Scientific Research Publishing

Journal of Computer and Communications, 2019, 7, 55-66
http://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Clustering Approach for Customer Billing Prediction in Mall: A Machine Learning Mechanism

Sriramakrishnan Chandrasekaran[1], Abhishek Kumar[2]

[1]Manager, KTech, Regional Delivery Center, KPMG LLP, Montvale, NJ, USA
[2]Computer Science Engineering Department, ACERC, Ajmer, India
Email: sriramakrishnanchand@kpmg.com

How to cite this paper: Chandrasekaran, S. and Kumar, A. (2019) A Clustering Approach for Customer Billing Prediction in Mall: A Machine Learning Mechanism. *Journal of Computer and Communications*, **7**, 55-66.
https://doi.org/10.4236/jcc.2019.73006

Received: January 9, 2019
Accepted: March 25, 2019
Published: March 28, 2019

Abstract

Machine learning implementations are being done in a long way in science and technology and especially in medical stream. In this article, we are focusing on machine learning implementation on mall customers and based on their income and how they can invest in the purchase in a mall. This explains the features like Customer ID, gender, age, income, and spending score. There, we mentioned a score in purchasing the goods in the mall. In this scenario, we are implementing clustering mechanisms, and here we apply the dataset of mall customers which is a public dataset and create clusters related to the customer purchase. We implement machine learning models for the prediction of whether the visited customer will purchase any product or not. For this kind of works, we require many of the inputs like the features mentioned in the paper. To maintain the features, we require a model with machine learning capability. We are performing K-Means clustering and Hierarchical clustering mechanisms, and finally, we implement a confusion matrix to achieve and identify the highest accuracy in those two algorithms. Here, we consider machine learning mechanisms to predict the category of the customer about whether they can buy a product or not based on the independent variables. This work presents you a simple machine learning prediction model based on which we can predict the category of the customer based on clustering. Before clustering, we don't know to what group they belong to. But after clustering, we can identify the category that data node belongs to. In this article, we are mentioning the process of determining the employee based information using machine learning clustering mechanisms.

Keywords

Clustering, Machine Learning, Category, Technology, Hierarchical, K-Means

1. Introduction

Machine learning mechanisms are widely used in a large number of applications related to science and technology, and we can implement those mechanisms even in employee-related things or student-related information. We need to predict something based on the information we have and with the past experiences. In this article, we are concentrating on the prediction of whether the customer will purchase any goods from the mall or not based on the gender and salary. Here, we have multiple independent variables and only one dependent variable which we need to predict whether the customer will make any bill on the product.

Hierarchical cluster mechanism is one algorithm we implement and the K-Means clustering mechanism we are implementing on the data we have. There are different scenarios for both of the clustering mechanisms, but the resultant work is common. The accuracy of the algorithm differs the most popular and most acceptable algorithm for the prediction model design and implementation. As per the clustering rules, we have two kinds of clustering mechanisms. One is hard clustering, and other is soft clustering.

1) Soft Clustering:

We need to identify whether the data point is belonged to any cluster instead of making every data point into the cluster we need to identify whether the current data point will fit into either of the existing data clusters.

2) Hard Clustering:

In this scenario, we need to find out whether the current dataset or data point belongs to the existing data set or not [1] [2] [3]. Consider if we have ten different datasets, we need to identify to which cluster the data point will belong.

There are different types of clustering mechanisms that are identified, and they are mentioned as follows. We need to learn about those, because we are utilizing two kinds of clusters in this mechanism [4] [5] [6] to identify whether the customer will purchase any product or not. Those are mentioned as follows.

Connectivity models are the first type which deals with the scenario of connecting the data points based on the category or the thing which is common in the relation. For supposition if one data point is lying far away and the new data point is related to the characteristic of the current data point, then there will be connectivity between the data point in the space.

Centroid models are another model which deals with similarity identification of the data point that will be done by how the data point is close to the centroid of the cluster. If the closeness from the centroid to the group is smaller, then there will be a good connection between then centroid, and the data point and the current data point will belong to the cluster which centroid belongs to [7] [8] [9].

The next model is a distribution model which deals with the probability of how the data points in the cluster belong to the same distribution. Based on the probability notation, the distribution will form. The distributions may be Gaussian or any other type.

The last model type is density type. It deals with the search for the density of the data point in the data space [7] [8] [9].

The difference between the previous researches and our model identification is a quite interesting thing. Because of models we implement and the plotting will be done in the form of required features instead of all the existing features. The feature extraction mechanism and the identification of what are the most prominent features in the model is most important thing we implemented. We use two kinds of classifiers and the plotting using the clustering mechanisms are main focus of gamification implementation and explanation. We focused on implementation of the models with those clustering mechanisms and the models will show the optimal model and the features to improve our model at any point of time of extension.

The following (**Figure 1**) [10] [11] [12] describes the structure of different clustering models in deals.

Here we are implementing the same mechanisms using two primary clustering mechanisms. In (**Figure 2**) we implement the distribution of clusters, in (**Figure 3**) we tried to project the Centroid model of clustering and in (**Figure 4**) we implemented density model of Clustering. They are K-Means (**Figure 5**) and hierarchical clustering mechanisms. In the K-Means clustering [13] [14], we are implementing the method with k = 2 as default value and identify the cluster to

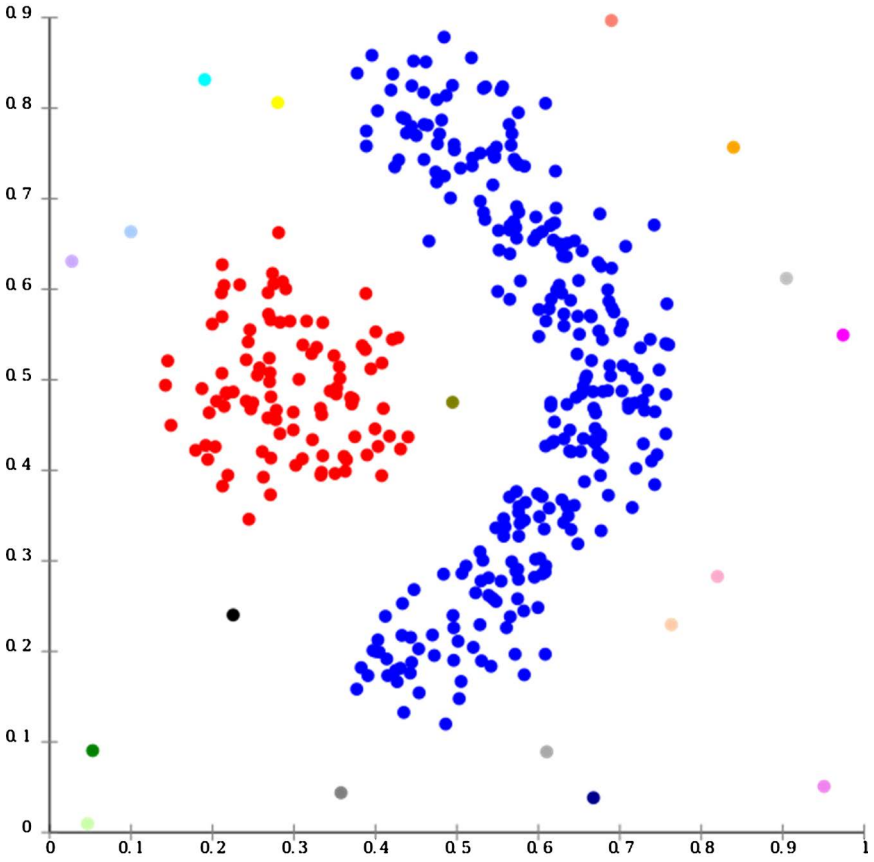

Figure 1. Connectivity model in clustering.

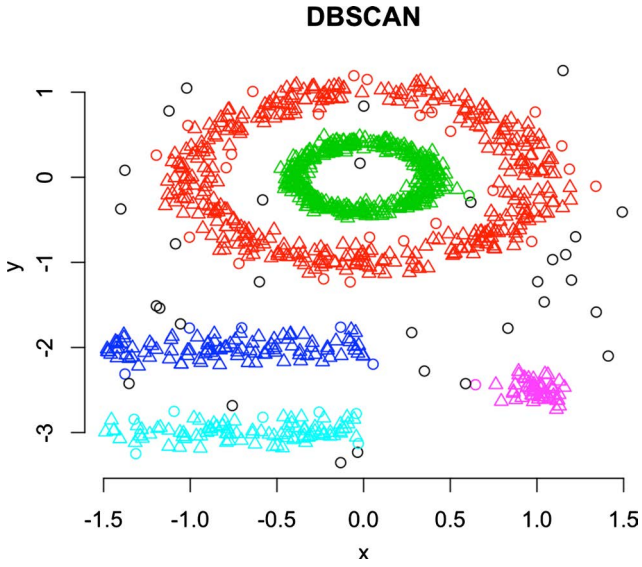

Figure 2. Distribution model of clustering.

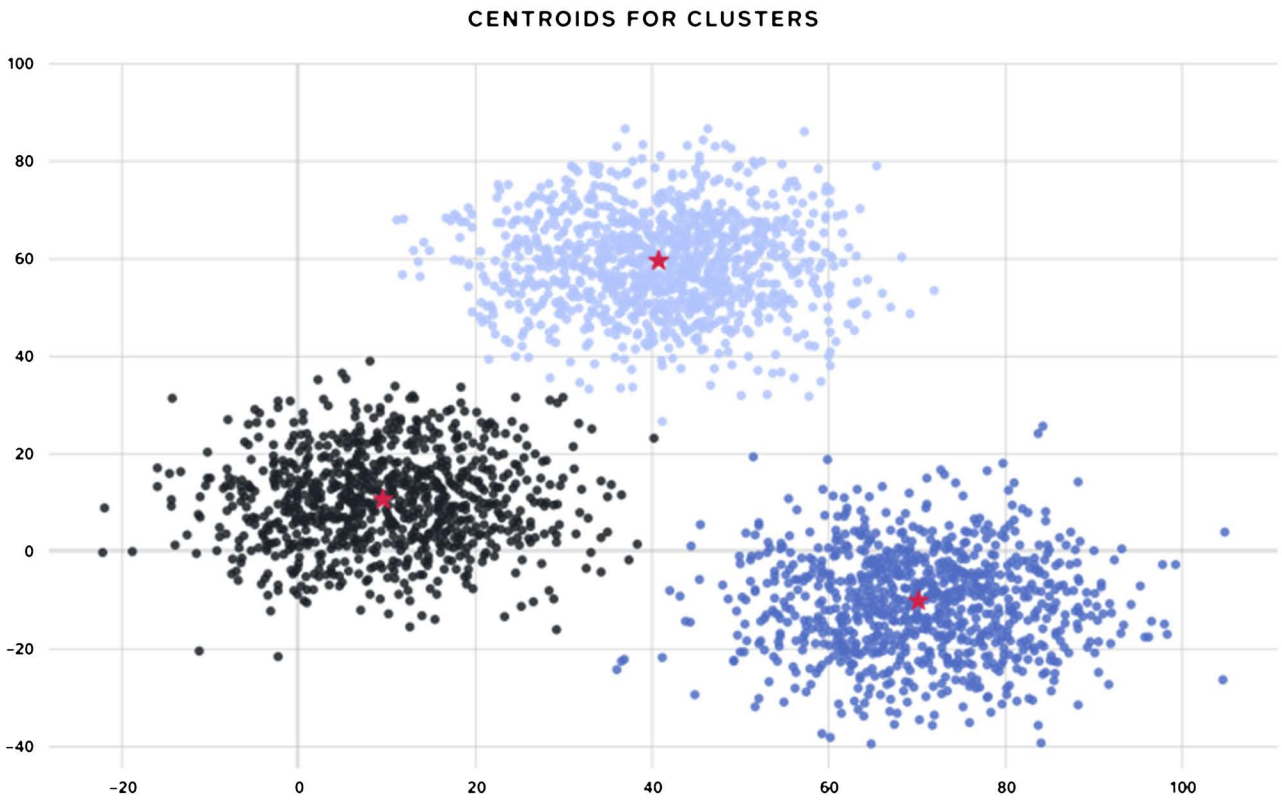

Figure 3. Centroid model of clustering.

which data point will belong to. In the hierarchical clustering, we form the dendrograms related to the group. Based on which we can identify the category [15] [16] [17]. The sample dendrogram and cluster as follows in the 2D plane.

The lateral part of this article will deal with the explanation of the K-Means and hierarchical clustering (**Figure 6**) mechanisms related this approach discussed in the article abstract, which is predicting whether the customer will make a bill in the mall or not based on his age and salary as main independent

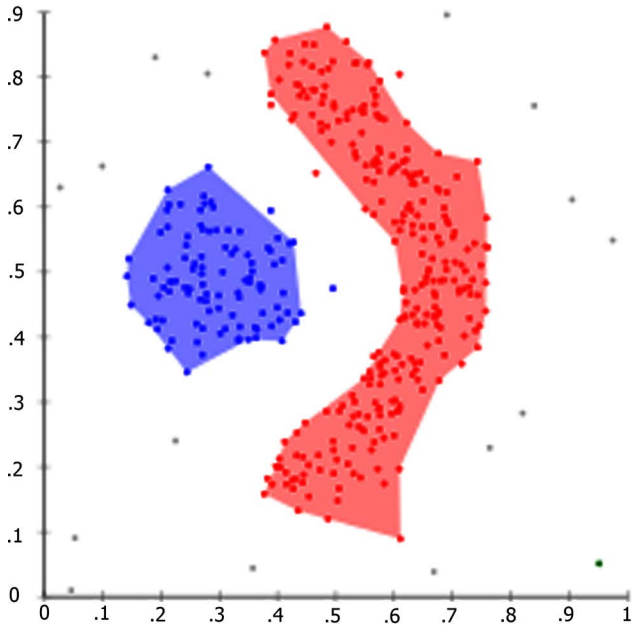

Figure 4. Density model of clustering.

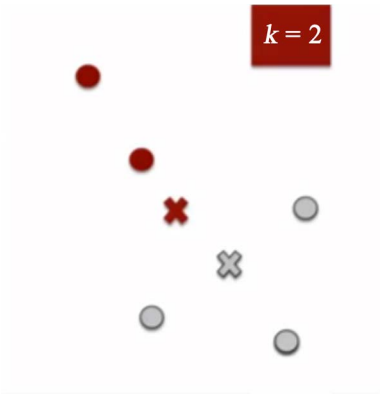

Figure 5. K-means for k = 2.

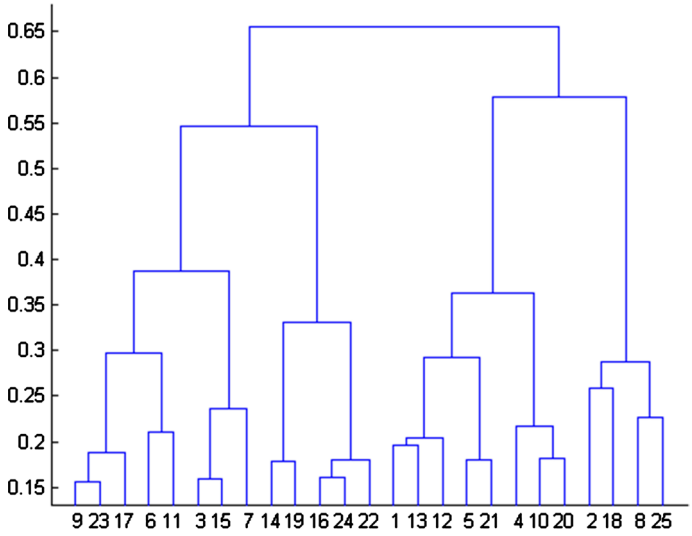

Figure 6. Dendrogram for hierarchical clustering.

variables. Next section will describe the flow of the process, next with sample results and plottings, next, we conclude the process with sample future scope of the work [15] [16] [17].

2. Mechanisms

2.1. K-Means

K-means works on the iterative process of the algorithm which aims for the local maxima in each of the iterations. There may be different iteration values based on the K Value considered. Here in this process, we found K value as 2. And the following will be the steps to be mentioned [18] [19].

Initially, we need to specify the number of clusters K in the 2 D space. In this regard, we are considering k as 2

In this above image, we can see that we considered two as the K value and the five different data point in the 2D plane space [18] [19].

We need to assign each data point to the cluster available. Suppose in this regard we are considering there are two clusters [19] [20] [21] which are mentioned in Red and White as indicated in (**Figure 7**).

Now we need to compute the centroid of the data points. They are mentioned in this below image as a cross symbol. For the red cluster red cross is mentioned, and for the white cluster, the white color cross is specified as below (**Figure 8**).

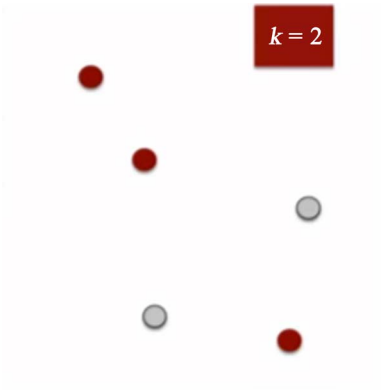

Figure 7. K value and the 5 data points in the 2D space.

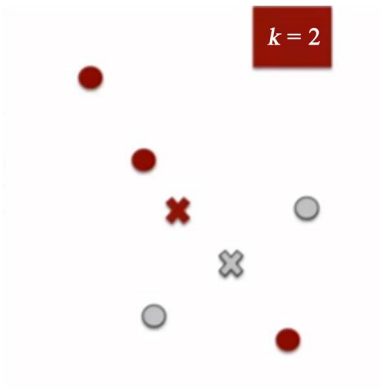

Figure 8. Centroid designing with the cross symbol.

1) Verify whether the newly created centroid is closest to the related category of the data points or not of the centroid is far from the data points of the same type then re-assign the centroid to the related data points in the cluster. The same mentioned in image nine below.

If we observer Figure 9, we can identify that there is an increase in white category data points and a decrease in red category data points. That happened when the centroid is far from the data points of the same category.

2) Re-compute the centroid based on the available data points if necessary as the new iteration of the data points. The following is the procedure of the centroid re-computing process [21].

Repeat the previous two steps until there are improvements identified in the cluster (Figure 10).

2.2. Hierarchical Clustering

As the name mentions there will be the hierarchy of the clusters based on the data pointsint he 2D plane or 2D space. In this regard, we design dendrograms which are related to the data points in few iterations as done in the K-means algorithm. First, the cluster starts with the data point assigned to it and then it will merge to the nearest data point in the space and forms the group. For every iteration, there will be a massive change in the cluster and the centroid of the cluster [21].

Dendrogram of the cluster will be formed for every iteration, and the best

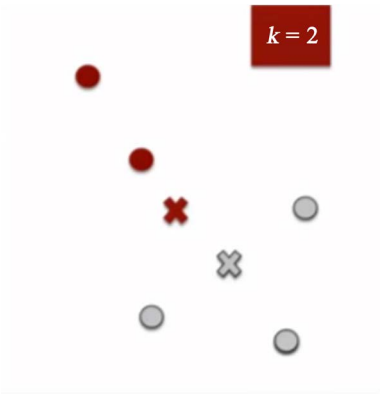

Figure 9. Re-Assigning the centroid.

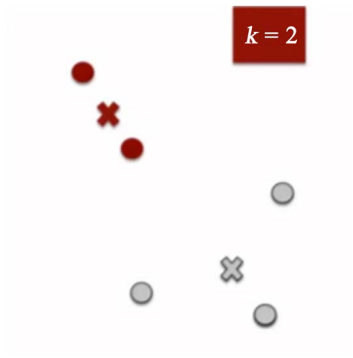

Figure 10. Re-Computing the centroid.

choice of the number of groups will be 4, and the red lines mentioned in Figure 11 defines the maximum vertical distance [22].

3. Process Flow

We are maintaining individual process flow for K-Means and Hierarchical clustering mechanism. They are described in this article with sample codes (Figure 12, Figure 13).

1) K-Means

The process consists of the following steps:

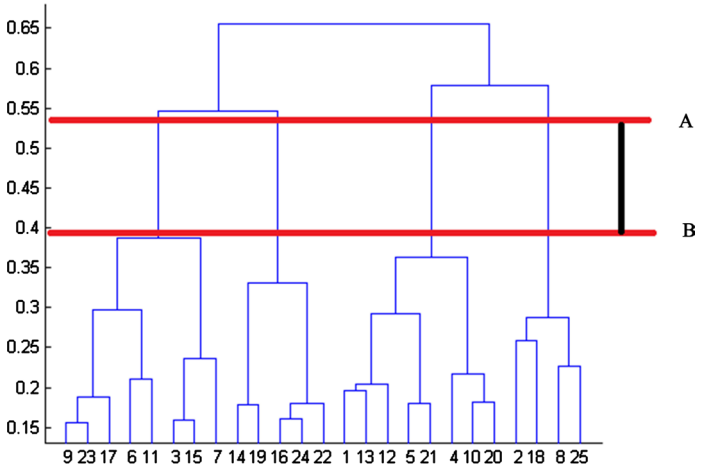

Figure 11. Hierarchical Clustering with the maximum vertical region.

```
# Fitting K-Means to the dataset
kmeans = KMeans(n_clusters = 5, init = 'k-means++', random_state = 42)
y_kmeans = kmeans.fit_predict(X)

# Visualising the clusters
plt.scatter(X[y_kmeans == 0, 0], X[y_kmeans == 0, 1], s = 100, c = 'red', label = 'Cluster 1')
plt.scatter(X[y_kmeans == 1, 0], X[y_kmeans == 1, 1], s = 100, c = 'blue', label = 'Cluster 2')
plt.scatter(X[y_kmeans == 2, 0], X[y_kmeans == 2, 1], s = 100, c = 'green', label = 'Cluster 3')
plt.scatter(X[y_kmeans == 3, 0], X[y_kmeans == 3, 1], s = 100, c = 'cyan', label = 'Cluster 4')
plt.scatter(X[y_kmeans == 4, 0], X[y_kmeans == 4, 1], s = 100, c = 'magenta', label = 'Cluster 5')
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s = 300, c = 'yellow', label = 'Centroids')
plt.title('Clusters of customers')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.legend()
plt.show()
```

Figure 12. Sample code snippet.

```
# Fitting Hierarchical Clustering to the dataset
from sklearn.cluster import AgglomerativeClustering
hc = AgglomerativeClustering(n_clusters = 5, affinity = 'euclidean', linkage = 'ward')
y_hc = hc.fit_predict(X)

# Visualising the clusters
plt.scatter(X[y_hc == 0, 0], X[y_hc == 0, 1], s = 100, c = 'red', label = 'Cluster 1')
plt.scatter(X[y_hc == 1, 0], X[y_hc == 1, 1], s = 100, c = 'blue', label = 'Cluster 2')
plt.scatter(X[y_hc == 2, 0], X[y_hc == 2, 1], s = 100, c = 'green', label = 'Cluster 3')
plt.scatter(X[y_hc == 3, 0], X[y_hc == 3, 1], s = 100, c = 'cyan', label = 'Cluster 4')
plt.scatter(X[y_hc == 4, 0], X[y_hc == 4, 1], s = 100, c = 'magenta', label = 'Cluster 5')
plt.title('Clusters of customers')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.legend()
plt.show()
```

Figure 13. Sample of hierarchical clustering.

a) Import the libraries
b) Import the related dataset in CSV or JSON format
c) Perform Feature scaling
d) Split the dataset into test and train set
e) Use the elbow method to identify the optimal number of clusters
f) Fit K-Means to the dataset
g) Visualizing the Cluster
2) Hierarchical Clustering
The process consists of the following steps:
a) Import the libraries
b) Import the related dataset
c) Perform feature scaling
d) Split the dataset
e) Using dendrogram find the optimal number of clusters
f) Fit hierarchical clustering to the dataset
g) Visualize the cluster

4. Results

The following are the results of the two mechanisms used in this architecture. The first one is K-Means and the second one is Hierarchical Clustering. As mentioned in previous discussions K-Means clustering here will identify whether the model designed with the features will identify the desired result is obtained or not. We use The Elbow method for this implementation to predict whether the customer will make bill in the mail or not based on his features.

1) K-Means

In this scenario, there are two sample plots which are consisting of identifying the number of possible clusters and then visualizing the number of groups. The resultant of the clusters is as follows in Figure 14, Figure 15.

2) Hierarchical Clustering

The following are the outputs we acquired for hierarchical clustering mechanisms.

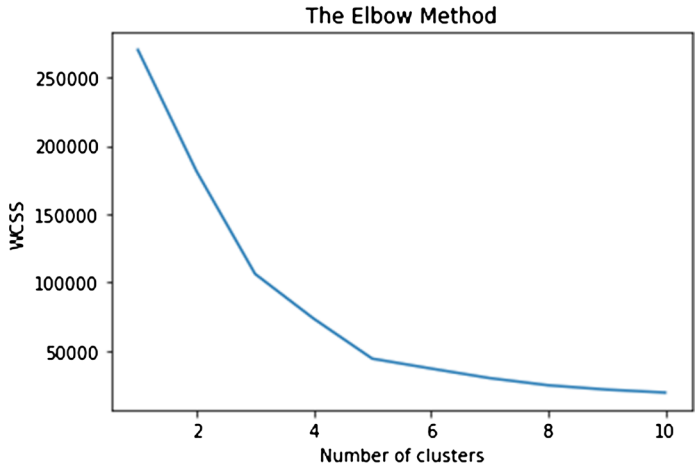

Figure 14. Number of possible clusters.

First one is sample dendrogram (Figure 16) and the second one (Figure 17) is the clusters of the customers based on the annual income.

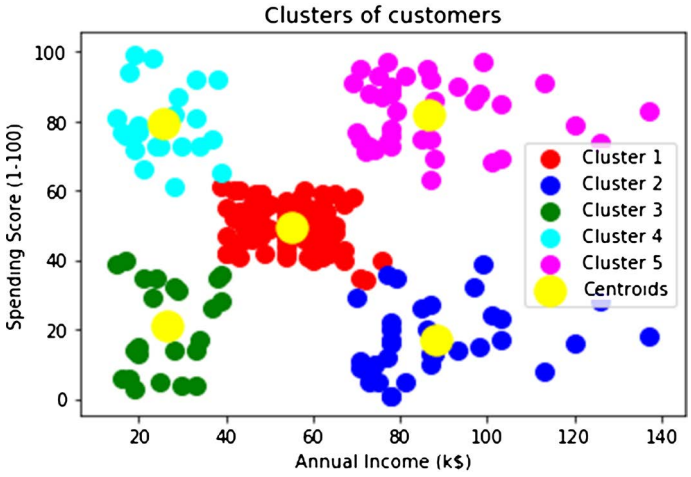

Figure 15. Clusters formed based on customers.

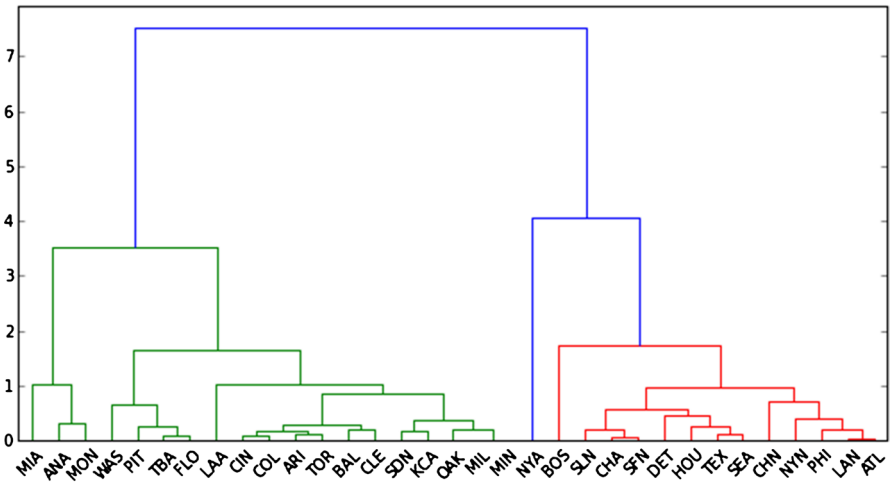

Figure 16. Sample dendrogram.

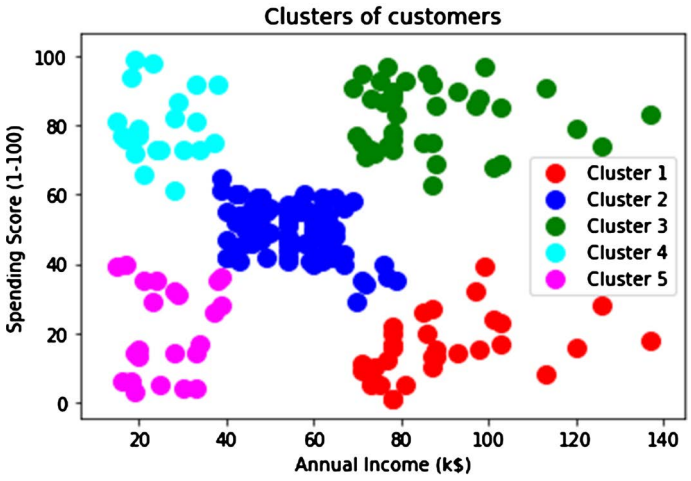

Figure 17. Clusters of the customers based on annual income.

5. Conclusion

We conclude the article with the sample outputs of the K-Means clustering and Hierarchical clustering. There are few scenarios in which we need to perform backward elimination process for identifying the best feature for the model to acquire the best accuracy in the models. As per the acquired results, we identified the best fit model for the addressed problem is k-means. The main reason behind highest accuracy of k-means is because of recurrent changes in Centroid based on the nodes modifications. The point of view of researchers is to identify whether there is a chance of identifying for the path purchasing the item in mall. But the model here requires a simple thing like feature extraction. More number of features will make the model wrong and not optimal. To find the optimal path of the model, we try to implement confusion matrix and identify the difference between obtained predicted result and actual result we want. The future scope of this research is to identify more optimal features to improve the model for better identification of the customer billing prediction.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] To, K.B. and Napolitano, L.M. (2012) Common Complications in the Critically Ill Patient. *Surgical Clinics of North America*, **92**, 1519-1557.

[2] Wollschlager, C.M. and Conrad, A.R. (1988) Common Complications in Critically Ill Patients. *Disease-a-Month*, **34**, 225-293.

[3] Desai, S.V., Law, T.J. and Needham, D.M. (2011) Long-Term Complications of Critical Care. *Critical Care Medicine*, **39**, 371-379.

[4] Perichappan, K.A., Sasubilli, S. and Khurshudyan, A.Z. (2018) Approximate Analytical Solution to Non-Linear Young-Laplace Equation with an Infinite Boundary Condition. 2018 *International Conference on Computing, Mathematics and Engineering Technologies* (*iCoMET*), Sukkur, 3-4 March 2018, 1-5. https://doi.org/10.1109/ICOMET.2018.8346349

[5] Johnson, A.E.W., Ghassemi, M.M., Nemati, S., Niehaus, K.E., Clifton, D.A. and Clifford, G.D. (2016) Machine Learning and Decision Support in Critical Care. *Proceedings of the IEEE*, **104**, 444-466. https://doi.org/10.1109/JPROC.2015.2501978

[6] Badawi, O., *et al.* (2014) Making Big Data Useful for Health Care: A Summary of the Inaugural MIT Critical Data Conference. *JMIR Medical Informatics*, **2**, e22. https://doi.org/10.2196/medinform.3447

[7] Reddy, C.K. and Aggarwal, C.C. (2015) Healthcare Data Analytics. Vol. 36, CRC Press, Boca Raton, FL. https://doi.org/10.1201/b18588

[8] Gotz, D., Stavropoulos, H., Sun, J. and Wang, F. (2012) ICDA: A Platform for Intelligent Care Delivery Analytics. *AMIA Annual Symposium Proceedings*, **2012**, 264-273.

[9] Perer, A. and Sun, J. (2012) MatrixFlow: Temporal Network Visual Analytics to

Track Symptom Evolution during Disease Progression. *AMIA Annual Symposium Proceedings*, **2012**, 716-725.

[10] Mao, Y., Chen, W., Chen, Y., Lu, C., Kollef, M. and Bailey, T. (2012) An Integrated Data Mining Approach to Real-Time Clinical Monitoring and Deterioration Warning. *Proceedings of the* 18*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing, 12-16 August 2012, 1140-1148. https://doi.org/10.1145/2339530.2339709

[11] Wiens, J., Horvitz, E. and Guttag, J.V. (2012) Patient Risk Stratification for Hospital-Associated C. Diff as a Time-Series Classification Task. *Advances in Neural Information Processing Systems*, 467-475.

[12] Saria, S., Koller, D. and Penn, A. (2010) Learning Individual and Population Level Traits from Clinical Temporal Data. *Neural Information Processing Systems* (*NIPS*), *Predictive Models Personalized Med. Workshop*.

[13] Dürichen, R., Pimentel, M.A.F., Clifton, L., Schweikard, A. and Clifton, D.A. (2015) Multitask Gaussian Processes for Multivariate Physiological Time-Series Analysis. *IEEE Transactions on Biomedical Engineering*, **62**, 314-322.

[14] Ghassemi, M., *et al.* (2015) A Multivariate Time Series Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data. *AAAI Conference on Artificial Intelligence*, **2015**, 446-453.

[15] Batal, I., Valizadegan, H., Cooper, G.F. and Hauskrecht, M. (2011) A Pattern Mining Approach for Classifying Multivariate Temporal Data. *Proceedings of the IEEE International Conference on Bioinformatics* (*BIBM*), **2011**, 358-365.

[16] Lasko, T.A. (2014) Efficient Inference of Gaussian-Process-Modulated Renewal Processes with Application to Medical Event Data. *Uncertainty in Artificial Intelligence*, **2014**, 469-476.

[17] Barajas, K.L.C. and Akella, R. (2015) Dynamically Modeling Patient's Health State from Electronic Medical Records: A Time Series Approach. 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, 10-13 August 2015, 69-78.

[18] Wang, X., Sontag, D. and Wang, F. (2014) Unsupervised Learning of Disease Progression Models. 20*th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, New York, 24-27 August 2014, 85-94. https://doi.org/10.1145/2623330.2623754

[19] Sasubilli, S., Perichappan, K.A.P., Kumar, P.S. and Kumar, A. (2018) An Approach towards Economical Hierarchic Search over Encrypted Cloud. *Annals of Computer Science and Information Systems*, **14**, 125-129.

[20] Zhou, J., Liu, J., Narayan, V.A. and Ye, J. (2012) Modeling Disease Progression via Fused Sparse Group Lasso. 18*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing, 12-16 August 2012, 1095-1103. https://doi.org/10.1145/2339530.2339702

[21] Choi, E., Du, N., Chen, R., Song, L. and Sun, J. (2015) Constructing Disease Network and Temporal Progression Model via Context-Sensitive Hawkes Process. *Proceedings of the IEEE International Conference on Data Mining Workshop*, 14-17 November 2015, 721-726. https://doi.org/10.1109/ICDM.2015.144

[22] Pivovarov, R., Perotte, A.J., Grave, E., Angiolillo, J., Wiggins, C.H. and Elhadad, N. (2015) Learning Probabilistic Phenotypes from Heterogeneous HER Data. *Journal of Biomedical Informatics*, **58**, 156-165.