Industry Internship on

A Novel framework for Business Intelligence Based on Dimensionality Reduction and Clustering Techniques.

Course Name: *Industry Internship*

Course Code: *ETII100*

Madhushree Sannigrahi

Enrollment No.: A910119819007

B.Tech(Artificial Intelligence)[2019-2023]

Under the Supervision of **Dr. Soumya Sen**

Submitted to: Mr Saubhik Bandyopadhyay

A.K.Choudhury School of Information Technology, CU

Amity University Kolkata (AUK)

Certificate

On the basis of the report submitted by **Madhushree Sannigrahi**, student of B.Tech (AI), I hereby certify that the report **"A Novel framework for Business Intelligence Based on Dimensionality Reduction and Clustering Techniques."** which is submitted to the Department of Computer Science, Amity School of Engineering and Technology(ASETK), Amity University Kolkata (AUK)in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology(Artificial Intelligence) is an original contribution with existing knowledge and a faithful record of work carried out by her under my guidance and supervision.

To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Dr. Soumya Sen
Assistant Professor
A.K.Choudhury School of Information Technology
University of Calcutta (CU)

Declaration

I, **Madhushree Sannigrahi**, student of B.Tech (AI) hereby declare that the project titled **"A Novel framework for Business Intelligence Based on Dimensionality Reduction and Clustering Techniques."** which is submitted by me to the Department of Computer Science, Amity School of Engineering and Technology(ASETK), Amity University Kolkata (AUK), in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology (Artificial Intelligence), has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition. This project presents the research conducted by the author under the supervision of **Dr. Soumya Sen**.

The Author attests that permission has been obtained for the use of any copyrighted material appearing in the report other than brief excerpts requiring only proper acknowledgement in scholarly writing and all such use is acknowledged.

Madhushree Sannigrahi
B.Tech(Artificial Intelligence)[2019-2023]
Amity School of Engineering and Technology(ASETK)
Amity University Kolkata (AUK)

Abstract

Name of the authors: **Abhiamnyu Bhowmik, Madhushree Sannigrahi, Prithwish Guha**

Organization: **A.K.Choudhury School of Information Technology, University of Calcutta (CU)**

Title: **A Novel framework for Business Intelligence Based on Dimensionality Reduction and Clustering Techniques.**

Name of the supervisor: **Dr. Soumya Sen**

Month and year of submission: **May 2023**

Feature ranking and feature subset selection is an important aspect of machine learning. Identifying the most significant feature or subset of features of any high-dimensional dataset that plays major role in clustering enables faster computation. This article proposes a method to rank features or dimensions and recognize a feature subset that primarily contributes in clustering datapoints. We have used Jenks Natural Breaks algorithm [9] to cluster categorical feature and K-Means++ algorithm [10] to cluster all possible combinations of selected categorical features from feature matrix. Silhouette score is used to determine the performance of clustering results. Clusters of datapoints belonging to categorical feature are compared with that of numerical based on entropy to determine the finally selected important feature subset.

Acknowledgements

Words cannot express my gratitude to my mentor and senior scientist Dr. Soumya Sen for his invaluable patience and feedback. Without the assistance of Amity University Kolkata, which provided me with the knowledge and experience I needed to conduct the study in this specific area, I would not have been able to go on this adventure. Additionally, this endeavour would not have been possible without the generous support from the University of Calcutta (CU) who provided me with this research opportunity.

I am also grateful to the professors of Amity University Kolkata, especially Dr. Semati Chakraborty and my classmates from my University, for their help, feedback sessions, and moral support. Lastly, I would be remiss in not mentioning my family, especially my parents. Their belief in me has kept my spirits and motivation high during this process.

Contents

Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
JNB	Jenks Natural Break
SVM	Support Vector Machine
KNN	K-Nearest Neighbour
RBFR	Relevant-Based Feature Ranking
MIC	Maximal Information Coefficient
IWFS	Interaction Weight based Feature Selection algorithm

Chapter 1

Introduction

In the world of Big Data [13], working with datasets having huge number of features and extracting meaning information from them is a popular demand of many business-oriented applications. According to researchers and data analysts, it has been found that feature selection and feature extraction [12] is a very important aspect of machine learning.

Machine Learning [11] domain has witnessed vast increases in dataset sizes including both sample size and number of dimensions. Availability of more data gives analysts more opportunities to create better models with better accuracy. However, processing such huge data creates computational burden in any learning process. Large datasets with high number of dimensions often face problems of slow computation. This brings the need to identify features or dimensions which can be marked to be important over others while applying any machine learning algorithm. Certain business applications require analysis based on some selected subset of features. For example, a sales company looking for prospective customers based on their buying pattern can push new sales offer. Such type of study needs to identify only those features which can be used for customer segmentation with minimum cost and effort as sales dataset are usually bulky in size.

Feature Selection [1] aims to obtain a binary decision about which dimensions to keep, in the original data space. Apart from reducing dataset size by discarding some dimensions, the approach also improves the learning model. This enables better learning of the data distribution by using less but more meaningful dimensions with less noise. The benefit of learning on a subset of the available features is two- model fitting and prediction can be both faster and more accurate. To select relevant dataset features, a wide variety of strategies exist. One is to assign a scoring to each dimension and keep only the most relevant ones - such a ranking is called a feature ranking [1]. There are already existing

methods for feature selection using variable ranking. In this paper, we propose a novel method to select subset of features which proves to be playing more important role over others in the decision of clustering data points.

Classically dataset contains both categorical and numerical data. Features that are of categorical type typically spreads through a limited domain of values. For example, smoker/non-smoker, gender, marital status, employment status, etc are all cases of categorical features. On the other hand, numerical data spreads across wide domain of values. While applying any clustering algorithm of machine learning, categorical data always overpowers numerical data. In this research work, we have emphasized on treating the dataset separately based on these two above categories of features. This is an important aspect of any machine learning model which highly impacts the performance of learning.

1.1 Key Contributions

- Proposing a system using JNB for efficient feature selection with more weightage.
- Effective feature reduction to reduce computational complexity drastically while maintaining the integrity of the output.
- Helping businesses and other similar sectors to take profitable decisions by clustering the most significant attributes.
- Emphasizing prospective future directions.

1.2 Organization of the Report

The structure of the report is as follows: Chapter II incorporates the literature review of similar research works while Chapter III confers the theoretical background behind the project. Chapter IV and Chapter V consist of the methodology and experimentation on different cases whereas, Chapter VI concludes the paper and highlights future research directions. Finally, Chapter VII includes the references vital for this research work.

Chapter 2

Literature Review

The domain of feature ranking and feature selection has a huge availability of literature, distributed over many subtopics. Feature selection is a pre-processing technique that identifies the key features of a given problem. [2] mentions the various feature selection techniques including Filter-based feature selection, Wrapper-based feature selection and embedded technique.

In a paper by Jasti et al. [3], a novel Relevant-Based Feature Ranking (RBFR) algorithm is proposed which identifies and selects smaller subsets of more relevant features in the feature space. The performance of the RBFR is also compared against other existing feature selection methods such as balanced accuracy measure, information gain, Gini index, and odds ratio on 3 datasets, namely, 20 newsgroups, Reuters, and WAP datasets. In this paper, 5 machine learning models (SVM, NB, KNN, RF, and LR) are used for testing and evaluation resulting in 25.4305% times more effective than the existing feature selection methods in terms of accuracy. Sasikala et al. [4] in their paper propose an adaptive feature selector based on game theory and optimization approach for an investigation on the improvement of the detection accuracy and optimal feature subset selection. An extensive experimental comparison of 22 benchmark datasets confirms that the proposed SVEGA strategy is effective and efficient in removing irrelevant and redundant features.

In this paper [5], a novel framework is proposed for selecting relevant features in supervised datasets based on a cascade of methods where speed and precision are in mind. This framework consists of a novel combination of Approximated and Simulate Annealing versions of the Maximal Information Coefficient (MIC) to generalize the simple linear relation between features. Sensitivity analysis is a popular feature selection approach employed to identify the important features in a dataset. This study [6] proposes a novel approach that

involves the perturbation of input features using a complex step. The implementation of complex-step perturbation in the framework of deep neural networks as a feature selection method is provided in this paper, and its efficacy in determining important features for real-world datasets is demonstrated.

Interacting features are those that appear to be irrelevant or weakly relevant to the class individually, but when it is combined with other features, they may highly correlate to the class. In this paper [7], the authors have proposed a novel feature selection algorithm considering features. Firstly, feature relevance, feature redundancy and feature interaction have been redefined in the framework of information theory. Then the interaction weight factor which can reflect the information of whether a feature is redundant or interactive is proposed. Finally, an Interaction Weight based Feature Selection algorithm (IWFS) is introduced.

A fast clustering-based feature selection algorithm, FAST, is proposed by Song et al. in their paper [8]. The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features.

Chapter 3

Theoretical Background

3.1 The Jenks natural breaks

The Jenks natural breaks algorithm is a method of data classification that is designed to minimize the variance within each class, while maximizing the differences between classes. This algorithm is often used to group data into a specified number of classes, known as "natural breaks," in order to create more interpretable and visually appealing maps.

To use the Jenks natural breaks algorithm, the user must first specify the number of classes into which they want to group the data. The algorithm then finds the class breaks that minimize the variance within each class and maximize the differences between classes. This is typically done using a heuristic search algorithm, such as the one proposed by Jenks and Caspall in 1971.

Once the class breaks have been determined, the data can be grouped into the specified number of classes using the natural breaks as the class boundaries. This can be useful for creating thematic maps, where different classes are represented by different colors or symbols on the map.

Overall, the Jenks natural breaks algorithm is a popular and effective method of data classification that can help to create more interpretable and visually appealing maps.

3.2 Silhouette Score

The silhouette score is a metric that can be used to evaluate the performance of a clustering algorithm. It measures how well each data point has been classified, with a higher score indicating a better classification.

To compute the silhouette score, we first need to calculate the silhouette coefficient for each data point. The silhouette coefficient is defined as the difference between the average distance to all other points in the same cluster (the intra-cluster distance) and the average distance to points in the next closest cluster (the inter-cluster distance), divided by the maximum of these two distances.

In other words, the silhouette coefficient for a point measures how well-matched that point is to the other points in its cluster, compared to the points in other clusters. If the silhouette coefficient for a point is high, it means that the point is well-matched to the other points in its cluster, and is very different from the points in other clusters. If the silhouette coefficient for a point is low, it means that the point is not well-matched to the other points in its cluster, and may be more similar to the points in other clusters.

The silhouette score is then calculated by taking the mean of the silhouette coefficients for all data points. This gives us a single score that can be used to evaluate the overall performance of the clustering algorithm.

In summary, the silhouette score is a measure of how well-matched the data points are to the clusters they have been assigned to. A high silhouette score indicates that the data points have been well-classified, while a low silhouette score indicates that the data points may not have been well-classified.

3.3 K-means

K-means is a type of unsupervised machine-learning algorithm used for clustering. It works by dividing a dataset into a specified number of clusters (k) based on the similarity of the data points.

1. The algorithm begins by randomly initializing k centroids, which are the center points of the clusters. Then, it iteratively performs the following steps until convergence.

2. Assign each data point to the cluster with the nearest centroid. Recalculate the centroids of each cluster by taking the mean of all the data points in the cluster.

These steps are repeated until the centroids no longer move, at which point the algorithm has converged and the clusters are considered to be finalized.

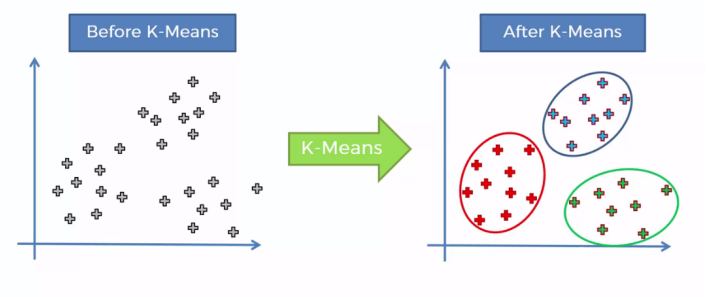

FIGURE 3.1: Entropy.

One of the advantages of the k-means algorithm is that it is relatively simple and easy to implement. It is also computationally efficient, making it a popular choice for clustering large datasets. However, the final clusters produced by the algorithm can be sensitive to the initialization of the centroids, and the algorithm may not always converge to the same solution.

3.4 K-means++

K-means++ is an algorithm for initializing the centroids in the K-means clustering algorithm. It is an improvement over the standard random initialization technique, which can sometimes result in poor performance.

The K-means++ algorithm works by first selecting a random centroid, and then repeating the following steps until all K centroids have been chosen:

Compute the distances between all data points and the current set of centroids Select the next centroid as the data point that is farthest from the current set of centroids Repeat until K centroids have been chosen By using this approach, the K-means++ algorithm ensures that the initial centroids are chosen in a way that will help the K-means algorithm converge quickly and produce good results. It is a widely used technique for improving the performance of K-means clustering.

3.5 Entropy

In the context of machine learning, entropy is a measure of the uncertainty or randomness in a dataset. It is often used in the field of information theory, where it measures the amount of information that is contained in a message or dataset.

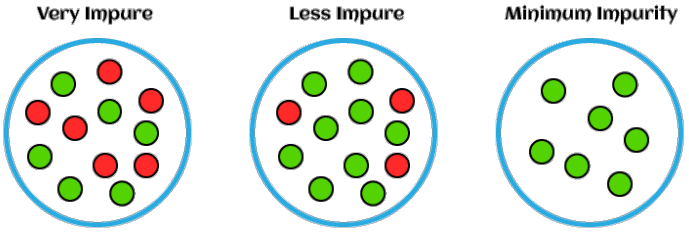

FIGURE 3.2: Entropy.

It is often used in decision tree algorithms. Decision trees are a type of machine learning algorithm that works by creating a tree-like model of decisions and their possible consequences. The goal of a decision tree is to make predictions about the target variable based on the values of the input features.

When information is processed in the system, then every piece of information has a specific value to make and can be used to draw conclusions from it. So if it is easier to draw a valuable conclusion from a piece of information, then entropy will be lower in Machine Learning, or if entropy is higher, then it will be difficult to draw any conclusion from that piece of information.

Entropy can be calculated using the following equation:

$$E = -\sum_{i=1}^{N} P_i log_2 P_i$$

In summary, entropy is a measure of uncertainty or randomness in a dataset, and it is often used in decision tree algorithms to help make decisions about which features to split on. A high entropy indicates that the data is unpredictable, while a low entropy indicates that the data is more predictable.

Chapter 4

Methodology

This section goes into further detail on the method used in this research project, which is depicted in Figure 4.1. the entire methodology can be divided into 5 sections. The first section discusses the datasets implemented over the model. In the second phase, the data is pre-processed, standardised, and split into numeric and categorical features. The third phase involves the k-means++ model along with Jenks Natural Break (JNB) for the clustering of numeric data samples. The features are ranked and sorted into a feature matrix in section 4. Finally, the categorical data is mixed with the best-ranking numeric feature and the final clusters are obtained.

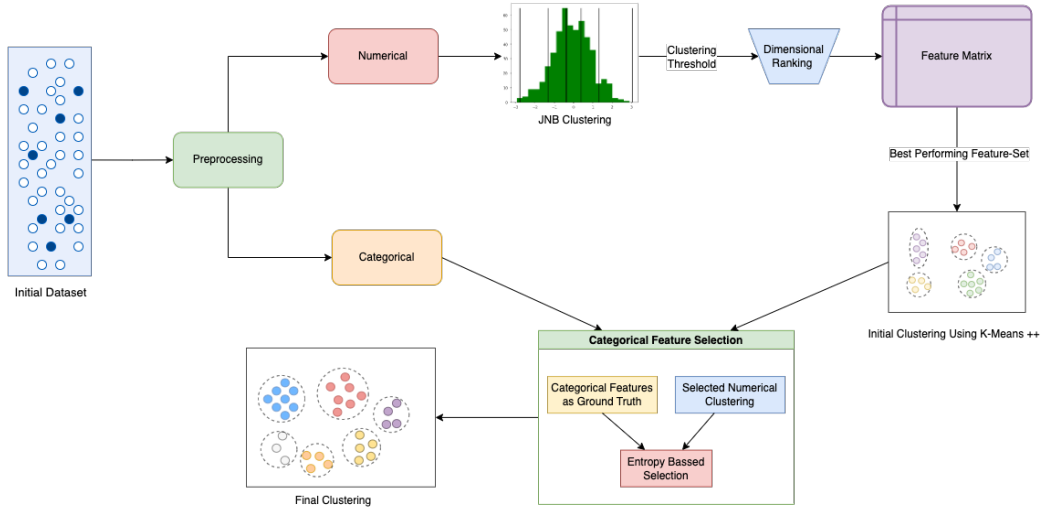

FIGURE 4.1: Overall view of the proposed model

4.1 Data Preprocessing

This stage involves processing the raw dataset to make it match the model. Firstly, all NaN values and irrelevant features are eliminated. The dataset is then normalised using LogTransformation to minimize its skewness. The data is then divided into numerical and category components. To do this, only characteristics with more than 10 distinct values were designated numerical features, while all other features were regarded as categorical. The resulting dataset is much more suited for further processing by the machine learning model.

4.2 Dimensional Ranking Using JNB:

4.2.1 Jenks Natural Breaks

The Jenks optimization method, also known as the Jenks natural breaks classification method, is a data clustering technique created to identify the ideal categorization of values into classes. This is achieved by aiming to maximise each class's divergence from the averages of the other classes while minimising the average deviation of each class from the class mean. To put it another way, the technique aims to increase the variation between classes while minimising the variance within classes.

4.2.2 Silhouette Score

Using JNB clustering, each numerical characteristic in the dataset is independently clustered. The features are ordered according to their scores for their silhouettes. JNB can produce both positive and negative silhouette scores. As a result, the user sets a threshold value (t = 0, by default) to identify the crucial traits based on the ranking. The threshold is expected to be greater than zero since points with negative silhouette ratings are better matched to nearby clusters than they are to their allocated cluster.

4.3 Feature Matrix:

All potential combinations from the chosen attributes are calculated and placed into a feature matrix for easy comparison. The matrix layout is described below.

If n (number of features) = odd, the features are divided into rows and columns,
where, the number of rows $(r) = (n+1)/2$
and the number of columns $(c) = (n-1)/2$

For example, if there are 3 features,(say A, B, and C), the number of rows would be 2 and number of columns would be 1 as shown in table 4.3.

	1	A	B	AB
1	1	A	B	AB
C	C	AC	BC	ABC

Here, we calculate,
$(A+1)(B+1) = 1 + A + B + AB$ and $1 + C$

On the other hand, if n = even, number of rows $(r) = n/2$ and number of columns $(c) = n/2$

For instance, if there are 4 features,(say A, B, C, and D), the number of rows and columns, would both be 2 as portrayed in table 4.3.

	1	A	B	AB
1	1	A	B	AB
C	C	AC	BC	ABC
D	D	AD	BD	ABD
CD	CD	ACD	BCD	ABCD

Following that, we calculate,
$(A+1)(B+1) = 1 + A + B + AB$ and
$(C+1)(D+1) = 1 + C + D + CD$

The silhouette scores are entered into the appropriate cells after the feature matrix has been created. The specific feature that would be most suited to represent the provided dataset is then selected from this.

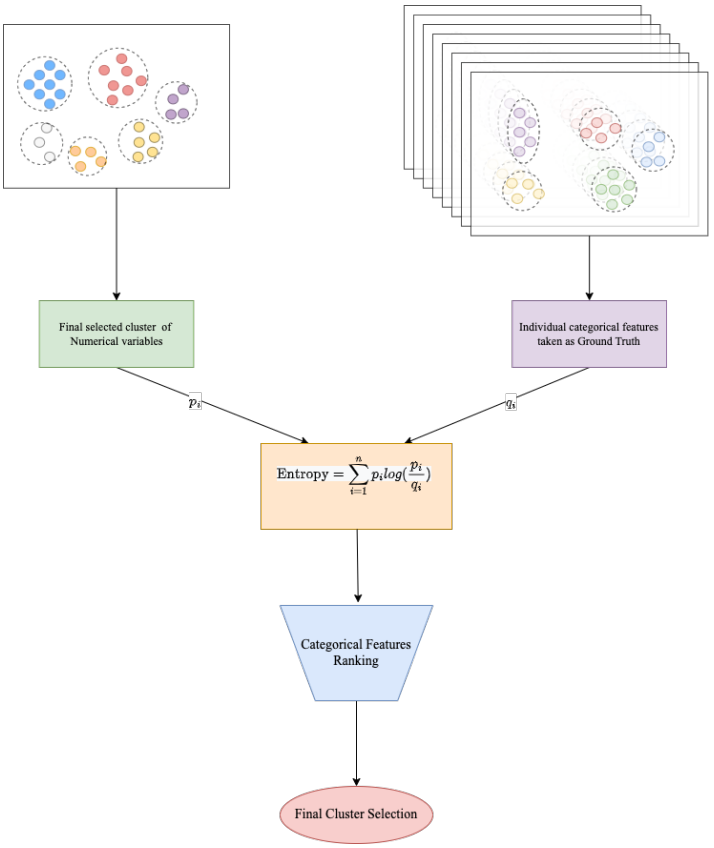

FIGURE 4.2: Categorical feature selection framework.

4.4 Categorical Feature Selection:

Based on entropy we have selected features that will most likely match our numerical features. As entropy is a measurement of disorders, we can use it to represent similarity by checking entropy between two clusters. If the clusters are similar they will generate less entropy compared to dissimilar clusters. Firstly, we consider individual categorical features as ground truth values and compare them against our numerically selected features. All the categorical features are considered individually and using the entropy distribution as in the equation below, their cluster entropy is calculated. Here, p = selected numerical cluster and q = categorical ground truth values. These entropy values are then ranked and the one with the topmost value is considered for the final clustering.

$$Entropy = \sum_{i=1}^{n} p_i log(\frac{p_i}{q_i})$$

Chapter 5

Case Studies

5.1 Customer Personality Analysis

Customer personality analysis is a thorough analysis of a business's ideal clients. It makes it simpler for businesses to adapt their goods to the unique wants, habits, and concerns of various consumer types. It also helps businesses better understand their clients.

For instance, a firm may assess which customer group is most likely to purchase the product and then promote the product only to that specific segment rather than investing money to

FIGURE 5.1: Elbow Method implemented over the Customer Personality dataset

The Silhouette scores of JNB Clustering based on the features:

	Features	Silhouette Score
0	NumDealsPurchases	-0.202392
1	NumWebVisitsMonth	-0.134018
2	Age	-0.112419
3	Gold	-0.0808994
4	NumWebPurchases	-0.0707335
5	Recency	-0.0484665
6	NumCatalogPurchases	-0.0376564
7	NumStorePurchases	-0.030623
8	Sweets	0.00230335
9	Fish	0.00653525
10	Spent	0.0159493
11	Fruits	0.0510502
12	Wines	0.0572965
13	Meat	0.0633605
14	Income	0.536214

market a new product to every consumer in the company's database. The dataset contains about 10,000 sample data with 29 attributes each. When JNB clustering is implemented over each feature of this dataset, the resultant silhouette score of top 5 attributes.

5.2 Medical Cost Personal Dataset

Medical insurance is a type of insurance coverage that covers the cost of medical expenses, including preventative care, diagnostic tests, and treatment for illnesses and injuries. It is an important part of a person's overall health care plan, as it can help to protect individuals and families from the financial burden of unexpected medical expenses. There are many different types of medical insurance, including employer-sponsored plans, individual plans, and government-sponsored programs like Medicare and Medicaid. The Medical Cost Personal Dataset consisted of around 7000 samples with 10 attributes such as age, smoking, previous medical instances, etc. The silhouette score of individual features based on JNB clustering is given in the figure.

The Entropy based on different subsets features:

	field	Entropy
0	Is_Parent	0.0665407
1	Complain	0.0682251
2	AcceptedCmp2	0.0708515
3	Kidhome	0.0809368
4	AcceptedCmp3	0.0864638
5	Children	0.0869668
6	AcceptedCmp4	0.0902929
7	Family_Size	0.0913386
8	AcceptedCmp1	0.0950207
9	AcceptedCmp5	0.100058
10	Response	0.111169
11	Teenhome	0.113369
12	Education	0.116706

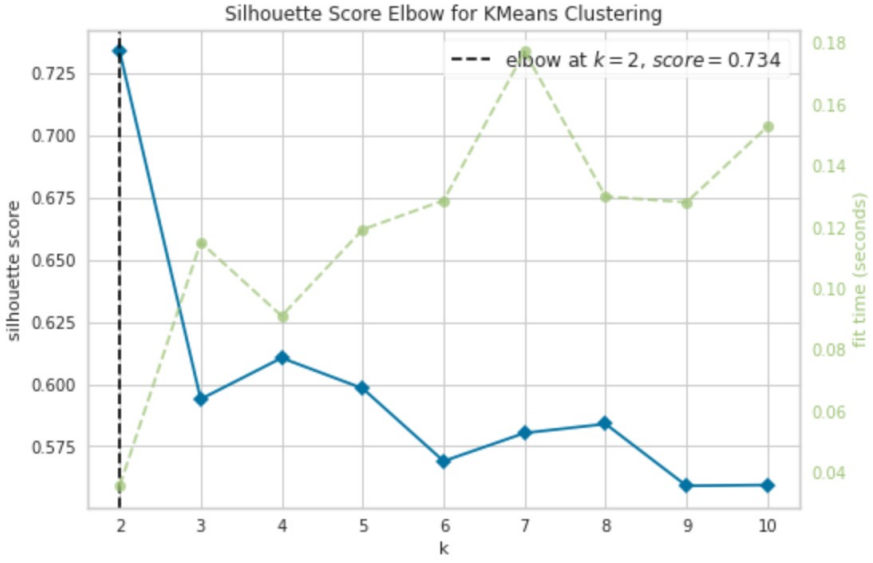

FIGURE 5.2: Elbow Method implemented over the Medical Cost Personal dataset

The Silhouette scores of JNB Clustering based on the features:

	Features	Silhouette Score
0	bmi	-0.212088
1	age	-0.0374214
2	charges	0.569111

The Entropy based on different subsets features:

	field	Entropy
0	children	0.278469

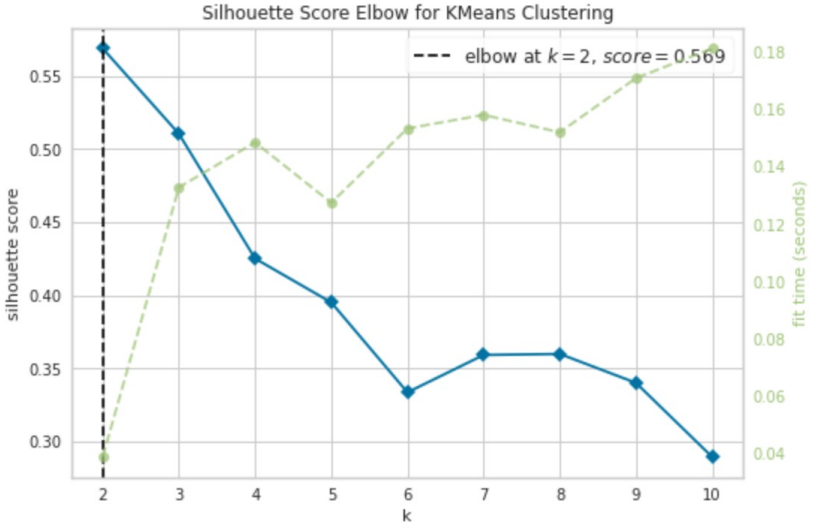

FIGURE 5.3: Elbow Method implemented over the Diabetes dataset

5.3 Diabetes Dataset

Diabetes is a chronic condition that affects the way the body processes blood sugar, or glucose. Glucose is the main source of energy for the body's cells, and it comes from the food we eat. In people with diabetes, the body either doesn't produce enough insulin (a hormone that helps regulate blood sugar) or doesn't effectively use the insulin it does produce. This can cause high levels of blood sugar, which can lead to a range of complications if left untreated.

The "Diabetes dataset" from the National Institute of Diabetes and Digestive and Kidney Diseases is utilised in this work. Based on diagnostic criteria, it is intended to identify

The Silhouette scores of JNB Clustering based on the features:

	Features	Silhouette Score
0	BMI	-0.146072
1	BloodPressure	-0.101593
2	Glucose	-0.086264
3	Pregnancies	-0.0625722
4	DiabetesPedigreeFunction	-0.0575446
5	Age	-0.0296686
6	SkinThickness	-0.00540127
7	Insulin	0.389149

The Entropy based on different subsets features:

	field	Entropy
0	Outcome	0.11173

whether a patient has diabetes. With 10 features and 768 occurrences, the implemented Diabetes dataset is a small portion of a much larger dataset. Female Pima Indians who are at least 21 years old make up all of the patients.

Chapter 6

Final Remarks

6.1 Discussion

The only problem is that it takes a huge amount of time(sometimes hours) to try every possible combination of the variables and the Goal for the dimensionality reduction is to create the clusters faster and more accurately and also finding out the main variables(key-features) responsible for creating this different Groups(Clusters). So, if it takes this much time, it's all for nothing. So, we have to find a way to reduce the time it takes to run the code. Maybe, considering all the combinations isn't necessary, only some will do, in that case, we have to find a way to check which ones are more essential. Or, maybe deducting the Categorical variables may be the Answer. Or, Maybe we have to calculate the categorical and non-categorical variables differently and then find a way to combine them.

6.2 Conclusion

This study developed a novel framework for effective dimensionality reduction. This novel methodology aims to create a positive impact on the decision-making process by facilitating data-driven decisions that are more reliable and robust. It also provides an efficient way to extract useful patterns of information from large datasets, which may be potentially beneficial for a range of businesses. This study tries to identify the most significant attributes or combinations of attributes for profitable business decisions. Overall, the proposed methodology successfully reduced the dimensions drastically without compromising

the quality of the final outcome. The model uses Jenks' Natural Break to sophistically cluster the more weighted attributes for overall better decision-making.

6.3 Future Works

1. The model is primarily addressing the clustering data problem as it is the most prominent use case in the business sector. It can easily be extended into other domains as well.

2. An interactive web application can also be a major step in making the proposed methodology user-friendly for the non-technical business owners.

3. Federated learning can be implemented for the privacy protection of its users by which the learning model can improve from individual clients' data feedback.

4. Implementing the system in a cloud platform like GCP can make it more accessible to a wide range of users while still keeping low resource costs.

Chapter 7

Bibliography

1. Overschie, J. G. (2022). A novel evaluation methodology for supervised Feature Ranking algorithms. arXiv preprint arXiv:2207.04258.

2. Khaire, U. M., & Dhanalakshmi, R. (2019). Stability of feature selection algorithm: A review. Journal of King Saud University-Computer and Information Sciences.

3. Jasti, V., Kumar, G. K., Kumar, M. S., Maheshwari, V., Jayagopal, P., Pant, B., ... & Muhibbullah, M. (2022). Relevant-based feature ranking (RBFR) method for text classification based on a machine learning algorithm. Journal of Nanomaterials, 2022.

4. Sasikala, S., alias Balamurugan, S. A., & Geetha, S. (2017). A novel adaptive feature selector for supervised classification. Information Processing Letters, 117, 25-34.

5. Garcia-Ramirez, I. A., Calderon-Mora, A., Mendez-Vazquez, A., Ortega-Cisneros, S., & Reyes-Amezcua, I. (2022). A Novel Framework for Fast Feature Selection Based on Multi-Stage Correlation Measures. Machine Learning and Knowledge Extraction, 4(1), 131-149.

6. Naik, D. L. (2021). A novel sensitivity-based method for feature selection. Journal of Big Data, 8(1), 1-16.

7. Zeng, Z., Zhang, H., Zhang, R., & Yin, C. (2015). A novel feature selection method considering feature interaction. Pattern Recognition, 48(8), 2656-2666.

8. Song, Q., Ni, J., & Wang, G. (2011). A fast clustering-based feature subset selection algorithm for high-dimensional data. IEEE transactions on knowledge and data engineering, 25(1), 1-14.

9. Khamis, N., Sin, T. C., & Hock, G. C. (2018, December). Segmentation of residential customer load profile in peninsular Malaysia using Jenks natural breaks. In 2018 IEEE 7th international conference on power and energy (PECon) (pp. 128-131). IEEE.

10. Bahmani, B., Moseley, B., Vattani, A., Kumar, R., & Vassilvitskii, S. (2012). Scalable k-means++. arXiv preprint arXiv:1203.6402.

11. Zhou, Z. H. (2021). Machine learning. Springer Nature.

12. Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (Eds.). (2008). Feature extraction: foundations and applications (Vol. 207). Springer.

13. Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In 2013 international conference on collaboration technologies and systems (CTS) (pp. 42-47). IEEE.