Avrupa Bilim ve Teknoloji Dergisi
Sayı 54, S. 127-143, Aralık 2024
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 54, pp. 127-143, December 2024
Copyright © 2024 EJOSAT
Research Article

Unsupervised Learning Methods and Application in Adult Census Income Dataset

Ayla Sayli[1*], Emel Ugurlu[2]

[1*] Yildiz Technical University, Faculty of Chemistry and Metallurgical, Department of Mathematical Engineering, Istanbul, Turkey, (ORCID: 0000-0003-0409-537X), sayli@yildiz.edu.tr

[2] Yildiz Technical University, Faculty of Chemistry and Metallurgical, Department of Mathematical Engineering, Istanbul, Turkey, (ORCID: 0009-0002-8055-0565), emel.ugurlu@yildiz.edu.tr

(İlk Geliş Tarihi 24 Kasım 2023 ve Kabul Tarihi 3 Nisan 2024)

(**DOI:** 10.5281/zenodo.14182242)

ATIF/REFERENCE: Sayli, A.& Ugurlu, E. (2024). Unsupervised Learning Methods and Application in Adult Census Income Dataset. *European Journal of Science and Technology*, (54), 127-143.

Abstract

Unsupervised Learning is a data analysis technique used to explorer latent structure in data. Unsupervised Learning is the process of concentrating together objects with similar properties and grouping different ones without supervising the elements of the data. In this study, K-Means, DBSCAN and BIRCH clustering algorithms, which are unsupervised learning methods, were applied to the Adult Census Income dataset with 14 attribute and the target attribute is the annual income target attribute on Jupyter Notebook using Python3. In general, this dataset was used for classification purposes according to the target attribute based on its values which are less than 50 thousand dollars (0, zero class) or not (1, one class). However the annual income may not be able to give the similar groups of people. The aim of this study is to find these groups not only based on the annual income, considering all the attibutes in the dataset, comparing the performances of the clustering algorithms to observe the effects of the optimal number of clusters on the results. We first preprocessed this dataset and named as the Preprocessed Dataset and then we solved the balancing problem in this preprocessed set by the SMOTE method and named as the SMOTED_Preprocessed Dataset. After applying the algorithms to the datasets, 2 and 3 clusters are found and the results of the clusters are evaluated and the features determining the clusters were interpreted.

Keywords: Machine Learning, Unsupervised Learning Methods, K-Means, DBSCAN, BIRCH.

Yetişkin Nüfus Sayımı Geliri Veri Setinde Denetimsiz Öğrenme Metotları ve Uygulaması

Öz

Denetimsiz Öğrenme, verilerdeki gizli yapıyı keşfetmek için kullanılan bir veri analizi tekniğidir. Denetimsiz Öğrenme, verinin elemanlarını denetlemeden, benzer özelliklere sahip nesneleri bir araya toplayıp farklı olanları gruplandırma işlemidir. Bu çalışmada denetimsiz öğrenme yöntemlerinden K-Means, DBSCAN ve BIRCH kümeleme algoritmaları Python3 kullanılarak Jupyter Notebook üzerinde 14 öznitelik ve hedef niteliği yıllık gelir olan Yetişkin Nüfus Sayımı Gelir veri setine uygulanmıştır. Genel olarak bu veri seti, 50 bin doların altında olan (0, sıfır sınıf) veya olmayan (1, bir sınıf) değerlerine dayanarak hedef niteliğine göre sınıflandırma amacıyla kullanılmıştır. Ancak yıllık gelir benzer insan gruplarını vermeyebilir. Bu çalışmanın amacı, bu grupları sadece yıllık gelire göre değil, veri setindeki tüm özellikleri dikkate alarak bulmak, kümeleme algoritmalarının performanslarını karşılaştırarak optimal küme sayısının sonuçlara etkisini gözlemlemektir. Bu veri setini ilk olarak ön işleme tabi tutarak Preprocessed Dataset adını verdik ve daha sonra bu ön işleme setindeki dengeleme problemini SMOTE yöntemi ile çözerek SMOTED_Preprocessed Dataset adını verdik. Algoritmalar veri setlerine uygulandıktan sonra 2'li ve 3'lü kümeler bulunarak kümelerin sonuçları değerlendirilerek kümeleri belirleyen özellikler yorumlanır.

Anahtar Kelimeler: Makine Öğrenmesi, Denetimsiz Öğrenme Metotları, K-Means, DBSCAN, BIRCH.

[*] Corresponding Author: sayli@yildiz.edu.tr

1. Introduction

With the rapid development of technology in recent years, the amount of data has also increased. With this change, big data are stored in databases and used in areas such as business intelligence, scientific research and project development. Machine Learning use data not only for knowledge discovery but also to determine hidden concepts for use in predicting future situation (Zou, H., 2020) (Chakrabarty, N., & Biswas, S., 2018, October). Unsupervised learning is one of the types of machine learning and unlike supervised learning or reinforcement learning, trains a model to find patterns in data without any labels for each input (Dayan, P., Sahani, M., & Deback, G., 1999). Unsupervised learning methods are generally divided into three such as partition based, hierarchical based and density based clustering as shown in Figure 1. Cluster analysis is an unsupervised learning method, which is one of the most important areas of machine learning. Clustering is a technique of finding the cluster structure in the dataset by maximizing the similarity of the elements in the same cluster and minimizing the similarity between the clusters (Zou, H., 2020) (Sinaga, K. P., & Yang, M. S., 2020). Clustering techniques is used in pattern recognition, image analysis, grouping of documents, bioinformatics and data compression (Bhattacharjee, P., & Mitra, P., 2021). There are many articles in the literature about clustering techniques in machine learning. Recent articles related this topic are given in Table 1 and Table 2.

In this study, K-Means, DBSCAN and BIRCH clustering algorithms from unsupervised machine learning on the Adult Census Income dataset (Becker, Barry and Kohavi, Ronny, 1996) were examined. The aim of this study is to compare the performance of clustering algorithms on the imbalanced dataset and the balanced dataset and to observe the effects of the optimal number of clusters on the results. This comparison was made using Silhouette Coefficient (SC), Davies Bouldin Index (DB), Calinski Harabasz Index (CH) metrics. The dataset used in this study is the Adult Census Income dataset, which can be accessed from the internet address. In the section 1, general information and objectives about the study are given and a literature study is conducted. In the section 2, the types of clustering techniques, K-Means, DBSCAN and BIRCH algorithms and clustering performance metrics are mentioned. In the section 3, the dataset is preprocessed, and in the section 4, four clustering algorithms are applied. The distribution of the effective features in clustering into clusters are interpreted. The performances of clustering algorithms applied to balanced and imbalanced datasets are given in tables. Finally, in the section 5, conclusions and recommendations are presented.

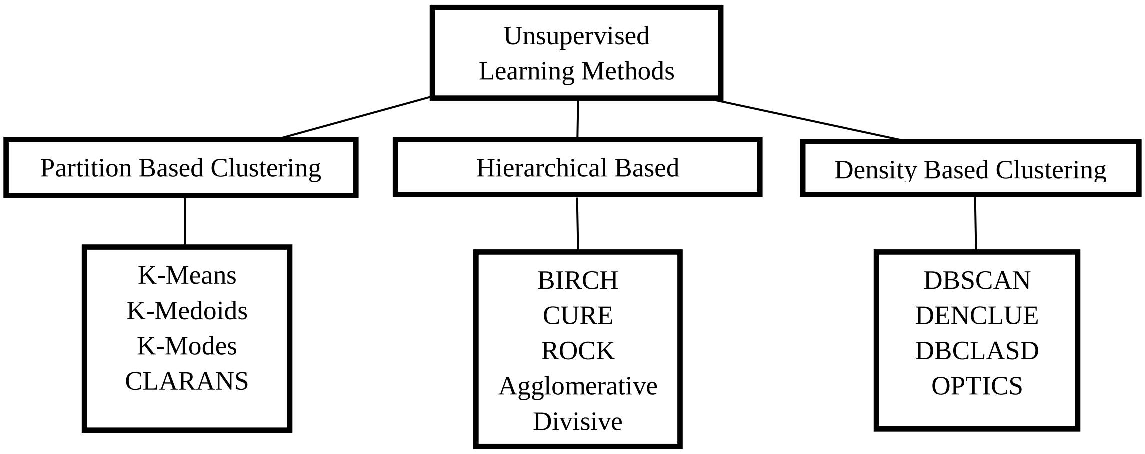


Figure 1. Unsupervised Learning Methods

Table 1. Literature Studies of Clustering Techniques

Study	Dataset	Clustering Algorithms	Summary
Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster (Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D., 2018, April)	Customer Profile	K-Means+ Elbow Method	By Sum of Square Error measurement, 3 clusters of 500 batik visitor data gave the best measurement.
A Hybrid Unsupervised Clustering-Based Anomaly Detection Method (Pu, G., Wang, L., Shen, J., & Dong, F., 2020)	NSL-KDD dataset	K-Means, DBSCAN, Sub-Space Clustering and One Class Support Vector Machine, SSC-EA	K-Means, DBSCAN and Space Clustering and One Class Support Vector Machine algorithms were applied to the dataset. It has been observed that SSC-OCSVM performs better than other methods.
A novel density-based clustering algorithm using nearest neighbor graph (Li, H., Liu, X., Li, T., & Gan, R., 2020)	Five artificial datasets, Wine, Irish, HTRU2, Seeds, Banknote, Ecoli, Leaf, NSL-KDD	DBSCAN, ADBS (Adaptive DBSCAN), HDBSCAN, DP, SNN (Shared Nearest Neighbors), CLUB, RNN (RNN-DBSCAN)	ADBSCAN clustering given better performance in Adjusted Rand Index and Normalized Mutual Information compared to other algorithm.
Unsupervised K-Means Clustering Algorithm (Sinaga, K. P., & Yang, M. S., 2020)	Four Synthetic Data, Iris, Seeds, Australian, Flowmeter D, Sonar, Wine, Horse, Waveform, SPECT, Parkinsons, WPBC, Colon, Lung and Nci9, CIFAR-10, Yale Face 32 x 32	Unsupervised k-means clustering, R-EM, Clustering by fast search, K-Means, Extended k-means, Robust-learning fuzzy c-means, k-means Gap-stat	The proposed Unsupervised k- means algorithm given better performance compared to other methods.
Improve BIRCH algorithm for big data clustering (Ramadhani, F., Zarlis, M., & Suwilo, S., 2020)	Online Retail Datasets	BIRCH, BIRCH (CF Leaf (modif))	The BIRCH (CF Leaf (modif)) algorithm given better performance in Silhouette Coefficient compared to other algorithm.
CSAL: Self-adaptive Labeling based Clustering Integrating Supervised Learning on Unlabeled Data (Li, F., Xu, G., & Cao, L., 2015)	Iris, Hearth Diseases, New Thyroid, Wine, Synthetic Gaussian data (GData1, GData2)	K-means, FCM, GMM, Proposed method: CSAL	In the proposed method, first it clusters unlabeled data and labels are given, and then classifies data by using a certain number of data for training. The proposed method gave better results than traditional methods.

Table 2. Literature Studies of Adult Census Income Dataset

Study	Dataset	Algorithms	Summary
A Statistical Approach to Adult Census Income Level Prediction (Chakrabarty, N., & Biswas, S., 2018, October)	Adult Census Income	Hyper Parameter Tuned Gradient Boosting Classifier	Gradient Boosting Classifier with Hyper-Parameter Tuning with Grid Search was applied to the dataset and gave the highest accuracy so far.
Adult Income Classification using Machine Learning Techniques (Moe, E. E., Win, S. S. M., & Khine, K. L. L., 2023, February)	Adult Census Income	Naïve Bayes, Decision Tree, J48 and Random Forest Classifiers	J48 Decision Tree algorithm given better performance than other algorithms.

This study can be used to group people according to their similarities in such as age, sex, relationship status, marital status, education, occupation, capital gain and capital loss features without annual income field and this study, like the study of Li et al. (Li, F., Xu, G., & Cao, L., 2015), can also be used to provide a target class.

2. Material and Method

2.1. Partition Based Clustering

Partition based clustering algorithms, the dataset is divided by the predefined number of clusters and all clusters are determined instantly. These clusters must satisfy some conditions: All cluster must contain at least one data point, and each data point must belong to the cluster. These algorithms iteratively replacement data points between clusters until they achieve the best clustering. Examples of this technique are the K-Means, K-Medoids, and CLARANS algorithms. (Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S. & Bouras, A., 2014). In this study, the K-Means algorithm from the 1.3.0 version scikit-learn library is used. This algorithm is the most used and easily implemented clustering method (Sinaga, K. P., & Yang, M. S., 2020) (Xu, R., & Wunsch, D., 2005). K-Means takes unlabelled data and divides the inputs into k clusters. The main purpose of this algorithm is to maximize the similarity of the data in the same cluster and to minimize the similarity between the clusters (Zou, H., 2020). Elbow method was used to determine the optimal number of clusters in K-Means (Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D., 2018, April).

2.2. Density Based Clustering

Density-based clustering algorithms use the idea of density to form clusters. It has features such as automatically discovering the number of clusters and identifying noisy data. Since data points located in low density areas are not included in any cluster, these points are treated as noisy observations. Clustering with density-based algorithms occurs in two steps: First, a method for determining the density of each sample is determined and used to find the core point. Second, samples that are connected to this core point are searched and then assigned to the same cluster (Bhattacharjee, P., & Mitra, P., 2021) (Hahsler, M., Piekenbrock, M., & Doran, D., 2019) (Li, H., Liu, X., Li, T., & Gan, R., 2020). In this study, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm from the 1.3.0 version scikit-learn library is used. The DBSCAN algorithm identifies high-density data as clusters and clusters with low-density data as outliers. (Hahsler, M., Piekenbrock, M., & Doran, D., 2019). DBSCAN algorithm takes two parameters Eps and Minpts. Eps specifies the neighborhood in the eps radius of each data point in the cluster. MinPts specifies the least data points around the cluster. If the Eps neighborhood of a data point contains at least MinPts of data points, this data point is a core point otherwise data point is border point. The algorithm assigns the boundary points to a set of nearest seed points. The data points that are not part of any cluster are called outliers. The DBSCAN algorithm is successful in detecting clusters that are separated in different ways (Bhattacharjee, P., & Mitra, P., 2021).

2.3. Hierarchical Based Clustering

Hierarchical based clustering algorithms groups similar data into given set. Agglomerative and divisive methods are hierarchical clustering technique types. In the agglomerative method, initially each of the data is a separate cluster and data that are close to each other are merged until all clusters merge into a single cluster. In the divisive method, all data points are in one cluster and each data is divided into smaller clusters until it becomes a separate cluster. In this study, the BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies) algorithm from the 1.3.0 version scikit-learn library is used. There are two concepts in the BIRCH algorithm, clustering features (CF) and cluster feature tree (CF Tree). BIRCH is a clustering algorithm that firstly creates a small summary of clustering useful information from a large data set. The data in the dataset is organized in a subset CF form (Ramadhani, F., Zarlis, M., & Suwilo, S., 2020). CF consists of three components with CF = (N, LS, SS). N is the number of data points, LS denotes their linear sum, and SS denotes their squared sum (Zhang, T., Ramakrishnan, R., & Livny, M., 1997). BIRCH incrementally calculates a summary of the CF subset. CF is an effective storage technique by summarizing information about the subset all data points (Ramadhani, F., Zarlis, M., & Suwilo, S., 2020).

2.4. Performance Metrics

Silhouette Coefficient (SC), Calinski Harabasz (CH) Index, Davies-Bouldin (DB) Index were used to evaluate the clustering results. In the SC calculation, the average of the distance of a data point to all other points in its closest cluster and the average of the distance to all other points in its cluster is used. SC takes values between -1 and 1 and a higher value indicate that the clustering algorithm performs better clustering (Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J., 2021). The DB Index measures the average similarity between cluster. The CH Index measures the sum of between-cluster distribution against the sum of within-cluster distribution. A lower value of the DB Index means better clustering and a higher value of the CH Index means better clustering (Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S. & Bouras, A., 2014) (Wang, X., & Xu, Y., 2019, July).

2.5. Principal Component Analysis (PCA)

The basic idea of PCA is to find important features in the data to express multidimensional data with fewer variables. PCA is a statistical method that uses an orthogonal transformation. PCA change a group of related variables into an unrelated group of variables. PCA can be used for exploratory data analysis and examination the relationships between variables. Therefore, PCA can be used for size reduction (Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T., 2020).

2.6. SMOTE

SMOTE (Synthetic Minority Oversampling Technique) produces artificial data from the minority class to create balanced datasets. To generate artificial data, first k-nearest neighbors are found for each sample and samples are selected from these neighbors. The difference between the examined feature and its nearest neighbor is found and multiplied by a value between 0 and 1 and added to the examined feature (Hussein, A. S., Li, T., Yohannese, C. W., & Bashir, K., 2019).

3. Data Collection, Pre-Processing and Visualization

3.1. Data Collection

Adult Census Income dataset was collected by Barry Becker in 1994 (Becker, Barry and Kohavi, Ronny, 1996). The dataset includes 32561 different entries and 14 attributes. As shown in Table 3 the dataset consists of 8 categorical, 6 continuous and 1 target attributes. The income field in the dataset is the dependent attributes that shows the annual income of a person is less than or more than 50 thousand Dollars.

Table 3. Data Definition (Becker, Barry and Kohavi, Ronny, 1996)

Data	Data Type	Description
age	Numeric	The age of individuals
workclass	Categorical	The work class of individuals
fnlwgt	Numeric	Description of final weight
education	Categorical	The education of individuals
education.num	Numeric	Years spent in education
marital.status	Categorical	The marital status of individuals
occupation	Categorical	The job of individuals
relationship	Categorical	The relationship of individuals
race	Categorical	The race of individuals
sex	Categorical	The sex of individuals
capital.gain	Numeric	The profit of individuals
capital.loss	Numeric	The loss of individuals
hours.per.week	Numeric	Working hours per week
native.country	Categorical	The native country of individuals
income	Categorical	The annual income of individuals

3.2. Data Pre-Processing

Data preprocessing is one of the data mining tasks that involves preparing data and converting it into a suitable form. The purpose of data preprocessing is to minimize the size of the data, solve the missing data problem, solve the outlier problem, and normalize the data (Alasadi, S. A., & Bhaya, W. S., 2017).

3.2.1. Handling Missing and Noisy Data

In the dataset, workclass, occupation and native.country fields contain values encoded as ‘?’. These values are missing values and were coded as ‘Nan’. The missing data in these three fields, which are categorical, were filled with the most frequently repeated value, mode. Since education.num is the numerical transformation of the education column, these two fields are related to each other, so the education.num field was removed from the dataset.

3.2.2. Detecting and Solving Outlier Problem

Outliers were detected in the numerical field of the dataset such as fnlwgt, hours.per.week fields to solve the outlier problem. The first quartile Q1values, third quartile Q3 values and IQR (Interquartile range) values which is the difference between Q3 and Q1 in this numerical fields were found. First quartile contains 25% of the observations in the data, third quartile contains 75% of the observations in the data (Najafabadi, M. Y., Heidari, A., & Rajcan, I., 2023). Then, the values of Q1-1.5*IQR and Q3+1.5*IQR were calculated and the values other than these values are called outliers. To solve the outlier problem, the winsorization method was applied to the outliers in the fnlwgt and hours.per.week fields. In this procedure, the outliers were converted into the values of Q1-1.5*IQR and Q3+1.5*IQR. Values 0 in the capital gain and capital loss fields were filled with the smallest value after 0. Formulas of Q1, Q3, and IQR are given below:

Lower (First)Quartile: $Q1 = (n+1)\frac{1}{4}$ (1)

Upper (Third)Quartile: $Q3 = (n+1)\frac{3}{4}$ (2)

Interquartile Range: $IQR = Q3 - Q1$ (3)

n: Size of data

3.2.3. Data Discretization

The age area was divided into 4 parts as 'Young Adult', 'Young-Middle-aged Adult', 'Middle-aged Adult', 'Old Adult' using the qcut method and labels were added. After discretization of the age field, the age field was dropped from the dataset.

3.2.4. Combining Categories in Features

Never-married, Divorced, Separated, Widowed properties in the marital.status field were changed to Single, and Married-civ-spouse, Married-spouse-absent, Married-AF-spouse properties were changed to Married. Black, Asian-Pac-Islander, Amer-Indian-Eskimo, Other properties in the race field were changed to Other because they are in the minority compared to the White race. Preschool, 1st- 4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th properties in the education field were changed to Primary or Secondary Schools. Not-in-family, Own-child, Unmarried, Other-relative properties in the relationship field were changed to Other.

3.2.5. Data Encoding

Label Encoding and One-hot Encoding are the process of converting categorical variables to numerical variables so that they can be used in predictions. The categorical data in the dataset such as race, sex, marital.status and income fields were converted into numerical data using Label Encoding and workclass, occupation, relationship, native.country, education and age fields were converted using One-hot Encoding. Results of data encoding 80 columns were created.

3.2.6. Normalization

Standardization scaling process was applied to fnlwgt, capital.gain, capital.loss, hours.per.week which are in numeric data type. The income field was removed from the dataset, since the clustering process does not require labelled data and after these processes, the dataset was named Preprocessed Dataset and saved.

3.2.7. Handling Imbalanced Data

In Adult Census Income dataset, the rate of people with annual income less than 50 thousand dollars is 76%, and the rate of people with annual income more than 50 thousand dollars is 24%, so this dataset is imbalanced. In the section 3.2.6. the dataset, which we named Preprocessed_Dataset, was balanced by sampling from the minority class using the Synthetic Minority Oversampling Technique (SMOTE) method and this balanced dataset was named SMOTED_Preprocessed Dataset and saved. As a result, two data sets were obtained: An imbalanced dataset called Preprocessed Dataset and a balanced data set called SMOTE_Preprocessed Dataset.

3.2.8. PCA

PCA was applied to the Preprocessed Dataset and SMOTED_Preprocessed Dataset for dimension reduction. In Figure 2 graph a cumulative explained variance graph for Preprocessed Dataset (blue one) and SMOTED_Preprocessed Dataset (red one). Afterwards, 5 components with the sum of the explained variance ratio of 0.6115 were selected for Preprocessed Dataset and 5 components with the sum of the explained variance ratio of 0.6748 were selected for SMOTED_Preprocessed Dataset. There is a 6% difference between cumulative explained variances ratio.

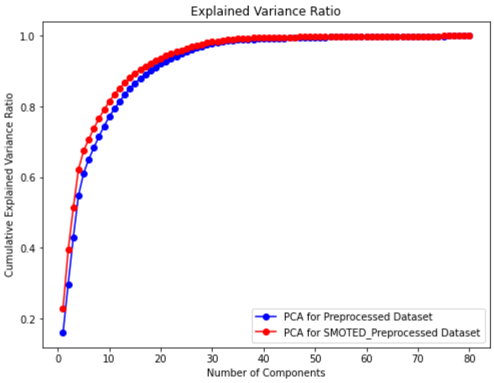


Figure 2. Cumulative Explained Variance Graph of Preprocessed Dataset and SMOTED_Preprocessed Dataset

4. Experiments Setup

4.1. Determining the Optimal Number of Clusters

Unsupervised learning methods were applied on the components determined by the PCA method. First, the K-Means algorithm was applied to the Preprocessed Dataset with PCA. Elbow and Silhouette Score methods were used to find the optimal number of clusters. In Figure 3, although the elbow point is not exactly clear according to the Elbow method, it reduced the rate of decrease in the sum of squares of error at point 3. According in Figure 4, the highest Silhouette Coefficient value was found in the 2 clusters and the optimal number of clusters was determined as 2 and 3.

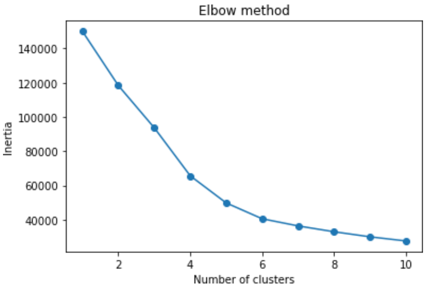


Figure 3. Elbow Method in Preprocessed Dataset with PCA for K-Means Clustering

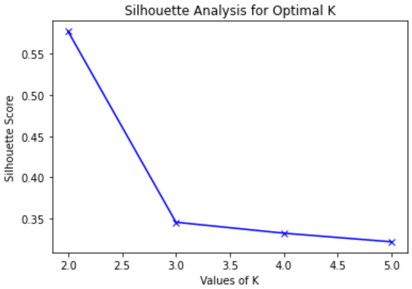


Figure 4. Silhouette Analysis in Preprocessed Dataset with PCA for K-Means Clustering

The K-Means algorithm was tested as both 2 and 3 clusters on Preprocessed Dataset with PCA. Figure 5 and Figure 6 show data points and cluster centers divided into 2 and 3 clusters.

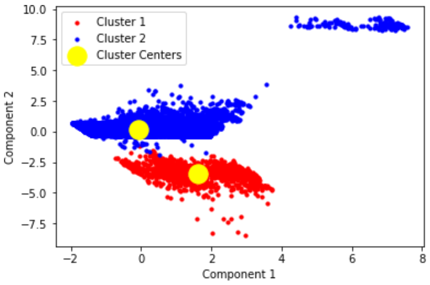


Figure 5. Two K-Means Clusters of Preprocessed Dataset with PCA

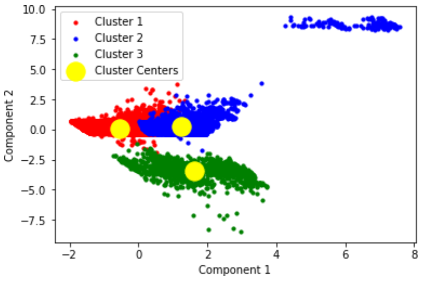


Figure 6. Three K-Means Clusters of Preprocessed Dataset with PCA

The K-Means algorithm was applied to the SMOTED_Preprocessed Dataset with PCA. The optimal number of clusters was determined as 3 according to the Elbow method by looking at the Figure 7 and the optimal number of clusters was determined as 2 according to Silhouette Score method because the highest Silhouette Coefficient value was found in the 2 clusters by looking at the Figure 8.

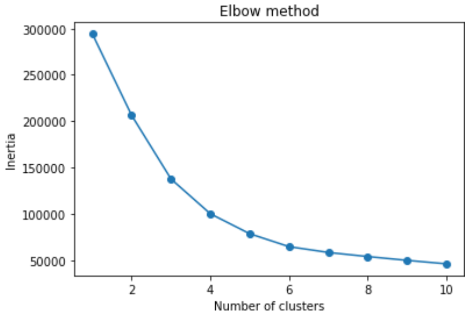


Figure 7. Elbow Method in SMOTED_Preprocessed Dataset with PCA for K-Means Clustering

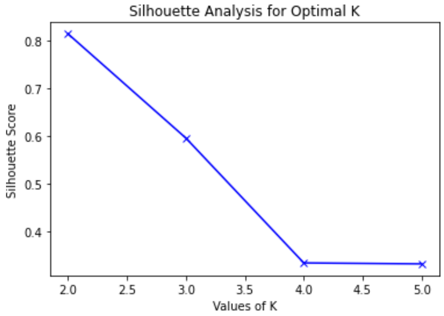


Figure 8. Silhouette Analysis in SMOTED_Preprocessed Dataset with PCA for K-Means Clustering

The K-Means algorithm was tested as both 2 and 3 clusters on SMOTED_Preprocessed Dataset with PCA. Figure 9 and Figure 10 show data points and cluster centers divided into 2 and 3 clusters.

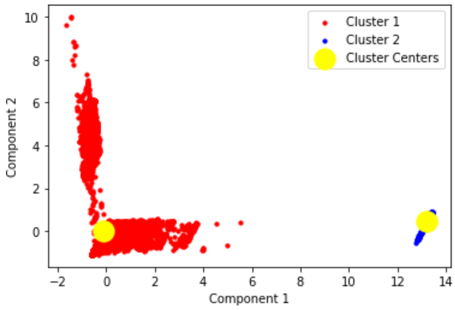


Figure 9. Two K-Means Clusters of SMOTED_Preprocessed Dataset with PCA

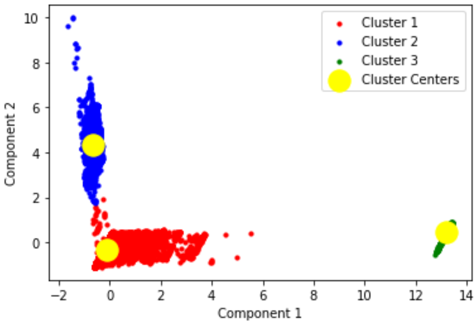


Figure 10. Three K-Means Clusters of SMOTED_Preprocessed Dataset with PCA

The DBSCAN algorithm was applied to the Preprocessed Dataset with PCA. In DBSCAN algorithm instead of experimenting with different epsilon values, the elbow point detection method was used to reach an appropriate epsilon value. In this method, the average distance is computed between each data point and its k nearest neighbors. The optimal value for Epsilon is the point with maximum bending. According to the graph in Figure 11, the Epsilon parameter was set to approximately 0.992. The min samples parameter was set to 50. As a result, the dataset was divided into 3 clusters according to the DBSCAN algorithm.

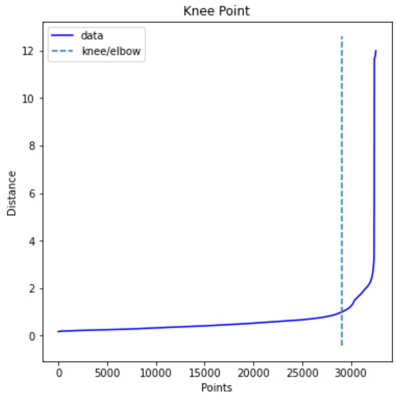


Figure 11. K Distance Graph

In Figure 12, the clusters were obtained results of the DBSCAN algorithm were shown in the graph. The 387 data points are labelled with -1 are noisy points.

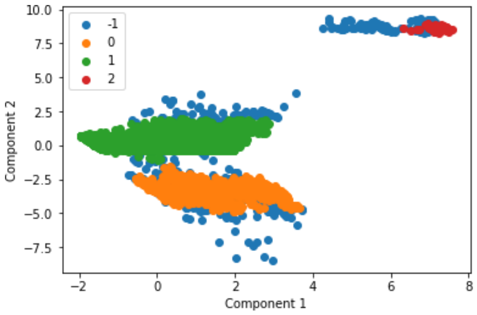


Figure 12. Three DBSCAN Clusters and Noises of Preprocessed Dataset with PCA

The Epsilon parameter was set to 1.5. The min samples parameter was set to 50. As a result, the dataset was divided into 2 clusters according to the DBSCAN algorithm. In Figure 13, the two clusters were obtained results of the DBSCAN algorithm was shown in the graph. The 56 data points are labelled with -1 are noisy points.

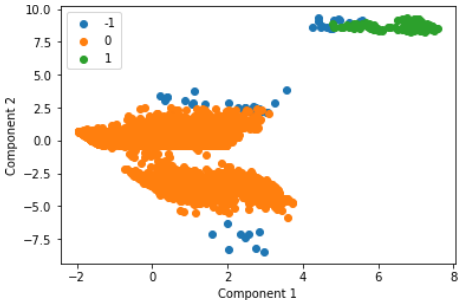


Figure 13. Two DBSCAN Clusters and Noises of Preprocessed Dataset with PCA

The DBSCAN algorithm was applied to the SMOTED_Preprocessed Dataset with PCA. According to the graph in Figure 14, the Epsilon parameter was set to approximately 1.03. The min samples parameter was set to 50. As a result, the dataset was divided into 3 clusters according to the DBSCAN algorithm.

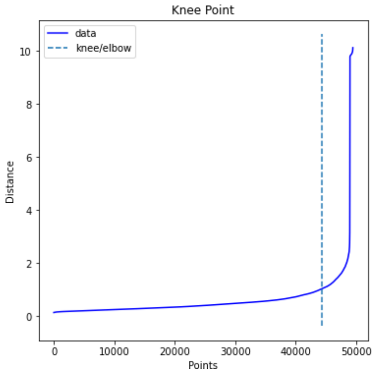


Figure 14. K Distance Graph

In Figure 15, the three clusters were obtained results of the DBSCAN algorithm was shown in the graph. The 146 data points are labelled with -1 are noisy points.

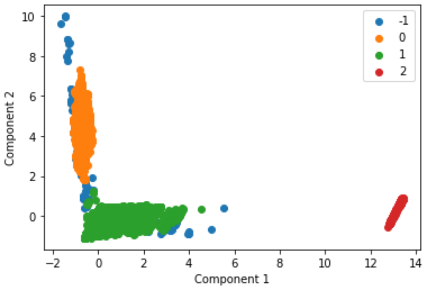


Figure 15. Three DBSCAN Clusters and Noises of SMOTED_Preprocessed Dataset with PCA

The Epsilon parameter was set to 2. The min samples parameter was set to 50. As a result, the dataset was divided into 2 clusters according to the DBSCAN algorithm. In Figure 16, the two clusters were obtained results of the DBSCAN algorithm was shown in the graph. The 11 data points are labelled with -1 are noisy points.

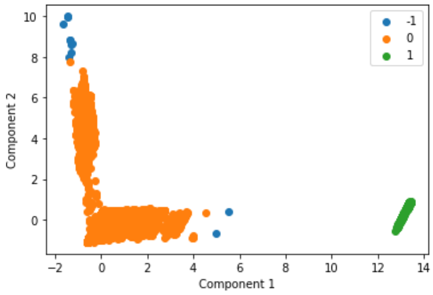


Figure 16. Two DBSCAN Clusters and Noises of SMOTED_Preprocessed Dataset with PCA

The BIRCH algorithm was applied to the Preprocessed Dataset with PCA. In order to determine the optimal number of clusters, SC, CH Index and DB Index Scores were examined. As explained in Section 2.4, a high value of SC and CH Index indicate better clustering performance. However, a low value of the DB Index indicates better clustering performance. According to the graph in Figure 17, the highest score in the SC was obtained when the dataset was divided into 2 clusters, the highest score in the CH Index was divided into 3 clusters, and the lowest score in the DB Index was obtained when it was divided into 2 clusters. As a result, the dataset was tried to be divided into both 2 and 3 clusters.

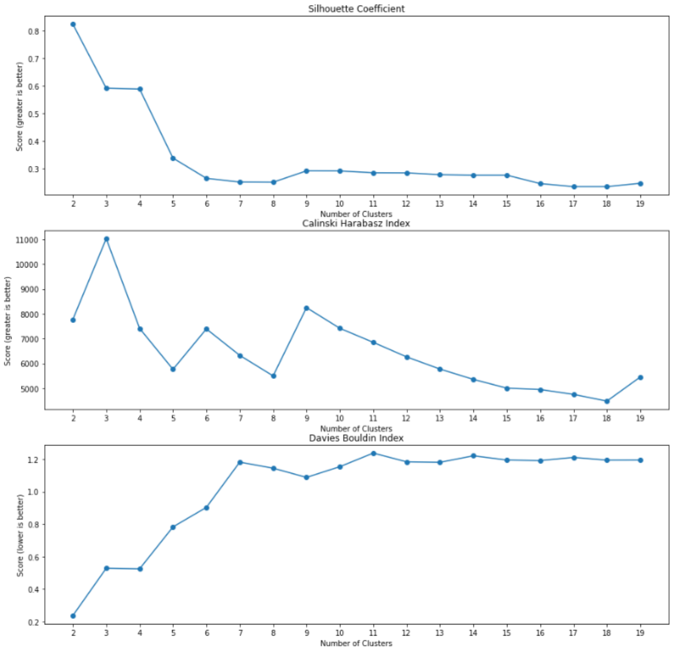


Figure 17. Silhouette Coefficient, Calinski Harabasz Index and Davies Bouldin Index Scores

The n_clusters parameter was set to 2, branching_factor and threshold were set to default values. In Figure 18, the two clusters were obtained results of the BIRCH algorithm was shown in the graph.

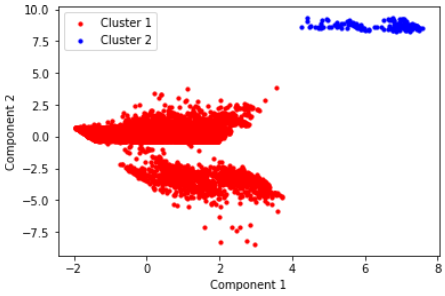


Figure 18. Two BIRCH Clusters of Preprocessed Dataset with PCA

The n_clusters parameter was set to 3, branching_factor and threshold were set to default values. In Figure 19, the two clusters were obtained results of the BIRCH algorithm was shown in the graph.

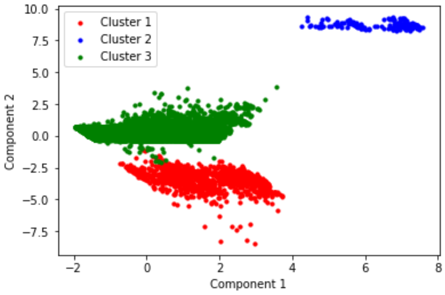


Figure 19. Three BIRCH Clusters of Preprocessed Dataset with PCA

The BIRCH algorithm was applied to the SMOTED_Preprocessed Dataset with PCA. As explained in Section 2.4, a high value of SC and CH Index indicate better clustering performance. However, a low value of the DB Index indicates better clustering performance. According to the graph in Figure 20, the highest score in the SC was obtained when the dataset was divided into 2 clusters, the highest score in the CH Index was divided into 3 clusters, and the lowest score in the DB Index was obtained when it was divided into 2 clusters. As a result, the dataset was tried to be divided into both 2 and 3 clusters.

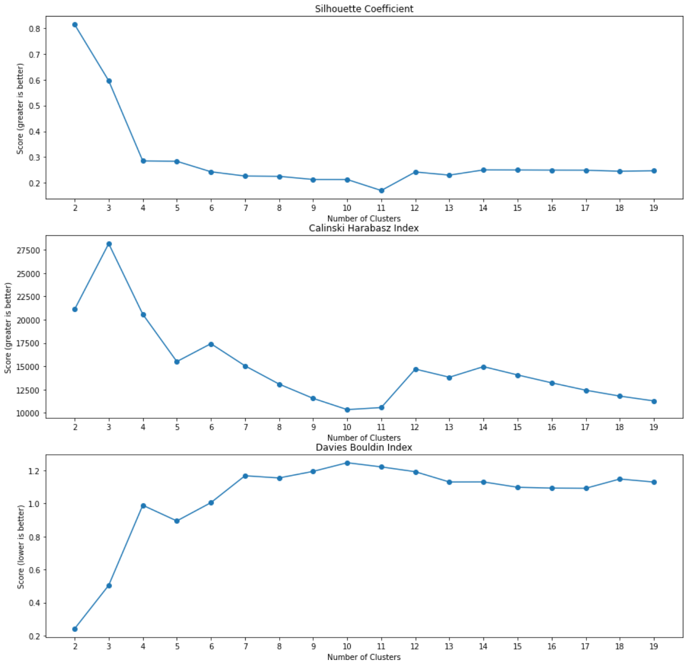


Figure 20. Silhouette Coefficient, Calinski Harabasz Index and Davies Bouldin Index Scores

The n_clusters parameter was set to 2, branching_factor and threshold were set to default values. In Figure 21, the two clusters were obtained results of the BIRCH algorithm was shown in the graph.

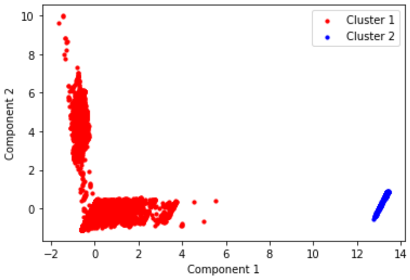


Figure 21. Two BIRCH Clusters of SMOTED_Preprocessed Dataset with PCA

The n_clusters parameter was set to 3, branching_factor and threshold were set to default values. In Figure 22, the two clusters were obtained results of the BIRCH algorithm was shown in the graph.

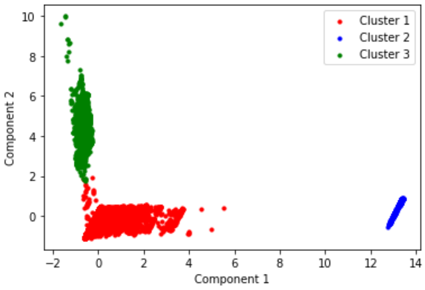


Figure 22. Three BIRCH Clusters of SMOTED_Preprocessed Dataset with PCA

4.2. Important Features in Clusters

The most important features affecting the distribution of clusters were determined by comparison of clusters and these features and features' categories are given in Table 4.

Table 4. The effective features in the clusters

Final weight, Marital Status, Sex, Capital Gain, Capital Loss, Hours Per Week
Education: Doctorate, HS-grad, Primary or Secondary Schools, Prof-school, Assoc-acdm, Some College
Workclass: Private, Self-emp-inc, Self-emp-not-inc, Federal-gov
Occupation: Handlers-cleaners, Farming-fishing, Craft-repair, Exec-managerial, Prof-specialty, Adm-clerical
Relationship: Husband, Wife, Other
Age: Young Adult, Young-Middle- aged Adult, Middle_aged Adult, Old Adult

In Figure 23, the cluster distribution of the 2-cluster BIRCH algorithm was given. The cluster distribution in Figure 23 is similar in other clustering algorithms. According to this graph, 99% of the people created Cluster 1 and 1% of the people created Cluster 2. Parameters such as age, sex, relationship status, marital status, education, occupation, capital gain and capital loss in the dataset cannot adequately explain the annual income.

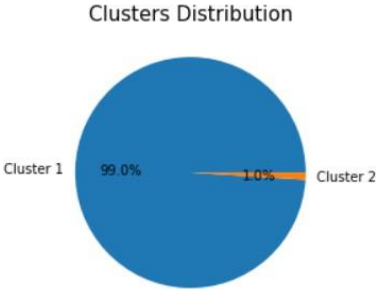
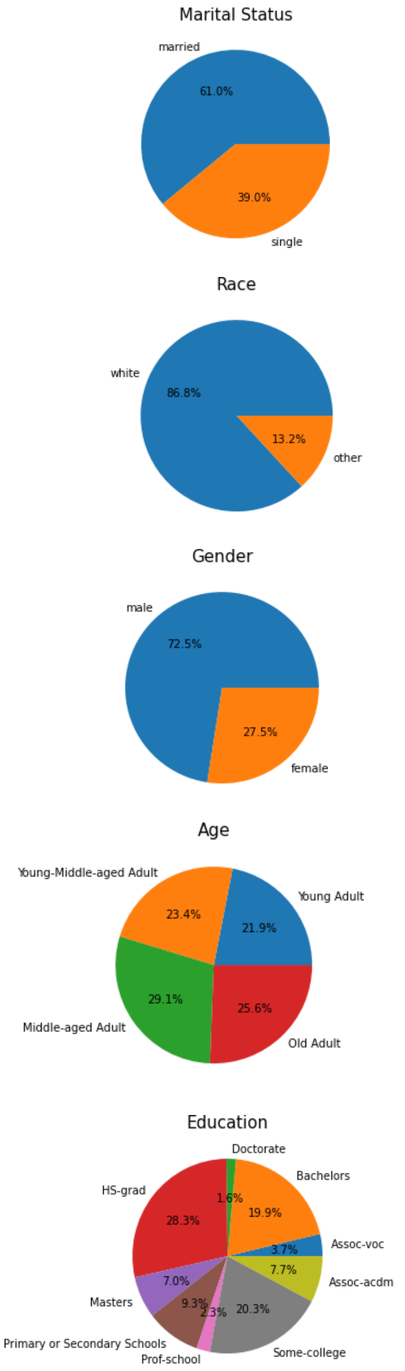
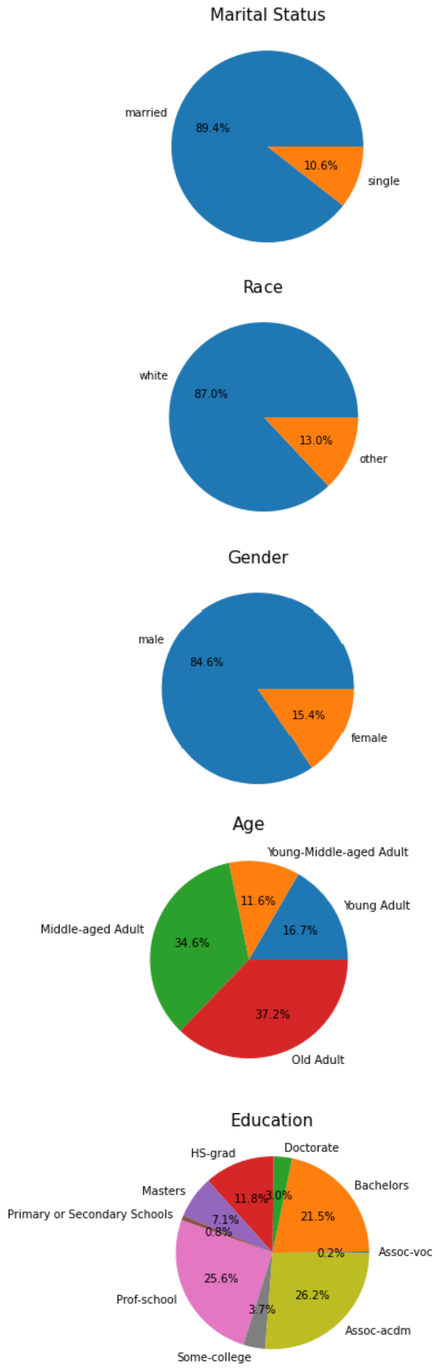
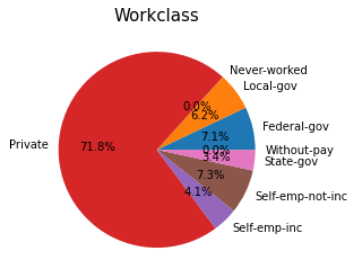


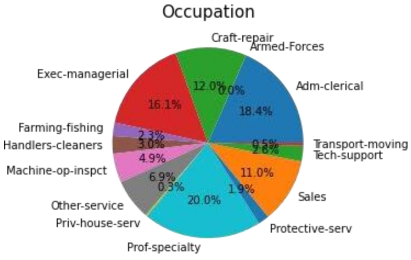
Figure 23. Clusters Distribution

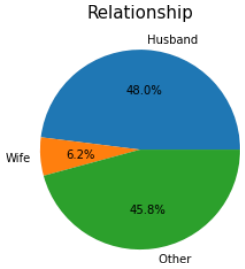
In the Table 5, the features of the clusters were visualized according to the 2-cluster BIRCH algorithm and the two clusters were compared according to these graphics. Clusters are graphed within themselves, so the vertical axis scale is different in bar graphs. According to Table 5, the number of married people in the cluster 1 and cluster 2 are higher than the singles and the married people in the cluster 2 are 28.4% more than the married people in the cluster 1. The race of people in the cluster 1 and cluster 2 are about the same. The number of males in the cluster 1 and cluster 2 are higher than the females and the males in the cluster 2 are 12.1% more than the males in the cluster 1. The Middle-aged Adult category and Old Adult category in age field is the higher in the cluster 2 than in the cluster 1. The Young-Middle-aged Adult and Young Adult category are higher than in the cluster 1 than in the cluster 2. In general, the average age of the cluster 2 is higher. The Primary and Secondary School graduates and High School graduates in cluster 1 is higher than in the cluster 2 and the bachelors, masters and doctorate graduates in cluster 2 is higher than cluster 1. As a result, the education level in cluster 2 is higher than in the cluster 1. The number of people work in private sector in the cluster 1 and cluster 2 are higher than the other sectors and the people work in the private sector in the cluster 1 are 19.8% more than the people work in private sector in the cluster 2. The number of people work in self-employed in own not incorporated business workers, self-employed in own incorporated business workers and federal government workers in the cluster 2 is higher than people in the cluster 1. The ratio of people who do occupations with a high level of education in cluster 2 is higher than in cluster one. The husband role in relationship field in the cluster 2 are 26% more than in the cluster 1. In the native country field, people whose native country is the United States are the most and this feature is not a very defining feature for the clusters. In the final weight area, in the range of 1000-2000, the number of people increases as the weight increases in cluster 1, but in cluster 2, the number of people decreases as the weight increases. According to the distribution of the capital gains field, there is a higher amount of earnings in the cluster 2 than in the cluster 1. In the field of capital loss, there is more loss in the cluster 1 than in the cluster 2. According to the weekly working hours graphs, it can be concluded that the second group's people work mean 4 hours longer than first cluster. Results of this comparison, it was determined that the people in the second group had a higher average age, higher education level and higher income compared to the first group.

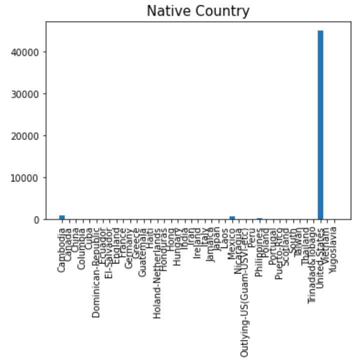
Table 5. Two BIRCH Clusters of SMOTED_Preprocessed Dataset with PCA

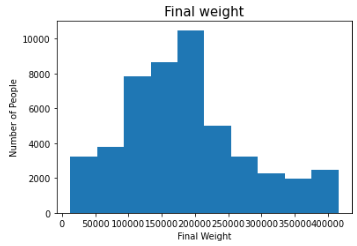
Cluster 1	**Cluster 2**
	

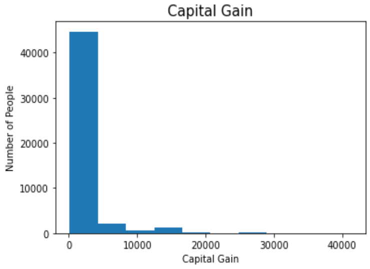

Workclass
Never-worked
Local-gov
Federal-gov
Without-pay
State-gov
Self-emp-not-inc
Self-emp-inc
Private
71.8%
0.0%
6.2%
7.1%
0.0%
3.4%
7.3%
4.1%

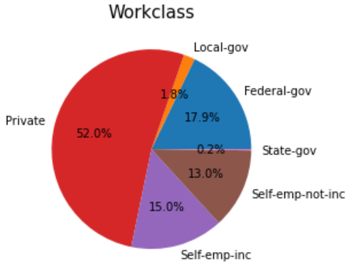

Occupation
Craft-repair
Armed-Forces
Exec-managerial
Adm-clerical
Farming-fishing
Handlers-cleaners
Machine-op-inspct
Transport-moving
Tech-support
Other-service
Priv-house-serv
Sales
Protective-serv
Prof-specialty
12.0%
0.0%
16.1%
18.4%
2.3%
3.0%
4.9%
0.5%
2.8%
6.9%
0.3%
11.0%
1.9%
20.0%

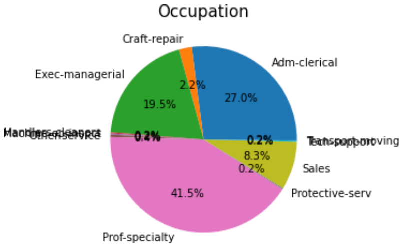

Relationship
Husband
48.0%
Wife
6.2%
45.8%
Other

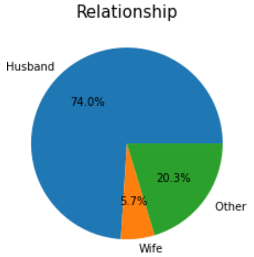

Native Country
40000
30000
20000
10000
0
Cambodia
Canada
China
Columbia
Cuba
Dominican-Republic
Ecuador
El-Salvador
England
France
Germany
Greece
Guatemala
Haiti
Holand-Netherlands
Honduras
Hong
Hungary
India
Iran
Ireland
Italy
Jamaica
Japan
Laos
Mexico
Nicaragua
Outlying-US(Guam-USVI-etc)
Peru
Philippines
Poland
Portugal
Puerto-Rico
Scotland
South
Taiwan
Thailand
Trinadad&Tobago
United-States
Vietnam
Yugoslavia

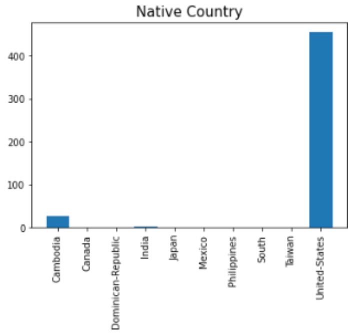

Final weight
Number of People
10000
8000
6000
4000
2000
0
0
50000
100000
150000
200000
250000
300000
350000
400000
Final Weight

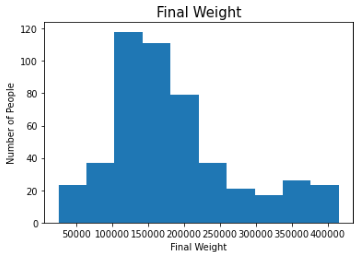

Capital Gain
Number of People
40000
30000
20000
10000
0
0
10000
20000
30000
40000
Capital Gain

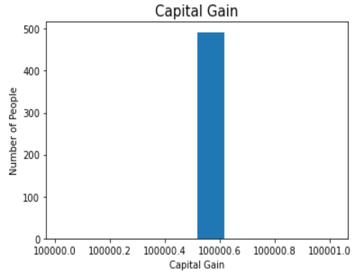

Workclass
Local-gov
Federal-gov
State-gov
Self-emp-not-inc
Self-emp-inc
Private
52.0%
1.8%
17.9%
0.2%
13.0%
15.0%

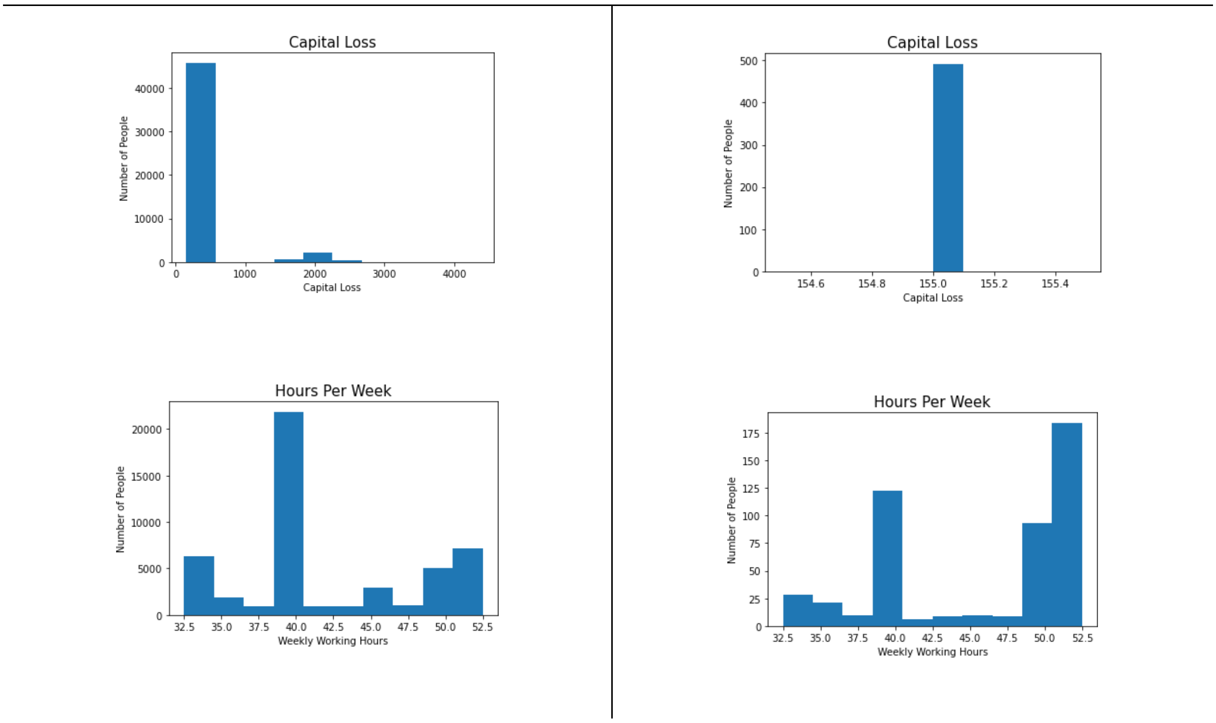

Occupation
Craft-repair
Adm-clerical
Exec-managerial
Handlers-cleaners
Machine-op-inspct
Other-service
Transport-moving
Tech-support
Sales
Protective-serv
Prof-specialty
2.2%
27.0%
19.5%
0.2%
0.1%
0.2%
8.3%
0.2%
41.5%


Relationship
Husband
74.0%
20.3%
5.7%
Other
Wife


Native Country
400
300
200
100
0
Cambodia
Canada
Dominican-Republic
India
Japan
Mexico
Philippines
South
Taiwan
United-States


Final Weight
Number of People
120
100
80
60
40
20
0
50000
100000
150000
200000
250000
300000
350000
400000
Final Weight


Capital Gain
Number of People
500
400
300
200
100
0
100000.0
100000.2
100000.4
100000.6
100000.8
100001.0
Capital Gain



4.3. Performance Metrics

The following tables, the performance metrics values of the clustering algorithms divided into 2 clusters and 3 clusters were given on the Preprocessed Dataset with PCA and SMOTED_Preprocessed Dataset with PCA. According to the tables, the areas marked in gray show which number of clusters an algorithm has the best performance metric. The bold marked indicate the on which dataset (Preprocessed Dataset and SMOTED_Preprocessed Dataset with PCA) the algorithms have the best performance metrics. The areas marked in yellow indicate performance metrics of the algorithms that give the best result in the Preprocessed Dataset and SMOTED_Preprocessed Dataset with PCA. According to Table 6, the algorithms with two clusters gave higher SC values than the algorithms with three clusters. The generally SMOTED_Preprocessed Dataset with PCA gave higher SC values than Preprocessed Dataset with PCA. The two-cluster BIRCH algorithm gave the best SC value with the ratios of 0.8242 in the Preprocessed Dataset with PCA and 0.8142 in the SMOTED_Preprocessed Dataset with PCA, the K-Means algorithm with two clusters gave the best SC value with the ratio of 0.8142 in SMOTED_Preprocessed Dataset with PCA.

Table 6. Silhouette Coefficient Results

Silhouette Coefficient	**Preprocessed Dataset with PCA**	**SMOTED_Preprocessed Dataset with PCA**
K-Means with 2 Cluster	0.5775	**0.8142**
K-Means with 3 Cluster	0.3456	**0.5958**
DBSCAN with 2 Cluster	0.7117	**0.7203**
DBSCAN with 3 Cluster	0.5714	**0.5910**
BIRCH with 2 Cluster	**0.8242**	0.8142
BIRCH with 3 Cluster	0.5921	**0.5959**

According to Table 7, the algorithms with two clusters gave lower DB Index values than the algorithms with three clusters. The generally SMOTED_Preprocessed Dataset with PCA gave the lowest DB Index values than Preprocessed Dataset with PCA. The BIRCH algorithm with two clusters gave the best DB Index value with the ratios of 0.2374 in Preprocessed Dataset with PCA and 0.2410 in the SMOTED_Preprocessed Dataset with PCA, the K-Means algorithm with two clusters gave the best DB Index value with the ratio of 0.2410 in SMOTED_Preprocessed Dataset with PCA.

Table 7. Davies Bouldin Index Results

Davies Bouldin Index	**Preprocessed Dataset with PCA**	**SMOTED_Preprocessed Dataset with PCA**
K-Means with 2 Cluster	0.6946	**0.2410**
K-Means with 3 Cluster	1.1675	**0.5046**
DBSCAN with 2 Cluster	1.1093	**0.5734**
DBSCAN with 3 Cluster	1.4235	**1.4099**
BIRCH with 2 Cluster	**0.2374**	0.2410
BIRCH with 3 Cluster	0.5284	**0.5042**

According to Table 8, the algorithms with 3 clusters gave higher CH Index values than the algorithms with 2 clusters. The generally SMOTED_Preprocessed Dataset with PCA gave the highest CH Index values than the Preprocessed Dataset with PCA. The BIRCH algorithm with 3 clusters gave the best CH Index value with the ratio of 11023.8333 in Preprocessed Dataset with PCA and in the K- Means algorithm with 3 clusters gave the best CH Index value with the ratio of 28251.7077 in SMOTED_Preprocessed Dataset with PCA.

Table 8. Calinski Harabasz Index Result

Calinski Harabasz Index	**Preprocessed Dataset with PCA**	**SMOTED_Preprocessed Dataset with PCA**
K-Means with 2 Cluster	8722.5345	**21232.2547**
K-Means with 3 Cluster	9807.7574	**28251.7077**
DBSCAN with 2 Cluster	3758.9191	**10743.2025**
DBSCAN with 3 Cluster	4989.5998	**18315.9073**
BIRCH with 2 Cluster	7770.9160	**21232.2547**
BIRCH with 3 Cluster	11023.8333	**28251.0707**

5. Conclusions and Recommendations

In this study, the best contribution is to group the people according to their similarities based on such as age, gender, relationship status, marital status, education, occupation, capital gain and capital loss attributes, not only considering the annual income and also this study can be used to provide a target class based on cluster numbers if any original set does not have this class for future researches. In detail, K-Means, DBSCAN, BIRCH clustering algorithms on the Adult Census Income dataset were examined. Other publications using this data set have worked with classification algorithms, and in this study, people were segmented with the clustering algorithms of this data set. In this study, it was examined whether these classes could be detected by clustering the two-class Adult Census Income data set. When Figures 3, 4, 17 are examined, the optimal number of clusters for Preprocessed Dataset with PCA was determined to be 2 and 3 and when Figures 7, 8, 20 are examined, the optimal number of clusters for SMOTED_Preprocessed Dataset with PCA was determined to be 2 and 3. The success of clustering algorithms was measured by performance metrics such as Silhouette Coefficient (SC), Davies Bouldin (DB) Index and Calinski Harabasz (CH) Index. In clustering algorithms, the effects of the optimal number of clusters on performance were investigated. According to the results, the algorithms in which the dataset is divided into 2 clusters gave better results. In general, the SMOTED_Preprocessed Dataset with PCA (balanced dataset) gave better results in performance metrics than the Preprocessed Dataset with PCA. The results show that, BIRCH and K-Means algorithms performed better on this dataset. Therefore, the clusters of the 2-cluster Birch algorithm are visualized in Table 5. With the help of this study, it has been seen that the clusters are determined by the characteristics such as the age, gender, relationship status, marital status, education, occupation, capital gain and loss of the individuals. Future research can achieve better clustering results by experimenting with other clustering algorithms.

References

Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences*, *12*(16), 4102-4107.

Becker,Barry and Kohavi,Ronny. (1996). Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20. Bhattacharjee, P., & Mitra, P. (2021). A survey of density based clustering algorithms. *Frontiers of Computer Science*, *15*, 1-27.

Chakrabarty, N., & Biswas, S. (2018, October). A statistical approach to adult census income level prediction. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* (pp. 207-212). IEEE.

Dayan, P., Sahani, M., & Deback, G. (1999). Unsupervised learning. *The MIT encyclopedia of the cognitive sciences*, 857-859.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., ... & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, *2*(3), 267-279.

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, *91*, 1-30.

Hussein, A. S., Li, T., Yohannese, C. W., & Bashir, K. (2019). A-SMOTE: A new preprocessing approach for highly imbalanced datasets by improving SMOTE. *International Journal of Computational Intelligence Systems*, *12*(2), 1412-1422.

Li, F., Xu, G., & Cao, L. (2015). CSAL: Self-adaptive Labeling based Clustering Integrating Supervised Learning on Unlabeled Data. *arXiv preprint arXiv:1502.05111*.

Li, H., Liu, X., Li, T., & Gan, R. (2020). A novel density-based clustering algorithm using nearest neighbor graph. *Pattern Recognition*, *102*, 107206.

Moe, E. E., Win, S. S. M., & Khine, K. L. L. (2023, February). Adult Income Classification using Machine Learning Techniques. In *2023 IEEE Conference on Computer Applications (ICCA)* (pp. 91-96). IEEE.

Najafabadi, M. Y., Heidari, A., & Rajcan, I. (2023). AllInOne Pre-processing: A comprehensive preprocessing framework in plant field phenotyping. *SoftwareX*, *23*, 101464.

Pu, G., Wang, L., Shen, J., & Dong, F. (2020). A hybrid unsupervised clustering-based anomaly detection method. *Tsinghua Science and Technology*, *26*(2), 146-153.

Ramadhani, F., Zarlis, M., & Suwilo, S. (2020). Improve BIRCH algorithm for big data clustering. In *IOP Conference Series: Materials Science and Engineering* (Vol. 725, No. 1, p. 012090). IOP Publishing.

Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *Ieee Access*, *8*, 54776-54788.

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *Eurasip Journal on Wireless Communications and Networking*, *2021*(1), 1-16.

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE access*, *8*, 80716-80727.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP conference series: materials science and engineering* (Vol. 336, p. 012017). IOP Publishing.

Wang, X., & Xu, Y. (2019, July). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. In IOP Conference Series: Materials Science and Engineering (Vol. 569, No. 5, p. 052024). IOP Publishing.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, *16*(3), 645-678.

Zhang, T., Ramakrishnan, R., & Livny, M. (1997). BIRCH: A new data clustering algorithm and its applications. *Data mining and knowledge discovery*, *1*, 141-182.

Zou, H. (2020). Clustering algorithm and its application in data mining. *Wireless Personal Communications*, *110*(1), 21-30.