ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Observation-Based Hybrid Classification Algorithms for Customer Segmentation

Kulkatechol Kanokngamwitroj[1] and Chetneti Srisaan[2]

ABSTRACT

The customer segmentation model aims to cluster customers based on their specific characteristics. Relying solely on a simple classification algorithm may not yield optimal results. Our research proposes an Observation-Based Hybrid Classification (OBHC) algorithm to enhance the customer segmentation model by utilizing customer segmentation data from a public source. Observation-based clustering methods differ from simple classification or clustering models by being a hybrid system specifically engineered to boost the performance of predictive models. Furthermore, the focus is on evaluating metric values after clustering to demonstrate performance improvement. The experiments demonstrate significant performance improvements across various classification algorithms. The most notable enhancement observed with the proposed algorithm is up to 43.86% on average accuracy score, 24.25% on average precision score, 20.25% on average recall score, and 32% on average F1-score, as shown in the experiment section. This research contributes by introducing a novel process for data scientists to tackle customer segmentation challenges, identifying higher-performing segments that meet business needs, and providing executives with the flexibility to adopt them. The research underscores the significance of employing hybrid models to classify customers better, providing valuable insights for advancing business development and improving customer service.

Article information:

Keywords: Classification, Clustering Algorithm, Hybrid Model Introduction, Machine Learning

Article history:
Received: March 17, 2024
Revised: May 2, 2024
Accepted: May 23, 2024
Published: June 8, 2024
(Online)

DOI: 10.37936/ecti-cit.2024183.256082

1. INTRODUCTION

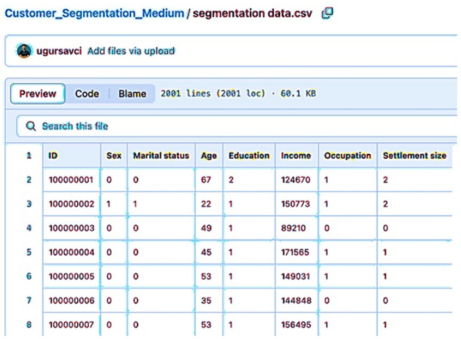
Customer segmentation can help understand customers and is essential in identifying potential future customers. It is also possible to define specific marketing strategies considering customer groups. For a data scientist, the main job is to find the best model to find the answer to a problem. Many researchers are exploring innovative approaches to tackle customer segmentation issues that exhibit unique characteristics compared to typical problems. The characteristics of each customer are not the same. Additionally, their needs and behaviors are different and can always change. As a result, customer segmentation has unique characteristics and is difficult to predict. The customer segmentation problems have a particular structure and need a new process or method for creating predictive models for executives. Consequently, segmentation is of great importance.

A classification model forms the foundation of a machine learning approach that aims to uncover latent patterns using class labels. While commonly used to create predictive models for customer segmentation, the limitations of existing classification techniques alone cannot provide the best results for the customer segmentation problem. As shown in Table 1, the performance values of the four algorithms are notably low. The problem arises because many segments are mixed in the same dataset, so separating the segments first is probably a good solution.

After reviewing the research papers [10,15,16,17,18] on the limitations of existing classification techniques in customer segmentation, a new algorithm is needed to overcome these limitations. The classification and clustering models are frequently combined within the same application. As detailed in multiple research papers [1, 2, 3, 4], hy-

[1,2] The authors are with the College of Digital Innovation Technology, Rangsit University, Thailand. E-mail: kulkatechol.k65@rsu.ac.th and chetneti@rsu.ac.th
[1] Corresponding author: kulkatechol.k65@rsu.ac.th

brid models blend clustering and classification techniques. Within the scope of Customer Segmentation, traditional classification algorithms might need to be revised, which is why this paper introduces the OBHC algorithm. The primary concept of the OBHC algorithm involves clustering the entire dataset into smaller clusters and renaming them based on their characteristics. A classification algorithm is subsequently used to create a predictive model for each new cluster.


Customer_Segmentation_Medium / segmentation data.csv

ugursavci Add files via upload

Preview Code Blame 2001 lines (2001 loc) · 60.1 KB

Search this file

	ID	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
2	100000001	0	0	67	2	124670	1	2
3	100000002	1	1	22	1	150773	1	2
4	100000003	0	0	49	1	89210	0	0
5	100000004	0	0	45	1	171565	1	1
6	100000005	0	0	53	1	149031	1	1
7	100000006	0	0	35	1	144848	0	0
8	100000007	0	0	53	1	156495	1	1

Fig.1: *A Selected Public Source from Github.com, which Contains Thousands of Records of Customer Segmentation Data (https://github.com/ugursavci/Customer_Segmentation_Medium/blob/main/segmentation%20data.csv).*

The objective of this research is to develop a newly invented algorithm that can increase the performance of the predictive model by first clustering into appropriate segments and then performing classification to get the best predictive model in each segment. The elbow method and K-Means algorithm were applied in this experiment to ascertain the best number of clusters.

1.1 Dataset and Tools

Python and Scikit-learn (Sklearn) are selected as the programming language and machine-learning tools, sequence, because of their ease of demonstration and implementation. The experiment section features the implementation of various Scikit-learn machine learning libraries, such as Gaussian Processes, K-Nearest Neighbours, Gradient Boosting, and Random Forest. Additionally, Decision Tree classification is applied to enhance data visualization, enabling the extraction of new insights from significant factors affecting the model.

The dataset, obtained from the widely recognized public data source Github.com, comprises thousands of entries related to customer segmentation, as shown in Figure 1.

1.2 Research Question

This experiment focused on three research questions about methods or processes for finding good customer segmentation.

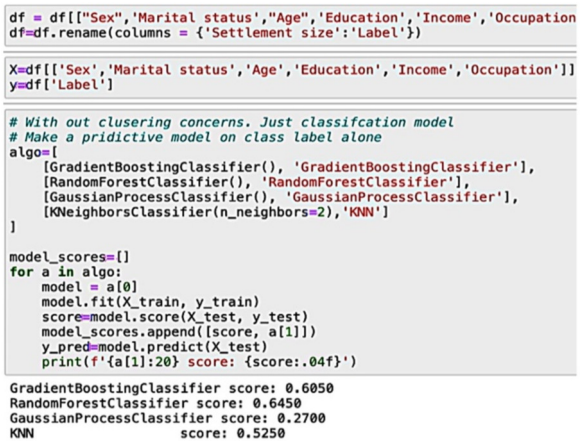
Question 1: Does customer segmentation have unique properties or unique structures?

Question 2: Are the techniques of clustering or classification sufficient for the process of finding customer segmentation?

Question 3: Want to find a method or process for finding customer segmentation that meets the needs of executives?

1.3 Performance of Simple Classification Model

This section demonstrates that a simple classification model is insufficient for discovering new knowledge. Four simple classification algorithms were used on the entire dataset.



```
df = df[["Sex",'Marital status',"Age",'Education','Income','Occupation
df=df.rename(columns = {'Settlement size':'Label'})

X=df[['Sex','Marital status','Age','Education','Income','Occupation']]
y=df['Label']

# With out clusering concerns. Just classifcation model
# Make a pridictive model on class label alone
algo=[
    [GradientBoostingClassifier(), 'GradientBoostingClassifier'],
    [RandomForestClassifier(), 'RandomForestClassifier'],
    [GaussianProcessClassifier(), 'GaussianProcessClassifier'],
    [KNeighborsClassifier(n_neighbors=2),'KNN']
]

model_scores=[]
for a in algo:
    model = a[0]
    model.fit(X_train, y_train)
    score=model.score(X_test, y_test)
    model_scores.append([score, a[1]])
    y_pred=model.predict(X_test)
    print(f'{a[1]:20} score: {score:.04f}')

GradientBoostingClassifier score: 0.6050
RandomForestClassifier score: 0.6450
GaussianProcessClassifier score: 0.2700
KNN                  score: 0.5250
```

Fig.2: *Define Class Label.*

Figure 2 displays a partial extraction of Python source code applied to the entire dataset. In the second line of Figure 2, the "pandas. DataFrame.rename()" function, 'Settlement size' is the class label for the entire paper. The class labels are assigned as follows: 'S' for small size city, 'M' for medium size city, and 'L' for big size city. The performance metrics, including Accuracy, Precision, Recall, and F1-score, are detailed in Table 1. Furthermore, the evaluation of The Receiver Operating Characteristic (ROC) curve and the Area Under the ROC Curve (AUC) metrics is covered in section 5 as part of the performance evaluation.

Table 1: *A Performance Matrix of The Original Datasets.*

Classification Algorithms	Accuracy	precision	recall	f1-score
Gaussian Process Classifier	0.2700	0.5900	0.6100	0.5700
K-Nearest Neighbours	0.5250	0.6500	0.6500	0.6500
Gradient Boosting Classifier	0.6050	0.5600	0.5600	0.5600
Random Forest Classifier	0.6450	0.4300	0.5000	0.3600

Table 1 displays a collection of low-performance metrics, without consideration for data distribution and structure. Considering these metric values, the model is deemed unacceptable.

The class label ('Settlement size') was kept consistent throughout the experiment to enable a performance comparison between the algorithm and the normal classification model. The dataset was split into four segments, each denoted by a number [0, 1, 2, 3], while maintaining the data structure. An additional Segment K-Means column, illustrated in Figure 3, was introduced. This supplementary column specifies the group to which each row belongs. The researcher then segmented the data into four datasets, each corresponding to a segment based on their respective names. To clarify, segment 0 was labeled 'Elderly_C', segment 1 was labeled 'Small_City_C', segment 2 was labeled 'White-collar_C', and segment 3 was labeled 'Labor_C'.

Unlike a standard classification model, an observation-based clustering method is a hybrid model designed to improve the performance of classification models. A customer segmentation model has been selected for demonstration purposes.

```
df_segm_kmeans= df.copy()
df['Segment K-means'] = kmeans.labels_
```

```
df_segm_analysis = df.groupby(['Segment K-means']).mean()
df_segm_analysis
```

Segment K-means	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
			Max				
0	0.501901	0.692015	55.703422	2.129278	158338.422053	1.129278	1.110266
1	0.352814	0.019481	35.577922	0.746753	97859.852814	0.329004	0.043290
			Avg			Min	
2	0.853901	0.997163	28.963121	1.068085	105759.119149	0.634043	0.422695
			Min			Max	
3	0.029825	0.173684	35.635088	0.733333	141218.249123	1.271930	1.522807
				Min	Max		

Fig.3: *Characteristic Findings on Data Analysis.*

Figure 3 illustrates that each cluster has different characteristics, as observed from the distribution statistics. For example, in Segment 0, individuals in this group tend to have higher levels of education and age than all other segments, which is a distinguishing feature of this group. Based on this observation, the researcher separated them into unique groups. This technique is termed the "observation-based clustering" method.

1.4 Research Objective and Contribution

The paper aims to develop more comprehensive predictive models for customer segmentation than previously achieved. In essence, the greater the diversity of new segments discovered, the broader the range of predictive models generated.

The contributions of this research are as follows:

1) Creating a novel process or guideline for data scientists to use as a new approach to solving customer segmentation problems.

2) The more segmentation, the more personalized the concept. In other words, the algorithm enhances customized services.

3) Highlighting the flexibility and enhanced performance of OBHC as a hybrid algorithm.

The rest of this research is structured as follows: Section 1 contains the introduction, while Section 2 provides a literature review. The methodology employed in the study is outlined in Section 3, followed by a description of the experiment in Section 4. Section 5 presents the experiment result, while Section 6 includes a discussion and considerations for future research. Lastly, Section 7 offers the conclusion.

2. RELATEWORK

Miškovic [1] examined both implicit and comprehensible knowledge, merging them into a hybrid classification model to harness the advantages of both types.

S. R. Gaddam, V. V. Phoha, and K. S. Balagani [2] proposed the first hybrid algorithm combining classifications and clustering. Their model is named K-Means+ID3.

Wong *et al.* [3] stated that the regular classification model provides low accuracy and precision scores and is impractical for real-life use. They introduced Hybrid Classification Algorithms designed to address instances of misclassification.

J. Xiao *et al.* [4] introduced the Hybrid Classification Framework using Clustering, which incorporates clustering techniques to boost classification accuracy and effectiveness through hybrid and one-step methods.

ZHANG, Libao, *et al.* [5] proposed using techniques such as K-Means to establish class labels automatically instead of manually. Previously, researchers manually assigned NBA class labels. Their approach improved upon traditional classification models by initially implementing clustering.

Deligiannis *et al.* [6] suggested a method for forecasting the ideal date and time for sending personalized marketing messages to returning customers. The methodology was based on data mining techniques and utilized historical transaction data to predict the future behavior of repeat buyers.

Tabianan *et al.* [7] proposed a clustering model based on customer purchase history records on e-commerce to analyze three groups, including event type, products, and categories, to assist vendors in identifying high-profitable segments.

Coelho, P. *et al.* [8] investigated the effect of service personalization on loyalty and measured some of the psychological dynamics involved in the process.

Kulkatechol, Kanokngamwitroj, *et al.* [9] demonstrated that machine learning can enhance understanding of the customer (student) needs to provide personalized services to individuals. They customized the Learning Management System using machine learning algorithms to better understand customer needs.

Erman, Jeffrey, Arlitt, Martin, and Mahanti [10] proposed a new classification model that effectively utilized a clustering model to identify groups of traffic as a class label. Their findings strongly supported the enhancement of classification models through clustering and marked a pioneering contribution to solving network traffic classification problems.

Lewaa, I [11] integrated machine learning and RFM analysis to predict customer churn using transactional data. They are using clustering methods such as K-Means and DBSCAN.

Li, Xiaotong, *et al.* [12] established the framework for a customer segmentation marketing strategy that combined support vector machines and clustering algorithms.

Altartouri, H., Tamimi, H., and Ashhab, Y. [13] demonstrated how a clustering technique could improve a classification model. Their work aimed to evaluate the improvement in protein classification by clustering proteins into sub-clusters at first. The findings demonstrated that the clustering algorithm could enhance classification accuracy.

Wu, J. and Lin, Z. [14] proposed a customer segmentation model using clustering algorithms. They used credit card consumption data as data samples to construct a modeling framework that utilizes pattern-based clustering approaches and signature discovery techniques.

Alghamdi, Abdullah [15] chose to use a hybrid method instead of relying solely on a classification model because each group of customers possesses unique characteristics. Considering these differences, relying on a single classification, or clustering model is inadequate for addressing real-life customer segmentation problems.

Gautam and Kumar [16] suggest using machine learning for customer segmentation and recommend exploring alternative methods beyond K-Means due to the complexity of customer segmentation problems in their article.

A. Hiziroglu [17] defines the term "soft computing" as a family of data mining techniques used for customer segmentation research. Various data mining techniques, including classification, and clustering algorithms, are employed to explore segmentation in this area.

M. Mosa *et al.* [18] opted for a hybrid model involving clustering, dividing the bank's customers into four clusters to comprehend unique features. The distinctiveness of each cluster enables executives to devise marketing strategies more effectively than traditional analysis methods.

3. METHODOLOGY

A research methodology is presented through the following Pseudocode, as shown in Figures 4 and 5.

Figure 4 shows all lines of code in Pseudo format for conciseness.

```
Function Personalized (C)
        BEGIN
        Classification (C)
        END;

Function Call_WCSS(D);
        Begin
        i=0,  i=Calculate_Number_cluster(D);
        reurn i;
        END;

BEGIN
Line1:  n=0, n = Call_WCSS(Dataset);
                n
Line 2:    C_i   =K-mean(n);
               i = 1
Line 3: For i=1 to n
                Tree_i = Personalized (C_i);
        End
END;
```

Fig.4: *Pseudo Code.*

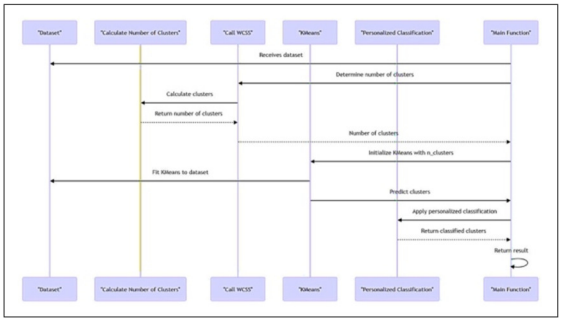


Fig.5: *Proposed Algorithm Diagram.*

Figures 4 and 5 show all lines of code in Pseudo format for conciseness and diagrammatic representation.

This section presents a hybrid model combining clustering and classification. The methodology is described and explained in the Pseudocode through the following steps.

1) Line 1: Calculate the number of clusters by initializing n=0 and calling the elbow method function. The function returns n, which is the optimal number of clusters.
2) Line 2: Call the K-Means function to cluster the dataset and assign names to each new cluster.
3) Line 3: Call classification algorithms to create prediction models, as discussed in section 4. This step involves tailoring each cluster based on its observed characteristics.

This research aims to formulate a new algorithm that can construct new sub-clusters of new knowledge. These new clusters are statistically at significant levels clustered into sub-clusters that are used for classification models. In other words, several new predictive models are based on these sub-clusters. These methods aim to personalize a classification model, creating new knowledge. Each classification model is utilized as a predictive model.

Lines 1, 2, and 3 in the Pseudocode represent the algorithms concisely. Lines 1 and 2 construct all clusters formulated by observation-based methods. Through observation, the researcher identifies

new clusters hidden in datasets. With each cluster, new classification models are developed. The greater the number of clusters, the more classification models or new knowledge can be obtained.

Line 3 is to personalize each cluster into a new knowledge (predictive model). Without clusters, it is impossible to discover those new predictive models.

Therefore, the OBHC algorithm is a groundbreaking method for uncovering and discovering new knowledge from the clusters.

4. EXPERIMENT

This section covers the data preparation phase, comprising data gathering, scrubbing, and refinement. During data collection, the source of all data collected is the customer loyalty cards.

4.1 Step 1: Pearson Correlation Coefficient and Correlation Heatmap Diagram

The formula for the Pearson coefficient is $\frac{cov(X,Y)}{\sigma_Y}$, where Cov (X, Y) denotes the covariance between variables X and Y, and σ_X and σ_Y are the standard deviations of variables X and Y, respectively. Significantly correlated variables with a positive association are (Age, Education) [0.65] and (Income, Occupation) [0.68], as shown in Figure 6.

The heatmap illustrates how the Seaborn Library can be utilized to create and visualize the covariance matrix across customer attribute variables. It also displays the correlation between each attribute.

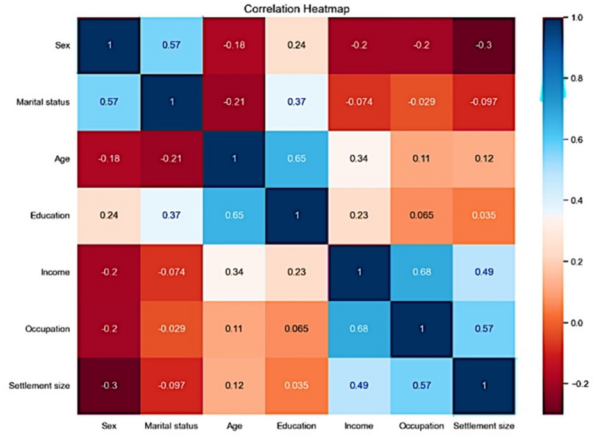


Fig.6: *Correlation Heatmap Generated from Customer Segmentation Data as A Model.*

4.2 Step 2: Find the Number Clusters Using the Elbow Method

The elbow method is a well-known technique used with K-Means to calculate the optimal number of clusters, as illustrated in Figure 7. To perform K-Means clustering, some features, such as income, may have much different scales than others. Sklearn's standardization is applied to all features to ensure all features are treated equally. The within-cluster sum of squares (WCSS) is employed to identify the optimal number of clusters for k values. The Sklearn kmeans.inertia_function is a built-in function utilized for this purpose.

```
plt.figure(figsize = (10,8))
plt.plot(range(1, 11), wcss, marker = 'o', linestyle = '--',color='red')
plt.xlabel('Number of Clusters')
plt.ylabel('Within Cluster Sum Square')
plt.title('K-means Clustering')
plt.show()
```

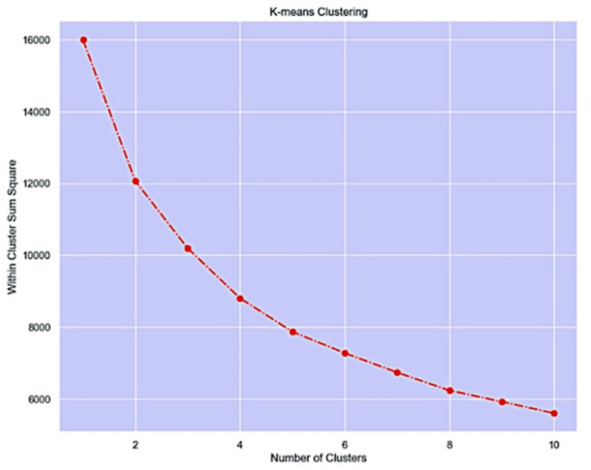


Fig.7: *Elbow Method.*

```
df['%change']=df['pct_chg'].pct_change() * 100 * -1
df
```

	k	inertia	pct_chg	%change
0	0	14000.000000	NaN	NaN
1	1	10514.558848	24.896008	NaN
2	2	8630.913217	17.914643	28.042109
3	3	7169.870822	16.928016	5.507373
4	4	6403.134168	10.693870	36.827389
5	5	5830.956303	8.935903	16.439014
6	6	5378.854705	7.753473	13.232356
7	7	5005.134610	6.947949	10.389194
8	8	4724.527780	5.606379	19.308863

Fig.8: *Optimal Point Determined by Elbow Method (N=4).*

Figures 7 and 8 illustrate finding the optimal point using the elbow method. Noticeably, the maximum percent change of inertia occurs at the point where N=4.

4.3 Step 3: Define New Clusters by Name

The classification performance scores in Table 1 are unacceptable and meaningless without a clustering model. Therefore, a clustering algorithm such as K-Means is applied first to uncover new insights. Each cluster is indexed based on its observations in Figure 3; hence, each cluster will never be identical. The steps of this experiment are described as follows.

Step 2, Create all new clusters: Based on the elbow method in Figures 7 and 8, the K-Means algorithm is applied with a value of k set to 4 for the number of clusters.

```
#Number of cluster is 4 from WCSS (elbow method)
kmeans = KMeans(n_clusters = 4, init = 'k-means++', random_state = 42)
kmeans.fit(df_std)
df_segm_kmeans= df_std.copy()
df_segm_kmeans = pd.DataFrame(data = df_std,columns = df.columns)
df_segm_kmeans['Segment K-means'] = kmeans.labels_
df_segm_analysis = df_segm_kmeans.groupby(['Segment K-means']).mean()
```

Fig.9: *Create All New Clusters.*

Figure 9 shows how to create four new clusters using the K-Means algorithm.

Step 3, Define and Assign cluster names: Figure 10 describes the characteristics of each segment, all of which are unique and clustered. Using human observations, each new cluster is named appropriately through demographic customer segmentation. Each new cluster is then separated into datasets for subsequent analysis using a classification model.

Cluster No 1: This cluster comprises young working professionals with an average age of 29. Refer to Figure 3 for details. They have good incomes and strong occupational records. Therefore, it is named "White-collar_C".

Cluster No 2: This cluster is characterized by the lowest salary values, suggesting residents living in small cities with a lower level of education than it is named "Small_City_C".

Cluster No 3: This cluster is identified as "Labor_C" and showcases relatively lower levels of education but higher income and occupational status. All individuals within this cluster reside exclusively in big or middle-sized cities. This characteristic tends to labor-intensive staff.

Cluster No 4: This cluster features the oldest average age and highest education attainment compared to all other clusters. Therefore, it is named "Elderly_C".

The new column labeled "Cluster_Name" serves as the designation for each cluster. The results of the assignment are shown in Figure 10.

```
df=df_segm_kmeans
df.head(4)
```

	Sex	Marital status	Age	Education	Income	Occupation	Settlement size	Cluster_Name
0	-0.917399	-0.993024	2.653614	1.604323	0.097524	0.296823	1.552326	Elderly_C
1	1.090038	1.007025	-1.187132	-0.063372	0.782654	0.296823	1.552326	White-collar_C
2	-0.917399	-0.993024	1.117316	-0.063372	-0.833202	-1.269525	-0.909730	Small_City_C
3	-0.917399	-0.993024	0.775916	-0.063372	1.328386	0.296823	0.321298	Labor_C

Fig.10: *Named Clusters After Data Normalization.*

Figure 10 shows four new clusters, including White-collar_C, Small_City_C, Labor_C, and Elderly_C.

Step 3: Separate the data into a new dataset. Figure 11 shows four new clusters:

Cluster 1 is the White_collar_DS dataset, consisting of 419 records.

Cluster 2 is the Small_City_DS dataset, consisting of 632 records.

Cluster 3 is the Labor_C_DS dataset, consisting of 270 records.

Cluster 4 is the Elderly_C_DS dataset, consisting of 679 records.

```
Labor_C_DS=df[df['Cluster_Name']=='Labor_C']
Labor_C_DS.shape
```

```
(270, 10)
```

```
Elderly_C_DS=df[df['Cluster_Name']=='Elderly_C']
Elderly_C_DS.shape
```

```
(679, 10)
```

```
White_collar_DS=df[df['Cluster_Name']=='White-collar_C']
White_collar_DS.shape
```

```
(419, 10)
```

```
Small_City_DS=df[df['Cluster_Name']=='Small_City_C']
Small_City_DS.shape
```

```
(632, 10)
```

Fig.11: *Four New Clusters.*

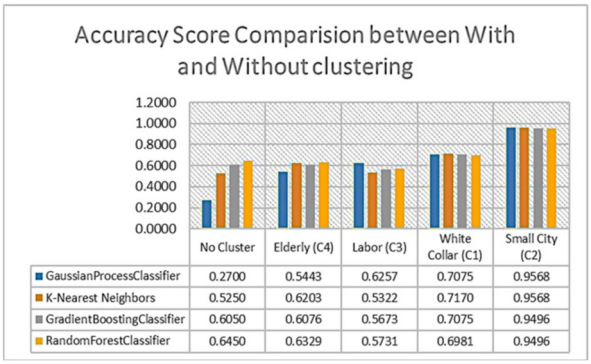


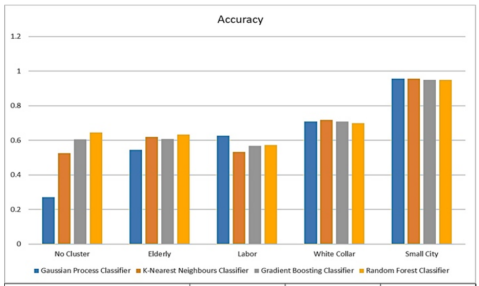
Fig.12: *Accuracy Score Comparison.*

Figure 12 demonstrates meaningful improvements in clusters C1 (White-collar) and C2 (Small City). Additionally, models for C4 (Elderly) also exhibit enhancements over simple classification (without clustering). However, the most notable improvement is seen in the small city cluster. Specifically, the Gaussian Process Classifier model shows the most considerable improvement, increasing from 0.27 to 0.9568, marking a substantial improvement of 68.68%. These increases in accuracy scores underscore the importance of clustering for model refinement. Nevertheless, determining the most suitable classifier for the predictive model depends on evaluating the Receiver Operating Characteristic (ROC) Curve and the Area Under the ROC Curve (AUC) values in Section 5.

5. PERFORMANCE EVALUATION

This research demonstrates how the new algorithm can enhance overall performance. Table 1 illustrates the original low performance of a simple classification. To improve this, the researcher proposes

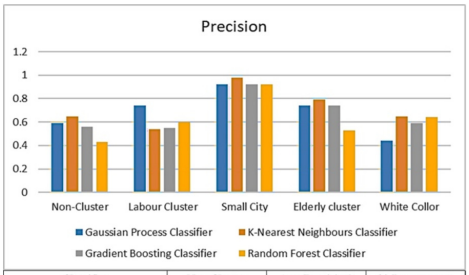
the OBHC algorithm, which utilizes clustering techniques. Specifically, the K-Means algorithm segments the dataset into smaller clusters based on its characteristics. This approach, termed 'Observation-based Clustering,' differs from traditional clustering methods by considering the dataset as a whole rather than clustering based on individual attributes and statistical values. These evaluate metric values after clustering to demonstrate the performance improvement.



Classifier	No Cluster	Avg (Accuracy)	% Improvment
Gaussian Process Classifier	0.2700	0.7086	43.86%
K-Nearest Neighbours Classifier	0.5250	0.7066	18.16%
Gradient Boosting Classifier	0.6050	0.7080	10.30%
Random Forest Classifier	0.6450	0.7134	6.84%

Fig.13: *Average Accuracy Score.*

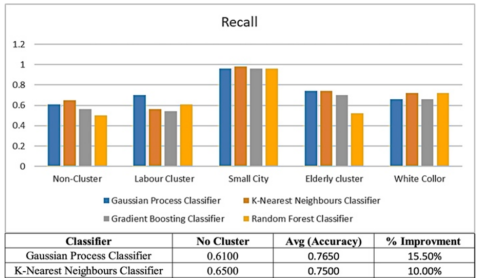
Figure 13 shows that all classifiers yield significantly higher average accuracy scores after clustering.



Classifier	Non-Cluster	Avg (Precision)	% Improvment
Gaussian Process Classifier	0.5900	0.7100	12.00%
K-Nearest Neighbours Classifier	0.6500	0.7400	9.00%
Gradient Boosting Classifier	0.5600	0.7000	14.00%
Random Forest Classifier	0.4300	0.6725	24.25%

Fig.14: *Average Precision Score.*

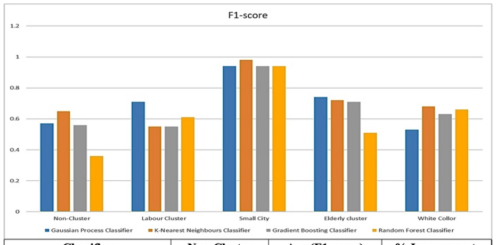
Figure 14 shows that all algorithms achieved a significantly higher average precision score after clustering.



Classifier	No Cluster	Avg (Accuracy)	% Improvment
Gaussian Process Classifier	0.6100	0.7650	15.50%
K-Nearest Neighbours Classifier	0.6500	0.7500	10.00%
Gradient Boosting Classifier	0.5600	0.7150	15.50%
Random Forest Classifier	0.5000	0.7025	20.25%

Fig.15: *Average Recall Score.*

Figure 15 shows that all algorithms attained significantly higher average recall scores after clustering.



Classifier	Non-Cluster	Avg (F1-score)	% Improvment
Gaussian Process Classifier	0.5700	0.7300	16.00%
K-Nearest Neighbours Classifier	0.6500	0.7325	8.25%
Gradient Boosting Classifier	0.5600	0.7075	14.75%
Random Forest Classifier	0.3600	0.6800	32.00%

Fig.16: *Average F1-Score.*

Figure 16 shows that all algorithms achieved significantly higher average F1 scores after clustering.

Based on the results, it can be concluded that organizing the data into four clusters led to improved performance.

In this section, the researcher focuses on assessing the effectiveness of classification models. The Receiver Operating Characteristic (ROC) curve and the Area Under the ROC Curve (AUC) are favored among other metrics due to their dual role as visualization tools and numerical indicators. Given that the class label comprises three values (S, M, L), the classification task falls into the Multiclass ROC type, necessitating the computation of an average ROC curve. Yellow brick, a widely used Machine Learning Visualization tool, is often employed for plotting ROC curves.

5.1 ROC Curves of The Original Dataset

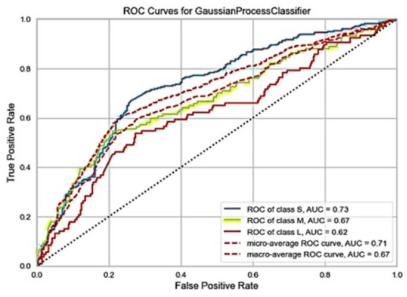


Fig.17: *Gaussian Process Classifier (Original Dataset).*

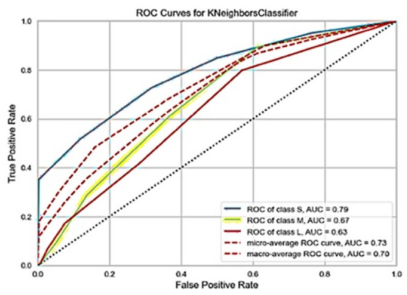


Fig.18: *K-Nearest Neighbours (Original Dataset).*

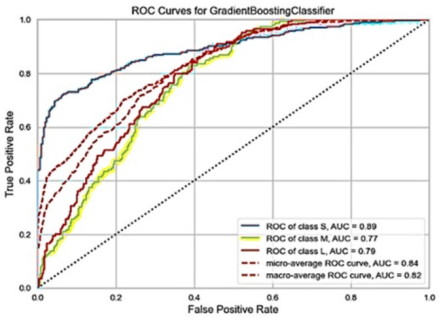


Fig.19: *Gradient Boosting Classifier (Original Dataset).*

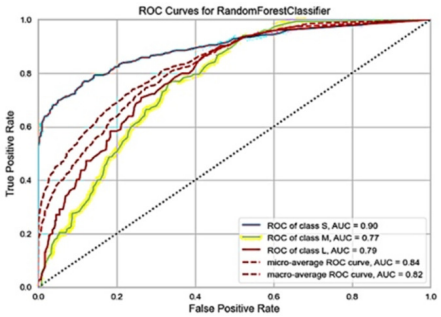


Fig.20: *Random Forest Classifier (Original Dataset).*

5.2 ROC Curves of the White-Collar Cluster

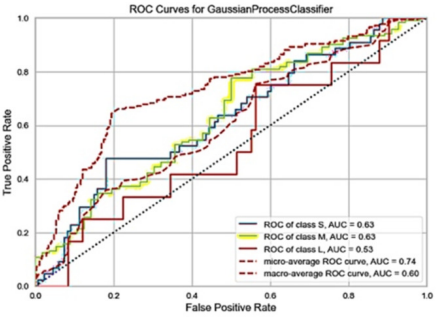


Fig.21: *Gaussian Process Classifier (White Collar Dataset).*

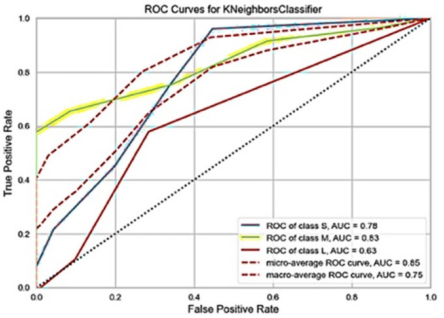


Fig.22: *K-Nearest Neighbours (Original Dataset).*

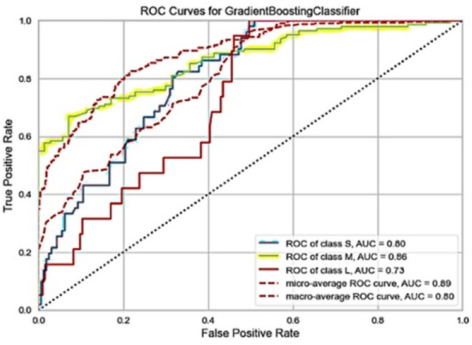


Fig.23: *Gradient Boosting Classifier (White Collar Dataset).*

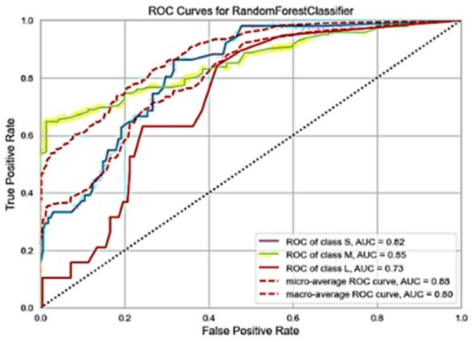


Fig.24: *Random Forest Classifier (White Collar Dataset).*

White Collar	precision	recall	f1-score
Gaussian Process Classifier	0.64	0.72	0.66
K-Nearest Neighbours Classifier	0.59	0.66	0.63
Gradient Boosting Classifier	**0.65**	**0.72**	**0.68**
Random Forest Classifier	0.44	0.66	0.53

Fig.25: *Performance Matrix (White Collar Dataset).*

In summary, within the white-collar cluster, the gradient-boosting classifier is identified as the top performer, owing to its achievement of the highest AUC in the micro-average ROC. The results from the performance matrix align with the ROC analysis, indicating that the Gradient Boosting Classifier is the preferred choice for white-collar clusters. Additionally, high-salaried workers and executives often choose to live in small and medium-sized cities because they may not be required to commute to the workplace frequently. Certain businesses provide adaptable work options such as remote work or flexible hours, reducing the need for strict time attendance.

5.3 ROC Curves of the Small City Cluster

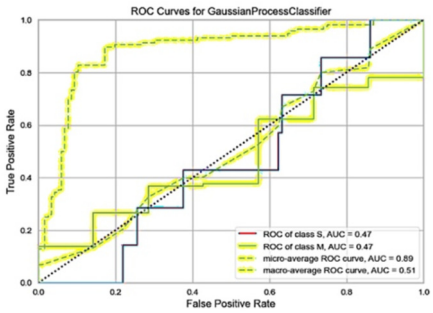


Fig.26: *Gaussian Process Classifier (Small City Cluster).*

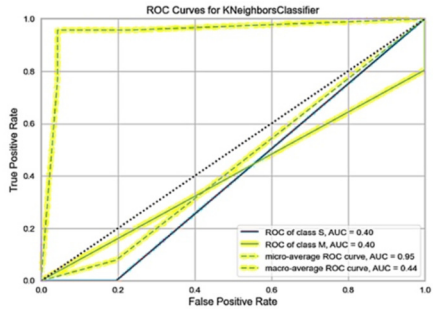


Fig.27: *K-Nearest Neighbours (Small City Cluster).*

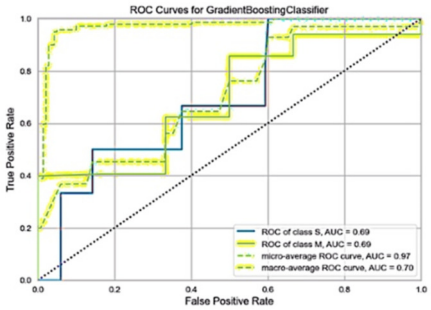


Fig.28: *Gradient Boosting Classifier (Small City Cluster).*

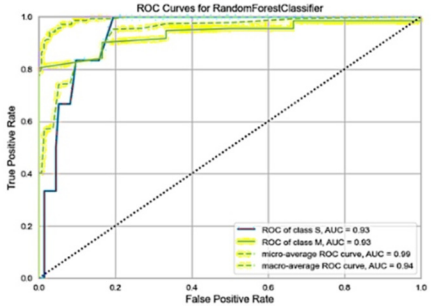


Fig.29: *Random Forest Classifier (Small City Cluster).*

Small City	precision	recall	f1-score
Gaussian Process Classifier	0.92	0.96	0.94
K-Nearest Neighbours Classifier	0.92	0.96	0.94
Gradient Boosting Classifier	**0.98**	**0.98**	**0.98**
Random Forest Classifier	0.92	0.96	0.94

Fig.30: *Performance Matrix (Small City Cluster).*

In summary, the Gradient Boosting Classifier is the top-performing classifier for the small-city cluster. The results from the performance matrix align with the ROC analysis, concluding that the Gradient Boosting Classifier is the optimal choice for the small-city cluster.

5.4 ROC Curves of the Labor Cluster

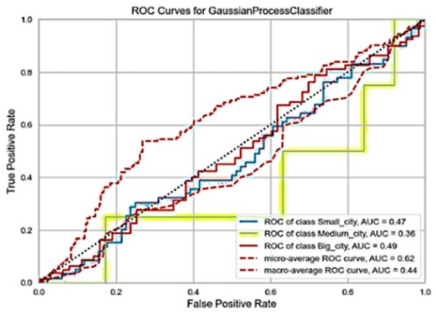


Fig.31: *Gaussian Process Classifier (Labor Cluster).*

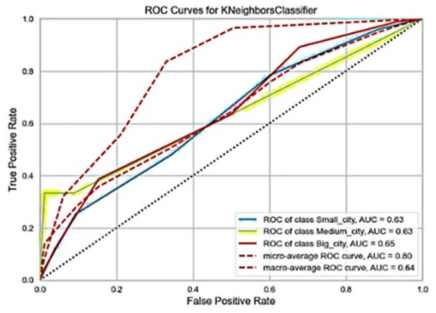


Fig.32: *K-Nearest Neighbours (Labor Cluster).*

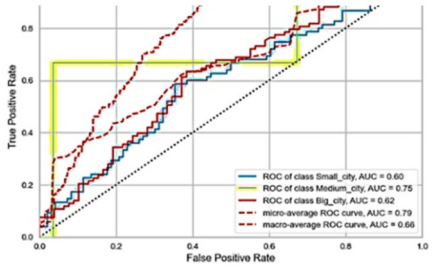


Fig.33: *Gradient Boosting Classifier (Labor Cluster).*

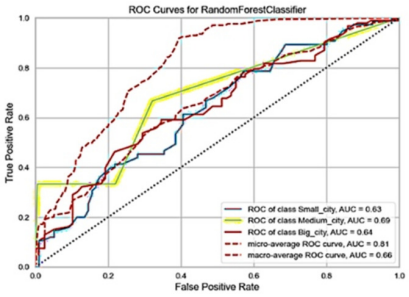


Fig.34: *Random Forest Classifier (Labor Dataset).*

White Collar	precision	recall	f1-score
Gaussian Process Classifier	0.64	0.72	0.66
K-Nearest Neighbours Classifier	0.59	0.66	0.63
Gradient Boosting Classifier	**0.65**	**0.72**	**0.68**

Fig.35: *Performance Matrix (Labor Dataset).*

In summary, within the labor cluster, the Random Forest Classifier emerges as the preferred option, with a micro-average ROC curve AUC of 0.71, the highest among the classifiers evaluated. The results from the performance matrix corroborate the ROC analysis, reinforcing the conclusion that the Random Forest Classifier is the optimal selection for the labor cluster. Additionally, labor or shift workers often prefer to reside in medium to large cities due to the greater availability of job opportunities than in smaller towns.

5.5 ROC Curves of the Elderly Cluster

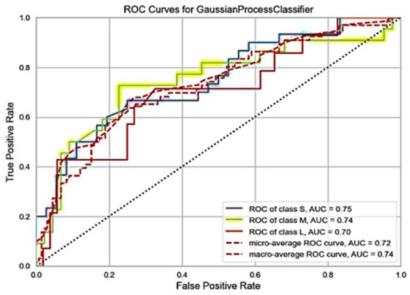


Fig.36: *Gaussian Process Classifier (Elderly Cluster).*

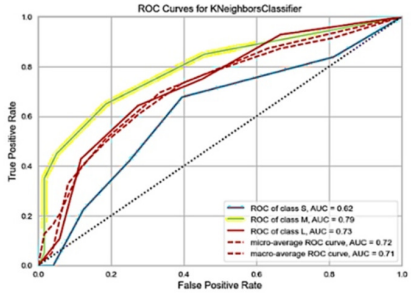


Fig.37: *K-Nearest Neighbours (Elderly Cluster).*

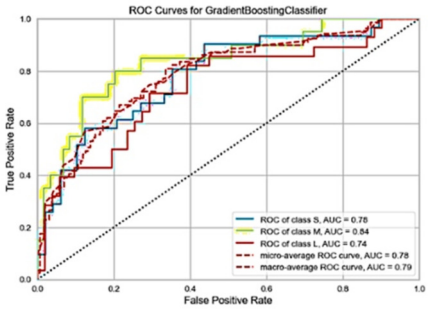


Fig.38: *Gradient Boosting Classifier (Labor Cluster).*

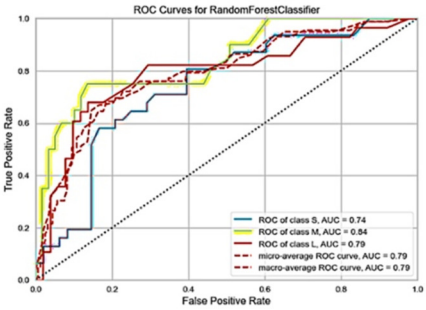


Fig.39: *Random Forest Classifier (Elderly Cluster).*

Elderly	precision	recall	f1-score
Gaussian Process Classifier	0.53	0.52	0.51
K-Nearest Neighbours Classifier	0.74	0.70	0.71
Gradient Boosting Classifier	0.79	0.74	0.72
Random Forest Classifier	**0.74**	**0.74**	**0.74**

Fig.40: *Performance Matrix (Elderly Cluster).*

In summary, for the elderly clusters, the Random Forest Classifier emerges as the best classifier due to the maximum AUC of the micro-average ROC. The results from the performance matrix align with the ROC analysis, indicating that the Random Forest Classifier is the preferred choice for elderly clusters.

5.6 Visualization Tree

A decision tree is a widely used graphical method for visualizing scenarios. In this study, the free software tool Graphviz was chosen. Figures 41 through 44 display four different representations of a decision tree. The new knowledge extracted from the tree indicates that factors such as occupation, income, age, and sex significantly influence the model, as demonstrated in the visualizations. Specifically, the visualization illustrates how income and age impact the decision tree model within the Small City cluster. In the Labor cluster, income stands out as the most influential factor. Additionally, education, income, and age are pinpointed as factors affecting settlement size for the Elderly clusters, while all aspects, including sex, are also considered for the White-collar cluster.

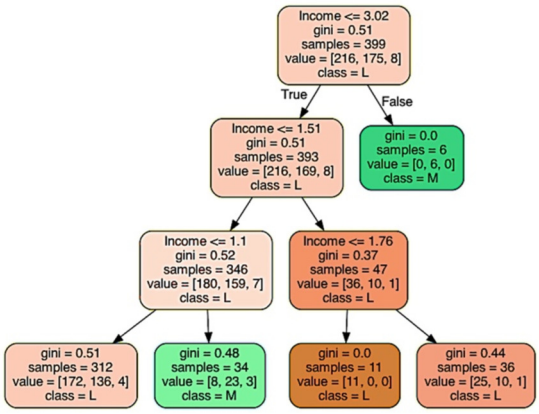


Fig.41: *The Graphic Visualization Tree for The Labor Cluster.*

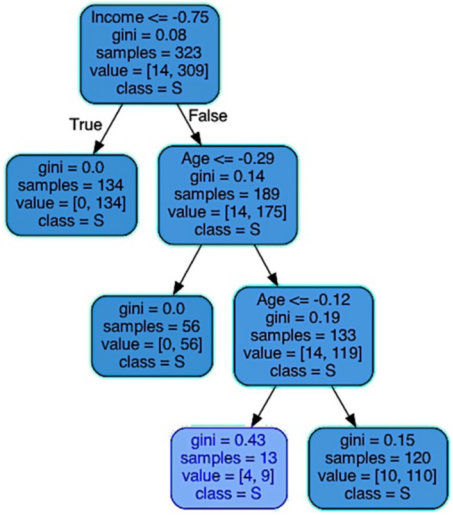


Fig.42: *The Graphic Visualization Tree for The Small City Cluster.*

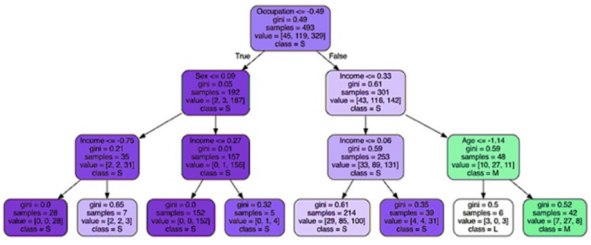


Fig.43: *The Graphic Visualization Tree for The White-Collar Cluster.*

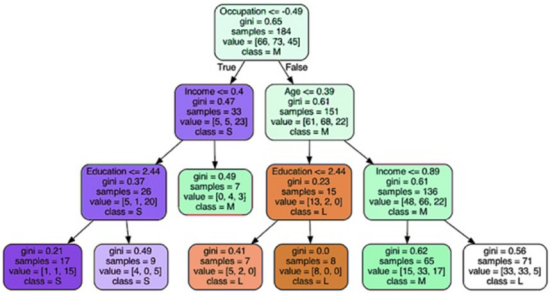


Fig.44: *The Graphic Visualization Tree for The Elderly Cluster.*

6. DISCUSSION AND FUTURE WORK

The experiment results show that performance metrics such as accuracy, precision, recall, and F1-score consistently improve when the data is clustered into smaller segments. This enables clearer personalization of customer groups, and the average score of each performance metric has significantly improved.

Customer segmentation can help understand customers, and it also enables the definition of specific marketing strategies tailored to individual customers. Future research aims to develop algorithms that can operate automatically without requiring programming, utilizing actual data to reflect real-world use cases. This approach will effectively benefit specific industries.

7. CONCLUSION

The research aims to demonstrate how a hybrid model can enhance the performance of basic classification. Additionally, it illustrates how new knowledge can be discovered by utilizing both clustering and classification models.

The evidence presented in section 5 demonstrates that the algorithm significantly enhances overall performance, as evidenced by notable improvements observed, including up to a 43.86% increase in average accuracy score (Gaussian Process Classifier), a 24.25% increase in average precision score (Random Forest Classifier), a 20.25% increase in average recall score (Random Forest Classifier), and a 32% increase in average F1-score (Random Forest Classifier). These metrics are essential for assessing the effectiveness of the OBHC algorithm in enhancing the customer segmentation model and improving the overall performance of classification algorithms, strongly advocating for the adoption of the OBHC hybrid algorithm.

A decision tree visualization was utilized to illustrate how income and age influence the decision tree model, particularly within the Small City cluster. In contrast, occupation and income emerged as significant factors within the Labor cluster.

In summary, the Observation-Based Hybrid Classification algorithm stands out from conventional classification methods in customer segmentation due to its integration of clustering, focus on uncovering new knowledge, customization of predictive models, performance enhancement, and provision of a unique and innovative approach to data analysis and segmentation within the realm of customer relationship management. Overall, leveraging the algorithm to improve data models with low performance across various customer segmentation scenarios can result in more precise targeting, enhanced customer relationships, optimized resource allocation, and a competitive edge in the market.

AUTHOR CONTRIBUTIONS

Kulkatechol Kanokngamwitroj: Conceptualization, methodology, software, field study, writing—original draft preparation, formal analysis, writing—review and editing. Chetneti Srisaan: Data curation, software, validation, visualization, investigation, and supervision. All authors have read and agreed to the published version of the manuscript.

References

[1] V. Mišković, "Machine Learning of Hybrid Classification Models for Decision Support," in *Proceedings of Sinteza 2014 - Impact of the Internet on Business Activities in Serbia and Worldwide*, Belgrade, Serbia, pp. 318-323, 2014.

[2] S. R. Gaddam, V. V. Phoha and K. S. Balagani, "K-Means+ID3: A novel method for supervised anomaly detection by cascading K-Means clustering and ID3 decision tree learning methods," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 345–354, Mar. 2007.

[3] T-T. Wong, N. Y. Yang and G-H. Chen, "Hybrid Classification Algorithms Based on Instance Filtering," *Information Sciences*, vol. 520, pp. 445-455, 2020.

[4] J. Xiao, Y. Tian, L. Xie and J. Huang, "A Hybrid Classification Framework Based on Clustering," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 7, pp. 4757-4768, Jul. 2020.

[5] L. Zhang, F. Lu, A. Liu, P. Guo and C. Liu, "Application of K-Means Clustering Algorithm for Classification of NBA Guards," *International Journal of Science and Engineering Applications*, vol. 5, no. 1, pp. 1-6, 2016.

[6] A. Deligiannis, C. Argyriou and D. Kourtesis, "Predicting the optimal date and time to send personalized marketing messages to repeat buyers," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, 2020.

[7] K. Tabianan, S. Velu and V. Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," *Sustainability*, vol. 14, no. 12, p. 7243, 2022.

[8] P. Coelho, M. Vilares, D. Ball, S. Coelho and M. Vilares, "Service Personalization and Loyalty," *Journal of Services Marketing*, vol. 20, no. 7, pp. 462-472, 2006.

[9] K. Kanokngamwitroj and C. Srisa-An, "Personalized Learning Management System using a Machine Learning Technique," *TEM Journal*, vol. 11, no. 4, pp. 1626-1633, 2022.

[10] J. Erman, M. Arlitt and A. Mahanti, "Traffic Classification Using Clustering Algorithms," in *Proceedings of the 2006 SIGCOMM Workshop on Mining Network Data, MineNet'06*, Pisa, Italy, pp. 281-286, 2006.

[11] I. Lewaa, "Customer Segmentation Using Machine Learning Model: An Application of RFM Analysis," *Journal of Data Science and Intelligent Systems*, vol. 2, no. 1, pp. 1-16, Sep. 2023.

[12] X. Li and Y. S. Lee, "Customer Segmentation Marketing Strategy Based on Big Data Analysis and Clustering Algorithm," *Journal of Cases on Information Technology*, vol. 26, no. 1, pp. 1-16, 2024.

[13] H. Altartouri, H. Tamimi, and Y. Ashhab, "The impact of pre-clustering on classification of heterogeneous protein data," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 11, no. 1, p. 3, 2022.

[14] J. Wu and Z. Lin, "Research on customer segmentation model by clustering," in *Proceedings of the 2005 International Conference on Management Science and Engineering*, Zhang Jia Jie, China, pp. 316-318, 2005.

[15] A. Alghamdi, "A Hybrid Method for Customer Segmentation in Saudi Arabia Restaurants Using Clustering, Neural Networks, and Optimization Learning Techniques," *Arabian Journal for Science and Engineering*, vol. 48, pp. 2031-2039, 2022.

[16] N. Gautam and N. Kumar, "Customer segmentation using k-means clustering for developing sustainable marketing strategies," *Business Informatics*, vol. 16, no. 1, pp. 72–82, 2022.

[17] A. Hiziroglu, "Soft computing applications in customer segmentation: State-of-art review and critique," *Expert Systems with Applications*, vol. 40, no. 16, pp. 6491–6507, 2013.

[18] M. Mosa, N. Agami, G. Elkhayat, and M. Kholief, "A novel hybrid segmentation approach for decision support: A case study in banking," *The Computer Journal*, vol. 66, no. 5, pp. 1228–1240, 2023.

Kulkatechol Kanokngamwitroj is a Ph.D. candidate in Information Technology at the School of Digital Innovation Technology, Rangsit University, Thailand. She has over 20 years of experience in the IT business and holds degrees in Business Administration from Assumption University, Thailand. Her areas of study interest include Big Data, Machine Learning, Artificial Intelligence, and Digital Transformation.

Chetneti Srisaan is an experienced practitioner currently serving as Vice President for Innovation at Rangsit University in Thailand. In addition to his various roles in Thailand, including President of the Association of Council of IT Deans and Dean of the College of Digital Innovation Technology at Rangsit University, he also worked for the Advanced Research Group and United Airlines in the USA. His current research interests include machine learning, big data analysis, deep learning, and data privacy.