Recognizing Fake Headlines Using Clustering Algorithms

Juthuka Arunadevi[1], A. Mary Sowjanya[2]
[1,2]Department of Computer Science and Systems Engineering,
Andhra University College of Engineering (A), Visakhapatnam, AP, India
[1]jarunadevi2003@gmail.com , [2]sowmaa@yahoo.com

Article Info
Page Number: 111 – 121
Publication Issue:
Vol 71 No. 2 (2022)

Abstract
The credibility of the news sources has hit a new low during the COVID-19. Hence, it is necessary to check for facts before trusting the news. Clustering is extremely important for analysing data, making predictions, and overcoming data abnormalities. So, in this work, the two most prominent clustering algorithms, K-Means and K-Medoids, are tested on a dataset, and K-Means outperforms k-Medoid. We now utilized supervised classification methods like Logistic Regression, K-Nearest Neighbours, and Support Vector classifier to train on the same news headlines we used for clustering with the 'Prediction' column, and then chosen the technique with the highest accuracy. The Support Vector Classifier had the maximum accuracy of 94.93 percent, according to the test. We have developed is a hybrid model consisting of an unsupervised K Means clustering algorithm and a supervised Support Vector classification algorithm**.** The K Means algorithm organizes the news headlines into clusters by capturing the usage of certain words and the support vector algorithm learns from those clusters to predict the categories into which the unseen news headlines belong to.

Article History
Article Received*: 28 December 2021*
Revised*: 25 January 2022*
Accepted*: 10 March 2022*
Publication*: 24 March 2022*

Keywords: k-means; k-Medoid; Logistic Regression; K-Nearest Neighbours; Support vector classification, Fake news, Fake headlines

1. INTRODUCTION

In today's environment, news may be found all over the internet. With a few mouse clicks, we can discover them on numerous news websites such as CNN, BBC News, India Today, NDTV, and Times of India, as well as social media sites such as Facebook, Twitter, and Instagram. The objective of news is to keep us informed about what is going on in the world. Unfortunately, the news we hear today has a dubious trustworthiness. During the COVID-19 epidemic, news sources' credibility has plummeted to new lows. During this time, several misinformation campaigns arose, mainly in the United States, to influence individuals against following COVID guidelines and getting vaccinated. As a result, it is vital to double-check. As a result, the credibility of these fact-checking websites is in doubt. As a result, we require a system with Artificial Intelligence for categorizing news headlines. Natural Language Processing makes this possible. We must recognize that the news is only certain to be genuine or incorrect in exceptional circumstances. We can't be certain in most circumstances. So, the chance of the news being genuine or false is what we should be searching for. We, as such designed a hybrid model that combines the unsupervised K Means clustering approach and the supervised Support Vector classification algorithm. The K Means algorithm groups news

headlines into clusters based on how frequently specific words are used whereas the support vector algorithm learns from those clusters and predicts the categories to which the unseen news headlines belong to.

2. LITERATURE SURVEY

Aslam and co investigators[1] tested K-means and K-Medoids algorithms for their ability to cluster data using various distance measurements. Methods such as scale, range, and Yeo-Johnson are used to alter the data. Ashwini and Sunitha [2] made a relative study on algorithmic rules specifying in the result of dataset on each algorithm. Vankayalapati Et. al [3] suggested that the grouping of data in the form of k-means can be used to evaluate student output. Machine learning can be employed in every system in a variety of fields, including education, pattern identification, sports, and industry. Its importance grows in tandem with the students' future prospects in the educational system. Evgeniou and Pontil[4] summarized the topics presented at a workshop on Support Vector Machines (SVM) Theory and Applications and presented a brief overview of background theory and current trends in SVM. They also discussed the research papers presented and concerns that arose during the workshop. Wang and co-investigators[5] proposed an innovative method for conducting service grouping with a medium or large category set employing descriptive data from categories in a large-scale taxonomy as sample data to support sample service documents. To enable efficient categorization utilizing this new form of sample data, a novel feature selection method has been introduced. Fernandes Et.al[6] proposed the most appropriate statistical technique for dealing with dichotomous dependent variables and presented in their work logistic regression to assess the impact of corruption scandals on the chances of reelection of candidates running for the Brazilian Chamber of Deputies (2014). They demonstrated the computational implementation in R and discussed the substantive interpretation of the results. Zewude and Ashine [7] have made an attempt to examine and determine the primary characteristics that influence student academic progress. Using a binary logistic regression model, they identified that students at Wolaita Sodo University's college of natural and computational science in Ethiopia's academic achievement are influenced by study time, peer influence, securing first choice of department, arranging study time outside of class, amount of money received from family, good life later on, and father's education level.

Reza Et.al [8] identified Random forests as a set of tree predictors in which each tree is dependent on the values of a randomly sampled vector with the same distribution across the forest.Guo and co-investigators [9] constructed a model of KNN for the data, which then served as the classification basis instead of the data ,The k value can be automatically calculated and fluctuated based on the data.. It is proved to be best in terms of accuracy of classification. The model's architecture decreased the model's reliance on k and speed up classification. Experiments were conducted on various publicly available datasets obtained from the UCI Machine Learning Repository. Ahmar, Ansari and co-investigators[10] suggested that the K-Means Clustering (KMC) technique can be used for data grouping. This data classification technique is based on each member's degree of membership. The goal of their study was to use K-Means Clustering to organize Indonesia's current provinces based on density of population, participation rate of school, human development index, and rate of open unemployment .In places like South Sumatra, DKI Jakarta, Lampung, Java Central, and West Kalimantan, the analysis revealed five big clusters in each center.

Chakraborty and Nagwani [11]studied the incremental behavior of splitting-based K-means clustering. The metadata gathered from the K-Means results has then be used to create incremental clusters. When the number of clusters increased, the number of objects also increased, whereas the length of the cluster radius reduced. Incremental clustering outperformed k-means clustering when a number of new data articles were introduced

into the current database. For this purpose The K-means clustering technique was applied to a active database with continuously updated data in the incremental approach. Wu et.al[12] discussed issues such as computational complexity and limited processing power due to the increasing development of data and the mass storage state. The distributed computing platform uses load balancing to dynamically configure a large number of virtual computing resources, effectively breaking the time and energy consumption bottleneck, and incorporating its unique benefits in massive data mining. As such parallel k-means algorithm has been extensively studied in their study. They have done cluster analysis in parallel by first initializing random sampling and then parallelizing the distance computation procedure, which ensures independence between the data objects. Li and Wu [13] proposed an improved K-Means clustering algorithm by combining the biggest minimum distance technique with the classic K-Means approach. The drawbacks of the standard K-Means technique for determining the initial focal point can be compensated by this revised algorithm. The modified K-Means algorithm effectively addressed two drawbacks of the original algorithm: the first is a larger reliance on choosing the starting focal point, and the second is that it is easy to become trapped in local minima. Newling and Fleuret[14] demonstrated that algorithm clarans finds better K-Medoids solutions than Hastie Voronoi ‘s iteration approach. This discovery, together with the Voronoi iteration algorithm's resemblance, clarans a K-means algorithm, as well as Lloyd's K-means algorithm, inspired them to employ Clarans algorithm as a K-means algorithm initializer. On 23 datasets, Clarans surpassed other methods. From them Clarans has undergone algorithmic enhancements that have improved its complexity and runtime, thereby allowing it to be used as a feasible initialization strategy for huge data sets. Balabantaray Et.al[15] found out that the massive increase in information in everyday life, has become challenging to mine useful information in a timely manner. Clustering was introduced to collect the necessary data in a group of people and has been done using a variety of algorithms like K-means and K-Medoids, then a comparison was made to determine which of algorithms is the most effective.

3. METHODOLOGY

ALGORITHM:

Input: NewsArticles Dataset

Output: Prediction of cluster to which the unseen headlines belong to

Step 1: pre-processing of data

Step 2: creation of corpus of words

Step 3: Feature Extraction

Step 4: Feature Scaling

Step 5: Dimensionality Reduction using PCA

Step 6: Clustering of headlines into appropriate clusters using k-means/k-Medoids

Step 7: Model Evaluation

Step 8: Predicting clusters of unseen headlines using SVM

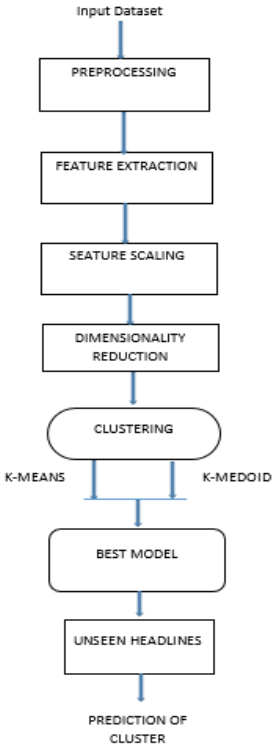

Fig.1 Proposed methodology

The proposed methodology for this work is presented in Fig.1

Input: Input given for our experiment is NewsArticles Dataset. The Times of India, News18, The Indian Express, and Republic World news headlines were used to construct the NewsArticles dataset. In addition, they were acquired from digital-only news sites such as Op India, News Punch, and Great Game India, which the International Fact-Checking Network considers to be fake news sites (IFCN). The dataset has 14,787 headlines from various sources in India and the United States for training purposes.

Then exploratory data analysis has to be done. The news articles number published by each source have been shown in Fig.2 as a bar plot.

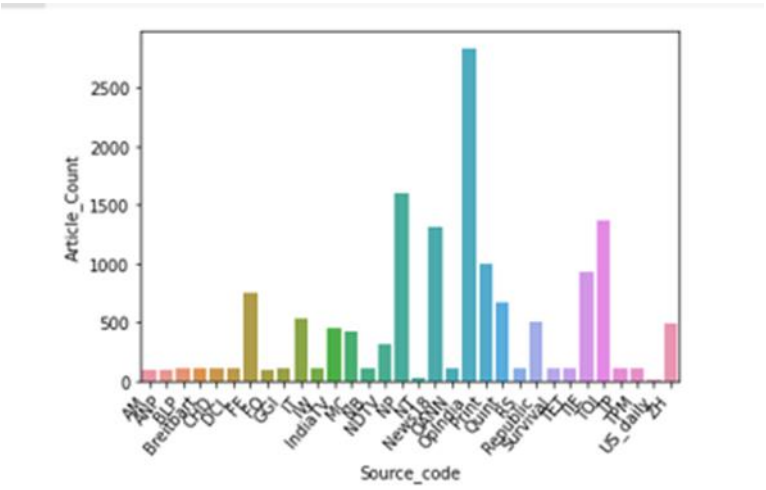

Fig.2: A bar plot showing news articles number published by each source.

Step1: **Text cleaning and preprocessing**

1.1. Removing stop words: Stop words are words that are typically filtered out before a natural language is processed. These are the most common words in any language (articles, prepositions, pronouns, conjunctions, and so on), and they don't add anything to the text. Stop words in English include "the," "a," "an," "so," and "what." We may conclude that removing such phrases has no detrimental repercussions for the model we are training for our assignment. Because there are fewer tokens involved in the training, removing stop words reduces the dataset size and hence reduces training time.

1.2 Removing special characters: Non-alphanumeric characters are known as special characters. These characters can be found in a variety of places, including comments, references, and currency numbers. These characters provide little benefit to text comprehension and cause algorithmic noise. Regular expressions (regex) can thankfully be used to remove these characters and integers. Removing special characters is done in our experiment as they are of very less use in our model.

Step2: Creating a corpus of words:

2.1 Tokenizing the headlines i.e., splitting the headlines into words.

2.2 Lemmatizing the words. It converts the words into meaningful base form, as shown in below Fig.3

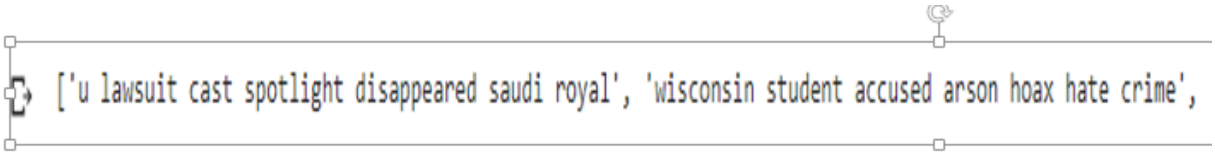

Fig.3 corpus of words

Step3: Feature extraction

We convert the headlines in the corpus of words into vectors by TF-IDF (Term Frequency-Inverse Document Frequency) vectorization**.** Also, the dimension of each vector is reduced to 1000.

Step4: Feature scaling

Machine learning methods that determine distances between data require feature scaling. When computing distances, if the feature with the higher value range is not scaled, the feature with the higher value range takes precedence. So, Feature scaling is a must before performing any unsupervised classification.

Step 5: Dimensionality Reduction using Principle Component Analysis (PCA):

PCA is a statistical process that converts a set of correlated variables to a set of uncorrelated variables using an orthogonal transformation. In exploratory data analysis and machine learning for predictive models, PCA is the most extensively used tool. Since the dimension of each vector formed is 1000, we apply PCA. This removes the curse of dimensionality.

Step 6: Clustering: k-means/k Medoid

Before performing the clustering, we need to find out the number of clusters required. This can be done using the method of Elbow. The Elbow curves obtained by K-Means clustering and k-Medoid clustering as shown in Figures 4 and 5.

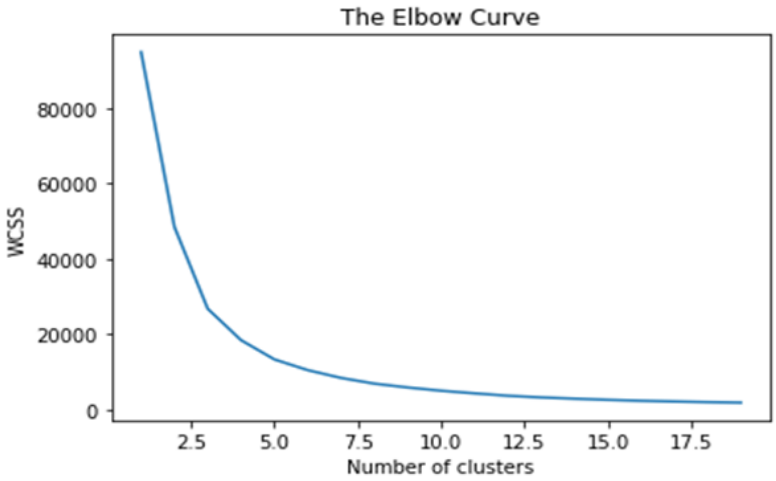

Fig.4 K-means elbow method

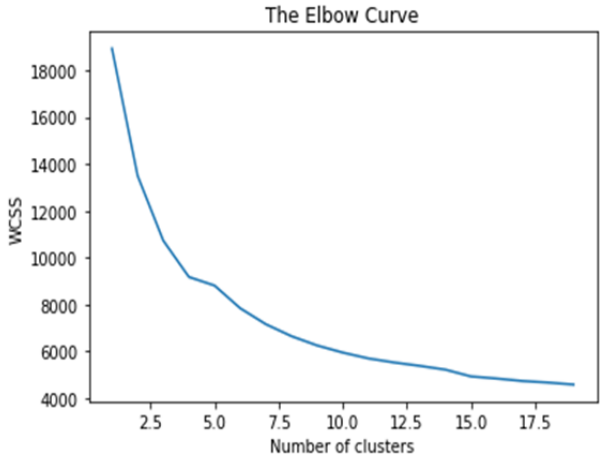

Fig.5 K-Medoid elbow method

From the above figures it can be seen that the number of clusters corresponding to the 'Elbow' in the graphs comes out to be 6 for both k-means and k-Medoid clustering. So, we need 6 clusters .Then, we performed both k-means and k-Medoid clustering. The scatter plots generated for the clusterings are shown in Figures 6 and 7.

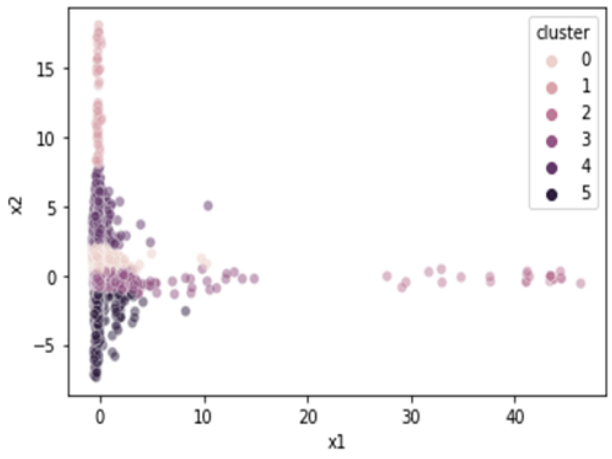

Fig.6 Scatter plot for k-means clustering

Step7: Model Evaluation:

Here we have compared the performance of both k-means and k-Medoids with The Calinski-Harabasz index, Davies-Bouldin Index, Silhoette Coefficients Average.

a**) The Calinski-Harabasz index** is commonly known as the Variance Ratio Criterion. It is the ratio of all clusters' between-cluster and inter-cluster dispersion. The higher the score, the better the performance [16]. The Calinski-Harabasz index for k-means is higher compared to k-Medoid as shown in Fig.8

	KMedoid	KMeans
Calinski-Harabasz Score	3061.351727	23621.005915

Fig .8 Calinski-Harabasz index for K-Medoid and Kmeans

b) Davies-Bouldin Index: A model with a lower Davies-Bouldin index has a better separation between the clusters. The lowest possible score is zero. A better partition is indicated by values that are closer to zero [17]. For k-means, the Davies-Bouldin Index is lower to k-Medoid as shown in Fig.9.

	KMedoid	KMeans
Davies-Bouldin Score	1.138118	0.542793

Fig.9 Davies-Bouldin Index for K-Medoid and K means

c) Silhoette Coefficients Average : Using Silhoette Coefficients Average value for silhouette coeffiencit. "-1" is the worst and "+1" is the optimal [18]. The Silhoette Coefficients Average for both k-means and k-Medoid are shown in fig.10

	KMedoid	KMeans
Avg Silhoutte Score	0.399168	0.454284

Fig.10 Silhoette Coefficients average for K-Medoid and K Means

From the above three evaluation metrics, it can be seen that k-means gives good results compared to k-Medoid. As such k-means can be confidently used for clustering headlines in our dataset.

Step8: Predicting the clusters of unseen headlines

8.1 Supervised Training

We now train the same news headlines with the 'Prediction' column using supervised classification algorithms like Logistic Regression, K-Nearest Neighbors**,** and Support Vector classifier and then choose the algorithm with the best accuracy. Two datasets namely NewsArticles dataset and ISOT dataset have been considered

Classifier	Accuracy	
	NewsArticles Dataset	*ISOT Dataset*
NB	85.85	87.93
KNN	72.61	74.09
SVM	88.32	89.68

Table1: Comparison of Accuracies for different classifiers on different datasets

From the table, it can be noted that the **Support Vector Machine** gives the highest accuracy of 94.9 %. Hence, we used SVM to train the dataset.

8.2 Predicting clusters for unseen headlines:

The table containing the unseen headlines, which were obtained by using the API key provided by Newsdata.io is shown in Fig.11.This is our test dataset.

In order to obtain accurate results, we combine both the test and train datasets and repeat pre-processing again. A pandas Data Frame is created with the unseen headlines and the predicted clusters. Then we map the cluster numbers to category numbers 1 to 6 for both train and test datasets to make the clustering more intuitive. This is shown in Fig.12

	Headlines	Sources	Prediction
0	PM Narendra Modi congratulates Indian archers ...	India TV	0
1	Samsung Galaxy M52 5G Camera Details Tipped, T...	NDTV	3
2	Dvara E-Dairy partners with IFFCO Tokio Genera...	Money Control	0
3	'Kumkum Bhagya' fame Zeeshan Khan reveals abou...	India TV	0
4	Google rolls out two new features to Messages ...	Times of India	3
5	Huawei to Offer Software Support for Honor Pho...	NDTV	0
6	AITA nominates Ankita Raina, Prajnesh for Arju...	India TV	0
7	T20 World Cup to be held from October 17 to No...	Times of India	3
8	T20 World Cup to be held from Oct 17-Nov 14 in...	India TV	3
9	Silver prices slide on dollar rebound; weaknes...	Money Control	4
10	Ranveer Singh to collaborate with Bear Grylls ...	The Indian Express	0
11	Tour de France 2021: Full schedule, dates, sta...	The Indian Express	3
12	T20 World Cup to be held from October 17-Novem...	The Indian Express	3
13	Puncch Beat 2 first impression: The curse of s...	The Indian Express	0
14	Sharwanand's 30th film titled Oke Oka Jeevitham	The Indian Express	0

Fig.11 Test Dataset

	Headlines	Sources	Source_code	Prediction	Category
0	Pipeline panic is preview of CYBER TAKEDOWN of...	Survival News	Survival	0	4
1	Infosys share price hits new record high; stoc...	Financial Express	FE	4	2
2	Fauci, Pfizer CEO, Big Tech Oligarchs, Chelsea...	News Punch	NP	3	3
3	COVID Vaccine Blood Clot Victims Demand Compen...	Great Game India	GGI	0	4
4	How Civilizations On Earth Will Function After...	All News Pipeline	ANP	0	4

Fig 12.Mapping of cluster numbers to category numbers

The description of the categories is given below:

Category 1:

The majority of the news is about the price of gasoline, gold, and silver. Fake news has a miniscule chance of being true. There are 53 headlines in the first category. Fuel, petrol, climbed, gold, silver, and dive are all commonly used words.

Category 2:

The likelihood of bogus news is slightly higher. The economy, financial markets, railways, airlines, automobiles, cellphones, and other topics dominate the headlines. There are a total of 798 headlines in this category. Market, nifty, bank, Samsung, train, Flip kart, and factory are all common terms.

Category 3:

There is a moderate chance that news is bogus. Health (primarily COVID-related) is a major topic in the news, as is the economy, infrastructure, and technology. This cluster also includes some political news headlines. There are 4089 headlines in this category. Bit coin, lockdown, COVID, vaccine, restrictions, unlock, projects, and sales are all common terms.

Category 4:

There's a good chance of coming across phony or altered news in this category. The majority of the headlines are about politics in India and around the world, as well as the COVID pandemic. This category encompasses the majority of news stories published in today's mainstream media in India and overseas. This category also includes propaganda and falsehoods that try to encourage people not to get the COVID vaccine. However, we may discover some sports and celebrity-related news here as well. This category contains a total of 7018 news headlines. COVID, vaccinations, sanitizer, actor, actress, tennis, match, Modi, Biden, hooligan, hostile, Leninism, stupidity, and insurgency are some of the most commonly used words.

Category 5:

There's a good chance of biased, distorted, or fraudulent news here as there is a lot of propaganda in this area. In this category, there are 2800 headlines. Modi, Rahul Gandhi, Congress, BJP, liberal, leftist, Muslim, Hindutva, riot, government, Pakistan, cult, scam, murder, and China are all commonly used terms.

Category 6:

This is a category with a lot of exceptions. This category contains 29 headlines. All of the headlines come from Alt-Market, a website infamous for spreading fake news. They are notices that the next issue of their newsletter will be released soon. As a result, we can disregard them.

Only the first, second, and third categories of news headlines are worth paying attention to. Except for sports, news headlines are often caustic and opinionated, and they are more likely to be forwarded on WhatsApp or shared on Facebook and Twitter. As a result, we must either take them with a grain of salt or disregard them.

4. CONCLUSION

In this paper, we tend to recognize fake headlines using two clustering algorithms k-means and k-Medoids. By recognizing the use of specific words, the K-Means groups the headlines into clusters. The TF-IDF value assigned to each word is proportional to how many times the term appears in a headline and inversely proportional to how many headlines contain the word. The quantity of news headlines with dubious reliability has increased, lowering the TF-IDF value of frequent words. As a result, on the K-Means scatter plot, the data points of those headlines are closer to the origin. Credible news headlines, on the other hand, are less in quantity, therefore the common terms have higher TF-IDF values. As a result, the data points in the scatter plot are further distant from the origin. Results show that k-means combined with SVM is quite good at recognizing fake news. Finally, employing machine learning to detect fake news is still a new and complex field. Despite the important findings of the proposed research, there is yet opportunity for improvement.

REFERENCES

[1] S. A. Abbas, A. Aslam, A. U. Rehman, W. A. Abbasi, S. Arif, and S. Z. H. Kazmi, "K-Means and K-Medoids: Cluster Analysis on Birth Data Collected in City Muzaffarabad, Kashmir," *IEEE Access*, vol. 8, pp. 151847–151855, 2020, doi: 10.1109/ACCESS.2020.3014021.

[2] L. P. Ashwini and M. R. Sunitha, "K-Means and K-Medoids Algorithms Comparision on TB Data," *Int. J. Eng. Res. Technol.*, vol. 5, no. 06, pp. 1–5, 2017, [Online]. Available: https://www.ijert.org/research/k-means-and-k-medoids-algorithms-comparision-on-tb-data-IJERTCONV5IS06025.pdf.

[3] R. Vankayalapati, K. B. Ghutugade, R. Vannapuram, and B. P. S. Prasanna, "K-means algorithm for clustering of learners performance levels using machine learning techniques," *Rev. d'Intelligence Artif.*, vol. 35, no. 1, pp. 99–104, 2021, doi: 10.18280/ria.350112.

[4] T. Evgeniou and M. Pontil, "Support vector machines: Theory and applications," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2049 LNAI, no. January 2001, pp. 249–257, 2001, doi: 10.1007/3-540-44673-7_12.

[5] H. Wang, Y. Shi, X. Zhou, Q. Zhou, S. Shao, and A. Bouguettaya, "Web service classification using support vector machine," *Proc. - Int. Conf. Tools with Artif. Intell. ICTAI*, vol. 1, pp. 3–6, 2010, doi: 10.1109/ICTAI.2010.9.

[6] A. A. T. Fernandes, D. B. F. Filho, E. C. da Rocha, and W. da Silva Nascimento, "Read this paper if you want to learn logistic regression," *Wang, Hongbing Shi, Yanqi Zhou, Xuan Zhou, Qianzhao Shao, Shizhi Bouguettaya, Athman*, vol. 28, no. 74, pp. 1/1-19/19, 2020, doi: 10.1590/1678-987320287406EN.

[7] B. T. Zewude and K. M. Ashine, "Binary Logistic Regression Analysis in Assessment and Identifying Factors That Influence Students ' Academic Achievement : The Case of College of Natural and Computational," *J. Educ. Pract.*, vol. 7, no. 25, pp. 1–6, 2016, [Online]. Available: https://eric.ed.gov/?id=EJ1115855.

[8] M. Reza, S. Miri, and R. Javidan, "A Hybrid Data Mining Approach for Intrusion Detection on Imbalanced NSL-KDD Dataset," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 6, pp. 1–33, 2016, doi: 10.14569/ijacsa.2016.070603.

[9] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2888, no. November 2012, pp. 986–996, 2003, doi: 10.1007/978-3-540-39964-3_62.

[10] A. S. Ahmar, D. Napitupulu, R. Rahim, R. Hidayat, Y. Sonatha, and M. Azmi, "Using K-Means Clustering to Cluster

Provinces in Indonesia," *J. Phys. Conf. Ser.*, vol. 1028, no. 1, 2018, doi: 10.1088/1742-6596/1028/1/012006.

[11] S. Chakraborty and N. K. Nagwani, "Analysis and study of incremental K-means clustering algorithm," *Commun. Comput. Inf. Sci.*, vol. 169 CCIS, no. 7, pp. 338–341, 2011, doi: 10.1007/978-3-642-22577-2_46.

[12] C. Wu Chunqiong Yan, "K -Means Clustering Algorithm and Its Simulation Based on Distributed Computing Platform," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/9446653.

[13] Y. Li and H. Wu, "A Clustering Method Based on K-Means Algorithm," *Phys. Procedia*, vol. 25, pp. 1104–1109, 2012, doi: 10.1016/j.phpro.2012.03.206.

[14] J. Newling and F. Fleuret, "K-medoids for K-means seeding," *Adv. Neural Inf. Process. Syst.*, vol. 2017-December, no. Nips, pp. 5196–5204, 2017.

[15] R. C. Balabantaray, C. Sarma, and M. Jha, "Document Clustering using K-Means and K-Medoids," 2015, [Online]. Available: http://arxiv.org/abs/1502.07938.

[16] https://medium.com/@haataa/how-to-measure-clustering-performances-when-there-are-no-ground-truth-db027e9a871c

[17] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html

[18] https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c