International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Machine Learning Based Ensemble Technique for Effective Clustering Of Registered Documents

K. Neelima[1], S.Vasundra[2]

Submitted: 06/06/2022 **Accepted:** 10/09/2022

Abstract: Data mining and machine learning techniques are very useful in different applications for performing predictions of useful patterns. Many Business Applications require the data to be prepared in a structured format so that it can help ease data validation, improve quality, performance, and handle exceptional data like Null Values, duplicates, unexpected data etc. Companies have applications that maintain various critical elements which require several mechanisms to present the data in the required format by applying several Business Rules. This work attempts to perform experimental evaluation of identifying an appropriate Business rule engine for data transformation of the critical element Document Number in the Land Registered documents by applying different data preprocessing techniques, like label encoding, one hot encoding, and Binary Encoding for data. Also, it aims to apply a clustering technique like K-Means clustering, to cluster the documents into buckets and Classify them into appropriate Labels. The distance measures such as Euclidean, Manhattan, Maximum, Binary, Minkowski and Canberra are used to calculate the number of inter and intra clusters. The appropriate clustering is derived using statistical techniques, namely, Elbow Curve Plot, Silhouette coefficient and ground truth labels. The clustering results are compared using a common metric called the Adjusted Rand Index(ARI). This work also applies the Principal Component Analysis (PCA) to confirm that the selected features are optimal. The proposed ensemble technique is evaluated and trained for effective derivation of clusters for Registered document numbers or a similar data set which contains mixed document number formats. The final objective of this work is to propose an unsupervised hybrid classification and clustering technique, which will enable users to identify and classify the appropriate business rules for any given data automatically.

Keywords: *Clustering, K-Means, Density, Machine Learning, Ensemble, DBScan, Outlier, Effective, Distance, Evaluation, Training, Testing, Cluster Center, Normalization, Statistics, Measures, Prediction, Registered document, Documents, ASCII Formats*.

1. Introduction

This section will present a brief introduction of various terminologies, techniques and other concepts applied in the proposed work. This section also provides a relative literature which is referenced for updating and applying to the proposed techniques. The first part of the introduction contains the chosen methodology and data preprocessing techniques. The second part contains a brief overview of the various clustering techniques considered and evaluated in the proposed work. The third part of this section provides a comprehensive review of various statistical measures and techniques used for comparing the effectiveness of the methods. Data Quality is of utmost importance for every domain and Registered document is not an exception.

Generally, the business rules are not known to the end user and would key the registered document number based on the image. Having the Raw data in the database may lead to many challenges in handling quality checking, performing various analytics, handling outliers, null values, invalid data etc. for the business. And thus, business would spend a lot in implementing the business rule for transforming the raw data into a business required formats. This may not be like a onetime effort as the data keeps changing always and it is a continuous effort. Thus, with the evolution of Machine Learning models, automatic business rule implementation on the raw data to transform the data into the business required format is going to save huge costs, Time, and resources in this ultra-competitive business world. Not just that, the model not only works for expected data but also effectively applies the most relevant business rules by itself. And, would also, unlearn in case of model not predicting the right rule.

To apply Machine Learning mechanism on the data, as a first step it requires to determine features that would play a vital role in predicting the result in this case it would be the formatted output or the business rule to be applied on the source pattern. Hence, a correction has been derived between each element and with the formatted output. In case, of multiple attribute correlation, PCA must be implemented. To confirm the quality of the given input data for further processing and analysis, there exists a need

to perform data preprocessing techniques such as encoding, standardization, and normalization.

Popular encoding techniques like Label, One Hot and Binary Encoding were evaluated for converting the features or variables of the source or input raw data set to a numeric form which can be easily processed by a machine learning or data mining algorithm. For example, similar features or labels present in the data set will be mapped into a uniform numeric identifier. Normalization is a technique for taking the varying range of values present in the data set to a standard scale for avoiding under fitting and over fitting problems and for improving the effectiveness of data processing. There are several normalization techniques which are useful for standardizing the given data. Some of the popular ones are Min-Max, Z-Score and Standard Scalar. This work applied the Standard Scalar technique for formatting the given document reference numbers.

For any unlabeled data, it is important to identify the similarity and dissimilarity between the featured data and the effective formulation of the number of clusters. In order to find the distance within the cluster and between the clusters it is required to find the similarity and dissimilarity matrix. The distance calculation can be done using several techniques, such as Euclidean, Manhattan, Maximum, Binary, Minkowski and Canberra. Euclidean distance is one of the simplest approaches based on the Pythagorean theorem; Manhattan distance uses the modulus values of two measures at the respective axis of the right angles. Varying numbers of clusters were identified and assigned to find the right combination and clustering of the tuples. Based on the data set and permutation and combination of similar patterns, the distance is calculated and assigned to the respective clustering group.

Clustering aids in the discovery of groups of things that are similar (or linked) to one another while being distinct (or unrelated) from those in other groups. High intra-class similarity and low inter-class similarity are produced by a successful clustering approach. The quality of a clustering result is determined by the method's similarity measure as well as its implementation. The ability of a clustering approach to discover some or all of the hidden patterns is also a criterion for its quality. The hierarchical and partitioned sets of clusters are fundamental distinctions among clustering types. Partitioned Clustering divides data objects into non-overlapping subgroups (clusters), with each data object belonging to exactly one subset.

A set of nested clusters is organised as a hierarchical tree in hierarchical clustering. A hierarchical cluster is a collection of nested clusters arranged in a hierarchical tree that can be seen as a dendrogram. A dendrogram is a tree-like structure that keeps track of merges and splits. Clustering is usually aimed to find groups related objects and also the objects which are unrelated to one another. For getting a better cluster, one should minimize the intra cluster distance (within the cluster data points) and also at the same time, aim to maximize the inter cluster distance (among the different cluster data points). The clusters can be formed using partitioning or the hierarchical approach.

Based on the approach and underlying technique, the clusters can be called as partitioned and hierarchical clusters. Hierarchical clustering can be either of the agglomerative type or divisive type. In partitioned clustering, for o number objects, C partitions of data will be constructed, so that each partition P represents a unique cluster and C <= o. To achieve good results and optimal number of clusters, K-means or K-medoids algorithms are considered. Each cluster is usually produced by the mean value of the data objects in the cluster in the K-means algorithm. Each cluster in the K-medoids algorithm is largely produced by one of the objects around the cluster's centre.

This work considered the application of a partitioned cluster and K-Means algorithm. The K-means algorithm is simple to construct and also very flexible to deal with different types of data. There are three main steps in the K-means algorithm. The first step is to take the mean value from the given data set or using a pseudo value which is similar to any of the values in the data set. The second step is to find the nearest number to select the mean value and partition the data set. The two steps will be continued until the same mean value is obtained.

2. Literature Review

This section presents a detailed literature review of the methods considered for performing the evaluation. Various research contributions on data mining, machine learning, clustering algorithms, distance measure calculation, deep learning and recent research directions were thoroughly analyzed and taken into consideration for the proposed work to fill the gaps.

Ravi et al, mentioned that the mining tool provides patterns of behavior, reflected in the data, to drive the accumulation of technical or process or business knowledge and the ability to foresee and shape future events. Existing legacy databases in organizations contain vast amounts of crucial data, such as customer / product / sales, sales statistics, etc. In the proposed work the existing database has domain related document with important features which are required for performing business intelligence using appropriate mining techniques [1].

Nair et al proposed a self-organizing map, an unsupervised neural network. SOM is used to generate a low dimensional grid from high-dimensional data without

affecting the topological order of the actual data. SOM is an outstanding tool in the data mining exploratory phase. SOM has been productively applied to a lot of fields like engineering as well as business domains. Investigational results show that the performance of clustering is enhanced with the newest approach of cloud computing [2].

Lidia et al, has presented how to develop a Background Knowledge by selecting a domain for each case and reducing the background Knowledge primitives. As a result, for autonomous data transformation, only domain-specific background knowledge and accompanying functions will be used [3].

Wolfgang Kratsch et al, have observed that Deep Learning outperformed in comparison with Machine learning for outcome based predictive modeling. It is also noted to look at a few essential characteristics such as the event-to-activity ratio, variant-to-instance ratio, and others that demonstrated the metrics, such as accuracy and F-score, for better Deep Learning-based works [4].

In the Title Insurance industry, Abhijit Guha et al [5] suggested a model employing Anomaly Detection by evaluating several classifications as Normal or Positive using high dimensional text data. This method combines the traditional one-class classification algorithm OSVM with the auto encoder, a deep learning-based, self-supervised, non-linear dimensionality reduction approach.

Rekha Nagar et al, gives a thorough overview of machine learning techniques, algorithms, and methodologies, as well as the usage of those algorithms and tools to implement various machine learning algorithms [6].

Sarker discussed several popular application areas based on ML techniques and discussed the key challenges and opportunities for applying ML algorithms for many application areas [7].

Alamuri et al provided a great comparison of the distance measures used for clustering and also suggested categorization of the data into context- free and context-sensitive to properly distinguish the differences [8].

Maclin, et al, suggested that ensemble techniques should use trained classified data sets for grouping some novel instances and could apply multiple methods like usage of decision trees along with neural networks or classification algorithms. Authors also emphasized that a single classifier may not be helpful and so accurate as to classify or cluster complex categorical data. Therefore, there exists a need to create an ensemble technique which is more accurate than a single classifier. [9].

Ganjarapalli et al, suggested the role of vector based classification and came up with a novel classifier using Boolean based naive factorization to improve the quality of location prediction. . Therefore, in order to improve the prediction, multiple document vectors and additional labeling can be used for proper clustering [10], [11], [12].

To address the dataset imbalance problem, Kaimuru et al. [11] colleagues created a heterogeneous ensemble model using the SMOTE - synthetic minority oversampling technique. The authors used Ada Boost and Random forest classifiers and used 10 fold stratified cross validation to validate the model. Apart from that, other measures like correlation, and information gain can be used to assess the proposed method. The authors suggested that it is important to remove irrelevant or redundant features that have an effect of reducing point-outliers, thus leading to improved classifier performance. This is one of the important considerations for the proposed work. Thus, this work considered the application of the PCA based technique as well as correlation to identify the right features which can provide the expected results.

The Artificial Neural Network-based strategy, according to Biswas, Saroj, et al. [13], [14] is a favoured approach because to its great performance. Using classed and misclassified data, the author presented a rule extraction approach using a neural network. The rules are derived from the neural network that has been trained. To determine relevant properties of the different classes, both classified and misclassified data ranges are used [15]. [16].

3. Proposed Hybrid Ensemble Method

Figure 1 represents the proposed methodology which supports effective rules classification using the machine learning technique. The keyed/raw data is first pre-processed and prepared by selecting the required associated attributes, applying data cleansing, and removal of duplicates, and then Patternized to bring the data into standardized representation.

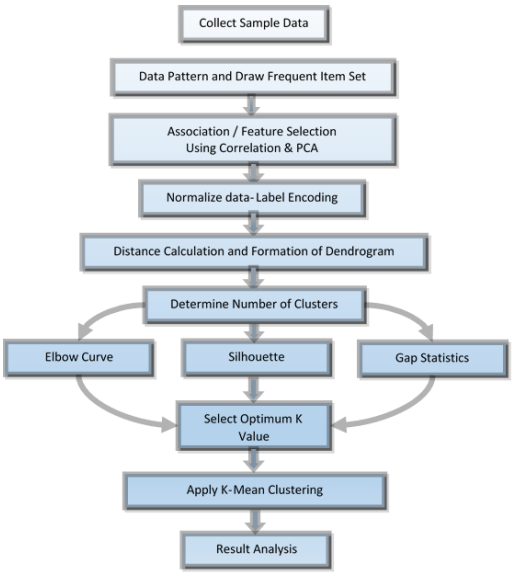

Figure 1: Proposed Ensemble Methodology

The Outliers and Data anomalies like missing information were identified in this layer and were reported to proceed further for research in advance, which can save time and efforts. The Patternized data is then classified using the Rule based method as a part of supervised learning. The Expected data are classified into the pertinent business rules in this method. Furthermore, For the labels partially known, some clustering techniques were used to form relative clusters or groups. KNN and Naïve Bayes clustering models and Decision Tree classification model have been applied to this data to identify the Label. For the unknown variables or features, there is a need to apply a neural network-based reinforcement and rewarding model.

3.1 Data Profiling

Registered document companies maintain the details available in the documents that are recorded or registered in the county, courthouse, and other register offices, like deed, Mortgage, assignment, judgement, divorce, lien etc. Applications would work effectively and facilitate easy maintenance, and data Manipulation when the data is stored in a structured manner in the Database. Though there are various elements in the documents like Party Names, Document Number, Recording Date, property details etc. The one critical element that keeps varying in course of time and location, and requires continuous modification in the business rules to format the raw data into a required structure is the Registered Document Number.

This defined structure is not the same across all locations. Hence, it is a bigger challenge and requires continuous efforts to implement new business logics to format the raw data into a defined format for the respective businesses based on various attributes. Thus, the need of the hour is for leveraging machine learning techniques and building a model that can automatically evaluate the apt business Rule whenever there is a change in the source format.

For this Research, the critical element (of Registered document that acts as a Key and requires continuous enhancements of the rule engine), which is nothing but the Document Number, has been chosen. The raw document numbers are extracted from the source using various data extraction methods like OCR (or) Keying. Since historical data was already available for the state and counties, as a first step, Samples from the Historical data for a few counties have been selected.

Since, the source and target document numbers are unique values, they are brought into a standardized format by Patternizing; i.e. replacing all the Numericals [1-9] by 'N' and Alphabets by 'A', while the special Characters remain the same. After Patternizing, a threshold value has been defined as 5% and the most Frequent Pattern set that qualifies this threshold value has been chosen to be part of the Document format dictionary (DFD). In Addition to it, the association between the various attributes of the sample data has been identified using the Apriori algorithm. The Location (Region), Registered Date, and Type Code (which represents the type of document e.g.: deed, mortgage) are recognized as the ones that are directly involved in identifying the defined format of the Document Number and thus become part of the DFD, making it a final dictionary which is ready to be used for further processes.

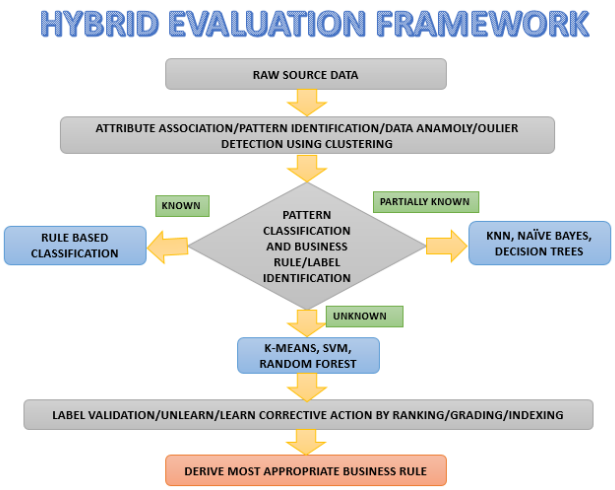

Figure 2: Hybrid Evaluation Framework

For the entered Source document pattern along with the obtained associated attributes, the registered date, Type code, and Region were used to find a match in the DFD to get the relevant target format, using the rule-based classification method. Clusters are formed for the data that doesn't fit the rule-based Model, using the K-Means algorithm. Also, a Decision Tree classification model has been built using the historical data by applying various encoding mechanisms to predict the output formatted pattern for any entered input raw registered document number along with the associated featured attributes and compare the accuracy.

For the experimental evaluation, the sample data has been selected by considering random registered documents of different regions. Two random documents per each month from the randomly selected region were selected from the oldest year from the historical database. The total number of records considered were 40K. The Number of columns were 8 and the key columns considered for the data analysis, clustering and classification are the raw registered document number.

The input data set has the following variables or features

1. REGION IN WHICH THE DOCUMENTS ARE REGISTERED
2. REGISTRATION DATE

3. TYPE CODE OF THE REGISTERED DOCUMENT
4. REGISTERED DOCUMENT NUMBER
5. REFERENCES
6. COMMENTS
7. PROPERTY DESCRIPTION
8. NAMES OF THE PARTIES

Among above eight attributes, 1-4 attributes are the relevant features that will be used by the model to predict or to cluster the data on which the output Document Number format is majorly dependent on. A Simple correlation has been applied in deriving these dependent Features. Using a Label encoder, the user entered the whose value has been converted to a label pattern to uniformly identify similar label patterns. This value will be used for forming clusters using the K-Means clustering algorithm. Using the sample data, four unique clusters have been formed which represent similar patterns based on the input data given.

Among the above 8 attributes, the first four attributes are found to be relevant, hence those relevant features will be used by the model to predict or to cluster the data on which the output Document Number format is majorly dependent. A Simple correlation has been applied in deriving these dependent Features. Using a Label encoder, the user entered value has been converted to a label pattern to uniformly identify similar label patterns. This value will be used for forming clusters using the K-Means clustering algorithm. Using the sample data, four unique clusters have been formed which represent similar patterns based on the input data given.

4. Experimental Evaluation

This section provides a detailed explanation of the experiments conducted using the data profiles for the insurance domain. Both R and Python programming tool kits have been used and the related packages and tools are used for performing many tasks mentioned in the previous sections such as data pre-processing, normalization, label encoding, distance calculation and cluster formation.

4.1 Encoding and Data Normalization Process

Since, the data that is being dealt with is non-Numeric and Categorical, it is required to encode it before applying the modelling mechanism. Label encoding, One Hot and Binary Encoding has been applied on a few features for better processing of data, using the machine learning code. For example, the document category with a pattern like N0NN0NN0NN is label coded as 770 and another one with a pattern like N0NN-0NNN00 is label coded as 391.Using a similar approach, the other three features were label encoded using sklearn APIs.

From the explorative data analysis, it is observed that label and One Hot Encoding has produced huge number of rows and column values compared to binary Encoding. Amongst three encoding techniques, an improved performance has been noticed with Binary encoding for decision tree Classification. Standardisation and Normalization are two important techniques which are applied to the given data as part of the data pre-processing to ensure that the data is ready and usable.

Normalization is a type of scaling method which normalizes the given data in the range of 0 and 1 or any other range. To normalize the data, one must use the data pre-processing package, such as MinMaxScalar or StandardScaler from the Machine Learning libraries. These methods are applied to a Registered document data set. The standard scalar method helps to standardise the unique features in the data set by scaling them to unit variance and removing the mean.

After normalizing, one of the document categories with 770 is normalized to 0.07826149. Similarly, the other document category with 391 is normalized to -0.813862. The next section illustrates one example of the application of the StandardScaler function on the input data and how the scaler fits the data in specific colors.

4.2 Distance Calculation and Cluster Formation

The Initial number of clusters can be assigned based on the business knowledge and by using the number of tuples and complexity of the document patterns. With the given sample data set of 2765 observations and based on the dendrogram, the data was clustered into four groups (K=4). The table below contains the number of tuples per cluster.

Table 1: Distance Comparison Table

METHOD	**GROUP-1 Tuples**	**GROUP-2 Tuples**	**GROUP-3 Tuples**	**GROUP-4 Tuples**
EUCLIDEAN	998	657	462	648
MANHATTAN	362	1173	268	962
MAXIMUM	759	954	469	583
MINKOWSKI	998	657	462	648
CANBERRA	1568	268	327	584

Cluster Formation - Using Elbow Curve method

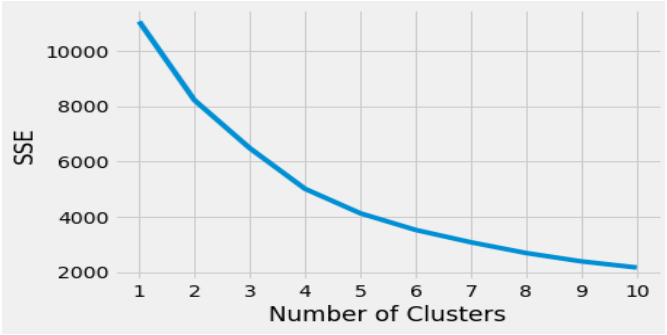

Figure 3: Cluster Formation using Elbow Curve with K-Means Clustering

From the above table, it is evident that either Euclidean or Minokowski provides the better distance calculation for the formulation of the number of clusters with the given input data. The Figure 3 depicts that effective number of clusters for the given sample dataset. The recommended cluster groups are within the range of 4 - 6. It is very imperative that effective data can be found using minimum four clusters and maximum six clusters.

Figure 4: Silhouette coefficient chart for K-MEANS clustering

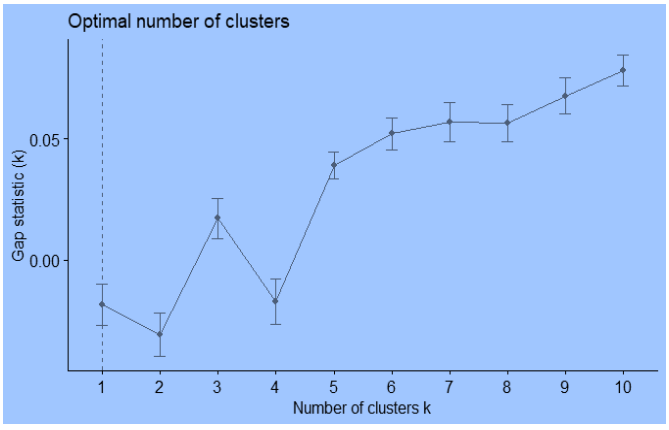

Figure 5: Cluster Formation using Gap Statistics Method from R

Figure 5 depicts that effective number of clusters for the given sample dataset and can be within the range of 4 - 7. The Gap statistics method is used to determine effective number of cluster. Figure 5 is the gap statistic diagram applied using R programming to confirm the statistical significance of the finding. As seen in the Figure 6 on the linearity with respect to average silhouette width, it is clear that K value can be from 4.

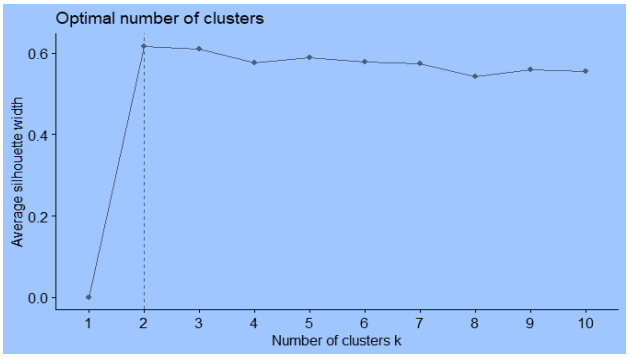

Figure 6: Optimal Number of Clusters - Using Silhouette Method

Without the use of ground truth labels, the elbow technique and silhouette coefficient evaluate clustering performance. Ground truth labels divide data points into categories based on a human's or an algorithm's classification. When used without context, these measures do their best to imply the correct number of clusters, but they can be deceptive. For best homogeneity within the group, choose the cluster value k which is lower within the sum of square values (withinss) and better heterogeneity between groups indicated by the higher between the sum of square values (betweenss). The following data has various cluster evaluation measures which would be helpful to decide on the number of best homogeneous clusters.

Intra/Inter Cluster similarity scores using k = 4

$ totss : num 11056

$ withinss : num [1:4] 2065 1098 1535 309

$ tot.withinss: num 5007

$ betweenss : num 6049

$ size : int [1:4] 1083 750 836 96

Intra/Inter Cluster similarity scores using k = 5

$ totss : num 11056

$ withinss : num [1:5] 1112 309 266 1041 1503

$ tot.withinss: num 4230

$ betweenss : num 6826

$ size : int [1:5] 666 96 369 813 821

Intra/Inter Cluster similarity scores using k = 6

$ totss : num 11056

$ withinss : num [1:6] 610 1512 303 563 414 ...

$ tot.withinss: num 3738

$ betweenss : num 7318

$ size : int [1:6] 553 662 95 388 446 621

Intra/Inter Cluster similarity scores using k = 7

$ totss : num 11056

$ withinss : num [1:7] 208 1052 309 566 461 ...

$ tot.withinss: num 3398

$ betweenss : num 7658

$ size : int [1:7] 71 754 96 538 315 354 637

A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. The total within the sum of squares (intra cluster similarity) of 3738 and between the sum of squares (inter cluster similarity) of 7318 seems to be optimal with 6 clusters for the given sample dataset. Even though the 7 cluster combination has a better similarity matrix, it introduces time and space complexity, so this evaluation continued with the 6 cluster formation. When we go for higher number for K value, it become difficult to label the data and would also increase the space and time complexity.

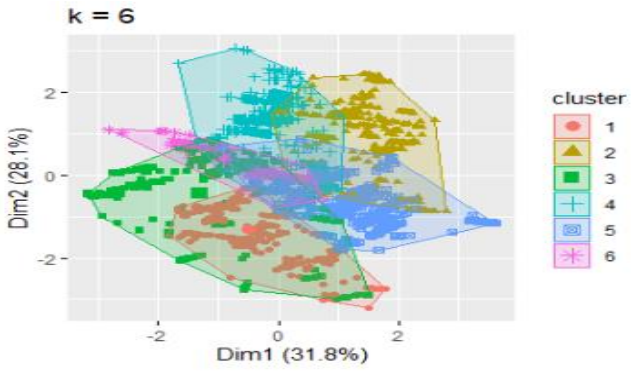

Figure 7: Cluster Plot Distribution Chart using PCA

Figure 7 portrays the virtualization classification plot. As there are more than two dimensions or features present in the data set, it is important to select the right dimensions or key factors for performing the cluster analysis. So, the **fviz_cluster** function in R has been used which will internally perform the Principal Component Analysis (PCA) for all the selected K-values (2 to 7). Dim 1 and Dim 2 in the plot show the first two principal components that explain the majority of the variance. The cluster plot above shows k =6 with high quality clusters having high intra-class similarity and low inter-class similarity. The total within the sum of squares (intra cluster similarity) of 3738 and between the sum of squares (inter cluster similarity) of 7318 seems to be optimal with 6 clusters. Based on the correlation analysis done on different attributes in the data set, it has been found that the region, date_encoded, and standardized pattern code provide the expected outcome for forming the clustering. For example, the suggested pattern code and region have a positive correlation value of 0.8 which has been considered as the dimension for forming the clusters.

Table 2: Clustering Groups and Document Patterns

Label	Group	Region	Document Category	Standardised Pattern Code	Date_Encoded	Suggested Pattern Code
TNVN	1	14.06456	590.8287	151.1568	449.0224	1422.3175
KL06	2	14.51282	344.6439	134.7379	375.3476	401.7407
TELAD	3	12.86853	789.8127	154.7849	1116.0199	1577.4622
MUM33	4	18.43340	1144.4015	152.5797	1728.9456	810.3039
KAR13	5	24.33935	652.5668	130.8899	1288.6606	456.1895
ASSA48	6	15.22713	940.0000	155.3155	1874.6782	1723.9621

Table 2 shows the classification and labelling of various data into multiple groups. The Clustering label was given based on the proximity and cluster means of the date encoded which represents the actual data keyed in by the users. There are six clusters formed which consist of multiple regions, different document patterns and system recognized patterns.

The following tables (Table 3 and Table 4) provide a sample list of cluster similarity with the closeness to data entered format in real time.

Table 3: Real time Mapping and Validation for Cluster-1

DATA ELEMENT	VALUE
REGION	MUM33
Raw Registered Document Pattern	AANN-00N000N
ML System Recognized pattern	AANN N000N
Actual user entered value is close to	KD19-0030002
ML System Recognized pattern value	KD19 30002

Table 4: Real time Mapping and Validation for Cluster-2

DATA ELEMENT	VALUE
REGION	KAR13
Document Entered pattern	N0NN - 0N0N00N
System recognized pattern	N0NN N0N00N
Actual userentered value is close to	3099 - 0304008
System Recognized pattern value	3099 304008

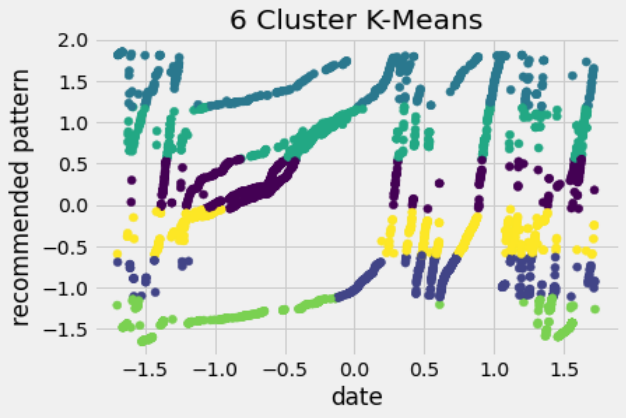

Figure 8: Clustering Using Date Vs Recommended Pattern

Figure 8 illustrates the cluster formation between two variables. The two dimensions (variables) Date Vs the Recommended code plot the data points according to the two components that explain most of the variance and classified as 6 clusters. Similar to these two dimensions the other related dimensions were analysed and plotted.

5. CONCLUSION AND FUTURE DIRECTIONS

This work not only proposed an ensemble technique which can be used for document clustering, but also performed an experimental evaluation by applying clustering algorithms like K-Mean. Prior to the application of the algorithm, different data preprocessing techniques, such as label encoding, one hot encoding, and standard scalar normalization were applied. A few distance measures were applied to identify the suitable distance measure which is valid for the documents with high variations.

The proposed ensemble technique is evaluated and trained with a good set of data to derive clusters for structured document numbers or a similar data set which contains mixed document number formats. This work concludes that an unsupervised hybrid clustering technique will be helpful to properly form the right groups for the given data set. For the given data and based on the experimental evaluation done using Python and R, it has been found that grouping of the data in clusters of six seems to provide the optimal classification of the data. Moreover, the high silhouette coefficient value suggests that it is better to proceed with K-Means clustering methodology.

It is always important to find and improve intra and inter cluster similarity scores for all the cluster groups for the related documents. One can choose the right cluster group, based on the lower intra cluster similarity score and higher inter cluster similarity scores. This work also applied three methods, namely Elbow curve, Silhouette coefficient as well as gap statistic methods to identify the optimal cluster value. Therefore, it is important to decide the optimal cluster value by using multiple evaluation methods, instead of going with random guess values or ad hoc values.

This work also concludes that getting statistical inferences and readability or interpretation results is much better in R studio when compared with Python IDE. From the experimental evaluation using both R and Python, it is imperative that, an optimal k-value should be selected based not only on the inter and intra similarity, but also on the business requirements and complexity of the data. More cluster groups could cause additional time and space complexity in the decision making process.

This proposed work applied the K-Means clustering algorithm because it is easier to use, efficient in execution, formulates tighter clusters, and is simple and better for centroid based documentation processing. In order to find the right features which, give improved accuracy, more detailed analysis of the features can be done using a Confirmatory factor analysis method like PCA. The future direction of this work would further focus on applying PCA and other feature engineering methods and neural network-based techniques. Once the cluster is formed, the clustered data, can be further taken and classified using the Decision Tree or Bayesian or SVM classifier for granular classification and performance. It is also important to try comparing different performance measures with the trained neural networks which can support automatic classification.

References

[1] Ravi Shankar, Sourish Acharia and Alok Baveja (2009), "Soft-system Knowledge Management Framework for New Product Development", Journal of Knowledge Management, Vol. 13, pp.135-153,

[2] Nair, S., & Mehta, J. (2011). Clustering with Apache Hadoop. Proceedings of the International Conference & Workshop on Emerging Trends in Technology - ICWET ’11, (Icwet), 505.

[3] Soni, D. K. ., M. , N. Kaushik, D. . Dhote, D. . Nigam, and K. G. . Krishna. “Website Redesign With Animation”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 2, Mar. 2022, pp. 01-10, doi:10.17762/ijritcc.v10i2.5499.

[4] Lidia Contreras-Ochando et al,(2020), "Automated Data Transformation with Inductive Programming and Dynamic Background Knowledge", Joint European Conference on Machine Learning and Knowledge Discovery in Databases

[5] Wolfgang Kratsch,Jonas , Manderscheid, Maximilian, Ro'glinger, Johannes Seyfried,(2020), "Machine Learning in Business Process Monitoring: A Comparison of

Deep Learning and Classical Approaches Used for Outcome Prediction",Business & Information Systems Engineering

[6] N. A. Libre. (2021). A Discussion Platform for Enhancing Students Interaction in the Online Education. Journal of Online Engineering Education, 12(2), 07–12. Retrieved from http://onlineengineeringeducation.com/index.php/joee/article/view/49

[7] Abhijit Guha Debabrata Samanta,(2020), "Hybrid Approach to Document Anomaly Detection:An Application to Facilitate RPA in Title Insurance",International Journal of Automation and Computing

[8] Rekha Nagar and Yudhvir Singh(2019), "A literature survey on Machine Learning Algorithms", Journal of Emerging Technologies and Innovative Research

[9] Sarker, I.H. Machine Learning: Algorithms(2021), Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160, https://doi.org/10.1007/s42979-021-00592-x

[10] Ghazaly, N. M. . (2022). Data Catalogue Approaches, Implementation and Adoption: A Study of Purpose of Data Catalogue. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(1), 01–04. https://doi.org/10.17762/ijfrcsce.v8i1.2063

[11] M. Alamuri, B. R. Surampudi and A. Negi(2014), "A survey of distance/similarity measures for categorical data," 2014 International Joint Conference on Neural Networks (IJCNN), pp. 1907-1914

[12] R. Maclin, D. Opitz (1999), Journal of Artificial Intelligence Research, Volume 11, pages 169-198

[13] Rudra Kumar, M., Rashmi Pathak, and Vinit Kumar Gunjan. "Diagnosis and Medicine Prediction for COVID-19 Using Machine Learning Approach." Computational Intelligence in Machine Learning. Springer, Singapore, 2022. 123-133

[14] Madapuri, Rudra Kumar, and P. C. Mahesh. "HBS-CRA: scaling impact of change request towards fault proneness: defining a heuristic and biases scale (HBS) of change request artifacts (CRA)." Cluster Computing 22.5 (2019): 11591-11599

[15] Ganjarapalli Manasa Divija Sree,S. Vasundra(2020). "Vector-Based ClassificationPrediction to Geographical Location", International Journal of Future Generation Communication and Networking.

[16] Dursun, M., & Goker, N. (2022). Evaluation of Project Management Methodologies Success Factors Using Fuzzy Cognitive Map Method: Waterfall, Agile, And Lean Six Sigma Cases. International Journal of Intelligent Systems and Applications in Engineering, 10(1), 35–43. https://doi.org/10.18201/ijisae.2022.265

[17] Kaimuru, Dalton & Mwangi, Waweru & Nderu, Lawrence. (2019). A Hybrid Ensemble Method for Multiclass Classification and Outlier Detection. International Journal of Sciences: Basic and Applied Research, Vol 45(1). 192-213.

[18] K, S., & srinivasulu, T. (2022). Design and Development of Novel Hybrid Precoder for Millimeter-Wave MIMO System. International Journal of Communication Networks and Information Security (IJCNIS), 13(3). https://doi.org/10.17762/ijcnis.v13i3.5096

[19] Chalapathi, M. M., et al. "Ensemble Learning by High-Dimensional Acoustic Features for Emotion Recognition from Speech Audio Signal." Security and Communication Networks 2022 (2022)

[20] M. N. Prasad* et al., “Reciprocal Repository for Decisive Data Access in Disruption Tolerant Networks,” International Journal of Innovative Technology and Exploring Engineering, 2019, 9(1), pp. 4430–443.

[21] Biswas, Saroj & Chakraborty, Manomita & Purkayastha, Biswajit & Roy, Pinki & Thounaojam, Dalton. (2017). Rule Extraction from Training Data Using Neural Network. International Journal of Artificial Intelligence Tools, World Scientific. 26. 10.1142/S0218213017500063