed in the incorrect cluster. This index is derived from the so-called silhouette width, which has the following expression: The silhouette coefficient of the data point X is denoted by S(x). The average distance c(x) is the difference between x and every other data point in the cluster that x is a part of. Whereas, avg distance to the closest other cluster is d(x). [13] [14].

$$S(x) = \frac{c(x) - d(x)}{max(c(x), d(x))} \tag{2}$$

D. Classification

The purpose of classification, a fundamental activity in machine learning, is to group data items according to their properties into preset classes or categories. Creating a model that recognizes patterns in labeled data and applies it to forecast the class labels of newly discovered data is the first step in the process [15].

1. *Random Forest Classifier:* A widely used approach for classification is the Random Forest Classifier. It is a component of the ensemble learning family of models, which merges several different models to improve prediction performance. Decision trees that have been randomly selected and trained on a fraction of the data and features make up random forests. The dataset is divided into two sections before the model is built: the training dataset and the test dataset. Most of the data (usually 70–80%) that is needed to train the model is included in the training dataset. The remaining fraction of the data, say 20–30%, is included in the test dataset, which is used to assess the model's performance [16].
2. *K-fold cross validation:* K-fold cross-validation is a method for evaluating a machine learning model's effectiveness and capacity for generalization. The dataset is divided into k equal-sized folds (or subsets), and the model is subsequently trained iteratively k times, utilizing k-1 folds for training and the remaining fold for validation each time. The performance measures (such as accuracy, precision, recall, F1-score, etc.) are calculated for each iteration of the model after it has been trained and evaluated on each fold. After that, an average of these measures over the course of all k iterations is calculated to provid e a single model performance estimate [17] [18].

VI. Experimental study and Result

This part presents experimental results of each section including a a descriptive statistics of the dataset and determining optimum number of clusters using elbow method and silhouette index for different clusters, scatter graph representation for those clusters, classification results using k-fold cross validation.

A. Descriptive statistics

The study comprised 127 participants, with a highest age range of 22 -26 years. Females represented 26.6% of the sample, and the majority were male students (72.4%). Notably, 59.1% identified as graduate students, while 32.3% are undergraduate. Among all participants, the majority of them are either in their first year or second year. 57.5% students found accommodation very challenging. Financially, only 11% reported self-funded, with 30.9% of family support and 21.3% bank loans. Approximately, more than half (60%) of the students are facing challenges for getting jobs and 59.1% reported, they are managing all their expenses though it is a bit tight. Due to these difficulties around 35% of students reported that they are stressed

B. Clustering Results

In Figure 1, Elbow method, the optimal number of clusters (k) is determined by observing a plot of Within-Cluster Sum of Squares (WCSS) against the number of clusters [19]. In this plot, it is clearly visible that the optimum no of clusters k = 3 , where WCSS sharply decreased initially before leveling off, indicating the appropriate number of clusters.
par

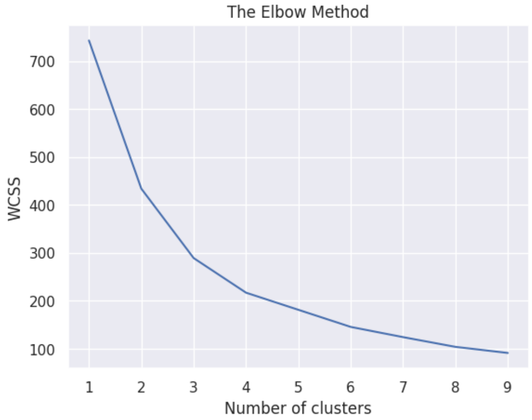

Fig. 1. Elbow graph (Clusters vs WCSS)

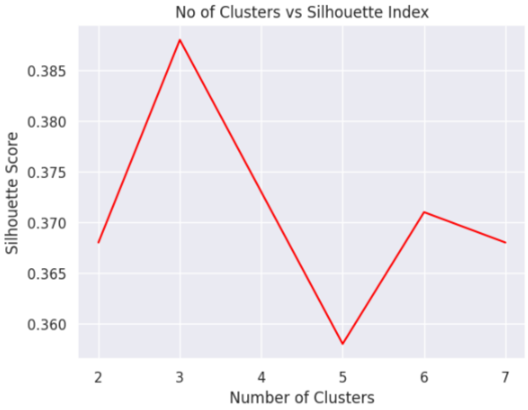

Fig. 2. Clusters vs Silhouette score

In Figure 2, Silhouette index vs. number of clusters graph has been analyzed to assess the quality of clustering solutions. The silhouette index calculates an object's cohesiveness (similarity to its own cluster) and separation (from other clusters) in relation to each other [20]. It offers a visual depiction of the effectiveness of clustering for varying numbers of clusters. From Figure 2, we can get the highest index value 0.388 and number of clusters = 3 which is the optimal cluster number.

Clustering scatter graphs have been provided in Figure 3 to get an visual representation of how data points are grouped into clusters based on their features for different numbers of clusters.as shown in Figure 3, after scaling the dataset and reducing the dimensions using PCA, subsequently, *k-means* clustering was applied, resulting in different scatter graphs for different number of k values . As shown in Figure 3, from the visual representation, it is obvious that output for k = 3 with a high Silhouette value of 0.388 is more acceptable

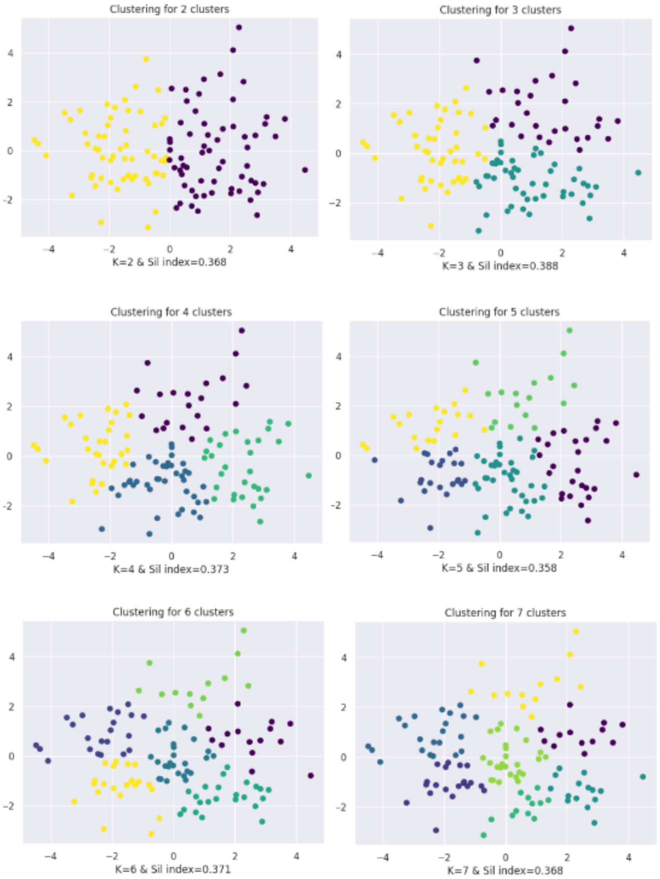

Fig. 3. Scatter graph for different k values

than others.

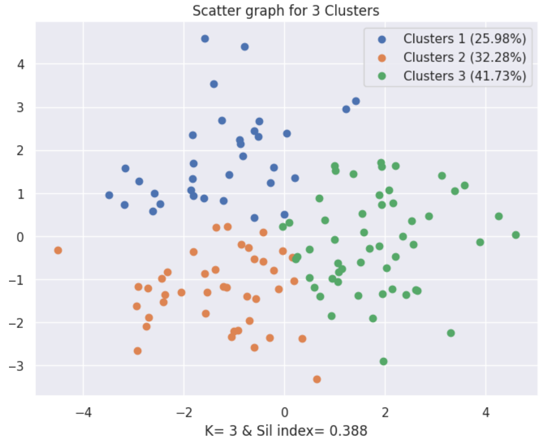

Fig. 4. K means Clustering for 3 clusters

Figure 4 illustrates the clustering outcome of the 127 students: Cluster 1 includes 33 (25.98%) students, Cluster 2 represents 41 (32.28%) participants and Cluster 3 indicates 53 (41.73%) students. This categorization provides valuable insights into anxiety levels among the sampled students, facilitating targeted interventions and support strategies.

TABLE I
K-FOLD VALIDATION

Fold Number	Accuracy
1	0.8846
2	0.7692
3	0.8462
4	0.8462
5	0.7692
6	0.8077
7	0.8462
8	0.8077
9	0.8462
10	0.8462
Mean Accuracy	0.8269

C. Classification on clustered feature

The mean accuracy of 0.8269, derived from 10-fold cross-validation applied on the number of cluster features after applying clustering considering the optimum number of clusters k =3, indicates the performance of a predictive model in classifying data points into clusters. In table 1, the accuracy represents the proportion of correctly classified data points over the total number of data points across all folds of the cross-validation process. A mean accuracy of 0.8262 suggests that the model, trained in the optimum number of cluster features, is reasonably effective in discerning patterns and groupings within the data. This level of accuracy implies that the model correctly predicts the cluster assignment for approximately 82.62% of the data points, indicating a relatively high level of predictive capability.

VII. Future Scope

Widely utilized unsupervised learning method K-means clustering has tremendous future potential for assessing and resolving a variety of issues encountered by foreign students studying abroad. Critical concerns like housing, employment crises, racism, and stress can be better understood by using clustering on data of international student's .When it comes to housing, clustering can reveal trends in the housing preferences of students, assisting authorities in more effectively allocating resources and customizing housing options to suit a range of demands. The development of focused support programs and career services by politicians and academic institutions can be guided by the patterns that clustering might reveal in employment opportunities and crises. Furthermore, clustering can reveal racist trends that foreign students encounter, allowing universities to take proactive steps to promote diversity and combat prejudice. Interventions like counseling services and stress management programs can be specifically designed to help students' mental health and well-being by identifying clusters that exhibit greater levels of stress.

VIII. Conclusion

In summary, this report focuses on the mental state of students facing various difficulties during their study and previous study. Parameters include age, degree, difficulties regarding accommodation, employment, housing, cultural barriers, homesickness, anxiety, loneliness and weather change etc. Among 127 students, while most of them are graduate students, the major findings of their mental health are a job crisis, hard to adapt with the workplace, accommodation, and cultural barriers. In conclusion, the highlighted studies underscore the intricate relationship between physical activity, mental health, and acculturation challenges among international students. It emphasizes the imperative for universities to tailor support programs, address mental health literacy, and enhance cultural sensitivity in counseling services. Implementing these recommendations can significantly contribute to the well-being and successful adaptation of the diverse international student population in Canada.

IX. Acknowledgement

I would like to sincerely thank everyone who helped me to finish this paper. First and foremost, i would like to express my sincere gratitude to our supervisor for all of his help, support, and encouragement during the research process. Their knowledge and perceptions were very helpful in determining the course of this endeavor. I also owe my team members a debt of gratitude for their constant support, cooperation, and teamwork during the entire process.

References

[1] A. G. Ryder, L. E. Alden, D. L. Paulhus, and J. Dere, "Does acculturation predict interpersonal adjustment? it depends on who you talk to," *International Journal of Intercultural Relations*, vol. 37, no. 4, pp. 502–506, 2013.

[2] G. Li, W. Chen, and J.-L. Duanmu, "Determinants of international students' academic performance: A comparison between chinese and other international students," *Journal of studies in international education*, vol. 14, no. 4, pp. 389–405, 2010.

[3] L. Stoll, R. Bolam, A. McMahon, M. Wallace, and S. Thomas, "Professional learning communities: A review of the literature," *Journal of educational change*, vol. 7, no. 4, pp. 221–258, 2006.

[4] H. Forbes-Mewett and A.-M. Sawyer, "International students and mental health," *Journal of International Students*, vol. 6, no. 3, pp. 661–677, 2019.

[5] D. E. Rosa, *Assessing Physical Activity, Mental Health, and Stress among International Students at the University of Toronto*. University of Toronto (Canada), 2019.

[6] N. Ekbote, P. Dhanshetti, and S. Sakhrekar, "Techniques of exploratory data analysis."

[7] D. Baghoori, "Mental health of international students studying at canadian universities," 2021.

[8] B. A. Clough, S. M. Nazareth, J. J. Day, and L. M. Casey, "A comparison of mental health literacy, attitudes, and help-seeking intentions among domestic and international tertiary students," *British Journal of Guidance & Counselling*, vol. 47, no. 1, pp. 123–135, 2019.

[9] S. C. Mori, "Addressing the mental health concerns of international students," *Journal of counseling & development*, vol. 78, no. 2, pp. 137–144, 2000.

[10] J. Ghosh and A. Liu, "K-means," in *The top ten algorithms in data mining*. Chapman and Hall/CRC, 2009, pp. 35–50.

[11] C. Yuan and H. Yang, "Research on k-value selection method of k-means clustering algorithm," *J*, vol. 2, no. 2, pp. 226–235, 2019.

[12] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *Eurasip Journal on Wireless Communications and Networking*, vol. 2021, pp. 1–16, 2021.

[13] A. Starczewski and A. Krzyżak, "Performance evaluation of the silhouette index," in *Artificial Intelligence and Soft Computing: 14th International Conference, ICAISC 2015, Zakopane, Poland, June 14-18, 2015, Proceedings, Part II 14*. Springer, 2015, pp. 49–58.

[14] X. Wang and Y. Xu, "An improved index for clustering validation based on silhouette index and calinski-harabasz index," in *IOP Conference Series: Materials Science and Engineering*, vol. 569, no. 5. IOP Publishing, 2019, p. 052024.

[15] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, J. Akinjobi *et al.*, "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128–138, 2017.

[16] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert systems with applications*, vol. 134, pp. 93–101, 2019.

[17] T.-T. Wong and P.-Y. Yeh, "Reliable accuracy estimates from k-fold cross validation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1586–1594, 2019.

[18] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto, S. Ridella *et al.*, "The'k'in k-fold cross validation." in *ESANN*, vol. 102, 2012, pp. 441–446.

[19] A. Kuraria, N. Jharbade, and M. Soni, "Centroid selection process using wcss and elbow method for k-mean clustering algorithm in data mining," *International Journal of Scientific Research in Science, Engineering and Technology*, pp. 190–195, 2018.

[20] L. Nitya Sai, M. Sai Shreya, A. Anjan Subudhi, B. Jaya Lakshmi, and K. Madhuri, "Optimal k-means clustering method using silhouette coefficient," 2017.