An Automated Python Script for Data Cleaning and Labeling using Machine Learning Technique

Matthew Abiola Oladipupo[1], Princewill Chima Obuzor[2], Babatunde Joseph Bamgbade[3], Kazeem M. Olagunju[4], Abidemi Emmanuel Adeniyi[5], Sunday Adeola Ajagbe [6*]
[1,2]Department of Data Science School of Science, Engineering and Environment, University of Salford, UK.
[3]Federal College of Forestry (FRIN), Jericho GRA, Ibadan, Nigeria.
[4,6]Department of Computer Engineering, Ladoke Akintola University of Technology LAUTECH, Ogbomoso, Nigeria
[5]Department of Computer Sciences, Precious Cornerstone University, Ibadan, Nigeria.
[6*]Department of Computer & Industrial Production Engineering, First Technical University, Ibadan, Nigeria
E-mail: m.a.oladipupo@edu.salford.ac.uk[1], p.obuzor@edu.salford.ac.uk[2], tundebamgbade@yahoo.com[3], kmolagunju@student.lautech.edu.ng[4], adeniyi.emmanuel@lmu.edu.ng[5], sunday.ajagbe@tech-u.edu.ng[6a*]

Keywords: data cleaning, information science, machine learning, financial dataset, automation, python script

Received: October 31, 2022

Every employee in the company who deals with data needs to have clean, noise-free data. Since data warehouses store and update enormous amounts of data from several sources, there is a potential that some of those references may contain inaccurate data. Due to the noise, inefficacy, and poor characterization of the vast amount of accessible data, as well as the ensuing insensitivity and inefficiencies of human data cleaning and labeling, the presentation of the data has become ambiguous, and the assessment of the information has become difficult. A hole in the creation of a better data analysis method was identified. This helped to guide the creation of a Python script for automatically cleaning and labeling data. The first step in the strategy used in this study to accomplish its goals and objectives was to obtain a financial dataset from the top database, "Kaggle". Create a machine learning (ML) approach in Python that intends to automate the financial dataset cleaning. This covers ingesting data, addressing incomplete data, addressing anomalies, one-hot wrapping and label encoding, extracting date and time values, and data normalization. Implementing an unsupervised machine learning method that attempts to automate financial dataset labeling (k-means). Using the method includes the elbow principle, k-means clustering, data modeling of "age" versus "arrival," dimensionality reductions, computer vision, and dataset categorizing using the groupings. An *empirical assessment of the cleaned and labeled automated trading dataset utilizing a comparison of the cleaned dataset before and after PCA adoption. The results show that the developed ML technique not only improved the performance of the audit data used in this study, but also classified the data after cleaning it and removing the unpleasant section and incomplete data, as shown by the k-means segmentation result and grouping by PCA.*

Povzetek: Razvili so skripto v Pythonu za avtomatsko čiščenje in označevanje finančnih podatkov ter podatke uporabili za strojno učenje za avtomatizacijo postopka.

1 Introduction

To prevent reaching the wrong conclusions, data cleaning is carried out to ensure the data is accurate. Data cleansing is an essential step in every operation using data. To enhance the outcomes of data mining, data purification is necessary. In a similar manner, data labeling guarantees that the dataset is accurately described. Firms are finding it less difficult to collect and retain enormous volumes of data. These huge datasets may help with better decision-making, greater comprehension, and, in some instances, training data for machine learning. However, data quality continues to be a significant issue, since flawed data can produce in incorrect conclusions and unreliable findings. Inadequate knowledge, errors, mismatched forms, numerous captures of the same genuine item, and transgressions of professional norms of regular errors are examples. Data cleansing has developed into a crucial area of database research because analysts must assess the effects of dirty data before reaching any conclusions. Databases can get corrupted for a number of reasons, such as missing, incorrect, or inconsistent data. ML techniques are increasingly being applied in current data analytics routes, and the effects of dirty data may be difficult to control. Simple sampling approaches are useless for elevated systems because dirty data is often of poor quality (Krishnan, et al., 2016). There has been a growth in interest in many aspects of data cleansing in latest years from both industry and academia, such as innovative abstractions (Beskales, et al., 2010; Fan, et al., 2010), interactions (Dallachiesa et al. 2013, Khayyat et al. 2015), robustness techniques, and crowdsourced techniques (Chen & Cafarella, 2014).

Information collecting is a major obstacle to learning algorithms and a popular research topic in many domains. The sudden rise in importance of data collecting may be attributed to mainly two factors. First, when machine learning becomes more widespread, new applications appear that might not always have enough tagged data (Roh, et al., 2019). Second, deep learning techniques build classification models as opposed to conventional ML algorithms, reducing feature engineering costs but requiring more labelled data (Adeniyi et al., 2022). Modern data exploration originates not only from machine learning, natural language processing, and object identification but also from controlling the data field due to the necessity to process enormous amounts of data (Roh, et al., 2019)

Machine learning has a significant influence on a wide range of applications, including textual analysis, picture and audio recognition, and care services genetics. We live in an exciting time of invention. For instance, deep learning algorithms are known to perform better than ophthalmologists in identifying diabetic eye issues in pictures (Phene, et al., 2019). Large amounts of training data and increased computer resources are largely to blame for the present success. Data collecting among other difficulties, has emerged as one of the main bottlenecks in machine learning. The majority of the time required to complete machine learning from start to finish is invested on data preparation, which involves data collection, cleanup, and analysis, presenting, and feature extraction. The goal of machine learning is to extract knowledge from data (Kubat, 2017). Supervised learning is the method of artificial intelligence that is most frequently used in stock market forecasting. This research trained the model using a number of machine learning techniques after properly cleaning and labeling the data (Ogunlese et al., 2022). KNN machine learning methods are used in this work to sanitize financial data. Using labelled data, the K-Nearest Neighbor (KNN) classifier employs supervised learning. In this instance, it was used to clean soiled financial datasets that were downloaded from the Kaggle database. Based on how similar its independent variables are to an existing instance, KNN determines the dependent variable.

The autonomous data cleansing and labeling (ADCL) used in this research aims to deliver the preciseness and accuracy of the user-provided dataset. By offering automatic cleansing and labeling, the unsupervised approach in this research aids in reducing the customer's labor, energy, and other guides. The productivity of the cleansed dataset was also evaluated and shown in comparison to the uncleaned customers records utilized in this experiment, which gives the user confidence in its efficacy. There are differences in the scope, discretization technique, imputed columns, and quantity of incomplete data. For Alqami Quant Data Analysis, an unsupervised clustering program based on client data and character was created. A user profile assessment is a thorough examination of a business' ideal customers. It improves a firm's comprehension of its clients and makes it simpler to customize items to the distinctive demands, habits, and problems of diverse clients.

This study consists of six sections. The next section describes the literature reviews. Section 3 presents the summary of a review of past work. The materials and methods used as described in section 4. Section 5 presents the result and discussion. Section 6 concludes the study.

2 Literature reviews

The most efficient way to gather, analyse, and analyze massive volumes of diverse data from many sources is through the use of big data. Information quality is impacted by the volume and pace of data generation and processing. At every level of the Big Data system, Quality of Big Data (QBD) must be used to guarantee data quality (Alkatheeri et al., 2015; Taleb et al 2020; Ajagbe & Adigun 2023). The pre-processing stage, which comprises sub-processes like cleansing and merging, mainly concentrates on data integrity. Massive volumes of data that are challenging to evaluate in typical data management methods are processed using big data platforms.

Toolan & Carthy (2010) looked at 40 characteristics that frequently occurred in the research. Four factors, Web address, specific topic, and script-based—were used to group the traits. Following the determination of the information obtained for each attribute throughout their inquiry, designs for each property were created and evaluated. The article's findings supported traits that are related to the body.

Advanced phishing detection characteristics were explored by Bergholz et al. in 2008 and in 2010. Despite the numerical pointlessness of improving detection by changing the classification method itself, the scientists found that adding enhanced characteristics significantly improved email phishing categorization. On the basis of an unsupervised algorithm, two sets of result in a significant were created to enhance the 27 often used criteria in phishing identification. The basic features included spam attributes, word list attributes, link functionalities, component attributes, and structure characteristics. Among the novel features were the dynamic Markov chain prototype, latent topic model characteristics, and subject phrase groupings founded on latent Dirichlet distribution. To resolve classification task and recognize phishing SVM, deep learning techniques or other techniques like naive Bayes and support vector machines are frequently used.

Recurrent convolutional neural network (RCNN)-based text classification model was proposed by Lai et al. (2015). To get context from the text, they created a repeating framework. In order to develop a written representation, they also used CNN. On four datasets, the model was put to the test, and its efficiency was assessed to that of a Convolution layer, a recursive neural network (RNN) concept, as well as other traditional models. They discovered that the RCNN prototype gave better results than all other tests conducted.

In phishing research, 20% of the population responded and visited the fake hyperlink in the messages, according to Benenson et al (2017) description. 34% of individuals who were asked why they visited the hyperlink said that they were curious. They recommended companies to take every precaution to prevent employees from viewing and responding to phishing emails. To combat this growing threat, automating of phishing email identification using body email content is necessary.

A brand-new text classification method utilizing graph neural networks was developed by Yao et al. in 2019. The main idea was to employ graph neural networks to train phrase and paragraph embeddings in tandem while representing the entire corpus as a heterogeneous graph. They put the system to the test on four text-size samples and contrasted the outcomes with those of existing state-of-the-art text classification and incorporate techniques. Having a 97% accuracy rate, this model was effective.

The message content and subject are the areas in which the THEMIS classification model, created by Fang et al. (2019), functions. For text, the scholars used deep learning rather than feature extraction. The word2vec tool was used to depict messages, and the char-level email headers, word-level email header, char-level message content, and word-level email body were all recovered. The RCNN deep learning technique was used to build the model. The THEMIS model's accuracy rate of 99% was encouraging, illuminating the value of using NLP for email phishing prediction.

Kulesza et al. (2014) found that annotators regularly changed their working framework of a baseline model and their supporting tags when they encountered more entries in a dataset. The ability to create specific frameworks for unclear items discovered during labeling helped annotators to gradually improve their overall understanding of the data and provide more regular ultimate descriptions.

Kairam & Heer (2016) used label conventions to classify crowdworkers (e.g., various numbers of entities were recognized by liberal and conservative labelers during an entity extraction job). The subjective examination of these groups was then used to enhance future challenge concepts. In contrast to previous research, we use public disagreements to find and clarify confusing concepts in data in order to give annotated information for machine learning.

Halgas et al., (2020) recommended an RNN-based classification to identify malicious email from legitimate emails based on the vocabulary they use. The classifier turned out to be reliable and helpful. It might also be used in combination with the existing classifiers. In order to increase the likelihood of accurately detecting the possibility that a message is a fraudulent message, this study develops an efficient phishing email identification classifiers employing NLP of email body features and deep learning techniques using GCN. In order to provide continuous and recurrent cleansing while keeping convergence assurances in statistical modeling issues, Krishnan et al. (2016) presented ActiveClean. ActiveClean focuses cleansing data that are likely to have an impact on the results and supports convex loss methods (such as regression analysis and SVMs). We evaluate ActiveClean using five real-world datasets: UCI Adult, UCI EEG, MNIST, IMDB, and Dollars for Docs, with both actual and fake problems. The results indicate that our suggested changes can increase model accuracy by up to 20% using the same volume of data that has been 2.5 times cleaned. Additionally, with a fixed cleansing expense and on all real datasets, ActiveClean builds more accurate estimates than regular selection and Active Learning.

Table 1 presents the summary of the reviewed literature

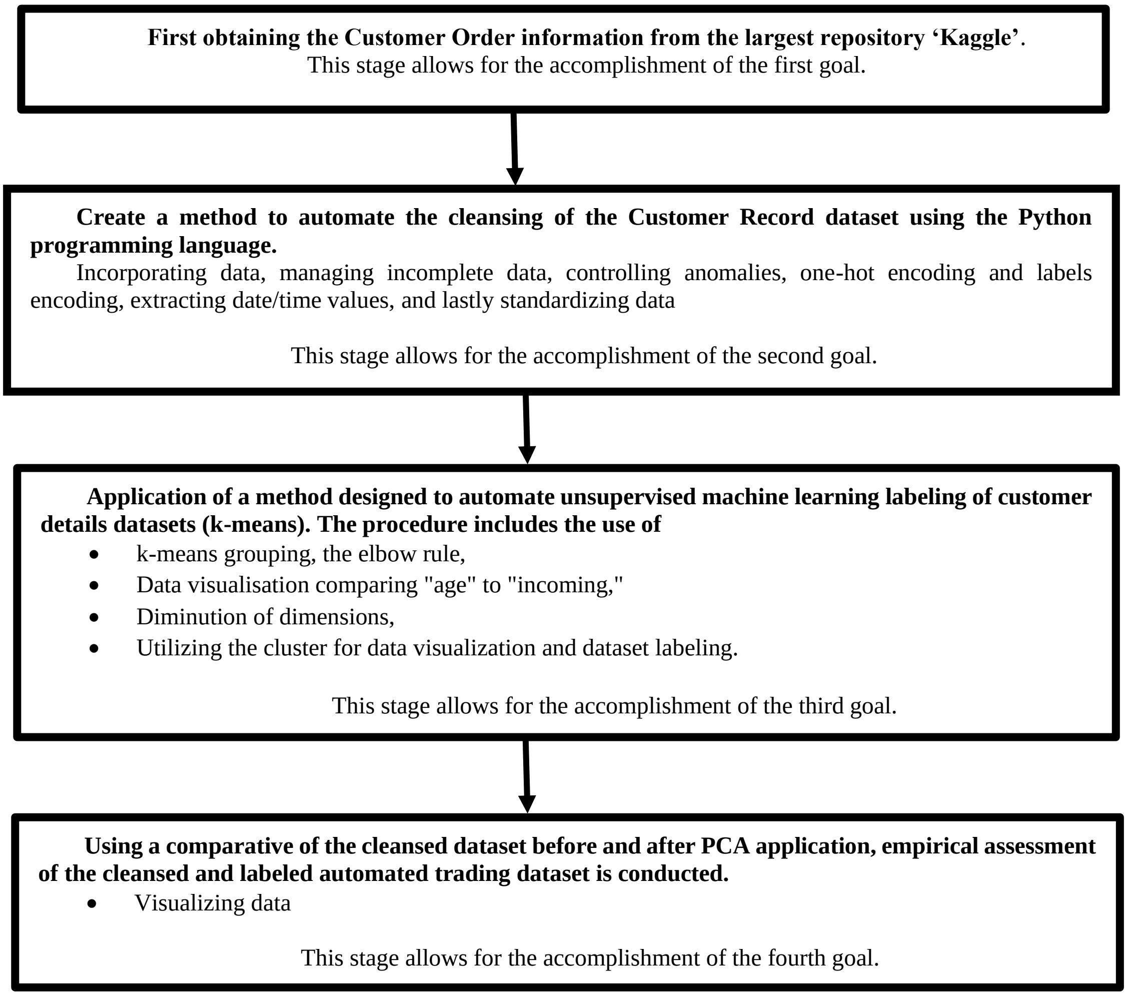
Table 1: Summary of review of literature

S /N	Author	Title	Methodology	Result
1	Benenson et al., (2017)	Unpacking spear phishing susceptibility	Audience questionnaire	Automation of phishing emails will help to address threats.
2	Bergholz et al., (2008) and (2010)	Novel phishing message segmentation techniques	Deep learning methods, dynamic Markov chain model	The upgraded features improved email phishing classification
3	Lai et al., (2015)	Text classification using repetitive convolutional neural systems	CNN and RCNN model	RCNN performed better than CNN
4	Yao et al., (2019)	Text classification using graph convolutional networks.	Graph Neural Networks	This system obtained 97% accuracy rate.
5	Fang et al., (2019)	Detecting phishing emails with an enhanced RCNN method with multilevel vectors and a probabilistic model	THEMIS categorization model	The model obtained 99% accuracy.
6	Kulesza et al., (2014)	Organized labeling in machine learning to aid idea transformation	Annotators structures for labelling.	The annotators allowed progrssively

				gain a global grasp of data.
7	Kairam & Heer, (2016)	Divergent explanations in crowdsourced tagging tasks: Parting the crowds	Machine Learning	The study analysis improved the future problem designs.
8	Halgas et al., (2020)	Catching the Phish: Using recurrent neural networks to discover malicious scams (RNNs)	RNN-based classifier. Deep learning using GCN.	The GCN used boost the chance of automatic recognition of potential email phishing
9	Krishnan et al., (2016)	ActiveClean is a visualization tools cleaning application for data analysis.	Linear Regression and SVM were supported with ActiveClean	The model improve accuracy by up to 20%.

3 Materials and methods

The approaches utilized in this research are summarized in this part. These include data collection via Kaggle, information retrieval, feature engineering, and assessment, among other things. The conceptual structure of the method employed in this research is depicted in Figure 1. The client details from the database of a grocery shop are included in the dataset, which was obtained through Kaggle. Each consumer who has visited the business is represented by their biometric information and purchase history. Client age, first purchase date, relationship status, gender, number of siblings, education level, and other factors are among the variables included in the dataset. Based on the summary statistics, the dataset comprises 26 quantitative columns and about 2240 rows. Figure 2 displays the omitted client record data from the study's perspective.



First obtaining the Customer Order information from the largest repository 'Kaggle'.
This stage allows for the accomplishment of the first goal.

Create a method to automate the cleansing of the Customer Record dataset using the Python programming language.
Incorporating data, managing incomplete data, controlling anomalies, one-hot encoding and labels encoding, extracting date/time values, and lastly standardizing data

This stage allows for the accomplishment of the second goal.

Application of a method designed to automate unsupervised machine learning labeling of customer details datasets (k-means). The procedure includes the use of

- k-means grouping, the elbow rule,
- Data visualisation comparing "age" to "incoming,"
- Diminution of dimensions,
- Utilizing the cluster for data visualization and dataset labeling.

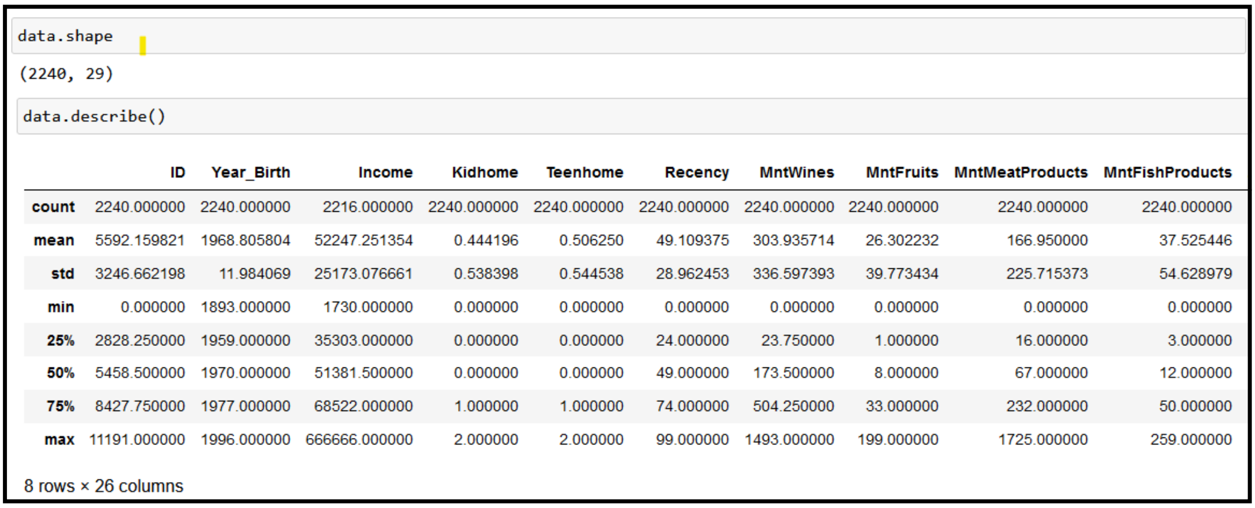
This stage allows for the accomplishment of the third goal.

Using a comparative of the cleansed dataset before and after PCA application, empirical assessment of the cleansed and labeled automated trading dataset is conducted.

- Visualizing data

This stage allows for the accomplishment of the fourth goal.

Figure 1: Cleansing and labeling of client data theoretical model.



```
data.shape
```

```
(2240, 29)
```

```
data.describe()
```

	ID	Year_Birth	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts
count	2240.000000	2240.000000	2216.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000
mean	5592.159821	1968.805804	52247.251354	0.444196	0.506250	49.109375	303.935714	26.302232	166.950000	37.525446
std	3246.662198	11.984069	25173.076661	0.538398	0.544538	28.962453	336.597393	39.773434	225.715373	54.628979
min	0.000000	1893.000000	1730.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2828.250000	1959.000000	35303.000000	0.000000	0.000000	24.000000	23.750000	1.000000	16.000000	3.000000
50%	5458.500000	1970.000000	51381.500000	0.000000	0.000000	49.000000	173.500000	8.000000	67.000000	12.000000
75%	8427.750000	1977.000000	68522.000000	1.000000	1.000000	74.000000	504.250000	33.000000	232.000000	50.000000
max	11191.000000	1996.000000	666666.000000	2.000000	2.000000	99.000000	1493.000000	199.000000	1725.000000	259.000000

8 rows × 26 columns

Figure 2: Highlight the customer record data exploration analysis.

3.1 Feature engineering

To create more functionalities, some information was manually orchestrated. As previously stated, this enhances the AI agent's ability to understand of the dataset. The newly added parameters were created by hand:

- Age was calculated by subtracting each client's birth date from the current date.
- Spent: To obtain this variable's value, add the amount expended on Vintages, Foods, Meat, Fish, Sweets, and so on. This represents the entire amount consumed at the supermarket.
- Living Conditions: The groups in this section were chosen for their likeness.
- Toddlers: This parameter was calculated by adding the quantity of kids and teenagers in the family.
- Is Caregiver: This block was employed to distinguish those who have had at least one child from those who have not.
- Schooling: The classifications in this block have been reorganized based on their resemblance.

Moreover, sections with obscure names like Mntwines, Mnths, and others were rebranded to a more instinctive and comprehensible.

```
Data Cleaning Algorithm
Algorithm 1: Customer Information Data preparation using k-means
Step 1: Start
Step 2: The user should initiate various data records or data sources to the Ml algorithm for cleanup.
Step 3: Fill in all of the blanks.
Step 4: Put feature engineering into action.
Step 5: Deal with anomalies by identifying them with the interquartile range (IQR)
Step 6: Used one-hot embedding data or attribute encoding to perform classification encoding.
Step 7: Requirements or situations for selecting either one-shot or label encoding data.
        If the attribute has ten distinct values, it will be one-hot engineered.
        If the attribute has 20 distinct values, it will be label-encoded.
        If the attribute has more than 20 distinct values, it will not be embedded.
Step 8: Datetime features extraction
Step 9: Use a classic scalar to standardize the client records scaler = StandardScaler ()
Step 10: Encoding labels with a label encoder LE=LabelEncoder ()
```

```
Data Labelling Algorithm
Algorithm 2: Customer Record Data labelling based on k-means with PCA
Step 1: Start
Step 2: Using the elbow rule, determine the optimal number of clusters.
Step 3: Use the k-means technique to group the dataset.
Step 4: To create a labelled dataset, attach the groupings to the original dataset.
Step 5: Use the cluster to label the data.
Step 6: Depict or reveal the clustering quality using a scatter graph from the Seaborn library.
Step 7: Used PCA to perform dimensional reduction.
Step 8: Repeat the preceding steps as shown below.
        using the elbow rule, determine the optimal number of clusters.
        cluster the dataset using the k-means technique groupings the data using the cluster
        grouping effectiveness can be visualized or displayed using a scatter graph from the Seaborn library.

Step 9: End
```

4 Results and discussion

The findings of the parametric data and statistics are shown below. After data profiling, verifying limitations, the comparison among sections, and finally asserting null values, the sterilised (pre-processed) and labeled measurements were reverted to the subscriber. Figure 3 depicts the cleanup of the clients' collect data before handling the anomalies. The figure includes three distinct visualization tools of customer details cleanup, equivalent to the visual analysis in (Ajagbe, et al., 2020). The visualization consists of:

i. Histogram (quantity of deals purchased vs count): Any dispersion that deviates from the sequence is an outcast; thus, anomalies were identified and addressed.
ii. Possibility plot (conceptual quartile vs RM quartile): the red line indicates the likelihood storyline; any findings that deviate from the line are considered outliers.

Boxplot (amount of deals acquired): Any data point outside the whiskers is an outcast. Anomalies are clearly present in the client record dataset obtained.

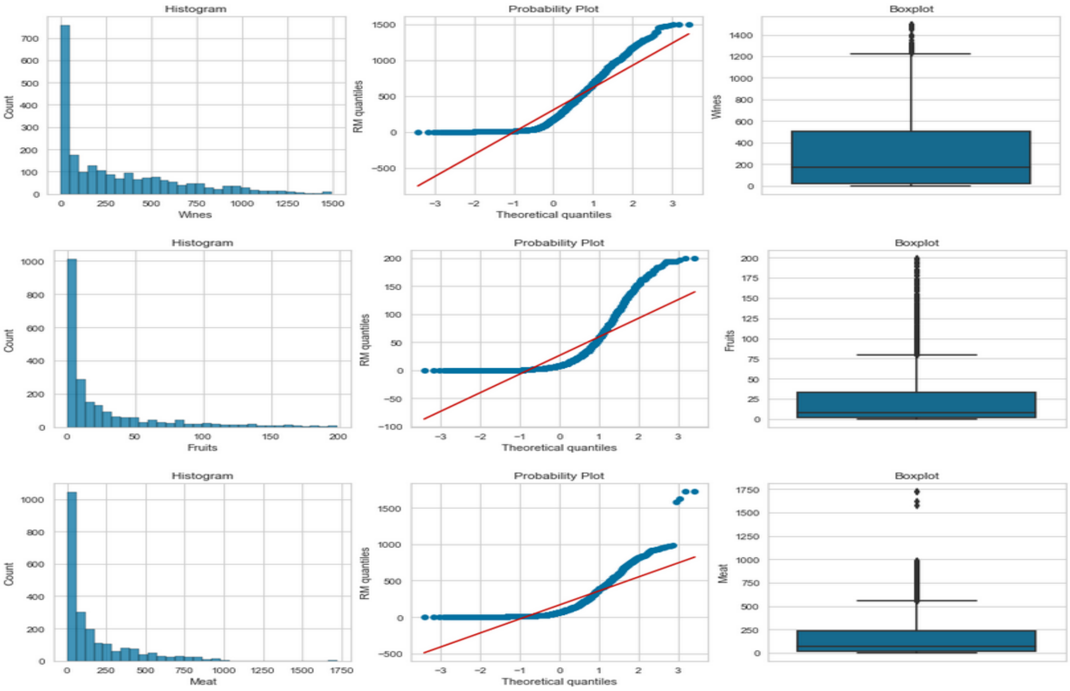


Figure 3: Before treating the outliers.

Figure 4 depicts the pictorial depiction of the user 's account after the anomalies have been removed. The interquartile range (IQR) (IQR). This demonstrates how the study second goal was met. Essentially;

i. Histogram: There are no observations that are outside the sequence, so the anomalies have been adequately dealt.

ii. Probability plot: There are no observations that are substantially removed from the likelihood plotline's red line. As a result, the anomalies have been appropriately handled.

iii. Boxplot: Because there are no data points outside the whiskers, the anomalies in the client recordings used in this study have been efficiently handled, achieving the goal of customer details data cleaning.

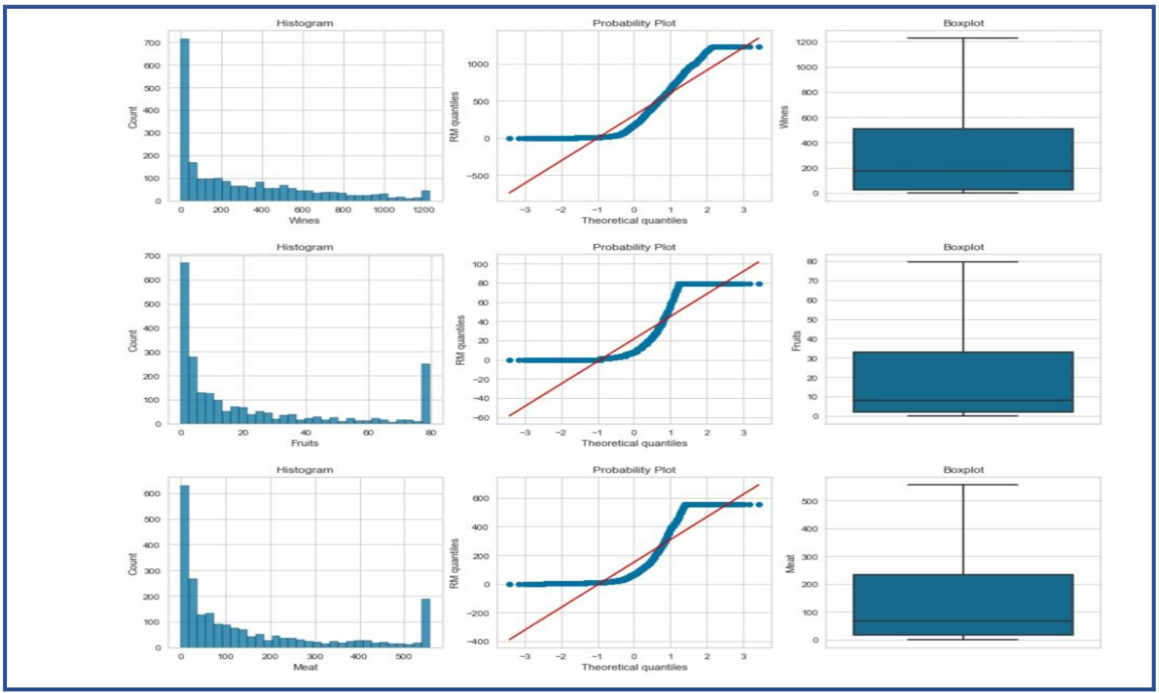


Figure 4: After treating the outliers.

4.1 Result of customer record data labelling technique developed

K-means was used in this study. Inertia was utilized to assess how well K-means performed on the study dataset. The quantity of clusters desired (n patterns), the quantity of initializations desired (n init), the highest number of rounds the technique will perform to selecting a sample of observations in order to decrease inertia (max iter), and the acceptance desired. The variety of clusters was charted against (WCSS) and depicted to demonstrate how the k-means algorithm works. WCSS is the total of the squared ranges between each spot and the centroid in a group. WCSS was mapped against the clustering. The number of groups utilized in the k-means concept and the best possible point selected for the elbow principal test is four. Figure 5 is the elbow rule that displays the outcomes of the data labeling investigation in this study with k-means prior to the application of PCA, and Figure 6 is the forearm principle of k-means after PCA.

Figure 7 shows that each group is quite comparable to the others, indicating that they are nearly equitably spread. It is clear that those in group 0 are elevated shoppers with limited wages, whereas those in group 2 are minimal shoppers with limited wages. Group 3 includes high purchasers with median earnings, whereas Group 1 comprises high purchasers with large salaries. Figure 8 depicts the breakdown of the client record dataset based on income and expenditures.

The dataset was clearly labeled by attaching the groupings to the original dataset and labeling it with the groupings. Figure 9 depicted additional details from this study's subcategories.

The K-means proposed technique was executed satisfactorily. The areas of overlap were not differentiated enough. Nevertheless, when dimensionality reduction was applied, the groupings were well isolated, as depicted in Figure 10, allowing for a more accurate and effective understanding of client records and datasets using k-means.

Table 2 presents the comparison of the existing work with our study and the SOTA of this work of this research.

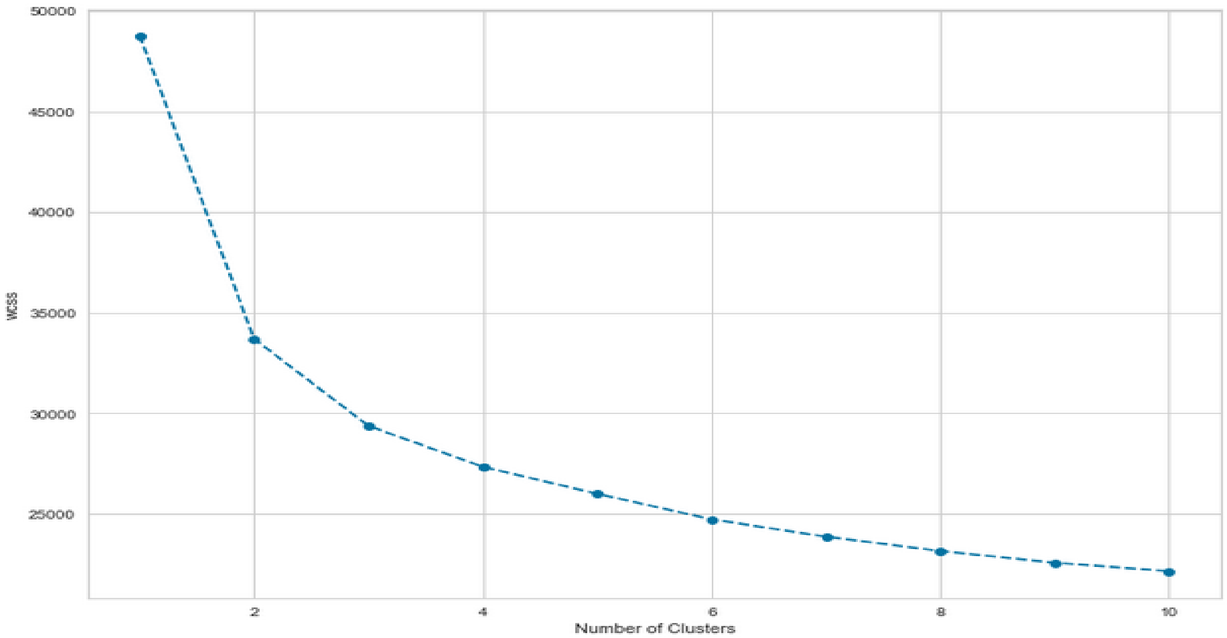


Figure 5: WCSS cluster analysis using K-means before application of PCA

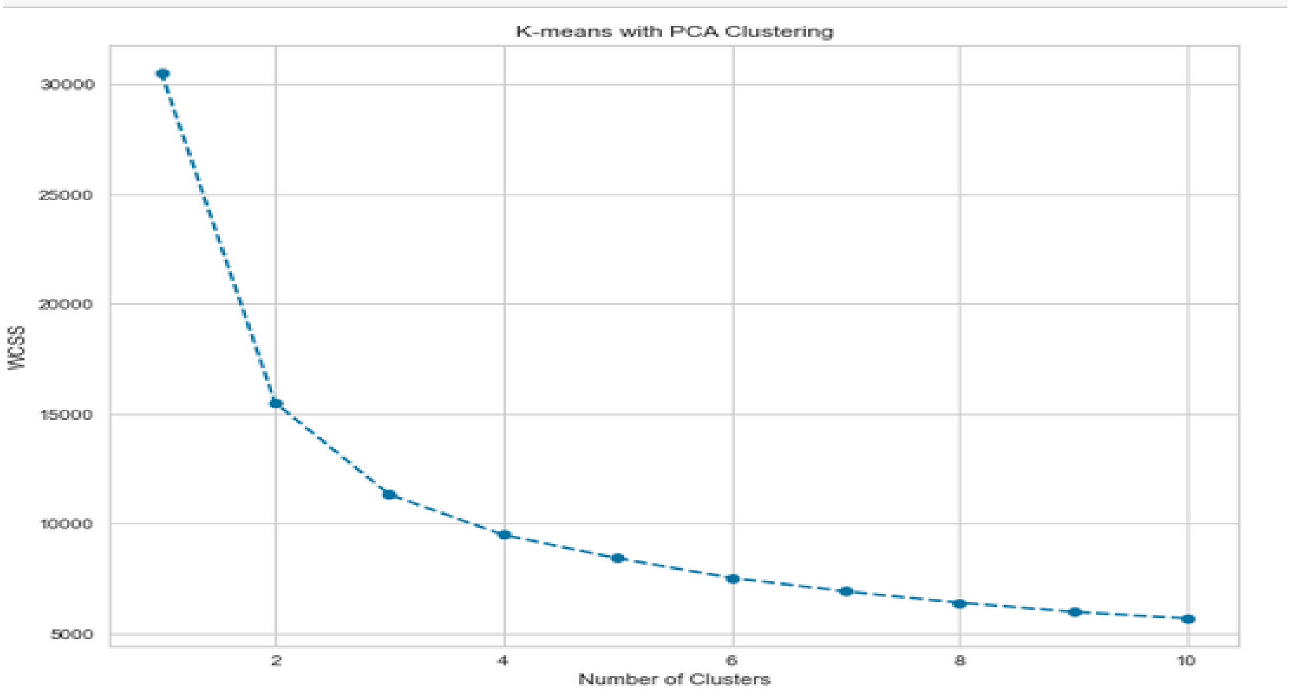


Figure 6: WCSS cluster analysis using K-means after application of PCA

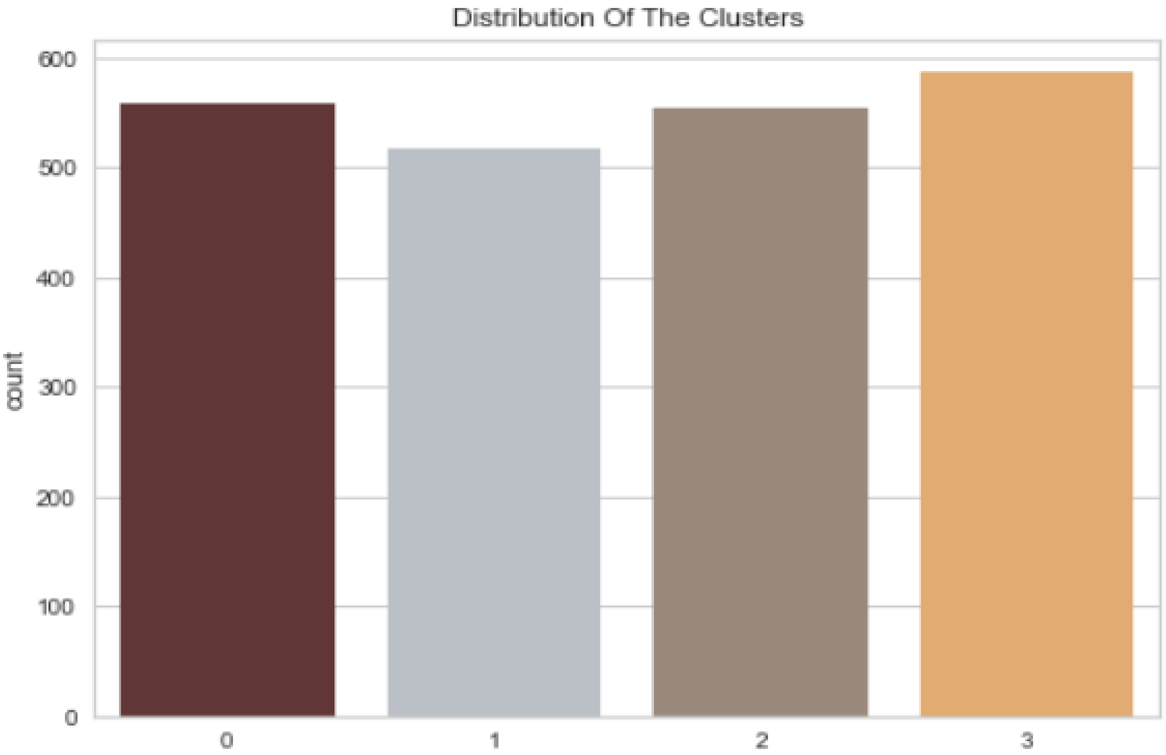


Figure 7: Distribution of the cluster.

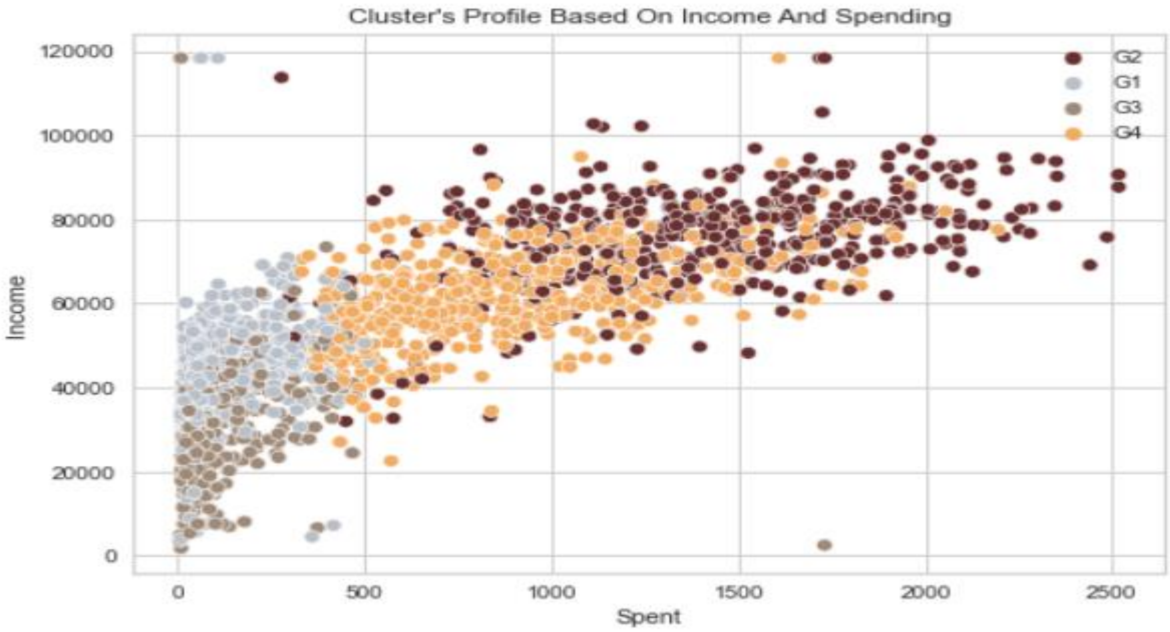
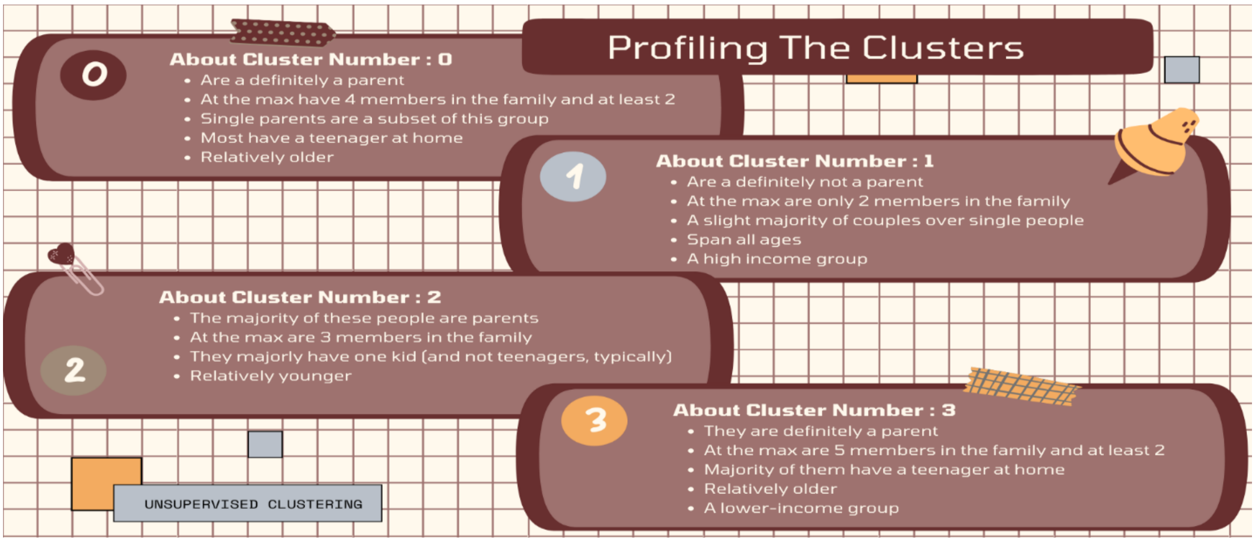


Figure 8: Revenue and expenditure dissemination.



Profiling The Clusters

0 About Cluster Number : 0

- Are a definitely a parent
- At the max have 4 members in the family and at least 2
- Single parents are a subset of this group
- Most have a teenager at home
- Relatively older

1 About Cluster Number : 1

- Are a definitely not a parent
- At the max are only 2 members in the family
- A slight majority of couples over single people
- Span all ages
- A high income group

2 About Cluster Number : 2

- The majority of these people are parents
- At the max are 3 members in the family
- They majorly have one kid (and not teenagers, typically)
- Relatively younger

3 About Cluster Number : 3

- They are definitely a parent
- At the max are 5 members in the family and at least 2
- Majority of them have a teenager at home
- Relatively older
- A lower-income group

UNSUPERVISED CLUSTERING

Figure 9: Profiling of the clusters

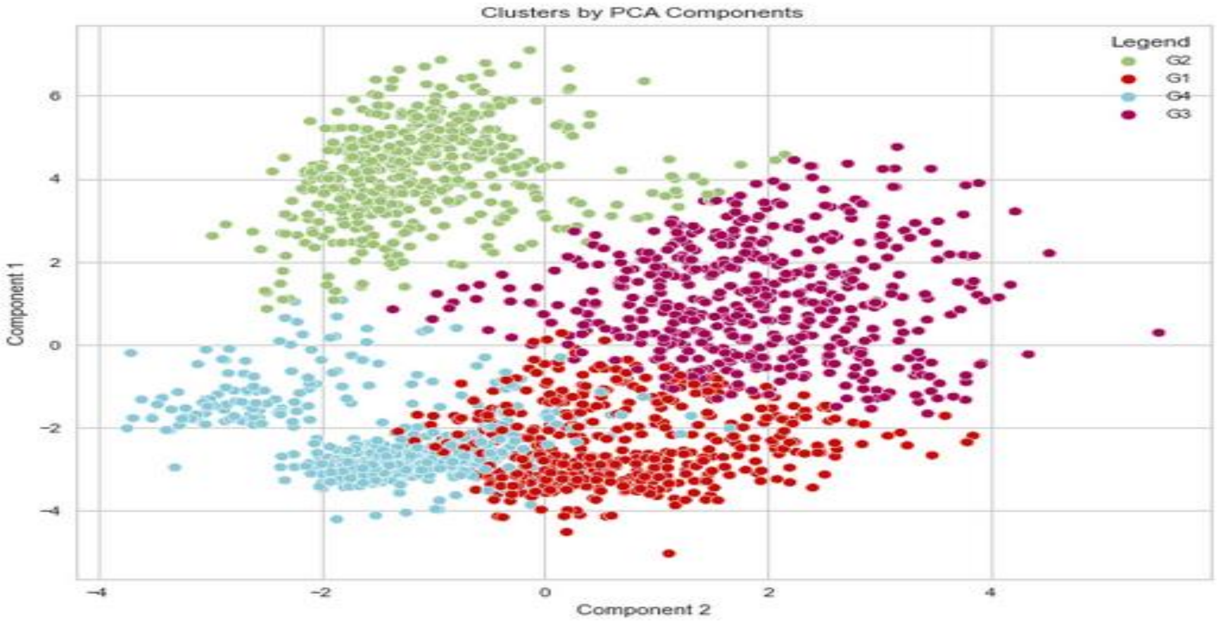


Figure 10: K-means clustering segmentation with PCA.

S/N	Author	Goals	Contribution
1	Lai et al., (2015)	Text classification using repetitive convolutional neural systems	The model was put to test on four datasets and the test result shows that RCNN model gave better results than all other test conducted for text classification.
2	Kairam & Heer, (2016)	To use label conventions to classify crowdsourced data.	The ability to create specific frameworks for unclear items discovered during labeling helped annotators to gradually improve their overall understanding of the data and provide more regular ultimate descriptions
3	Krishnan et al., (2016)	Is to use ActiveClean as a visualization tool for cleaning application for data analysis.	The study was evaluated using five real world datasets. The results indicate that the suggested changes can increase model accuracy by up to 20% using the same volume of data that has been 2.5 times cleaned. Linear Regression and SVM were supported with ActiveClean
4	Fang et al., (2019)	Detecting phishing emails with an enhanced RCNN method with multilevel vectors and a probabilistic model	THEMIS categorization model was created using the word2Vec tool to depict messages.
5	Halgas et al., (2020)	Catching the Phish: Using recurrent neural networks to discover malicious scams (RNNs)	The study develops an efficient phishing email identification classifier employing NLP of email body features and deep learning techniques using GCN. In order to provide continuous and recurrent cleansing while keeping convergence assurances in statistical modeling issues.
6	Proposed study	To use an automated python script for data cleaning and labeling with machine learning technique	The developed ML technique not only improved the performance of the audit data used in this study, but it also classified the data after cleaning it and removing the unpleasant section and incomplete data, as shown by the k-means segmentation result and grouping by PCA.

Table 2: Comparison of the existing work with our study.

5 Conclusion

ADCL (autonomous data cleanup and labeling) attempts to ensure the preciseness and accuracy of the dataset supplied by the user. By providing automated cleanup and labeling, the unsupervised method employed in this research assists in reducing the customer's time, commitment, and other manuals. There are differences in the amount of omitted variables, sections ascribed, discretization approach, and variety. All of these factors were considered when assessing the experiment's effectiveness and ability to sanitize a dataset. The schemes proffered here are utilized to choose the data that produces the most effective and optimal findings for the raw data provided. This goal was accomplished because it improved the quality of information provided by clients by utilizing their ideal cleaning solution. The study obtained a customer record dataset from Kaggle, and information gathering revealed that the client log includes information such as customer age, original purchase period, family status, sex, number of dependents, skills training, and other variables that demonstrate the dataset's appropriateness. To obtain a better grasp of the data, an overview statistic, incomplete data recognition, and a dataset tally were created. This resulted in the use of the elbow rule to determine the optimal number of clusters, K-means grouping, visual analytics, and the realization that improved effectiveness of k-means, dataset labelling using the groupings. The study's effectiveness was determined by comparing the clean dataset before and after PCA implementation. According to the analytical outcomes, principal component analysis delivered a reasonable outcome. Additional research could look into other types of Principal component analysis methods, like iterative PCA, sparsity PCA, and single attribute decomposing.

References

[1] Adeniyi, E. A., Oguns, Y. J., Egbedokun, G. O., Ajagbe, K. D., Obuzor, P. C., Ajagbe, S. A. (2022), “Comparative Analysis of Machine Learning Techniques for the Prediction of Employee Performance”, *Paradigmplus*, vol. 3, no. 3, pp. 1-15, https://doi.org/10.55969/paradigmplus.v3n3a1

[2] Ajagbe, S. A., Adigun, M. O. (2023) Deep learning techniques for detection and prediction of pandemic diseases: a systematic literature review. *Multimedia Tools Application* (2023). https://doi.org/10.1007/s11042-023-15805-z

[3] Ajagbe, S. A., Oladipupo, M. A. & Balogun, E. O., 2020. Crime Belt Monitoring Via Data Visualization: A Case Study of Folium. *International Journal of Information Security, Privacy and Digital Forensic,* 4(2), pp. 35-44.

[4] Alkatheeri, Y. et al., 2020. The effect of big data on the quality of decision-making in Abu Dhabi Government organisations. In: *Data management, analytics and innovation* . s.l.:Springer, Singapore.

[5] Alwert, K., Bornemann, M. & Will, M., 2009. Does intellectual capital reporting matter to financial analysts?. *Journal of intellectual capital.,* Volume 10, pp. 354-368.

[6] Bansal, S. K., 2014. *Towards a semantic extract-transform-load (ETL) framework for big data integration.* s.l., IEEE, pp. 522-529.

[7] Bansal, S. K. & Kagemann, S., 2015. Integrating big data: A semantic extract-transform-load framework. *Computer*, 48(3), pp. 42-50.

[8] Benenson, Z., Gassmann, F. & Landwirth, R., 2017. *Unpacking spear phishing susceptibility.* s.l., Cham: Springer, p. 610–627.

[9] Bergholz, A. et al., 2010. New filtering approaches for phishing email. *Journal of Computer Security,* 18(1), pp. 7-35.

[10] Bergholz, A. et al., 2008. *Improved Phishing Detection using Model-Based Features.* Mountain View, California, USA, s.n., pp. 1-10.

[11] Beskales, G., Ilyas, I. F. & L., G., 2010. Sampling the repairs of functional dependency violations under hard constraints. *PVLDB,* 3(1-2), pp. 197-207.

[12] Chang, J. C., Amershi, S. & Kamar, E., 2017. *Revolt: Collaborative crowdsourcing for labeling machine learning datasets.* s.l., s.n., pp. 2334-2346.

[13] Chen, Z. & Cafarella, M., 2014. *Integrating spreadsheet data via accurate and low-effort extraction.* s.l., ACM, p. 1126–1135.

[14] Chicco, D., 2017. Ten quick tips for machine learning in computational biology. *Bio Data mining,* 10(1), pp. 1-17.

[15] Dallachiesa, M. et al., 2013. Nadeef: a commodity data cleaning system. *SIGMOD,* pp. 541-552.

[16] Fang, Y. et al., 2019. Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism. *IEEE Access,* Volume 7, pp. 56329-56340.

[17] Fan, W. et al., 2010. Towards certain fixes with editing rules and master data. *PVLDB,* 3(1-2), pp. 173-184.

[18] Halgaš, L., Agrafiotis, I. & Nurse, J. R. C., 2020. *Catching the Phish: detecting Phishing Attacks Using Recurrent Neural Networks RNNs.* s.l., Springer, pp. 219-233.

[19] Hellerstein, J. M., 2008. *Quantitative data cleaning for large databases*, s.l.: United Nations Economic Commission for Europe (UNECE).

[20] Johnson, G. M., 2021. Algorithmic bias: on the implicit biases of social technology. *Synthese,* 198(10), pp. 9941-9961.

[21] Kairam, S. & Heer, J., 2016. *Parting Crowds: Characterizing Divergent Interpretations in Crowdsourced Annotation Tasks.* s.l., ACM, pp. 1637-1648.

[22] Khayyat, Z. et al., 2015. *Bigdansing: A system for big data cleansing.* s.l., ACM, pp. 1215-1230.

[23] Kostopoulos, G., Kotsiantis, S. & Pintelas, P., 2015. *Estimating student dropout in distance higher education using semi-supervised techniques.* s.l., s.n., pp. 38-43.

[24] Krishnan, S. et al., 2016. *ActiveClean: interactive data cleaning for statistical modeling.* s.l., ACM, p. 948.

[25] Kubat, M., 2017. *An introduction to machine learning (2nd Ed.).* s.l.:Springer Publishing Company, Incorporated.

[26] Kulesza, T. et al., 2014. *Structured labeling for facilitating concept evolution in machine learning.* s.l., ACM, p. 3075–3084.

[27] Lai, S., Xu, L., Liu, K. & Zhau, J., 2015. *Recurrent convolutional neural networks for text classification.* s.l., ACM, p. 2267–2273.

[28] Liebchen, G. A. & Shepper, M., 2005. *Gernot Armin Liebchen, Martin Shepper, “Software Productivity Analysis of a Large Data Set and Issues of Confidentiality and Data Quality” 11th IEEE International Software Metrics Symposium (METRICS 2005).*. s.l., ACM.

[29] Madanagopal, K., Ragan, E. D. & Benjamin, P., 2019. Analytic provenance in practice: The role of provenance in real-world visualization and data analysis environments. *IEEE Computer Graphics and Applications,* 39(6), pp. 30-45.

[30] Myklebust, T. et al., 2021. Data safety, sources, and data flow in the offshore industry. *ESREL, Angers.*

[31] Ogunseye, E. O., Adenusi, C. A., Nwanakwaugwu, A. C., Ajagbe, S. A., Akinola, S. O. (2022) Predictive Analysis of Mental Health Conditions Using AdaBoost Algorithm”, *Paradigmplus*, vol. 3, no. 2, pp. 11-26, Aug.

[32] Phene, S. et al., 2019. Deep Learning and Glaucoma Specialists: The Relative Importance of Optic Disc Features to Predict Glaucoma Referral in Fundus Photographs. *Ophthalmology,* 126(12), pp. 1627-1639.

[33] Pisani, M., 2020. CHAPTER 1 – Introduction. In: *MACHINE LEARNING* . s.l.:Rootstrap, pp. 1-10.

[34] Rajasekar, S. P., Philominathan, P. & Chinnathambi, V., 2019. Research Methodology. Knowledge Management Techniques for Risk Management in IT Projects.. *Knowledge Management Techniques for Risk Management in IT Projects*, pp. 1-53.

[35] Reddy, U. S., Thota, A. V. & Dharun, A., 2018. *Machine learning techniques for stress prediction in working employees.* s.l., IEEE, pp. 1-4.

[36] Roh, Y., Heo, G. & Whang, S. E., 2019. A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. *IEEE Transactions on Knowledge and Data Engineering,* pp. 1-1.

[37] Sadique, F., Kaul, R. & Badsha, S. S. S., 2020. *An Automated Framework for Real-time Phishing URL Detection.* s.l., IEEE, pp. 335-341.

[38] Sidi, F. et al., 2012. *Data Quality: A Survey of Data Quality Dimensions.* s.l., IEEE, pp. 300-304.

[39] Taleb, I., Dssouli, R. & Serhani, M. A., 2015. *Big data pre-processing: A quality framework.* s.l., IEEE, pp. 191-198.

[40] Tang, N., 2014. Big Data Cleaning. *International Journal of Database Theory and Application,* pp. 13-24.

[41] Thadson, K., Visitsattapongse, S. & Pechprasarn, S., 2021. Deep learning-based single-shot phase retrieval algorithm for surface plasmon resonance microscope based refractive index sensing application. *Scientific Reports*, 11(1), pp. 1-14.

[42] Tomar, D. & Agarwal, S., 2014. A Survey on Pre-processing and Post-processing Techniques in Data Mining. *International Journal of Database Theory and Application* , 7(4), pp. 99-128.

[43] Toolan, F. & Carthy, J., 2010. *Feature selection for Spam and Phishing detection.* s.l., IEEE, pp. 1-12.

[44] Yao, L., Mao, C. & Luo, Y., 2019. *Graph convolutional networks for text classification.* s.l., ACM, p. 7370–7377