The Journal of Supercomputing
https://doi.org/10.1007/s11227-022-04907-4

# Improve customer churn prediction through the proposed PCA-PSO-K means algorithm in the communication industry

**Maryam Sadeghi[1] · Mohammad Naderi Dehkordi[2] · Behrang Barekatain[2] · Naser Khani**

Accepted: 21 October 2022


**Abstract**
Customer churn prediction is one of the areas in Customer Relationship Management that differentiates loyal customers from factors that have a negative impact on business growth. Hence, various machine learning-based methods have been developed by researchers to accurately predict customer churn. However, high dimensionality and low prediction accuracy are problems in identifying averse customers. This paper presents a new system called PCA-PSO-K Means algorithm, which combines three algorithms: principal component analysis (PCA) for data set feature reduction, K Means algorithm for classification, and particle swarm optimization (PSO) algorithm to optimize K Means in providing initial centroids. The experimental results in the data set of one of the fixed internet providers in Isfahan Province show the improvement of the accuracy of customer churn prediction. The proposed system has an accuracy of 99.77%, a sensitivity of 75%, a specificity of 99.81% and a correlation coefficient of $0.443 \pm 0.271$. Found.

**Keyword** CRM · Customer churn prediction · Data mining · Principal component analysis · K Means algorithm · PSO algorithm

## 1 Introduction

Customer Relationship Management (CRM) is the strategy of how to interact and actively engage with customers. Successful CRM sees the business through the eyes of the customer and takes the customer experience into account in its

✉ Mohammad Naderi Dehkordi
m_naderi_d@yahoo.com

1 Department of Management, Najafabad Branch, Islamic Azad University, Najafabad, Iran

2 Department of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran

Published online: 17 November 2022

planning. Customer vision helps you to see the gaps and opportunities in your business and to consider more effective strategies and processes for the company. Churn prediction is very important task for CRM. Churn prediction can be obtained with many data mining techniques. However, when data mining, dimension reduction and data reduction are two major steps in data preprocessing. Due to the advancement of computer and network technologies, telecommunications have become the number one need in today's environment. As a result, the telecommunications industry has become a rapidly growing market. The increase in the number of telecommunications companies has led to fierce competition. Therefore, creating problems for the customer is currently the number one concern of the telecommunications company. Identifying the customer in the near future is important because attracting new customers is more expensive than retaining existing customers [1]. In addition, customers can actively use their rights to change from one company to another that meets their needs and causes resentment.

Subscriber-based companies prefer customer retention to new customer acquisition because of the higher cost of customer acquisition [2]. The telecommunications industry is witnessing highly saturated markets around the world. Therefore, retaining current customers is one of the most important concerns of companies to maintain product stability. The telecommunications industry has two characteristics that make it a competitive program to analyze data to predict a churn.

First, companies store and use different types of customer data such as personal behavior, demographics, billing and usage. Second, due to the saturation of the telecommunications market, there is fierce competition to attract customers. Many studies focus on improving the accuracy of churn prediction for telecom customers. Jafari et al. have reported an accuracy of over 90% for telecommunication companies' prediction [3]. While much of this paper focuses on improving the accuracy of the factor in predicting churn, it is imperative that telecommunications companies seek to apply financial reversal models [4]. Prediction of churn in different disciplines depends on different statistical and analytical methods of data. Although research efforts to develop more accurate data analytics models are worthwhile, it is also important to understand that data analysis can only detect patterns in data. The predictive performance of a learning system depends on the quality of learning resulting from the training set. If there is no pattern that leads to complete prediction in the training set, it is impossible to achieve 100% accuracy.

Learning from large amount of data is a very challenging issue that faces most of feature-based clustering and other data analysis techniques, as data analysis becomes more difficult due to the "curse of dimensionality." Therefore, processing high-dimensional data requires the use of appropriate techniques and methods. The goal of dimensionality reduction methods is to learn suitable and simplified data representations from the original data set to gain more insight from big data [5].

In this paper, a method based on PSO optimization algorithm and K Means algorithm for data clustering is presented. Accordingly, the properties of each of these algorithms are used. The PSO algorithm can provide good initial answers, but it cannot find the universal optimization. But K Means clustering can provide an acceptable answer. The main disadvantage of the K Means algorithm is that the user must first determine the number of clusters in advance. Several solutions have been proposed to solve this problem.

This paper proposes an algorithm that systematically extracts data from a non-supervised learning method of principal component analysis (PCA) while reducing the characteristics of the data set to calculate the value of each client and using these values in supervised learning. Use a cycle classification task and combine the two methods of classification and particle optimization algorithm to optimize K Means in providing primary centers of gravity. A hybrid algorithm called (PCA-PSO-K Means) is used to predict churn to cost incorrectly classify customers in method training. The new algorithm does not depend on the initial clusters and can avoid being trapped in a local optimal solution. The general structure of the article is as follows. Section 2 provides the research background as well as a background for the proposed PCA-PSO-K Means system. Section 3 examines the research methodology, and Sect. 4 will refer to the research findings, and finally, we will conclude in Sect. 5.

## 2 Research background

There are two different approaches to predicting churn in the literature. The most commonly used approach uses machine learning and data mining techniques, where various user features are used as input, for example, demographic information, usage history and payment order. An extensive review of advanced classification techniques was provided by Verberk et al. [4].

An important part of the literature is introducing and improving various data mining techniques for better prediction [6]. K Means clustering is widely used to minimize the square spacing between two-point feature values in a cluster. Particle swarm optimization is an evolutionary computational technique that finds the optimal solution in many applications. PSO-optimized component clustering is used to achieve more accurate clustering performance [7]. Garvishkumar K. Patel et al. applied a combination of particle swarm optimization and K Means for data clustering. Their proposed approach attempts to improve the performance of traditional partition clustering techniques such as K Means by avoiding the initial requirement of the number of clusters or centers for clustering [8].

Kamrasamy and Amitabh Vahi describe the improvement of clustering performance by combining particle swarm optimization (PSO) and K Means algorithm; their new algorithm does not need a specific number of given clusters before performing the clustering process and can find the local optimal number of clusters in the length of the clustering process. In each iteration process, the inertia weight is changed based on the current iteration and the best fit [9]. Naji Al-Rizwan et al. have used the adaptive firefly optimization algorithm, which is a nature-inspired algorithm, to improve K Means clustering [10].

Data reduction techniques aim to fill in irrelevant features and noisy data samples. To reduce the high-dimensional data, we represented them into a subspace using principal component analysis (PCA) and a new approach based on auto-encoder neural network, which in this way reduces the dimensionality of the original data.

K Means clustering is then applied to the original and reduced data sets. Different internal measures were performed to evaluate the clustering for different number of dimensions, and then, we evaluated how the reduction method impacts the clustering task [5].

### 2.1 Recommended model PCA-PSO-K Means

Clustering is an important topic in data mining. Clustering is one of the techniques that can be used to perform classification [11]. The K Means algorithm is one of the most popular clustering techniques with performance with good accuracy, which is considered due to its easy and very efficient implementation. However, the K Means algorithm has several drawbacks. The K Means objective function has many local minimums. As a result, in the process of minimizing the objective function, there is a possibility of getting stuck in local external points. Therefore, the result of K Means algorithm strongly depends on the initial selection of cluster centers. Convergence to the solution is guaranteed by choosing the initial centers close to the optimal solution. In the PSO algorithm, potential solutions, called particles, are obtained by flowing in the problem space following the current optimal particles.

In general, the PSO algorithm has a strong ability to find the best result, but is incapable of converging to a local optimization. By properly modifying the PSO parameters, convergence can be accelerated and the ability to find the optimal global result can be increased. However, since the PSO algorithm has several parameters that must be adjusted experimentally, if these parameters are not adjusted properly, the search near the global optimal will be very slow. In contrast, the K Means algorithm has a strong ability to find the local optimal result, but its ability to find the global optimal is weak. For this reason, in this paper, an attempt has been made to provide an efficient method for data clustering by combining these two algorithms and also with the help of principal component analysis (Fig. 1).

#### 2.1.1 Principal component analysis

The proposed algorithm must first be accelerated. The probable direction is data processing. The principal component analysis method (PCA) can be used. This is a feature extraction method that extracts a new set of variables from a large set of variables in the data set through linear conversion. The following is presented:

*Step 1*: Build the Covariance Matrix Database.
*Step 2*: Calculate the Eigen values in this matrix (Table 1).
*Step 3*: Eigen vectors with high Eigen values are used to reconstruct the original data set (Table 2).
*Step 4*: High variance properties are selected as the main components.

The first step of PCA is to standardize the data set matrix, and it is important to note that the features or columns on a larger scale than the other columns ultimately outperform the principal component matrix. The second step is to calculate the

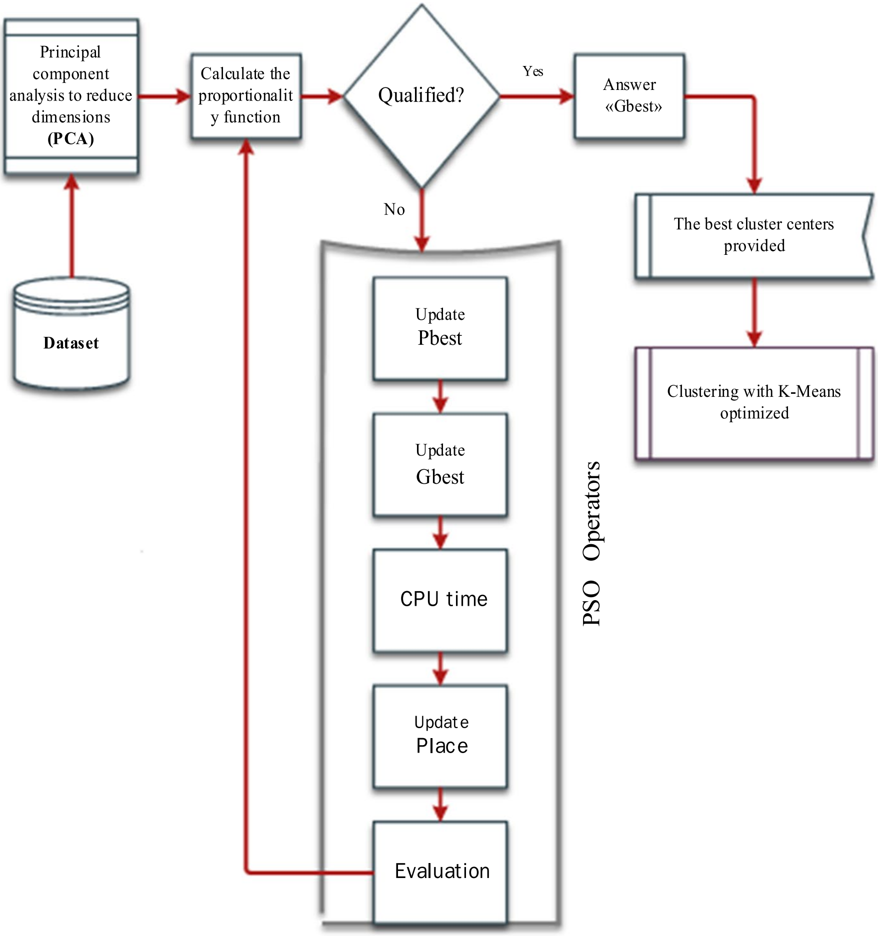


**Fig. 1** Proposed framework of PCA-PSO-K Means algorithm

**Table 1** Eigen values

| Component | Standard deviation | Proportion of variance | Cumulative variance |
|---|---|---|---|
| PC1 | 978,193,293.763 | 0.979 | 0.979 |
| PC2 | 144,301,530.284 | 0.021 | 1000 |
| PC3 | 17,733 | 0.000 | 1000 |
| PC4 | 1388 | 0.000 | 1000 |
| PC5 | 1002 | 0.000 | 1000 |
| PC6 | 0.994 | 0.000 | 1000 |
| PC7 | 791,000 | 0.000 | 1000 |
| PC8 | 0.663 | 0.000 | 1000 |

**Table 2** Eigen vectors

| Attribute | PC 1 | PC 2 | PC 3 | PC 4 | PC 5 | PC 6 | PC 7 | PC 8 |
|---|---|---|---|---|---|---|---|---|
| F | 0.000 | 0.000 | −0.001 | −0.610 | 0.031 | 0.000 | −0.007 | 0.792 |
| Ccr | 0.000 | 0.000 | 0.000 | −0.042 | −0.948 | 0.312 | 0.045 | 0.005 |
| L | 0.000 | 0.000 | −0.005 | 0.022 | 0.311 | 0.950 | −0.014 | 0.005 |
| M | 0.000 | 0.000 | −0.001 | −0.560 | 0.058 | 0.007 | 0.707 | −0.428 |
| R | 0.000 | 0.000 | −0.001 | −0.558 | −0.009 | 0.008 | −0.705 | −0.437 |
| CodePosty | 1000 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Age | 0.000 | 0.000 | −1000 | 0.002 | −0.002 | −0.005 | 0.000 | 0.000 |
| mobile | 0.002 | −1000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

standardized covariance of the matrix. The correlation between the variables can be obtained by multiplying the matrix between the standardized matrix and its displacement. The resulting matrix must be a symmetric matrix. The third step is to discover the special vectors of the covariance matrix. The matrix of special vectors has columns that represent the main components (new dimensions); each component is orthogonal to each other and is explained by reducing the order of variance. The fourth step is to use the special vector matrix, by multiplying it in the data scale matrix to calculate the scores of the principal components in order to find the optimal number of components that get the most variance of the data [5].

#### 2.1.2 PSO evolutionary algorithm process

PSO is an optimization algorithm (Fig. 2). Optimization means that there are one or more variables that need to find their optimal values. First, the fuzzy system must be written in such a way that it takes the variables and gives a score in the output proportional to the value of the variables. This function can be used as a cost function in the PSO algorithm. In evolutionary algorithms, two points are very important, the definition of the objective function and how to formulate the problem, which this section addresses these two issues.

To formulate the problem in the PSO algorithm, it is done in such a way that each particle or row of customer information represents a solution or the selection of the primary centers of the K Means algorithm. Particle position and velocity equations that are proposed in the particle optimization algorithm to update the coordinates and position of the particles in order to achieve the optimal state are:

$$V_{i,t+1} = W.V_{i,t-1} + C_1 r_1 \left(\text{Pbest}_i - P_{i,t}\right) + C_2 r_2 \left(\text{Gbest}_i - P_{i,t}\right) \tag{1}$$

$$P_{t+1} = P_t + V_t \tag{2}$$

The variables of Eqs. (1) and (2) are presented in Table 3.

PSO is used to search for optimal parameters that can lead to minimal errors. The parameters obtained by the PSO are then connected to the K Means algorithm

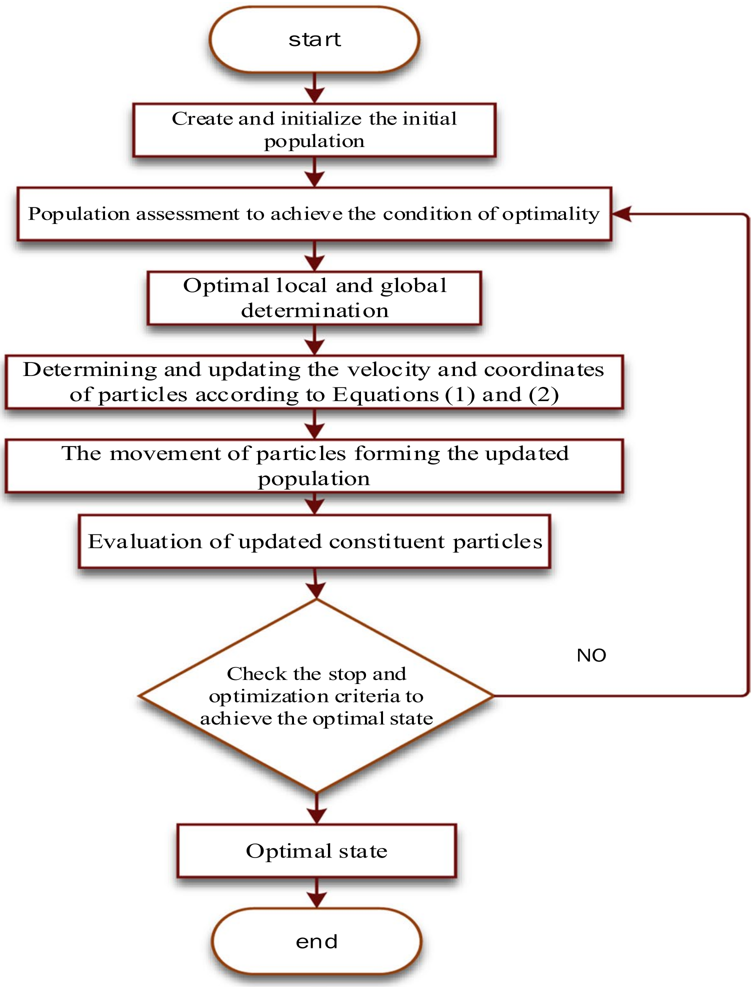


**Fig. 2** PSO algorithm

(Fig. 1). Experiments are tested on different data sets in different initial states, and the results (Fig. 8) show that the proposed algorithm is efficient in most conditions.

**Table 3** Definition of equation variables (1) and (2)

| Definition | Variable |
|---|---|
| Particle speed*i*I repeat*t* + *1* | $V_{i,t+1}$ |
| Particle position *i*In repetition *t* | $P_{i,t}$ |
| Represents the best particle position *i* Um during his move | $Pbest_i$ |
| Best group position with best current population | $Gbest_i$ |
| Random numbers with uniform distribution between 0 and 1 | $r_2, r_1$ |
| There are learning parameters that are adjusted according to the type and form of the problem to improve the performance and control the behavior of the algorithm | $W, C_1, C_2$ |
| Position of the particle in repetition $t+1$ | $P_{t+1}$ |
| Position of the particle in repetition $t$ | $P_t$ |
| Particle speed *i* In repetition$t-1$ | $V_{i,t-1}$ |
| Particle velocity in repetition $t$ | $V_t$ |

***PCA-PSO-K Means Pseudo-algorithm:***

```
Compute mean
Center the data
Compute covariance matrix
Compute eigenvalues
Compute eigenvectors
Fraction of total variance
Choose dimensionality
Reduce basis
Reduce dimensionality data
For each particle
  Initialize particle
END
Do
  For each particle
    Calculate fitness value
    If the fitness value is better than the best fitness value (PBest) in history
      set current value as the new PBest
  End
  Choose the particle with the best fitness value of all the particles as the GBest
  For each particle
    Calculate particle velocity according equation (1)
    Update particle position according equation (2)
  End
kmeans = KMeans(n_clusters=2) # You want cluster the passenger records into 2: Survived or Not survived
kmeans.fit(P)
```

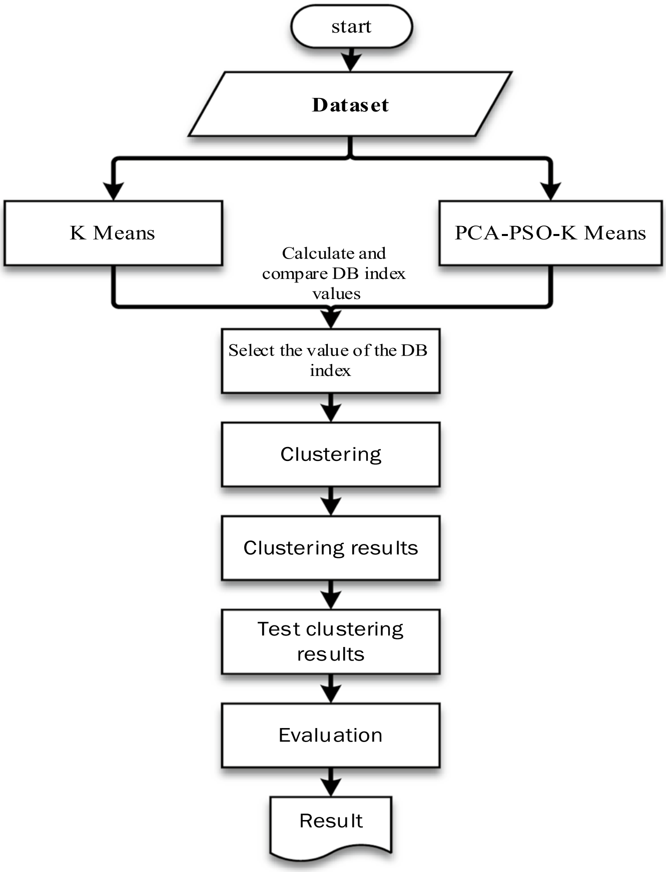


**Fig. 3** General research framework

## 3 Research methods

This section describes the data set used in this paper, including feature extraction as well as the clustering process of the high-dimensional data process presented in Fig. 3. Perform clustering on the main data set by performing the data reduction method. Thus, the comparison between the results allows to understand whether the reduced data sets can perform the clustering accuracy better than the result without reducing the data.

PCA is first applied to the original data. Data, as a combination of key features in a new space, are compressed in such a way that the most important information is retained and converted into new features. Then, the combination of K Means and PSO algorithms on the converted data set was used.

The method chosen in this paper to predict customer churn is based on the standard CRISP-DM method. The data mining process is defined in various formats, the

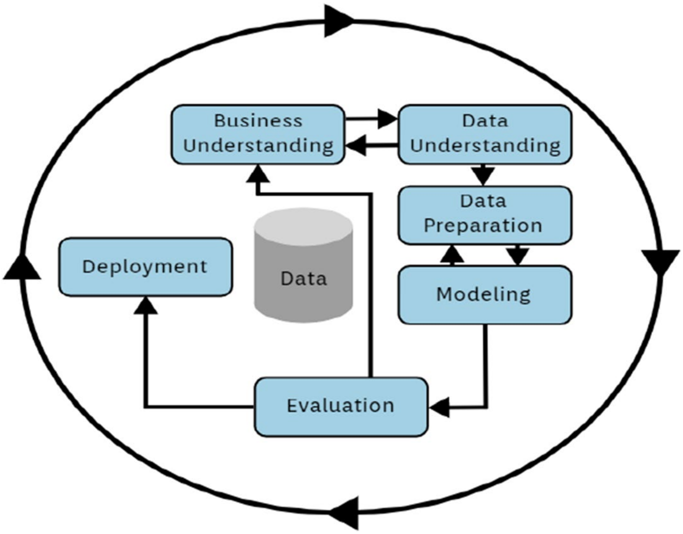


**Fig. 4** CRISP-DM life cycle (nimblecoding.com)

most extensive of which is the CRISP-DM format [12]. CRISP-DM as a process model provides an overview of the data mining process life cycle. The life cycle is a data mining process consisting of six phases, as shown in Fig. 4.

### 3.1 Bullet Business understanding

This company, which provides fixed internet services, is currently providing services in Isfahan Province. The company's fixed internet customer base reaches more than 600,000 subscribers. These customers are mainly home, office, commercial and customers. Therefore, the customer clustering approach is very important for recognizing and analyzing the behavior of active customers in this industry. For this reason, this information is used to evaluate and analyze the characteristics of customers and, finally, to formulate a marketing strategy tailored to each department and achieve the desired results in the field of CRM.

– Understand data

The data set used in this article is obtained from CRM, one of the companies providing fixed internet in Isfahan Province. This data set contains the exact information of 10,000 customers, each of which has 17 variables. Each row of this data set belongs to a client, and when the predictor model is executed, the *churn* variable will be used as the dependent variable and the other variables as the independent variables.

Before using the PCA algorithm and implementing the prediction model, the data set must be prepared in a format suitable for analytical modeling. The first step involves examining the data to check for missing values for each available variable.

### 3.2 Data set information

This data set was collected randomly during 12 months from the database of one of the companies providing fixed internet services in Isfahan Province. A total of 10,000 rows of data, each representing a client, contain information for 17 columns. The features in this data set are as shown in Table 4. All features except the *churn* feature are customer status change tags at the end of 12 months.

– Data preparation

Using a large number of independent variables to model relationships with a dependent variable may complicate the interpretation of the analysis, which states that the number of independent variables must be reduced so that the result can be easily interpreted. Likewise, keeping too many independent variables may lead to over-installation. Therefore, to avoid multiple lines and reduce the number of independent variables, the principal component analysis (PCA) method in the data preparation phase of the process data mining is used. The PCA model is a dimensional reduction technique that uses the correlation structure of independent variables.

– Modeling

In the field of machine learning, modeling techniques can be classified into supervised and unsupervised learning techniques. Unsupervised learning technique is a method that is used without the need for educational data. Unsupervised learning makes observations or data without any labels/classes/decisions. Some of the algorithms that can be used in unsupervised learning are: K Means, K Medoids, Hierarchical Clustering, dbscan [13, 14]. The data set used in this paper includes a Is a dependent variable that is already known. Thus, the goal is to implement a monitored model that predicts the event, that is, customers who are annoyed by the services provided by a fixed internet service provider. In this prediction model, the selected machine learning algorithm learns which values of the dependent variables are related to different values of the independent variables.

The data set are used as an input process using RapidMiner software in Table 4. The file used is an output Excel file (.xls) under the DATA subprocess. Using the Set Role operator, the data is used as input by specifying *churn* as the label. The Set Role operator then connects to the Multiply operator to compare the results of K Means and PCA-PSO-K Means.

In addition, the multiplier operator is connected to two processes, K Means and PCA-PSO-K Means, the output of which is used by the cluster operation operator. The number of clusters tested ($k$) is between 2 and 9. Based on the design process to select the number of clusters ($k$), the number of clusters ($k$) was tested using K Means and PCA-PSO-K Means methods. The following are the results of the Davies Bouldin index of each $k$ value processed in Table 5 and also the value chart of the DB index in Fig. 5.

**Table 4** Customer data set of a fixed internet provider in Isfahan Province

| A symbol | Variable | Definition | The unit |
|---|---|---|---|
| id | Customer ID | The number assigned to each customer | Number |
| Age | Customer age | Calculated from the customer's date of birth | Number |
| Sex | Gender | Gender type of each client | Male/female |
| Code Posty | Postal code | Number 8 digits | Number |
| Mobile | Mobile number | Each customer's mobile number | Number |
| L | Length of customer relationship | Number of days from the first purchase to the last purchase | Number of days |
| R | Recently bought | The number of days since the last purchase until now | Number of days |
| F | Number of repeat purchases | Number of purchases during the period | Times |
| M | Value for money | Money volume purchased during the period | Rials |
| Ccr | Customer consumption records | The amount of fixed internet consumed by the customer during the period | The amount of |
| Date | Time of purchase | Fixed internet charge history | Date |
| Price | Value for money | The amount of money to buy each time you buy | Rials |
| Charg | Customer consumption | Fixed internet rate per purchase | The amount of |
| BirthDay | Date of birth | Every customer's date of birth | Date |
| Regdate | Client registry time | Contract start time with a fixed internet service provider | Date |
| Customertype | Customer payment type | Contract( prepayment or postpaid) | Post/Pre |
| *Churn* | Class label | Determining the customer's downfall | False/True |

**Table 5** DB value for each number of clusters ($k$)

| PCA-PSO-K Means DB | K Means DB | $k$ |
|---|---|---|
| 0.112 | 0.924 | 2 |
| 0.141 | 0.922 | 3 |
| 0.216 | 0.992 | 4 |
| 0.188 | 0.996 | 5 |
| 0.254 | 1.007 | 6 |
| 0.255 | 0.95 | 7 |
| 0.259 | 0.903 | 8 |
| 0.247 | 0.885 | 9 |

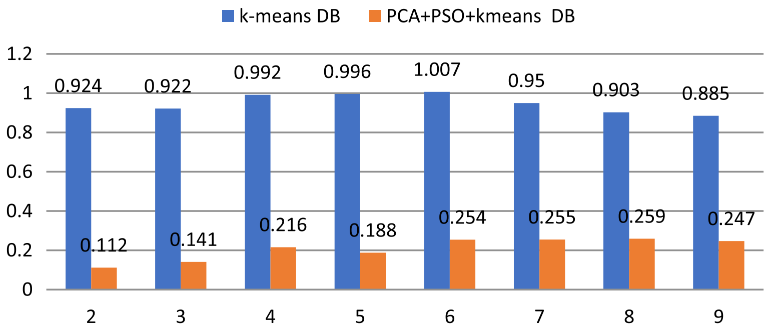


**Fig. 5** DB value chart in K Means and PCA-PSO-K Means

The data set in Table 4 is tested using the Correlation Matrix, Weight by PCA, Optimize Weights (PSO), Cross-Validation and Performance operators, where the results are formed with precision and cluster values. The Weight by PCA operator generates attribute weights from the data set using components created by PCA (Fig. 6). This operator behaves exactly as if a PCA model gave the Weight by operator a component model. The Optimize Weights (PSO) operator also weighs the properties with the particle swarm optimization approach (Fig. 7).

The parameters obtained by the PSO will then be connected to the K Means algorithm. The test results of PCA-PSO-K Means clusters using RapidMiner software are shown in Fig. 8.

### 3.3 Bullet evaluation

The predictive performance of the model can be observed using the confusion matrix. The confusion matrix is a two-row, two-column table in which diagonal cells show the correct predictions of the categorized items and those in the opposite diagonal represent incorrect predictions of the training and experimental data sets, respectively. Shown in Table 6.

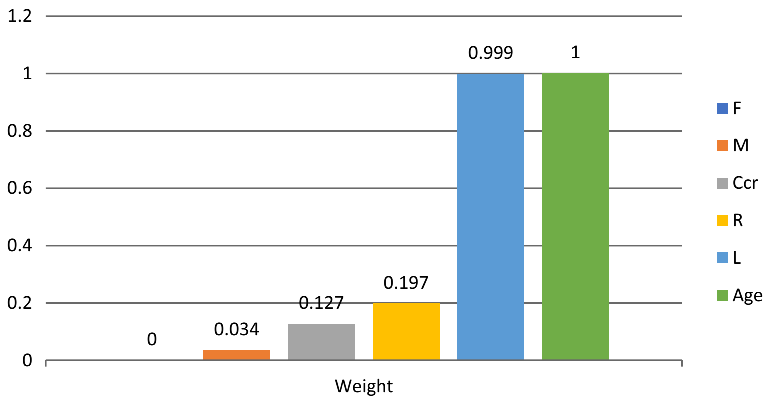


**Fig. 6** Weight diagram of correlation matrix variables

Confusion matrix cells are called true negative (TN), false positive (FP), false negative (FN) and true positive (TP) (Table 6). Other criteria in Table 6, such as TNR (real or specific negative rate), TPR (positive predictive rate or sensitivity), NPV (negative predictive value), PPV (positive predictive value) and ACC (accuracy), are calculated using the following equations:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{4}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

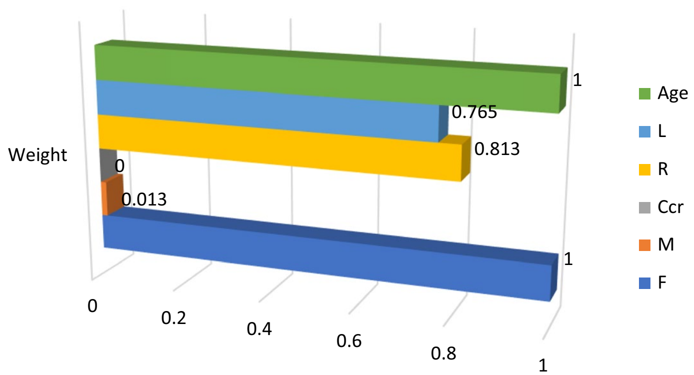


**Fig. 7** Optimized weight chart of variables with Optimize Weights (PSO) operator

```
PerformanceVector:
accuracy: 99.77% +/- 0.26% (micro average: 99.77%)
ConfusionMatrix:
True:    false    true
false:   2642     5
true:    1        3
kappa: 0.499
weighted_mean_recall: 64.98% +/- 24.12% (micro average: 68.73%), weights: 1, 1
weighted_mean_precision: 62.41% +/- 21.31% (micro average: 87.41%), weights: 1, 1
spearman_rho: 0.271 +/- 0.443 (micro average: 2.706)
correlation: 0.271 +/- 0.443 (micro average: 0.529)
```

**Fig. 8** Test results of PCA-PSO-K Means clusters

**Table 6** Confusion matrix

| Data set | Observed | Predicted | | |
|---|---|---|---|---|
| | | No | Yes | %Correct |
| Training/Test | No | TN | FP | TNR |
| | Yes | FN | TP | TPR |
| | Total % | NPV | PPV | ACC |

**Table 7** Educational data confusion matrix accuracy: 99.77% ± 0.19% (micro average: 99.77%)

| | True false | True true | Class precision |
|---|---|---|---|
| pred. false | **6125** | *14* | 99.77% |
| pred. true | *0* | **17** | 100.00% |
| class recall | 100.00% | 54.84% | |

**Table 8** Experimental data confusion matrix accuracy: 99.77% ± 0.26% (micro average: 99.77%)

| | True false | True true | Class precision |
|---|---|---|---|
| pred. false | **2642** | *5* | 99.81% |
| pred. true | *1* | **3** | 75.00% |
| class recall | 99.96% | 37.50% | |

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{6}$$

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{P + N} \tag{7}$$

In the training phase, 6125 customers are correctly classified and offer 99.77% specificity and (Table 7) among the 17 customers who have already left the services provided by the fixed internet provider, the forecasting model correctly classified 17 customers with a 100% real positive rate.

In the experimental data set (Table 8), out of the 4 customers who have stopped using the services provided by the fixed internet service provider, 3 customers are

**PerformanceVector:**
accuracy: 95.88% +/- 1.54% (micro average: 95.88%)
ConfusionMatrix:

| True: | cluster_0 | cluster_8 | cluster_3 | cluster_1 | cluster_7 | cluster_6 | cluster_5 | cluster_2 | cluster_4 |
|---|---|---|---|---|---|---|---|---|---|
| cluster_0: | 530 | 1 | 7 | 1 | 0 | 0 | 0 | 0 | 0 |
| cluster_8: | 4 | 755 | 14 | 5 | 0 | 0 | 0 | 0 | 0 |
| cluster_3: | 5 | 3 | 418 | 13 | 4 | 1 | 0 | 0 | 0 |
| cluster_1: | 0 | 23 | 18 | 807 | 0 | 0 | 0 | 0 | 0 |
| cluster_7: | 1 | 0 | 4 | 0 | 84 | 3 | 0 | 0 | 0 |
| cluster_6: | 0 | 0 | 1 | 0 | 3 | 56 | 1 | 0 | 1 |
| cluster_5: | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| cluster_2: | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| cluster_4: | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

kappa: 0.946 +/- 0.020 (micro average: 0.946)
weighted_mean_recall: 64.28% +/- 4.00% (micro average: 68.72%), weights: 1, 1, 1, 1, 1, 1, 1, 1, 1
weighted_mean_precision: 64.20% +/- 4.19% (micro average: 68.48%), weights: 1, 1, 1, 1, 1, 1, 1, 1, 1
spearman_rho: 0.967 +/- 0.011 (micro average: 9.675)
correlation: 0.967 +/- 0.012 (micro average: 0.967)

**Fig. 9** Test result of formed clusters of K Means

**Table 9** Final classification results

| Metrics | Weighted mean accuracy | Weighted mean recall | Weighted mean precision | Correlation |
|---|---|---|---|---|
| PCA-PSO-K Means | 99.77 | 64.98 | 62.41 | $0.271 \pm 0.443$ |
| K Means | 95.88 | 64.28 | 64.20 | $0.967 \pm 0.012$ |

correctly classified (true positive rate 75%). And out of 2647 customers who were still using the service, 2642 customers were correctly classified (feature99.81%).

Evaluating the results obtained using the confusion matrix, it can be seen that for the True class of the *Churn* dependent variable, the K Means model has a predicted performance of 95.88% (Fig. 9). When examining the results generated for each class of the *Churn* dependent variable, the proposed PCA-PSO-K Means algorithm achieves 99.77% predictive performance (Fig. 8). It also shows a real positive rate or sensitivity of 75% and a real or specific negative rate of 99.81% (Table 8) and a correlation coefficient of $+0.443/-0.271$ (Fig. 8).

## 4 Findings

The results of the final classification in Table 9 show the proposed model with 99.77% accuracy and 64.98% average call weight, as well as 75% sensitivity and 99.81% specificity (Table 8) and a correlation coefficient of $0.271 \pm 0.443$ (Fig. 8), better performance than it has the K Means algorithm.

## 5 Discussion and conclusion

In this study, a customer segmentation model for data analysis is presented and validated through standard evaluation criteria. A proposed model based on PSO algorithm principal component analysis is proposed to improve the K Means algorithm. The results show that performing data reduction and feature selection by principal component analysis and optimization of the K Means algorithm can be pre-modeled. The nose allows for the highest accuracy of prediction. In addition, this priority allows the predictive model to be reduced by 50% of core features and 12% of data samples for more efficient learning, and also results in efficiency and accuracy when K Means is used in combination with PSO, because both algorithms fix their shortcomings. In other words, PCA transforms the data by linearly combining the input features into features that have minimum correlation, and the combination of PSO and K Means provides good performance compared to the K Means algorithm. The new algorithm improves the convergence speed of PSO and helps K Means to become independent in the initial clusters.

The limitation of this work is the collection of reliable data from random customers (only about 10,000 customers). Due to following direct methods and limited human resources, this process took a long time (about 7 months), as well as limitations on availability. The existence of this data applies, which has been used under the specific conditions of the present study and is therefore not available to the public. For future work, we plan to achieve the PSO process time efficiency, as many iterations of the algorithm may involve iterative calculations.

**Acknowledgements** The authors thank the data provider for their extensive support.

## References

1. Aziz AA, Ismail NH, Ahmad F (2013) Mining students' academic performance. J Theor Appl Inform Technol 53(3):485–495
2. Gupta S, Lehmann DR, Stuart JA (2004) Valuing customers. J Mark Res 41(1):7–18
3. Ruholla Jafari-Marandi, Joshua Denton, Adnan Idris, Brian K. Smith, Abbas Keramati (2020) Optimum profit-driven churn decision making: innovative artificial neural networks in telecom industry. Springer-Verlag London Ltd., part of Springer Nature, Cham
4. Verbeke W, Dejaeger K, Martens D, Hur J, Baesens B (2012) New insights into churn prediction in the telecommunication sector: a profit driven data mining approach. Eur J Oper Res 218(1):211–229. https://doi.org/10.1016/j.ejor.2011.09.031
5. Alkhayrat M, Aljnidi M, Aljoumaa K (2020) A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. J Big Data. https://doi.org/10.1186/s40537-020-0286-0
6. García DL, Nebot A, Vellido A (2017) Intelligent data analysis approaches to churn as a business problem: a survey. Knowl Inf Syst 51(3):719–774
7. Pritesh Vora, Bhavesh Oza (2013) A survey on K-mean clustering and particle swarm optimization. Int J Sci Modern Eng (IJISME) 1(3). ISSN: 2319-6386.
8. Garvishkumar K Patel, Vipul K. Dabhi, Harshadkumar B. Prajapati (2017) Clustering using a combination of particle swarm optimization and K Means. J Intell Syst 26(3):457–469

9. Komarasamy G, Amitabh Wahi (2011) Improving the cluster performance by combining pso and K means algorithm. https://doi.org/10.21917/ijsc.2011.0032
10. Radhwani AMNA, Algamal ZY (2021) Improving K Means clustering based on firefly algorithm. J Phys Conference Ser. https://doi.org/10.1088/1742-6596/1897/1/012004
11. Buchori A, Prasetyowati D, Wijayanto (2020) Design of the magic book math media based on augmented reality. Test Engineering & Management. http://www.testmagzine.biz/index.php/testmagzine/issue/view/7
12. Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer CEA (2000) CRISP-DM 1.0: Step-by-Step Data Mining Guide
13. Pratiwi H, Windarto AP, Susliansyah S, Aria RR, Susilowati S, Rahayu LK, Fitriani Y, Merdekawati A, Rahadjeng IR (2020) Sigmoid activation function in selecting the best model of artificial neural networks. J Phys Conf Ser. http://iopscience.iop.org/article/10.1088/1742-6596/1471/1/012010
14. Sudipa IGI, Astria C, Irnanda KF, Perdana A (2020) Application of MCDM using PROMETHEE II Technique in the Case of Social Media Selection for Online Businesses. Application of MCDM using PROMETHEE II Technique in the Case of Social Media Selection for Online Businesses